Promoting User Engagement and Learning in Search Tasks by Effective Document Representation

Piyush Arora

B.Tech. (Hons.), and MS by Research in Computer Science and Engineering

A dissertation submitted in fulfillment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisors:
Prof. Gareth J. F. Jones
Dr. Jennifer Foster

June 2018

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

(Candidate) ID No.: 12113409

Date:

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisors Gareth Jones and Jennifer Foster, they have played an important role in my PhD journey. The interactions and discussion with them helped me to grow, learn and develop as a better researcher and person, leading to my overall development.

I would like to thank CNGL/ADAPT centre which provided a wonderful opportunity to pursue my PhD and meet great people working in different research areas over the last couple of years. Special thanks to all my colleagues and friends, who made the time spent at Dublin City University and in general, in Dublin, quite memorable and amazing. The wonderful time spent together will be cherished for years to come.

Finally, I would like to express my gratitude towards my parents, my sisters and my extended family who have been constantly encouraging and supportive. The credit of all my work and efforts goes to them. I am deeply thankful to them for their eternal love and support.

Some lines by William Ernest Henley which has always motivated me and are my inspiration source:

"Out of the night that covers me, black as the pit from pole to pole,
I thank whatever gods may be, for my unconquerable soul.
In the fell clutch of circumstance, I have not winced nor cried aloud.
Under the bludgeoning of chance, my head is bloody, but unbowed.
Beyond this place of wrath and tears, looms but the horror of the shade,
And yet the menace of the years, finds and shall find me unafraid.
It matters not how strait the gate, how charged with punishments the scroll,
I am the master of my fate: I am the captain of my soul."

Promoting User Engagement and Learning in Search Tasks by Effective Document Representation

Piyush Arora

Abstract

Much research in information retrieval (IR) focuses on optimisation of the rank of relevant retrieval results for single shot ad hoc IR tasks. Relatively little research has been carried out on supporting and promoting user engagement within search tasks. We seek to improve user experience by use of enhanced document snippets to be presented during the search process to promote user engagement with retrieved information. The primary role of document snippets within search has traditionally been to indicate the potential relevance of retrieved items to the user's information need. Beyond the relevance of an item, it is generally not possible to infer the contents of individual ranked results just by reading the current snippets. We hypothesise that the creation of richer document snippets and summaries, and effective presentation of this information to users will promote effective search and greater user engagement, and support emerging areas such as learning through search.

We generate document summaries for a given query by extracting top relevant sentences from retrieved documents. Creation of these summaries goes beyond existing snippet creation methods by comparing content between documents to take into account novelty when selecting content for inclusion in individual document summaries. Further, we investigate the readability of the generated summaries with the overall goal of generating snippets which not only help a user to identify document relevance, but are also designed to increase the user's understanding and knowledge of a topic gained while inspecting the snippets.

We perform a task-based user study to record the user's interactions, search behaviour and feedback to evaluate the effectiveness of our snippets using qualitative and quantitative measures. In our user study, we found that richer snippets generated in this work improved the user experience and topical knowledge, and helped users to learn about the topic effectively.

Chapter 1

Introduction

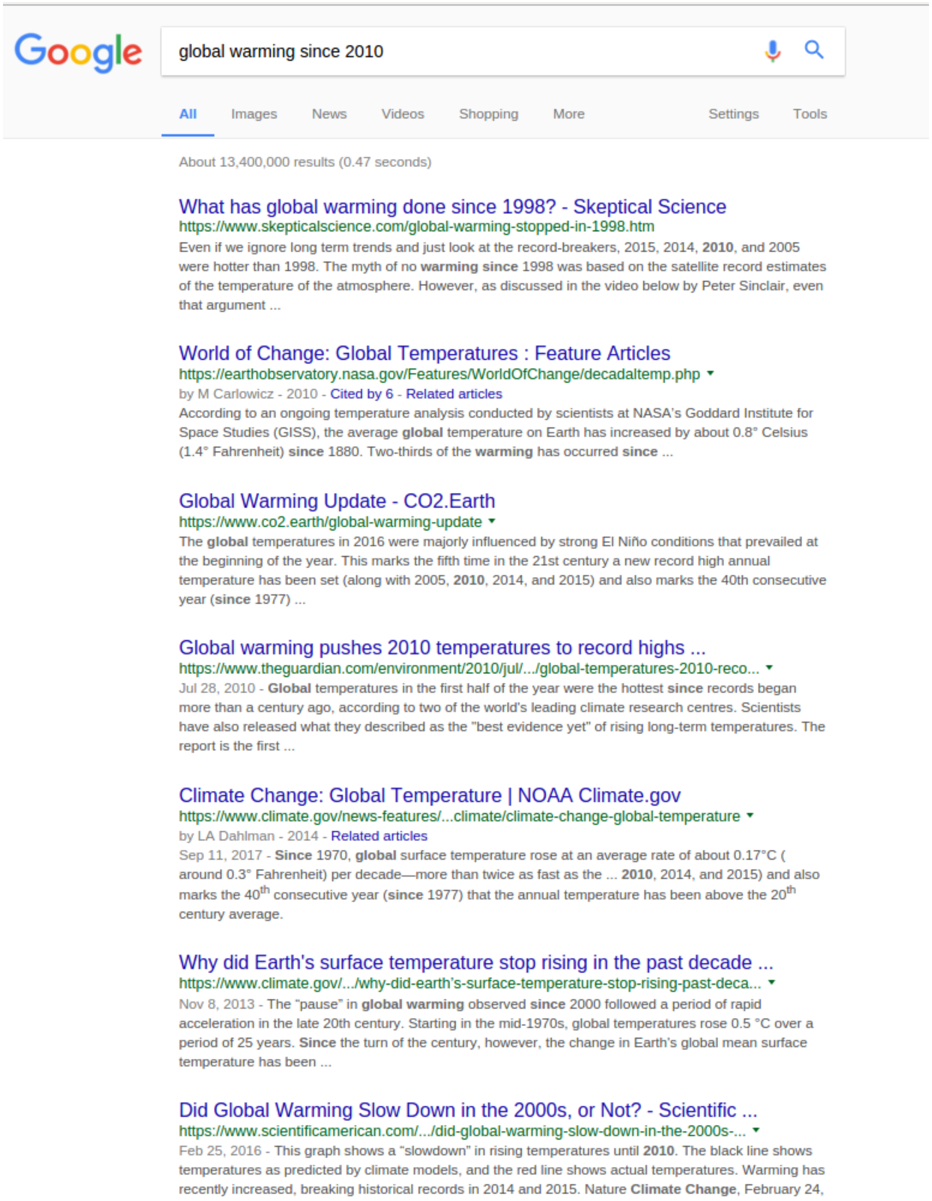
Searching for information online has become an indispensable part of our daily routine. Search activities range from finding answers to questions to satisfy our curiosity, and increasingly, for educational purposes. Most commonly users look towards the web to learn about different topics through MOOCs, Wikipedia, and general web content. Search engines such as Google[1] and Bing[2] which store, crawl and index a vast amount of data, provide near instantaneous access to information able to address a huge range of information needs. A key challenge is finding the relevant information amongst the vast amount available. The information found by search engines is generally returned to the searchers as a list of ranked documents. Once the documents containing potentially relevant information are returned, how to represent it and present to the user is a complex task which is the main focus of this thesis.

Traditionally retrieved documents are returned in a search engine result page (SERP) where each document is represented as a snippet. A standard SERP is shown in Figure 1.1, where a list of returned documents is sorted by decreasing order of relevance scores. Document snippets consist mainly of a title, url, and a short summary of the document. Document snippets are the primary way in which users interact with potentially interesting documents in current IR applications.

[1] `www.google.com`

[2] `www.bing.com`

This thesis focuses on an investigation of the generation of more *effective snippets* to present in a SERP which will enhance the depth and potentially improve the efficiency of access to returned information. In this thesis, we focus specifically on summaries represented in the snippets. We focus only on textual information associated with retrieved documents. Along with measuring the utility of document snippets, we focus on measuring changes in user learning and knowledge gain as users interact with SERPs in a task-based setting. Our overarching goal in pursuing this work is to create better informed and generally more satisfied searchers.



About 13,400,000 results (0.47 seconds)

What has global warming done since 1998? - Skeptical Science
https://www.skepticalscience.com/global-warming-stopped-in-1998.htm
Even if we ignore long term trends and just look at the record-breakers, 2015, 2014, **2010**, and 2005 were hotter than 1998. The myth of no **warming since** 1998 was based on the satellite record estimates of the temperature of the atmosphere. However, as discussed in the video below by Peter Sinclair, even that argument ...

World of Change: Global Temperatures : Feature Articles
https://earthobservatory.nasa.gov/Features/WorldOfChange/decadaltemp.php ▾
by M Carlowicz - 2010 - Cited by 6 - Related articles
According to an ongoing temperature analysis conducted by scientists at NASA's Goddard Institute for Space Studies (GISS), the average **global** temperature on Earth has increased by about 0.8° Celsius (1.4° Fahrenheit) **since** 1880. Two-thirds of the **warming** has occurred **since** ...

Global Warming Update - CO2.Earth
https://www.co2.earth/global-warming-update ▾
The **global** temperatures in 2016 were majorly influenced by strong El Niño conditions that prevailed at the beginning of the year. This marks the fifth time in the 21st century a new record high annual temperature has been set (along with 2005, **2010**, 2014, and 2015) and also marks the 40th consecutive year (**since** 1977) ...

Global warming pushes 2010 temperatures to record highs ...
https://www.theguardian.com/environment/2010/jul/.../global-temperatures-2010-reco... ▾
Jul 28, 2010 - **Global** temperatures in the first half of the year were the hottest **since** records began more than a century ago, according to two of the world's leading climate research centres. Scientists have also released what they described as the "best evidence yet" of rising long-term temperatures. The report is the first ...

Climate Change: Global Temperature | NOAA Climate.gov
https://www.climate.gov/news-features/...climate/climate-change-global-temperature ▾
by LA Dahlman - 2014 - Related articles
Sep 11, 2017 - **Since** 1970, **global** surface temperature rose at an average rate of about 0.17°C (around 0.3° Fahrenheit) per decade—more than twice as fast as the ... **2010**, 2014, and 2015) and also marks the 40th consecutive year (**since** 1977) that the annual temperature has been above the 20th century average.

Why did Earth's surface temperature stop rising in the past decade ...
https://www.climate.gov/.../why-did-earth's-surface-temperature-stop-rising-past-deca... ▾
Nov 8, 2013 - The “pause” in **global warming** observed **since** 2000 followed a period of rapid acceleration in the late 20th century. Starting in the mid-1970s, global temperatures rose 0.5 °C over a period of 25 years. **Since** the turn of the century, however, the change in Earth's global mean surface temperature has been ...

Did Global Warming Slow Down in the 2000s, or Not? - Scientific ...
https://www.scientificamerican.com/.../did-global-warming-slow-down-in-the-2000s-... ▾
Feb 25, 2016 - This graph shows a “slowdown” in rising temperatures until **2010**. The black line shows temperatures as predicted by climate models, and the red line shows actual temperatures. Warming has recently increased, breaking historical records in 2014 and 2015. Nature **Climate Change**, February 24,

Figure 1.1: Example of a Standard Search Engine Result Page (SERP)

The rest of the chapter is structured as follows: Section 1.1 describes a general overview of an information retrieval (IR) system and illustrate how our work presented in this thesis relates to an IR system and its different components. Section 1.2 presents an overview of document snippets, prior work on snippet generation and the main challenges and open problems in the area of snippet generation and evaluation. Section 1.3 describes an overview of our work and reviews the research questions explored in this thesis. Section 1.4 presents the contributions of this work, and finally Section 1.5 describes the structure and the layout of this thesis.

1.1 Information Retrieval (IR) Systems - an Overview

This section presents a formal definition of IR and provides an overview of a standard IR system. We describe the different steps, and processes involved in the working of an IR system, and its main goals.

1.1.1 Conceptual model of an IR System

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) (Manning et al., 2008). An overview of an IR system is presented in Figure 1.2. A query representing a user's information need is used to retrieve documents from the collection, and the top retrieved ranked documents are returned and displayed to the user in a SERP. A key element of the IR process is that the relevance of retrieved content is determined by its level of usefulness in addressing the user information need, rather than whether it matches the information need as represented by the query.

Traditionally, IR systems are considered to comprise of two main components: *Users* and *Systems* (Saracevic, 1975, 1997). In this section we describe different aspects and processes involved in user interactions with the system and the user-centric components of an IR system.

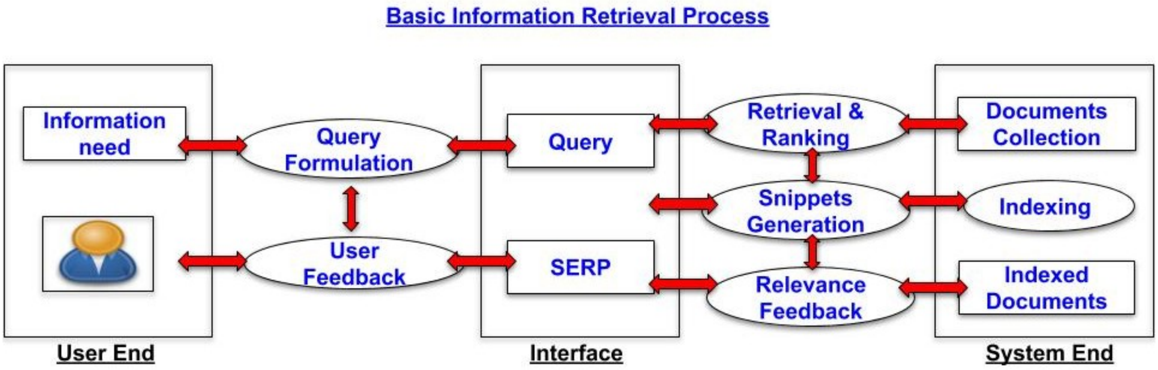


Figure 1.2: Example of an Information Retrieval process

The system-centric component focus on the following aspects of IR:

- *Indexing*: The study of the structure and the manner in which data related to the document collection to be searched is stored, to enable efficient retrieval of potentially relevant documents based on a user's request.

- *Retrieval and Ranking*: Is concerned with how the documents that match a user's input query are effectively ranked for return to the user.

- *Snippet Generation and SERP representation*: Focuses on the generation of effective document representations to help a user to gauge the potential usefulness and relevance of a document.

- *Relevance Feedback*: Focuses on potential incorporation of the user's feedback on the current retrieval operation to improve document ranking in a subsequent retrieval pass.

The user-centric component focuses on the following main aspects:

- *Query Formulation*: The main focus is on the representation of the user's information need as a query for input to the IR system.

- *SERP Interaction and Behaviour*: The main focus involves studying and investigating user search behaviour and experience in their interaction with the results of a search operation.

- *User Feedback*: The focus is on considering how users judge the usefulness and relevance of retrieved content. The two most common mechanisms used for this are:

  a) Users evaluate document usefulness by inspecting the snippet representation created for a given query.

  b) Users can evaluate the document usefulness by examining the complete document.

  Ideally, it would be desirable that the user provides feedback on the relevance of document returned by the search system. However, in practice users are reluctant to provide such feedback (Manning et al., 2008). Thus, indirect feedback (also called as implicit feedback) such as clicks made by the user, their dwell time examining individual documents is recorded and analysed as inferred measure of the user's feedback, and used as a proxy for document relevance and user interest (Ruthven, 2008).

- *Query Reformulation*: Concerned with studying the reformulation of queries as users learn more about a search topic and refine queries previously issued to the search system for earlier search.

The system and the user-centric components and operations within and their associated challenges are complex research topics. In this thesis we focus on snippet generation and its presentation in a SERP to measure changes in users interactions, behaviour and knowledge gain as they interact with document snippets in task-based setting.

As highlighted previously and shown in Figure 1.2, an important aspect of the whole search process is the underlying user information need. An information need is the topic about which the user desires to know more (Manning et al., 2008). Information needs can vary a lot, some examples from the TREC Session track (Kanoulas et al., 2012) are shown in Table 1.1.

Well defined information needs with a well defined answer are categorised as known item (KI) and known subject (KS) needs, whereas those which are open ended and do not have a well defined answer and are investigative and learning oriented in nature are classified as interpretive (IP) and exploratory (EX). Next, we describe the goals of an IR system.

Type of Information Need	**Examples**
Known Items (KI)	Where is Bollywood located? From what foreign city did Bollywood derive its name?
Known Subject (KS)	You think that one of your friends may have depression, and you want to search information about the depression symptoms and possible treatments.
Interpretive (IP)	A friend from Kenya is visiting you and and you would like to surprise him with by cooking a traditional swahili dish. You would like to search online to decide which dish you will cook at home.
Exploratory (EX)	You would like to buy a dehumidifier. On what basis should you compare different dehumidifiers?

Table 1.1: Few examples of different types of information need

1.1.2 Goals of an IR system

IR systems provide a way to access information from very large unstructured collections in an effective and an efficient manner to satisfy a user's information need. Broadly, the goals of an IR system can be categorised into following different types:

- **Utility**: provide an answer to a user's query when it exists in the collection searched, help the user to gather useful information to satisfy their information need and complete their search task. Utility captures the notion of usefulness, satisfaction and effectiveness (Belkin, 2010; Manning et al., 2008).

- **Efficiency**: users should obtain the information they want quickly, and thus be able to complete a task in less time than other means by which they might find the information that they need (Manning et al., 2008).

- **Knowledge Gain**: improve user understating of the topic and help them to gain topical knowledge during the search process (Belkin et al., 1982a; Freund et al., 2014b).

- **Adaptation**: adapt to user knowledge and search behaviour, provide results based on user prior topical knowledge and search interactions with the system (Ruotsalo et al., 2015; Belkin, 2010).

The goals *Utility and Efficiency* are well defined and are achieved by the current state-of-the-art models and systems, for KI, KS types of information needs where the need is well defined and have a specific answer. However, for *Knowledge Gain and Adaptation* and complex information needs (e.g. EX and IP), it is hard to set-up experimental investigations, to design collections, to define evaluation measures and metrics and are open research problems in the IR field. Recently there has been increased interest in pursuing and trying to address these open research problems for complex search topics (Hassan Awadallah et al., 2014; Belkin et al., 2017; Kanoulas et al., 2012). Our work is also in this direction to investigate and address the goals of *Utility and Knowledge Gain* for information needs which are exploratory in nature. Next, we present an overview of document snippets and their goals, and commonly used snippet generation approaches.

1.2 Overview of Document Snippets

In this section, first we discuss the goals of document representation, then we present prior work on snippet generation and investigations on user interaction with SERPs and document snippets. Finally we present the challenges and open problems in the area of snippets generation.

1.2.1 Document snippets and their roles

A snippet seeks to represent the potentially relevant information from a document to assist a user to gauge the potential usefulness of the document to satisfy their information need (Ruthven, 2008). In creating a snippet, it can be challenging to identify potential relevant information within a document since the query statement is often a poor expression of the user's information need (Salton and Buckley, 1990). The words expressed in the query often does not match the words used in the relevant information to be retrieved, this word mismatch issue also called as a vocabulary mismatch problem is a major issue in IR. The vocabulary mismatch challenge causes relevant information to be retrieved at lower ranks, or not to be retrieved at all. We investigate and address the vocabulary mismatch issue to capture relevant information to generate effective snippets in this work.

Generating document snippets is also challenging since the information that is displayed to a user in a SERP poses a trade-off. Even if a snippet does not contain the potential relevant information itself, it seems that increasing the size of the snippet may help users to better decide whether the given document is likely to contain useful information they want, before they navigate to it and determine its usefulness for same. However, additional snippet length can bear substantial costs. Irrelevant search results would also include more information, and too much information could be misleading and lead to wrongly interpreting non-relevant and non-useful results as relevant and useful (also called as false hits) which can make the user disinterested and affect the user search experience.

Ideally, document snippets in a SERP can be expected to help an IR system to achieve the goals of *utility* and *knowledge gain* as discussed in Section 1.1.2 for different types of information need. In addition to indicating the potential usefulness of a document, snippets should enable users to identify duplicate information in different retrieved documents to avoid spending time reading similar information, thus improving efficiency and the user search experience. Our definition of **effective snippets** within the scope of this work is the snippets which are: i) clear and easy to read, ii) useful in satisfying a user's information need or to judge the usefulness of a document, and iii) helpful in learning about a topic and improving the user's topical knowledge. Prior work on snippet generation and evaluation has focused majorly on usefulness and readability aspect of the snippets, and not on the knowledge gain aspect that we explore in this work.

Next, we present an overview of earlier work on the generation of document snippets and investigations on user interaction with SERPs and document snippets.

1.2.2 Prior work on snippet generation

In this section, we review alternative approaches and models used for snippet generation and the general problems and challenges in snippet generation.

- **Static methods:** In static methods each sentence from individual document is scored according to some metric such as sentence position, sentence length and the top scoring sentences are selected for inclusion in the snippet, and are shown to the users (Kupiec et al., 1995; Manning et al., 2008). The main problem with the static methods is that irrespective of the query the document snippet generated remains same, thus it becomes difficult to judge the usefulness of a document with respect to a specific information need and determine why a particular document was retrieved for a given user query.

- **Query Biased models:** Initial investigation in late 1990's and early 2000's (Tombros and Sanderson, 1998; White et al., 2003) found that when query-dependent summaries were presented in a SERP, participants were better able to identify relevant documents without reading through to the full text, than the static methods for snippet generation.

  Following these initial findings, research on snippet generation and evaluation has focused on query biased approaches where the top scoring sentences matching a user's query are retrieved from the documents and are combined to form snippets to be shown to the users in a SERP (Leal Bando et al., 2015; Metzler and Kanungo, 2008; Tsegay et al., 2009).

There has been research investigating the length of snippets for different types of information needs (Cutrell and Guan, 2007; Maxwell et al., 2017), the number of document snippets to be presented in a SERP (Kelly and Azzopardi, 2015), and how user behaviour and interaction varies in a SERP (Clarke et al., 2007; Joachims et al., 2005; Pan et al., 2007; Jiang et al., 2014). We review background work on snippet generation and SERP presentation in more detail in Chapter 3. We now turn attention to the open problems and challenges of snippet generation and evaluation.

1.2.3 Problems and challenges of current snippet generation and evaluation approaches

Based on our literature survey and background work on snippet generation and evaluation, in this section we review the challenges and limitations of generation and evaluation of snippet and SERPs methods.

Snippets not being informative enough

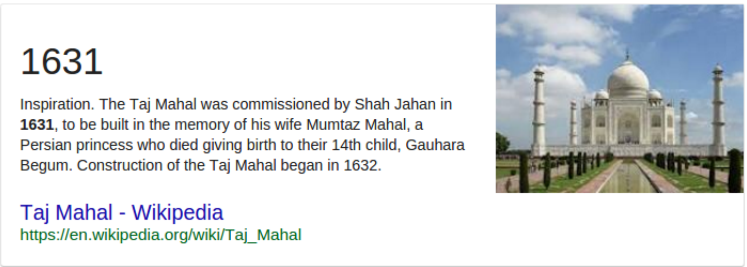
The document snippets may not contain appropriate information to help user to infer whether the information contained in the document will address their information need (Turpin et al., 2009; Jiang et al., 2014). Investigation done in the INEX Snippet retrieval benchmark task found that poor snippets cause users to miss more than half of all relevant results (Trappett et al., 2011, 2013). The analysis of the results showed that users were generally able to identify most non relevant results, but missed significant numbers of relevant results indicating that there is still substantial work to be done in the area of snippet generation to capture relevance information effectively. In a user-based study, Turpin et al. (2009) found that 14% of highly relevant and 31% of relevant documents were never examined because they are judged to be non-relevant based on their snippet summary.

Information in SERPs being repetitive and similar

Jiang et al. (2014) reports an experimental study for KI, KS, EX and IP information needs. This study showed that during a search - documents provided in SERP were often exactly similar or had similar information. When inspecting a ranked list, the relevant document results gradually became less interesting for the users, since the results overlapped significantly or included very similar information. They state the need to remove redundant and repetitive information while presenting information in a SERP to improve the user experience and increase the likelihood of including relevant information in the SERP.

Need for effective SERP representation for different information needs

Search engines are used to support a wide range of different information needs ranging from simple fact finding to complex topic exploration. Current search systems work well for known-item (KI) and known-subject (KS) tasks and factoid questions where the searcher is generally looking for a specific piece of information such as *When was the Taj Mahal built?* and *Average temperature in Dublin in summer*. For KI and KS types of information needs, current web systems (e.g. Google) present information cards typically generated using information extracted from knowledge bases as shown in Figure 1.3. Whereas, for EX and IP needs, e.g. *Global warming since 2010* as shown in Figure 1.2, general SERP representation faces the challenges of information not being informative, repetitive and having low readability. The construction of SERPs thus needs to be investigated with the goal of improving user experience and satisfaction (Jiang et al., 2014).



1631

Inspiration. The Taj Mahal was commissioned by Shah Jahan in **1631**, to be built in the memory of his wife Mumtaz Mahal, a Persian princess who died giving birth to their 14th child, Gauhara Begum. Construction of the Taj Mahal began in 1632.

Taj Mahal - Wikipedia
https://en.wikipedia.org/wiki/Taj_Mahal

Figure 1.3: Example of a featured snippet or information card

The challenges of snippet evaluation

Evaluation of snippets is a complex problem (Savenkov et al., 2011). First, the notion of a “good snippet” is multifaceted and subjective (Ruthven, 2008). It is often hard to balance the different requirements for a snippet, e.g. a snippet containing many query terms from different fragments of the original document is, in general, less readable (Kanungo and Orr, 2009). Longer snippets contain more information about the retrieved document, but hinder overall comprehension of the SERP, and can lead to misleading information (Cutrell and Guan, 2007). The complex and diverse task of snippet evaluation is addressed with a range of different methods: user task-based studies (Maxwell et al., 2017; White et al., 2003; Ageev et al., 2011), automated measures measuring Precision and ROUGE (Keikha et al., 2014b; Leal Bando et al., 2015; Yulianti et al., 2016), manual pairwise comparative evaluation (Leal Bando et al., 2015; Ageev et al., 2013; Kanungo and Orr, 2009; Mishra and Berberich, 2017).

Performing user-based IR evaluation can be expensive and difficult to do correctly, since a properly designed user-based evaluation must use a sufficiently large, representative sample of actual users (Voorhees, 2001). Thus instead of evaluating snippets when presented in a SERP in a task-based setting most prior work performed pairwise evaluation of snippets where two alternative snippets are placed next to each other and are manually compared. The pairwise evaluation approach leaves a number of open questions – How well these enhanced snippets evaluated in a pairwise setting, perform and compare when presented in a SERP? How does user behaviour and experience vary when users interact with SERP's of varying snippet quality?

In the next section we introduce the work described in this thesis which aims to address some of these open problems and challenges of snippet generation and evaluation.

1.3 Overview of our work

We present an overview of our work described in this thesis in which we address the challenges and research gaps in the area of snippet generation, as discussed in Section 1.2.3. Then we discuss the main objectives of our work leading to the main research questions that we investigate in this thesis.

In this work, we develop a framework to generate snippets focusing on three main components: i) **relevance**, ii) **novelty** and iii) **readability**. We focus on the *relevance* aspect to address the challenges of snippets not being informative enough to capture topical information effectively to present in a snippet. We focus on the *novelty* aspect to address the challenges of information being repetitive in a SERP, to provide more diverse and new topical-information which we anticipate will improve user experience. We focus on the *readability* aspect to generate snippets which are clear and easy to read. Using relevance models and improved text similarity techniques, we aim to identify information which is *topically relevant, novel and easily readable.* We perform task-based user studies to investigate changes in user behaviour, learning and experience when snippets of varying quality are represented in a SERP. We measure knowledge gain when users interact with SERPs in a search-task by measuring changes in pre- and post-tests questionnaire in this work.

The main objective of our work is to address and support: *Creation of effective document snippets, to promote search and engagement, and support emerging areas such as improving learning through search.* Next, we describe the specific research questions (RQ's) that we investigate in this thesis.

- **RQ-1:** *Can we develop effective models to address the vocabulary mismatch issues of sentence-level relevance prediction?*

  As discussed in Section 1.2.1 and 1.2.3, finding relevant information is a significant challenge which may lead a user to miss useful documents or wrongly interpret a non-relevant document as relevant. Thus it is important to identify relevant and useful information from the documents to generate effective snippets. The main goal is to address vocabulary mismatch issues to find topically relevant information to generate effective snippets.

- **RQ-2:** *Can we find novel information by comparing information within and across documents effectively?*

We investigate the identification and removal of repetitive and redundant information to select novel content to be shown to the users. The main goal is to investigate sentence-level novelty detection within and across documents to generate effective snippets.

- **RQ-3:** *How to combine sentence-level relevance, novelty and readability features to generate effective snippets?*
  We investigate the combination of output of sentence-level relevance, novelty and readability prediction models to generate effective snippets. We explore the quality of the summaries obtained after combining different proportions of relevance, novelty and readability information.

- **RQ-4:** *How does user search behaviour and gain of topical knowledge vary using snippets generated by our framework?*
  We investigate how user behaviour and experience vary while users interact with a SERP comprising of snippets generated by our framework as compared to snippets generated by a standard BM25-based probabilistic relevance model. We perform a comparative detailed analysis to study changes in *User behaviour*, *User experience*, and *Gain in topical knowledge* with different types of snippets. We perform user-centred task-based evaluation to determine how effective the snippets presented in a SERP are in helping users to learn details of a topic and to satisfy their information need.

1.4 Thesis Contributions

The work described in this thesis makes the following contributions to user engagement with retrieved information and the creation of SERPs and document snippets.

We provide a focused list of contributions from our research:

- **Novelty based snippet generation**: How to include novelty information while generating snippets to be presented in a SERP has not been explored and investigated before. To the best of our knowledge, this is the first experimental study on the generation of snippets for inclusion in a SERP to include novelty information and to take account of how user behaviour varies when novelty information is incorporated. Results measuring changes in user experience and knowledge gain are positive indicating that the SERP presented with snippets generated using novelty and relevance-based information are more effective than snippets generated using only relevance-based information.

- **Knowledge Gain**: Measurement of knowledge gain in a search task has not been widely explored. How to design, frame questions for measuring and evaluating knowledge gain in a search task is an open problem. Based on our experimental investigation we contribute to the study of the measurement of knowledge gain and design of experiments to examine this. We measure the change in knowledge gain as users interact with SERPs presented using different snippets. To the best of our knowledge this is the first published study on how knowledge gain changes when snippets of varying quality are presented in a SERP in task-based setting.

- **Models for sentence-level relevance prediction and novelty detection**: We propose and experimentally investigate novel models for sentence-level relevance and novelty prediction. Our proposed models perform statistically significantly better than our baseline state-of-the-art models. For relevance prediction, the best model comprises a combination of embeddings and a traditional pseudo relevance feedback (PRF)-based query expansion approach. For novelty prediction, the best model compares sentences using syntactic information and embedding-based approaches.

- **Distributed Representation – Embeddings Exploration**: We explore a novel approach of using distributed dense representations of words and sentences for the problem of sentence-level relevance and novelty prediction. We experimentally demonstrate: how we can capture and handle semantically similar words and phrases, and incorporate them into our model of relevance and novelty prediction.

- **Snippet generation framework development**: We develop a framework for generating snippets by varying the threshold of relevance, novelty and readability aspects.

- **Comprehensive evaluation of query biased summaries**: We perform a comprehensive evaluation of summaries and snippets generated using our framework. We measure precision, recall and F-score for sentence-level relevance and novelty prediction, individually assess summaries generated by snippet combination models explored in our work, and perform task-based evaluation where best snippet generation approaches are presented in a SERP.

- **User log collection, analysis and findings**: We collect user logs for a task-based user study, and report our analysis and findings. We explore how user behaviour, knowledge gain and experience varies when users interact with SERPs generated by our framework and a standard BM25-based retrieval model. Our findings suggest that richer document snippets (capturing novelty and relevance) can help to improve the user's knowledge gain and overall user experience effectively as compared to using only relevance-based approaches.

1.5 Thesis Structure

This thesis is organised into following chapters which describe the details of our study as follows.

Chapter 2 describes the related background work. We present an overview of the development of the field of information science, focusing on information seeking, information retrieval, and interactive information retrieval within the scope of this work.

Chapter 3 reviews work on summarization, snippet generation and sentence selection topics. We compare and describe how our work on snippet generation relates to previous work on summarization and snippet generation.

Chapter 4 describes background work on the topic of distributional semantics and its application for learning different dense representations for words and sentences (embeddings), that we explore in this thesis for relevance and novelty prediction.

Chapter 5 discusses our initial investigations that are conducted to understand the system-centric and user centric-components of IR system, to equip ourselves with sufficient knowledge to address the research questions described in Section 1.3.

Chapter 6 discusses our work on sentence-level relevance prediction. We present the experimental investigation done using our proposed novel approaches for topical relevance-based sentence retrieval. We explore traditional retrieval models and embedding-based approaches for relevance prediction and query expansion.

Chapter 7 describes our work on sentence-level novelty prediction. We present the experimental investigation done using our proposed novel approaches for novelty prediction. We explore different sentence-level similarity models, and embedding-based word and sentence representation approaches for comparing sentences within and across documents for finding novel information.

Chapter 8 discusses our work on snippet generation. We present the framework developed for generating snippets to be presented to the user. We combine sentence-level relevance, novelty and readability information to generate effective snippets. We study the quality of snippets developed in this work.

Chapter 9 describes a pilot-study and crowdsource-based study conducted to measure the effectiveness of snippets generated using our framework. We present the study design, experimental setup, detailed analysis and findings from these studies.

Chapter 10 presents the summary of our work. We discuss in brief the summary and novelty of our work. We conclude with some open challenges that remain to be explored in future research.

Appendix A presents a list of publications from our investigations carried out as part of this thesis.

Chapter 2

Related Work

In this chapter we introduce the motivation for information retrieval (IR) systems. We introduce the topic of information seeking which attempts to model the processes involved in satisfying an information need via an IR process, the types of information need typically addressed using an IR system, the models developed to facilitate IR in operational systems and the methods developed to enable the comparative evaluation of alternative approaches to IR.

2.1 Overview of Information Seeking

Information seeking (IS) is "a process, in which humans purposefully engage in order to change their state of knowledge" (Marchionini, 1997). Some examples of information seeking behaviours are asking a colleague for advice, browsing through journals to keep up-to-date, searching in a library for some specific information etc. IS is a fundamental process related to learning and problem solving in which humans engage to change their state of knowledge. Figure 2.1 illustrates the relation between difference processes related to IS and shows that IR is one way to support people in their information-seeking behaviours.

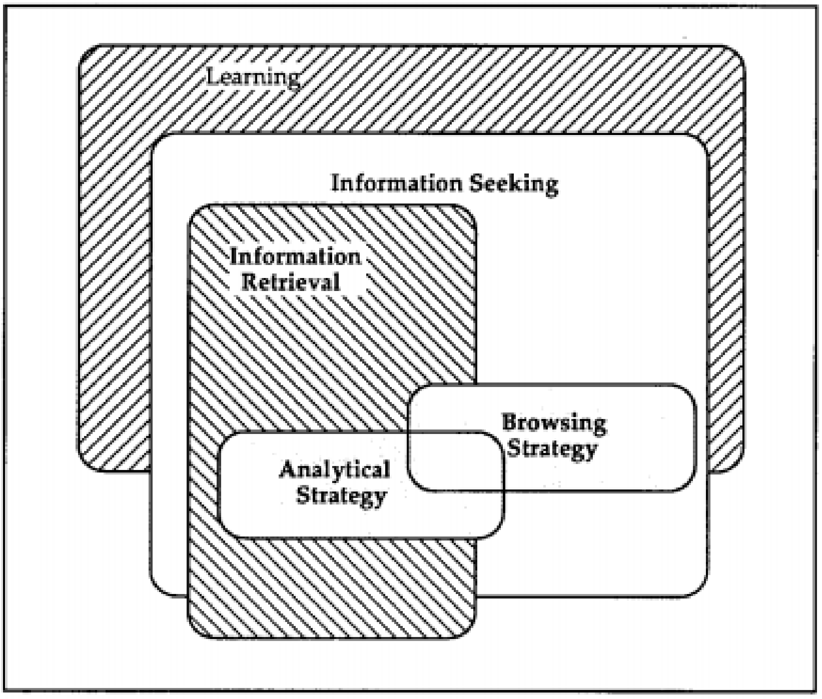


Figure 2.1: Relationship between different processes from Marchionini (1997)

2.1.1 Different IS models

Major focus in IS research has been on the development of theoretical and conceptual models to explain how and why users engage with information. Most of the earlier work in IS studied the processes users go through while interacting with information systems. In this section we review key IS models relevant to this thesis which are commonly used to study users engagement and interactions with information. We review these specific IS models as they helped us to design and plan our user studies to measure user search behaviour and knowledge gain when subjects interact with information in a task-based setting, described later in Chapter 5 and Chapter 9.

Sense-making model: Dervin's sense-making model is shown in Figure 2.2 (Dervin, 1992). This model consists of four main aspects: i) a situation in time and space, which defines the context in which information problems arise; ii) a gap, which identifies the difference between the contextual situation and the desired situation (e.g. uncertainty); iii) an outcome, that is, the consequences of the sense-making process, and iv) a bridge (represented by the triangle), that is a mean of closing the gap between situation and outcome. The strength of Dervin's model lies in its methodological consequences, in relation to information behaviour, which at an abstract level describes how when we encounter problems, can indicate a gap in our knowledge. Further, we can use information systems to fill that gap and complete the task at hand or solve the problem we encounter which can then be measured or assessed by the outcome of the sense-making process.

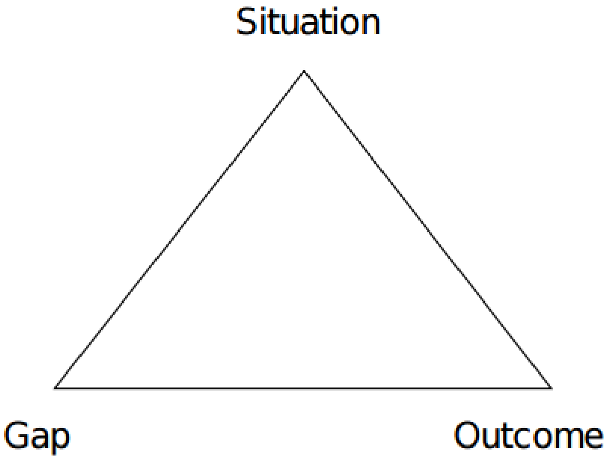
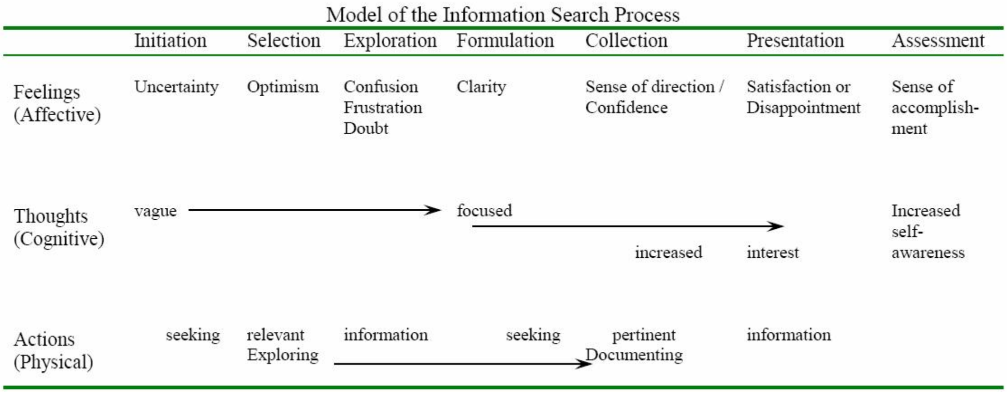


Figure 2.2: Dervin sense-making model

Information Search Process (ISP) model: Kuhlthau (1991) presented a seven stage model of the user's familiarity and use of information seeking progress which was developed based on five user studies. This was the first model to investigate affective feelings of the person in the process of information seeking as well as cognitive (thoughts) and physical (actions) aspects of the process. The ISP Model relates to developing informational skills where the initial search process is broken to several stages of initiation, selection, exploration, formulation, collection, presentation and assessment. Further how the feelings and thoughts vary as person goes through these different stages of a search process is assessed and mapped as shown in Figure 2.3. Understanding these different stages can help to identify the stage an individual is at to provide better support to a user.

Ellis model: Ellis's elaborated different steps and processes a person goes through when involved in an information seeking activity (Ellis, 1989). These steps



Model of the Information Search Process

	Initiation	Selection	Exploration	Formulation	Collection	Presentation	Assessment
Feelings (Affective)	Uncertainty	Optimism	Confusion Frustration Doubt	Clarity	Sense of direction / Confidence	Satisfaction or Disappointment	Sense of accomplish-ment
Thoughts (Cognitive)	vague →			focused →	increased	interest	Increased self-awareness
Actions (Physical)	seeking	relevant Exploring →	information	seeking	pertinent Documenting	information	

Figure 2.3: Stages of ISP model

are defined as: i) starting: the means employed by the user to begin seeking information, for example, asking a colleague; ii) chaining: following footnotes and citations in known material; iii) browsing: "semi-directed or semi-structured searching", iv) differentiating: using known differences in information sources as a way of filter the amount of information obtained; v) monitoring: keeping up-to-date with the information vi) extracting: selectively identifying relevant material in an information source; vii) verifying: checking the accuracy of information; viii) ending: which may be defined as "finishing the task" as shown in Figure 2.4.

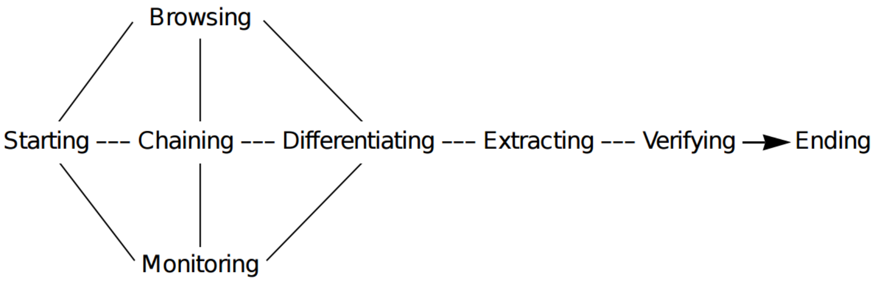

Browsing
Starting --- Chaining --- Differentiating --- Extracting --- Verifying → Ending
Monitoring

Figure 2.4: Stages of Ellis's behavioural model

Bystrom and Jarvelin Model: Byström and Järvelin (1995) developed a qualitative method for task-level analysis of the effects of task complexity on information seeking. They studied how task complexity affects the user-behaviours and the type of information people sought in a finnish public administration context. They found that, as task complexity increased from simple to complex, so did the needs for domain information and problem solving information, the sources being referred (experts, literature, personal collections) also increased for complex tasks as compared to simple tasks, and the success of information seeking decreased for complex tasks. This contrast between simple and complex tasks indicated the importance and consequences of task complexity, and the need to model these complexities and understand different factors that affects the task performance as shown in Figure 2.5. These findings led to a focus on task-based information seeking research.

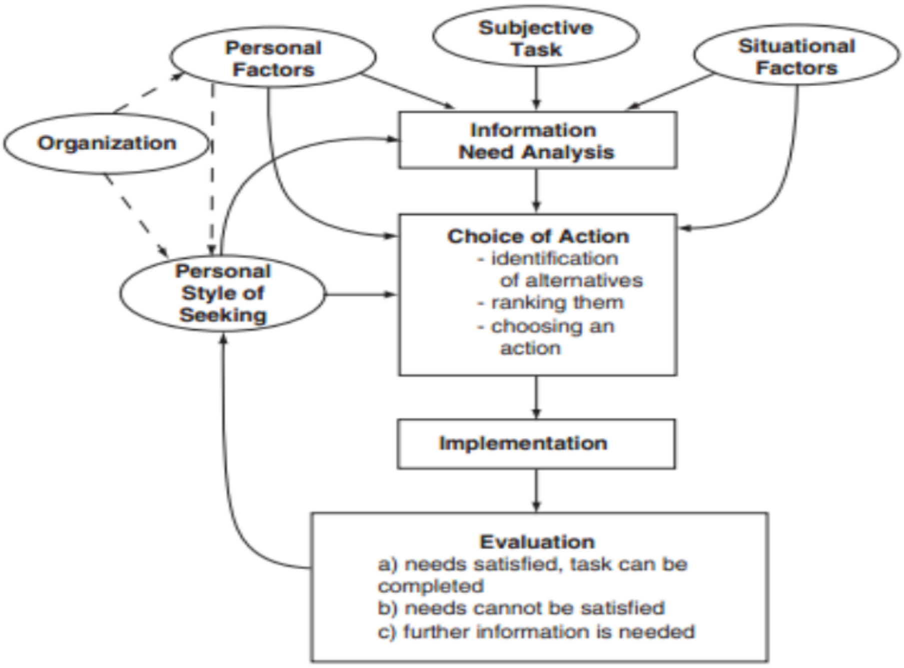

Personal Factors
Subjective Task
Situational Factors
Organization
Information Need Analysis
Choice of Action
- identification of alternatives
- ranking them
- choosing an action
Personal Style of Seeking
Implementation
Evaluation
a) needs satisfied, task can be completed
b) needs cannot be satisfied
c) further information is needed

Figure 2.5: Bystrom and Jarvelin Information Seeking Surface model

ASK Model: The "anomalous state of knowledge (ASK)" hypothesis is that an information need arises from a recognised anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly (Belkin, 1980). As Belkin et al. (1982a) describe for the purposes of IR, it is more suitable to attempt to describe that ASK, than to ask the user to specify her/his need as a request to the system. They proposed to address the information needs of the user from a cognitive viewpoint, which suggests that interactions of humans with one another, with the physical world and with themselves are always mediated by their states of knowledge about themselves, and about that with which or whom they interact. Each time a person interacts, they learn and add knowledge and change their state of knowledge (Belkin et al., 1982a,b) as shown in Figure 2.6. They proposed to incorporate the user state of knowledge and changes in knowledge while modelling user interactions with an IR system to support user interactions effectively.

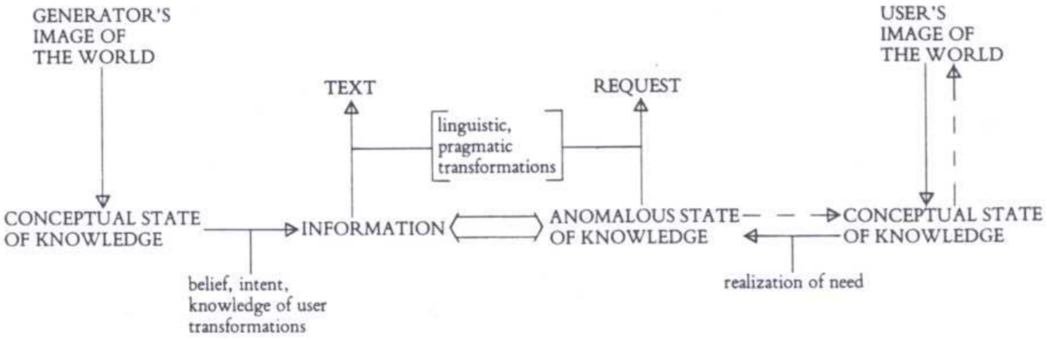

GENERATOR'S IMAGE OF THE WORLD
TEXT
linguistic, pragmatic transformations
REQUEST
USER'S IMAGE OF THE WORLD
CONCEPTUAL STATE OF KNOWLEDGE
INFORMATION
ANOMALOUS STATE OF KNOWLEDGE
CONCEPTUAL STATE OF KNOWLEDGE
belief, intent, knowledge of user transformations
realization of need

Figure 2.6: A cognitive communication system for information retrieval (Belkin et al., 1982a)

We refer interested reader to Wilson (1999); Ruthven and Kelly (2011); Marchionini (1997) for more detailed descriptions of these IS models, their comparison and analysis.

Overall, significant progress has been made in understanding how users go about performing search, and how their feelings, emotions and behaviour changes while doing so. However, most of these models are more conceptual in nature and face challenges while developing systems and evaluating them in experimental setting. Thus most of this work has remained as lab-based or classroom-based investigations. IS theories and models have regained focus to measure user-system interactions in the research conducted in the area of interactive information retrieval (described later in Section 2.3). The IS models described in Section 2.1.1 helped us to design user studies and set up experiments to measure how users knowledge changes in a search session, and how to capture subjective and objective elements of user experience and engagement effectively.

Next we review the specific topic of information needs and the types of tasks commonly used in IR and IS investigations.

2.2 Information Needs and Task Types

Information seeking to address an information need typically takes place within a user task. In this context tasks have commonly been used in IR & IS research as an object of study, where the researcher is interested in how different task types or properties impact search experiences (Toms, 2011). Different topics have been explored to study users in IS research and the effectiveness of search systems in IR research. It has been found that the user behaviour varies a lot depending on the nature of task and user information needs (Byström and Järvelin, 1995). Thus work started in the direction of understanding different types of information needs and tasks which we describe in brief next.

Recapping on Chapter 1, formally, an "information need" is the topic about which the user desires to know more, and is differentiated from a query, which is what the user conveys to the computer in an attempt to communicate the information need (Manning et al., 2008). For system-based evaluation, topics representing information needs are provided in a test collection, to enable test of the effectiveness of an IR system.

For user-based evaluation, the use of search and work task is prevalent. "Search tasks" are goal-directed activities carried out using search systems (Peter et al., 2014). Li and Belkin (2008) define an information search task as "a task that users need to accomplish through effective interaction with information systems". Both of these definitions restrict search tasks to activities done with information systems. " Work tasks" are defined as an "activity to be performed" in order to accomplish a goal (Vakkari, 2003). Toms (2011) defined a work task as having "a defined objective or goal with an intended and potentially known outcome or result, and may have known conditional and unconditional requirements". Both Vakkari's and Toms's definitions go beyond an individual search task and focus instead on the larger goals of the user. Further Borlund (2003) promoted the use of simulated work-task situations in order to create more realistic search tasks. Simulated work-task situations are short search narratives that describe not only the need for information but also the situation – the work task – that led to the need for information. In our user-based investigations for measuring the utility of document snippets when presented in a SERP, users are given a simulated work task where we gave scenarios that are real life information needs for e.g. to "write a report for your college project" or "gather information on the given topic to prepare for a college quiz" (described later in Chapter 9). Next, we review the topic on interactive information retrieval (IIR), where the main goal is to study the user-system interactions.

2.3 Interactive Information Retrieval

The incorporation of users into IR system evaluation and the study of users' information search behaviours and interactions are important concerns for IR researchers. As described by Belkin (2010), the real issue in IR system design are not whether its recall-precision performance goes up by a statistically significant percentage, rather, it is whether it helps the user solve the search task more effectively or efficiently. Thus with the development of the world wide web (WWW) and the availability of user search logs, IS investigations on task-based search have become a major area of focus which led to the development of IIR combining aspects from many fields including traditional IR, information and library science, psychology, and human–computer interaction (HCI). IIR focuses on users' behaviours and experiences including physical, cognitive and affective aspects and the interactions that occur between users and systems, and users and information (Kelly, 2009).

Next we give an overview of different topics on user engagement, user learning and knowledge gain, designing IIR experiments which has been an area of major focus in the field of IIR and is related to the work presented in this thesis.

2.3.1 User Engagement

User Engagement refers to the quality of the "user experience" that emphasises the positive aspects of the interaction, and in particular the phenomena associated with being captivated (engaged) by technology (Attfield et al., 2011). User experience is an important aspect of user engagement. Different metrics used for evaluating user experience can be divided into two broad types: subjective and objective. Subjective measures record a user's perception, generally self reported as a detailed in person interview (O'Brien and Toms, 2008) or as open questions in a questionnaire. Objective measures focus on specific aspect of engagement which can address a range of variables for example physiological measures such as mouse movements and eye movements which can indicate user's attention, or closed questions in a questionnaire to measure specific aspects such as user satisfaction.

Different signals which are commonly captured to measure user experience in a web search session includes: click behaviour, time spent (dwell time) interacting with different parts such as text, multimedia content (Lalmas et al., 2014). Context can heavily influence the search behaviour, where different interpretation of same signals can be made depending on the context, for example for mouse hover features over a document– a short dwell time indicates *not relevant*, but a long dwell time can have different interpretations: it might indicate user is engaged and thus the document is relevant or the user is having a hard time understanding the content (document difficulty) thus spend more time finding useful information and thus can be treated as non-relevant. Overall for careful deductions to be made it is recommended to follow the approach of triangulation (Lazar et al., 2017), where different signals are measured for making effective conclusion and deductions from the studies. In our user-based investigation we record user interactions with the system (mouse movements, clicks made, time spent) and measure changes in user experience using a mixture of objective and subjective measures (described in detail in Chapter 9).

For interested readers we refer the book *Measuring user engagement* by Lalmas et al. (2014) which describes in detail approaches commonly used to study and measure user engagement in a web search.

2.3.2 User learning and knowledge gain

Supporting and measuring learning through search has been an area of active research in recent years. Some factors that contributed to this emerging interest are the workshop at SWIRL 2012 (Allan et al., 2012) where main focus was laid to support learning in a web search, and the recent workshops on *SearchingAsLearning (SAL)* (Freund et al., 2014a) and (Gwizdka et al., 2016). These workshops were conducted to share initial results and findings in the area of learning through search, and bring people working in this area together to discuss ideas and design a road-map to better support gain in knowledge and learning as users interact with search systems in a web search. In this section we discuss the work done on measuring learning and knowledge gain in the education settings and its application and development within IR settings.

2.3.2.1 Overview of learning

Human learning is a complex combination of processes which takes place when a person encounters an experience or situation. Humans are all unique and learn in different ways. People learn by different ways of interactions: listening, observing, sharing in groups, and discussion. Next, we discuss Bloom's taxonomy that we use in our work for measuring user learning and knowledge gain in a task-based setting.

Bloom's taxonomy (Bloom, 1956) is a set of three hierarchical models used to classify educational learning objectives into levels of complexity and specificity. The three lists cover the learning objectives in cognitive, affective and sensory domains. Initial bloom's model was modified and further developed by Anderson et al. (2001); Krathwohl (2002), which lead to creation of six stages with respect to the cognitive dimension and complexity domain as indicated in Figure 2.7. The cognitive domain list has been the primary focus of most traditional education and is frequently used to structure curriculum learning objectives, assessments and activities. These different six stages of Bloom's revised taxonomy (remember, understand, apply, analyse, evaluate and create) is commonly used for designing questionnaires, tests and search-tasks to measure user knowledge gain and learning in education and web search settings (Jansen et al., 2009; Kelly et al., 2015).

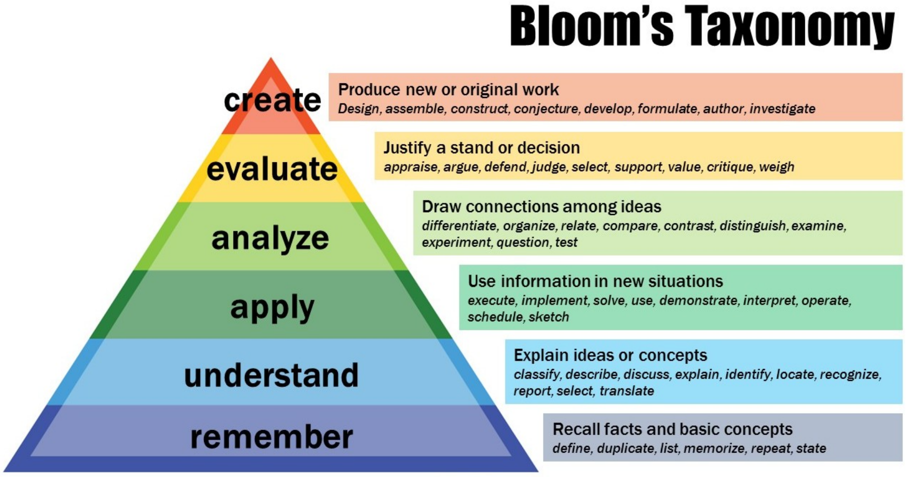

Bloom's Taxonomy
create
Produce new or original work
Design, assemble, construct, conjecture, develop, formulate, author, investigate
evaluate
Justify a stand or decision
appraise, argue, defend, judge, select, support, value, critique, weigh
analyze
Draw connections among ideas
differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test
apply
Use information in new situations
execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch
understand
Explain ideas or concepts
classify, describe, discuss, explain, identify, locate, recognize, report, select, translate
remember
Recall facts and basic concepts
define, duplicate, list, memorize, repeat, state

Figure 2.7: Bloom's Taxonomy

Measuring learning and knowledge gain in education settings such as schools and colleges (Kopainsky et al., 2011; Vakkari, 2000; Pennanen and Vakkari, 2003; Reynolds, 2016) provides the mechanism to measure constant development and regular progress of an individual by conducting tests, examination or quizzes; evaluating presentations or projects etc. However evaluation in online settings poses quite some challenges because of the short interaction time and limited input signals being captured and recorded online. Next, we describe some of the main challenges while measuring learning and knowledge gain in a search-based setup.

2.3.2.2 Challenges in measuring learning

The main challenges in measuring learning and knowledge gain in search can be categorised into four main areas:

- **Supporting diverse type of learning**: There are different learning objectives and knowledge levels as discussed in the Bloom's taxonomy paradigm. Knowledge can be created fresh or is transferred. Further there are two fundamental types of knowledge: i) procedural knowledge – knowing how to do something, and ii) declarative knowledge – knowing about something (Anderson, 1976), Designing experiments for multiple types of learning objectives and different types of knowledge is complex. However there has been some work which study specific knowledge types or knowledge levels described later in this section (Eickhoff et al., 2014; Syed and Collins-Thompson, 2017).

- **Capturing signals and input**: User behaviour and experience for a search task varies based on different variables such as user interest, experience, perceived-difficulty, task knowledge, user background etc. It is challenging to find which signals relate to user's knowledge gain to effectively design studies to measure knowledge gain keeping other variables constant in the experiments. As described before another challenge is to capture signals of learning effectively in a search session where users interact with system for a short span of time.

- **Measuring pre-knowledge**: Most of the time detailed user information is not available thus it is challenging to measure what users already know on a given topic in a single session. Most studies use a questionnaire to assess pre-knowledge of a user on a topic where they are asked to self report on their perceived topical knowledge, or to write a summary on what they already know on the topic or are given a test to measure their topical knowledge (Hunt, 2003; Collins-Thompson et al., 2016).

- **Evaluating gain in knowledge**: How to evaluate knowledge gain in a search session is the most difficult problem while modelling learning in a web search (Hunt, 2003). General methods involve: i) measuring differences in Pre-task, Post-task ratings collected using self-report, ii) knowledge gain measured by evaluating the changes in user's summary and iii) assessing the post- and pre-test scores (Collins-Thompson et al., 2016). Further there is a challenge of knowledge acquisition and retention which effects the user learning, and is hard to incorporate in a search setup.

Next we describe some experimental work done on measuring learning in a search task-based setting for the WWW.

2.3.2.3 Measuring learning in web search

We describe three types of study: i) web search logs-based analysis, ii) lab-based study, and iii) crowdsourcing-based study for measuring learning and knowledge gain.

- *Logs-based analysis*: Eickhoff et al. (2014) investigated evidence of users' within session knowledge acquisition based on the log files of Bing[1] search engine. They focused on two specific types of knowledge acquisition: procedural knowledge (how to do something) and declarative knowledge (knowing facts about something). The authors found evidence both for learning progress within single session, and for persistence of learning across sessions. They found that people behaviour changed over the course of a search session in a way that suggested they learn as they search, and observed that what they learnt appeared to persist across sessions. They found significant proportions of new query terms came from result page snippets and recently visited pages, showing that the search process itself contributed to augmenting the user's domain knowledge.

- *Lab-based user studies*: Collins-Thompson et al. (2016) conducted a lab-based user study in which they investigated potential indicators of learning in web searching, effective query strategies for learning, and the relationship between search behaviour and learning outcomes. Using questionnaires, analysis of written responses to knowledge prompts, and search log data, they found that searchers' perceived learning outcomes closely matched their actual learning outcomes; that the amount searchers wrote in post-search questionnaire responses was highly correlated with their cognitive learning scores; and that the time searchers spent per document while searching was also highly and consistently correlated with higher-level cognitive learning scores.

[1] `https://www.bing.com/`

- *Crowdsource-based studies*: (Syed and Collins-Thompson, 2017) introduced and evaluated a retrieval algorithm designed to maximise educational utility for a vocabulary learning task, in which users learnt a set of important keywords for a given topic by reading representative documents on diverse aspects of the topic. Using a crowdsourced study, they compared the learning outcomes of users by measuring difference in pre- and post reading vocabulary test to measure learning outcome. They focused on the remembering learning outcome based on the Bloom's taxonomy. They found that re-ranking documents based on keyword density helped people to learn words and their definitions effectively.

In our work, we aim to improve user topical knowledge by development of effective snippets and their presentation in a SERP in a task-based setting. We investigate to select potentially relevant, novel and readable information from the document to present it to the users to support engagement and learning. Thus we design specific topics-based pre- and post-tests to effectively measure users knowledge gain based on Anderson et al. (2001) categorisation of Blooms's taxonomy for measuring different learning objectives following the work of (Collins-Thompson et al., 2016; Syed and Collins-Thompson, 2017) in our user-based investigation (described later in Chapter 9).

We refer interested readers to (Hansen and Young Rieh, 2016; Rieh et al., 2016; Vakkari, 2016) which provide a detailed overview of the area of search as learning focusing on the definitions, challenges and proposed methodologies that can be adopted to model learning through search.

2.3.3 IIR experimental design and setup

Designing an experimental investigation to measure user behaviour and experience is a challenging problem. As described in Kelly (2009), users vary based on their prior-knowledge, motivation, prior-search experience, interest and other factors. Individual variations in these factors mean that it is difficult to create an experimental situation that all people will experience the same, which in turn, might makes it difficult to establish causal relationships between the variables being studied. However, due to the increased interest in the area of IIR & HCI there is some literature (Kelly, 2009; Kelly and Gyllstrom, 2011; Kelly et al., 2008; Crescenzi et al., 2016) on designing and evaluating IIR systems with users which we refer for our work, to learn more about the topics of experimental study design (between group, withing group, factorial design), data collection and analysis. Next, we describe some studies which answer design questions related to our investigation.

- **Are lab-based studies better than remote-based studies?**

  Kelly and Gyllstrom (2011), compare two delivery modes for interactive search system (ISS) experiments: remote and laboratory. Their study was completed by two groups of subjects from the same population. The first group completed the study remotely and the second group completed the study in the laboratory. They compared differences in participants, participation behaviours, search behaviours and evaluation behaviours. Overall, for most measures there was no significant differences, but there were some notable differences. Lab subjects provided more favourable responses to exit questionnaire items and reported significantly higher satisfaction. Lab subjects also provided significantly longer responses to open questions, while remote subjects provided more null responses. Their results suggested that many behaviours do not change significantly according to study mode and that results from remote ISS experiments are similar to those from laboratory experiments. Following these findings in our work we perform a mixture of lab-based study with small number of participants and crowdsourcing-based investigation to capture data from larger number of participants (discussed later in Chapter 9).

- **How does time-limit constraint affects user-behaviour and search experience?**

Crescenzi et al. (2016) conducted a study with forty-five participants in which they investigated how time constraints and system delays impacted the user experience during information search. They randomly assigned half of their study participants to a treatment condition where they were only allowed five minutes per search task (the other half were given no time limits). They found time constraints made the search more stressful. Participants with time constraint experienced more pressure, reported that tasks were more difficult, and were not satisfied with their search performance.

Following their findings we do not keep time-constraints in our user-based investigations so that users can perform the given task without any time-pressure.

2.4 Retrieval Models

In this section, we introduce the most widely used models to support the operation of IR system. This is a crucial topic in IR, addressing information needs requires effective IR engines, which need to be constructed using well focused methods. Early IR systems adopted a Boolean model, these have been replaced in many situations by ranked IR models which seek to deliver documents to the user in decreasing order of likely interest.

Boolean Model: In Boolean models terms in a user query are combined using one or either of the Boolean operators "OR", "AND" or "NOT" and are searched within the indexed collection to retrieve documents matching the input query. Documents are represented as set of words. The main problem using Boolean model for search is that results are returned as a set of documents with no scores indicating the varying level of potential relevance to a user, which led to the development of ranked retrieval models such as VSM, BM25, LM which are described next.

Vector Space model: The VSM (Salton et al., 1975) was the first widely used ranked IR model. IR is conceived in a vector space where the axes are defined by terms (typically words), and each document and each query is represented by a vector of terms – a point in the vector space as shown in Figure 2.8 where $d1$, $d2$, and q represents document-1, document-2 and a query respectively. Document similarity, or the similarity between a document D and a query Q is seen as (the reverse of) a distance measure in the space. The scoring function most commonly used in VSM is cosine similarity as shown in Equation 2.1 where $D = d_1, d_2....d_n$ and $Q = q_1, q_2....q_n$, d_i and q_i are weights associated with the i^{th} term in the document and the query respectively (typically represented by the term frequency), n is the number of terms in the collection. Cosine similarity measures angles between vectors as shown in Figure 2.8, where α represents the angle between the input query and document-2, and β represents the angle between the input query and document-1.

$$\frac{\sum_{i=1}^{n} D_i Q_i}{\sqrt{\sum_{i=1}^{n} D_i^2}\sqrt{\sum_{i=1}^{n} Q_i^2}} \tag{2.1}$$

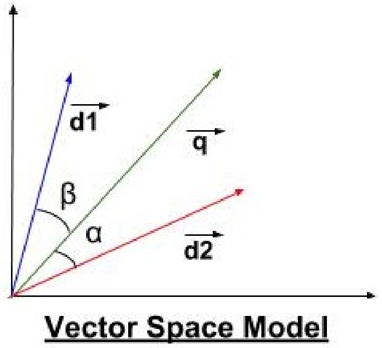


Figure 2.8: Vector Space Model

BM25 model: The VSM was largely replaced by BM25 model developed by Robertson et al. (1995). This is a theoretically motivated probabilistic model that assigns a probability score to each document indicating its relevance to a given query. The probabilistic model for IR is based on the Probability Ranking Principle (PRP) (Robertson, 1977), which states that optimal retrieval effectiveness can be obtained if documents are ranked in decreasing order of their probability of relevance to the user's information need.

Robertson and Sparck Jones (1976) proposed weighting functions to score query terms to retrieve potentially relevant documents. A commonly used weighting function also known as Robertson-Jones relevance weight is described in Equation 2.2. This weighting function which was derived probabilistically assumes distinct query terms distributions in relevant and non-relevant documents, and modelled both query terms presence and query terms absence information from the documents.

$$rw(q_i) = log\frac{(r(q_i) + 0.5)(N - n(q_i) - R + r(q_i) + 0.5)}{(n(q_i) - r(q_i) + 0.5)(R - r(q_i) + 0.5)} \tag{2.2}$$

where $n(q_i)$ is total number of documents containing term q_i, R is total number of assumed relevant document for this query, N is total number of documents in the collection, $r(q_i)$ is number of assumed relevant documents containing term q_i.

In practice the exact values of R and r(q_i) are unknown, an approximation of the rw(q_i) weight for a term-document pair can be obtained by assuming r(q_i)=0, R=0, then the weighting function reduces to Equation 2.3. This function is also known as *BM1* function which is an inverse collection frequency weight of a query term. (Robertson and Sparck Jones, 1976; Robertson et al., 1995).

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{2.3}$$

BM1 model was initially used for document retrieval task at TREC-1, it was found that BM1 model favoured long documents and did not retrieve documents effectively (Robertson et al., 1995). BM1 model had two main limitations it did not incorporate the length of the document and the query term-frequency in the document. These major limitations led to the revision of the initial weighting function to incorporate document length and term frequency factors which led to the development of BM25 weighing function shown in Equation 2.4.

$$score(d,q) = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,d).(k_1+1)}{f(q_i,d)+k_1.(1-b+b.\frac{|d|}{avgdl)})} \tag{2.4}$$

In Equation 2.4, f(q_i, d) is the term frequency of q_i in the document d, $|d|$ is the length of the document d in words, and $avgdl$ is the average document length in the text collection from which documents are drawn, k_1 and b are parameters to weight term frequency and normalise document length variations, and IDF(q_i) is represented in Equation 2.3.

The BM25 model is sensitive to parameters like term frequency and document length which can be easily varied by changing the parameters b and k_1, the optimum values for these parameters are generally calculated by performing grid search over a fixed range of values for a test collection. The BM25 term weighting formula has been quite widely and successfully used across a range of collections and search tasks, especially in the TREC evaluations.

Language Model (LM): An alternative to the VSM and BM25 probabilistic model, is provided by Language models in IR, originally introduced by Ponte and Croft (1998). This works on the theory that a document is a good match to a query if the document model is likely to generate the query, which will in turn happen if the document contains the query words often. The basic language modelling approach builds a probabilistic language model M_d from each document d in the collection C, and ranks documents based on the probability of the model generating the query: $P(q|M_d)$, it is also called as Query-likelihood model.

Using LM for document retrieval, the goal is the calculation of the probability of a document being generated given a query q, which is represented as $P(d_j|q)$ as shown in Equation 2.5. Using the Bayes theorem, the prior probability of $P(d_j|q)$ can be calculated using the likelihood model, and is reduced to $P(q|d_j)$, as $P(q)$ is constant and documents are assumed to come from a uniform distribution thus we can ignore $P(d_j)$. Further, under words independent assumption, $P(q|d_j)$ gets further reduced to relative count of query terms $(w_i, ...w_n)$, in a document as shown in Equations [2.5 - 2.7].

$$P(d_j/q) = \frac{P(q/d_j) * P(d_j)}{P(q)} \tag{2.5}$$

$$P(d_j/q) \cong P(q/d_j) = \prod_{i}^{n} \frac{count(w_i, d_j)}{count(d_j)} \tag{2.6}$$

$$P(q/d_j) = (1 - \lambda) * p(q/d_j) + \lambda * p(q/C) \tag{2.7}$$

To avoid the problem of zero probability when a query term does not occur in a document as shown in Equation 2.6, smoothing process is performed. Jelinek-Mercer smoothing method proposed by Zhai and Lafferty (2001) to perform language model based document retrieval is shown in Equation 2.7. In Jelinek-Mercer smoothing, λ is a parameter which is used to learn weight distributions of a word, thus effectively combining the query term occurrence in a document and collection. The main intuition behind using the collection count of a term is to assign a non-zero probability to the unseen words and improve the word probability estimates.

Query Formulation: A key consideration in the behaviour of an IR system is the query entered describing the user's information need. Often the initial query issued is not a good representation of the user's need since users may not be familiar with terminologies relating to the topic, and thus may not know how to frame effective queries (Salton and Buckley, 1990) or they may not make the effort to form a meaningful query. To address the challenge of the initial query not being useful, query enhancement methods can be used based on relevance feedback. Relevance Feedback is an automatic process for query reformulation, where the main idea consists in selecting important terms, or expressions, attached to documents retrieved in an earlier retrieval pass that have been identified as relevant. Adding these terms to the initial query is intended to make the query a better description of the information need. However, generally users are reluctant to provide feedback thus a common approach of Pseudo Relevance feedback (also called as blind relevance feedback) is used to expand query without user's input.

In our work, to find potential relevant sentences to be used for generation of effective snippets, we explore the techniques of pseudo relevance feedback. Initial query issued to the system being short might not be informative and may not represent the information need effectively, thus to address the query-sentence vocabulary mismatch issues we use relevance feedback approach in our work. We use Okapi's Pseudo Relevance Feedback expansion approach that we describe next.

Pseudo Relevance Feedback (PRF): Robertson (1990) proposed term selection techniques for performing query expansion, where initial search performed using initial query is refined using top ranked documents. In this approach the terms from top retrieved documents which can act as good clues or representative terms to capture a user's query intent are selected using Robertson selection value i.e. rsv scores to expand the initial query to boost the retrieval effectiveness as shown in Equation 2.8 & 2.9.

$$rsv(i) = r(i) * rw(i) \tag{2.8}$$

where $r(i)$ is number of assumed relevant documents containing term i, and $rw(i)$ is the standard Robertson-Jones relevance weight (Robertson and Sparck Jones, 1976) as introduced earlier in this section.

$$rw(i) = log\frac{(r(i)+0.5)(N-n(i)-R+r(i)+0.5)}{(n(i)-r(i)+0.5)(R-r(i)+0.5)} \tag{2.9}$$

where $n(i)$ = total number of documents containing term i, R = total number of assumed relevant document for this query, N = total number of documents in the collection.

An important challenge while performing PRF is determining the assumed potentially relevance documents R and the number of terms to be used for query expansion, which are generally explored using a grid search for a given test collection.

2.5 Evaluation of IR systems

The evaluation of information retrieval (IR) system is the process of assessing how well a system meets the information needs of its users. Broadly there are two classes of IR evaluation user-based evaluation and system evaluation. User-based evaluation measures the user's satisfaction with the system, while system evaluation focuses on how well the system ranks the documents. Although we would prefer user-based evaluation to see the utility of the system (assess whether a user is happy or not), but user-based evaluation is extremely expensive and difficult to do correctly as it must use a sufficiently large, representative sample of actual users and further each systems to be compared must be equally well developed and completed with an appropriate user interface (Voorhees, 2001). Thus a less expensive system evaluation is commonly used for IR experiments which is an abstraction of the retrieval performance that equates good performance with good document rankings.

The Text REtrieval Conference (TREC[2]) which started in 1991, by US National Institute of Standards and Technology (NIST), organises benchmarks campaigns (commonly called as tracks) with an evaluation paradigm for building collection, datasets, tools and resources. TREC has lead to major development in the field of IR towards better retrieval models, statistical models such as best-match (BM25), language model (LM) and better techniques for indexing and storing collections. Our experimental investigation on sentence-level relevance and novelty prediction focuses on standard TREC datasets, described later in detail in Chapter 6 and 7.

An important aspect within the topic of evaluation of IR systems is *Relevance*. There have been different definitions and interpretation of relevance. *Relevance* in general is conceptualised as the user's judgement of the strength of the relationship between a document and their information need (Saracevic, 1975). Similarly,

[2] `https://trec.nist.gov/`

Manning et al. (2008) states that a document is *relevant* if it is one that the user perceives as containing information of value with respect to their personal information need. However, under the TREC definition: a document is considered relevant if it contains any relevant information (topically related information) (Soboroff and Harman, 2005), which is usually judged by a human assessor. It is assumed that the relevance judgement by an assessor will be indicative of the usefulness of a document for an end-user using the system which might not be the case always. For our work relevance measure the extent to which an information (document or sentence) is related to the information need expressed as a query.

In system-based evaluation the retrieval effectiveness of an IR system is measured using a test collection following the cranfield paradigm consisting of three things: i) A document collection, ii) A test suite of information needs, expressible as queries, and iii) A set of relevance judgements, for each of the information need.

With respect to an information need, a document in the test collection can be judged either at binary-level: Relevant or Non-Relevant or at graded level such as Relevant, Partially Relevant, or Non-Relevant.

Evaluation measures commonly used in IR are:

- **Precision (P)**: It is the fraction of retrieved documents that are relevant as shown in Equation 2.10. For information needs which are quite specific such as “finding the website of Dublin City University”, or for our use case to find top relevant sentences from a document it seems more apt to look at just top k ranked results, where Precision at rank k (P@k) measure is commonly used, $P@k = \frac{relevant\ results\ in\ top\ k\ results}{k}$

$$Precision = \frac{relevant\ items\ retrieved}{retrieved\ items} = P(relevant|retrieved) \tag{2.10}$$

- **Recall (R)**: It is the fraction of relevant documents that are retrieved as shown in Equation 2.11.

$$Recall = \frac{relevant\ items\ retrieved}{relevant\ items} = P(retrieved|relevant) \tag{2.11}$$

- **F-Measure**: A single measure that trades off precision versus recall is the F measure, which is the weighted harmonic mean of precision and recall as shown in Equation 2.12.

$$\begin{aligned} F - score &= \frac{(1+\beta^2) * (Precision * Recall)}{(\beta^2 * Precision) + Recall} \\ F - score &= \frac{2 * Precision * Recall}{Precision + Recall} \quad (\beta = 1) \end{aligned} \tag{2.12}$$

- **Normalised Discounted Cumulative Gain (NDCG)**: It is designed for situations of non-binary notions of relevance. Like precision at k, it is evaluated over k top search results. For a query q , let R(d) be the relevance score assessors gave to document d then NDCG scores is calculated as shown in Equation 2.13, where Z_k is a normalisation factor.

$$NDCG(k) = Z_k \sum_{d=1}^{k} \frac{2^{R(d)} - 1}{\log_2(1+d)} \tag{2.13}$$

Next, we describe the statistical tests that we use in our work to measure if the difference between two systems and models performance for system-based experiments, and data captured from different sample group in user-based experiments is statistically significant or not. T-test is quite robust and suitable for IR experiments in comparing systems performance (Hull, 1993), thus we conduct t-test in our experiments for finding whether differences between two systems and settings is statistically significant or not.

- *Student t-test, paired*: Paired samples t-tests typically consist of one group of units that has been tested twice. This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures). The t statistic to test whether the means are different can be calculated as shown in Equation 2.14.

$$t = \frac{\overline{x}}{\frac{s}{\sqrt{n}}} \tag{2.14}$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation and n is the sample size, where for each sample x_i the difference of the two observed values $\overline{x_i}$ =x_{i_1} - x_{i_2} is used for calculating sample mean and sample standard deviation.

- *Student t-test, independent*: The independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared. The t statistic to test whether the means are different can be calculated as shown in Equation 2.15.

$$\begin{aligned} t &= \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{2}{n}}} \\ s_p &= \sqrt{\frac{s_{x1}^2 + s_{x2}^2}{2}} \end{aligned} \tag{2.15}$$

where $\overline{x_1}$, $\overline{x_2}$ are the sample means, s_{x1}, s_{x2}, are the sample standard deviations and s_p is the pooled standard deviation and n is the sample size

- *Pearson Correlation*: This is a measure of the linear correlation between two variables x and y as shown in Equation 2.16. It indicates the degrees of relationship between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.16}$$

where n is the sample size, x_i,y_i are the individual sample points indexed with i, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

2.6 Main aspects and topics of this research

The investigations described in this thesis aim to advance user interaction with IR system, to improve the user experience, to improve learning and understanding of the topic under investigation, while enabling resolution of user information needs. In addition to the search technologies reviewed so far in this chapter, our investigations require us to incorporate technologies not at present generally used in interactive IR systems. Specifically we need to investigate methods to effective sentence-level relevance prediction and estimation of the novelty of a new item compared to those seen so far. In addition to these factors, we also need to be able to measure the expected readability of retrieved material. In this section we review background work on sentence-level relevance, novelty and readability prediction within the scope of this thesis.

2.6.1 Sentence-level relevance prediction

Sentence-level relevance prediction is a common task in IR applications, where given an information need expressed as a query, sentences within a document topically related (matching) to the query are retrieved (Harman, 2002; Soboroff and Harman, 2005). This task is similar to a passage retrieval (Trotman and Geva, 2006) where a passage consists of one sentence. As described by Harman (2002) passages can be hard to work with as the size of a passage is not easily defined and paragraphs are not always available for the documents thus working at sentence-level granularity is more appropriate to work with. Sentence-level relevance prediction is similar to the task of *sentence selection* (passage selection) for Question-Answering applications (Yulianti et al., 2016; Yang et al., 2016; Chen et al., 2015, 2017) where the goal is to return top ranked sentences potentially answering the given question. It is also similar to the task of *sentence extraction* for query biased summarization applications (Goldstein et al., 2000; Metzler and Kanungo, 2008) where the goal is to extract top sentences from a single document using features such as sentence position, keywords.

We review these different topics of summarization, sentence selection and snippet generation in detail in Chapter 3.

Sentence-level relevance in this thesis focuses on the “topical aspect” of relevance. Sentence-level relevance prediction is a more challenging problem than document-level relevance prediction since sentences are typically shorter than the documents, thus there is less textual information to work with to determine the relevance of a sentence. Traditional approaches to sentence-level relevance prediction can be categorised into following three types based on the focus of the work:

- *Scoring sentences using Keywords* (Dkaki et al., 2002): In this method potential key terms from the topic are selected using term frequency or syntactic analysis (extracting named entities), the key terms are looked for calculating the potential relevance of a sentence for a given topic. This method does not perform well in comparison to relevance-based and query expansion-based approaches which are described next.

- *Retrieval Models* (Allan et al., 2003; Losada, 2010; Zhang et al., 2003): The retrieval models reviewed in Section 2.4, can be used for sentence retrieval where instead of retrieving documents the top sentences from a document matching a user query are returned.

- *Query Expansion approaches* (Zhang et al., 2003; Losada, 2010): As described in Section 2.4, the initial query might not be a good representation of user information need. Thus query expansion approaches using wordnet and PRF-based approaches have been commonly used for sentence retrieval.

To rank sentences for potential inclusion in advanced snippets, we investigate alternative ranking models incorporating QE. We propose novel techniques to perform query expansion using: i) a semantic-based embedding approach (overview of embedding is presented in Chapter 4), and ii) combining embedding and PRF-based approaches described in Chapter 6.

2.6.2 Novelty prediction

Initial work on sentence-level novelty detection began in early 2000's. There were two major works that influenced the development of TREC task on Novelty detection (Harman, 2002): i) the benchmark campaign on First story detection (FSD) at Topic Detection and Tracking task (TDT) (Allan et al., 1998, 2000) where the task was to monitor a stream of arriving news stories and to mark each story as a "first" or "not first" story depending on whether it discussed a new topic, and ii) the seminal paper on maximal marginal relevance (MMR), in which Carbonell and Goldstein (1998) developed the MMR technique for ranking documents in a retrieved list based on the combination of relevance and novelty (anti-redundancy) measures, to select documents that are relevant and also diverse from the other already retrieved documents.

For sentence-level novelty prediction, *novelty* is defined as topically relevant information which is *new*, and has not appeared previously in a set of ranked documents on a given topic (Soboroff and Harman, 2005). Novel sentences are determined by identifying which relevant sentences add new information as users read sentences from top to bottom in a linear fashion. Similar to ranked sentence retrieval sentence-level novelty prediction is inherently more challenging in nature compared to document-level novelty detection since i) sentences are short and thus there is less information to work with, and ii) each sentence needs to be compared effectively with all other sentences occurring before it in a ranked list of sentences from a set of ranked documents. The approaches which have been used for novelty prediction can be categorised into four types:

- *Using statistical and distance metrics approach* (Tsai et al., 2010; Tang et al., 2010; Allan et al., 2003; Zhang et al., 2003): In this method sentences are represented as a bag-of-words and are compared using distance metrics such as cosine similarity, to calculate the similarity scores between two sentences. If a sentence has a similarity below a predefined threshold as compared to all other sentences occurring before it, then it is classified as *novel*. In general, distance-based metrics have shown to perform better than other cluster and linguistic-based approaches for novelty detection.

- *Using linguistic and syntactic processing of information* (Li and Croft, 2005; Schiffman and McKeown, 2004; Abdul-jaleel et al., 2004): This method uses syntactic processing of information for novelty detection, where named entities such as person, location, date, organisation etc., are extracted from the sentences. Instead of comparing complete sentences only named entities within the sentences are compared and scored to identify novel information.

- *Supervised models* (Lee, 2015): A recent approach to sentence-level novelty detection focuses on supervised techniques combining multiple features such as: i) sentence position, ii) distance metric based scores, iii) semantic-similarity based scores and others. Lee (2015) found in his investigation of sentence-level novelty prediction that a supervised approach performs better than using only distance metrics-based and syntactic information-based approaches for novelty prediction.

- *Cluster based approaches* (Zhang et al., 2003): In this method instead of comparing sentences with all the sentences occurring above it from a ranked list of documents, documents cluster are made to compare sentences within the documents from the same cluster. A topic is divided into multiple sub-topics and documents are categorised into different sub-topics based clusters, each sentence is compared with the sentences from a document in the same cluster. The main challenge lies in effectively building the clusters to be used for novelty detection.

In our work to present snippets in a SERP which are non-redundant and provide new topical information to the user, we examine novelty prediction using the following main approaches: i) distance metric-based sentence comparison, ii) syntactic information-based sentence comparison, iii) embedding-based sentence comparison, and iv) combining distance metric, syntactic information and embedding-based approaches, described in Chapter 7.

2.6.3 Readability of textual information

Reading is a form of attention to the document itself that modifies our understanding and knowledge structures (Levy, 1997). Readability prediction provides a measure of the accessibility of the information in a document to the reader. Initial work on readability prediction scored the textual content in relation to education levels and different computational models were developed in reference to US school grade levels to measure the ease of reading a text. For example the *FOG* index (Gunning, 1952) which estimates the years of formal education a person needs to understand a text on first reading.

How people interact and read information on the web has been an area of active interest to support comprehension and learning (Dodson et al., 2017). Predicting readability of a textual document or an essay has been an area of active focus in natural language processing (NLP) (Napolitano et al., 2015; Sheehan et al., 2014). Most models for scoring essays and computing readability commonly use syntactic and statistical features such as the number of sentences, average length of sentences, parts-of-speech-based information, along with readability computational model scores e.g. FOG score. Most of these readability prediction methods assume long, well written texts, whereas assessing readability of a snippet or document summary which interests us is challenging because of their short length (typically 2-4 sentences). Next, we describe some of the prior work on measuring readability of document snippets.

Clarke et al. (2007) conducted a clickthrough log analysis of logs gathered from Windows Live Search Engine to study user behaviour when interacting with document snippets. They found that the readability of a snippet can significantly influence the user's web search behaviour and has a direct impact on the click behaviour on the SERP (which is considered as a proxy for relevance). In their log analysis they found that readability of the snippets was statistically correlated with the click through rates. This investigation showed that the readability of web summaries affects clickthrough behaviour. This finding motivated Kanungo and Orr (2009) to work on the task of predicting the readability of short web summaries. They collected a training data consisting of 5382 judgements of readability done by seven human editors over about a year, where document summaries were taken from Yahoo! and Google search results. Each result was rated on a scale of 1-5, where 1 was the least readable (poor), and 5 was the most like written English (good). They extracted various features such as size of snippets, readability models (e.g FOG scores), fraction of capital letters and other features from the summaries and modelled the judgements as function of the features. They found that a supervised model combining multiple features performed significantly better than using only readability measure such as FOG.

Readability level information has also been used for re-ranking of web search results (Collins-Thompson et al., 2011). In their clicklog analysis, the authors found a strong relationship between the difference in the predicted reading levels of a SERP snippet and the full text of the web document, and the average user dwell time (in seconds) for that document. They found that the more difficult the underlying page is, compared to the clicked snippet for that page, the more likely it became that the user would be unsatisfied and leave that page quickly (e.g., spend less than 30 seconds reading it). The findings from Clarke et al. (2007); Collins-Thompson et al. (2011) that readability features impact on the user behaviour and interactions, motivated us to investigate sentence-level readability prediction to generate snippets which are easy to read and clear.

After exploring sentence-level relevance, novelty and readability prediction, we investigate how these features might be combined to generate effective snippets, described in Chapter 8.

The next chapter, reviews existing work on summarization, sentence selection, snippet generation and SERP presentation, Chapter 4 then introduces background work on distributed representation of words and sentences (embeddings), use of embedding for IR applications and an overview of their use in our work for the task of relevance and novelty prediction.

Chapter 3

Background on Summarization and Snippet Generation

In this chapter, we introduce the main concepts of summarization, snippet generation and sentence extraction in IR applications and their relation to our thesis work. We review previous work done in the areas of summarization, snippet generation and sentence extraction. We provide an overview of our work introduced in this thesis on snippet generation and its relation to prior work on summarization and snippet generation.

3.1 Introduction

Summarization is a process of selectively reducing the amount of information contained in an original piece of information (text, video, audio). The summary created seeks to capture the most important content of the original information source relevant to a particular application. Summarization can take the form of generalisation or specialisation. In the former case, the summary seeks to provide an overview of the whole item being summarised, while, in the latter case, summarization is achieved by focusing on elements of source items which are of particular interest to a user for which the summary is intended. An overview of the topic of summarization is contained in (Radev et al., 2002). *Topic-oriented summaries* concentrate on the reader's desired topic(s) of interest, whereas *generic summaries* reflect the author's point of view. *Extracts* are summaries created by reusing portions (words, sentences, etc.) of the input text verbatim, while *abstracts* are created by regenerating the extracted content. Depending on the type of summary being constructed summary generation consists of the following steps:

- *Extraction* – the process of identifying important material in the text,
- *Abstraction* – the process of reformulating important material in novel terms,
- *Fusion* – the process of combining extracted portions, and
- *Compression* – the process of removing out unimportant material.

Further, it is also important to maintain grammatical correctness and coherence at all stages in order to deliver effective summaries (Radev et al., 2002; Mani, 2001a; Goldstein et al., 2000).

Initial work on summarization

Summarization dates back to early research in the 1950-70's (Edmundson, 1969; Luhn, 1958). Early techniques for sentence extraction computed a score for each sentence based on features such as position in the text, word and phrase frequency, important cues such as key phrases (Edmundson, 1969; Luhn, 1958). While summarization continued to attract some research interest over the years, interest in robust summarization methods increased significantly following the emergence of the world wide web (WWW) in the 1990's. Tremendous progress has been made in the last twenty years with the advent of datasets, tools, interactive technologies and systems, with benchmarks competitions such as *SUMMAC* (Mani et al., 1999) and *DUC* (Harman and Over, 2002), and summarization tasks organised by Text Analysis Conference (TAC)[1], Document Understanding Conferences (DUC)[2] and

[1] http://tac.nist.gov/

[2] https://duc.nist.gov//

Text Retrieval Conference (TREC)[3]. These benchmarks campaigns led to the development of evaluation techniques for measuring summarization systems and models.

3.1.1 Types of Summarization

In this section we discuss different types of summarization as been commonly studied in the area of NLP and IR applications.

Single Document Extractive Summarization

Most of the work on summarization relies on extraction of sentences from the original document to form a summary. Each sentence in a document is scored based on a set of features such as words positions and key phrases and summaries were then generated by normalising sentence scores and including high-scoring sentences (Luhn, 1958; Edmundson, 1969). Recent approaches focus more on machine learning (ML) techniques where the emphasis is on developing effective summaries from a set of document-summaries pair (training data), which are represented using a vector of features such as sentence length, sentence position, similar to the features used in earlier approaches (Metzler and Kanungo, 2008). Other approaches rely more on natural language analysis and understanding, where the goal is to capture relation and structure between passages, capture discourse, and rhetorical structure to form more coherent and comprehensive summaries. For more details we direct readers to an overview survey by Radev et al. (2002) which presents different approaches followed for automatic summarization in the late 1990's and early 2000's.

Single Document Abstractive Summarization

The process of abstractive summarization involves multiple steps including: selection, reduction and reformulation (Radev et al., 2002). In general, abstraction involves recognising that a set of extracted passages together constitute something new, something that is not explicitly mentioned in the source, and then replacing them in the summary with the (ideally more concise) new concepts (Rush et al.,

[3] https://trec.nist.gov

2015). Further, the information selected can be compressed, and abstracted based on a knowledge base or a set of concepts, with the main focus being to understand, interpret and infer information. In information extraction applications, many approaches to abstractive summarization involve predefined template or categories that needs to be searched for within a document (Genest and Lapalme, 2012).

Multi-Document Summarization

Generating single summary from a set of related source documents has been an area of active research from last two decades. Apart from the general challenges of single document summarization multi-document summarization involves additional challenges (Radev et al., 2002; Goldstein et al., 2000):

- Recognising and coping with redundancy of information.
- Identifying important differences between documents.
- Ensuring summary coherence level is maintained when sentences come from different documents.

General approaches to multi-document summarization involve identifying novel information from a set of documents, where the focus is on sentence and passage similarity comparison (Goldstein et al., 2000; Carbonell and Goldstein, 1998). In most systems, the sentences are ranked by combining statistical and linguistic features. Then domain independent techniques based on fast, statistical processing, for reducing redundancy and maximising diversity in the selected passages are used. Most of the previous work use cosine similarity to match fragments and passages across documents (Goldstein et al., 2000; Allan et al., 2003; Carbonell and Goldstein, 1998). To generate coherent summaries by combining sentences or passages, most approaches try to maintain the initial order in which the sentences occur in the documents. For news related documents, some approaches add time stamps to the sentences while generating summaries. Recent work on multi-document summarization in IR applications focus on generating answer summaries from a set of documents on a given event, or topic (Keikha et al., 2014a). Next, we discuss the task of evaluation of summarization models.

3.1.2 Evaluation of Summarization Models

Evaluating summaries is a complex problem. The main challenges, as discussed by Mani (2001a,b) are as follows:

1. **Measuring Correctness**: Summarization involves a machine producing output that results in natural language communication. In cases where the output is an answer to a factoid question, there may be a correct answer, but in other cases it is hard to arrive at a notion of what the correct output is. There is always the possibility of a system generating a good summary that is quite different from any human summary used as an approximation to the correct output.

2. **Creating Manual judgements**: Since humans may be required to judge the system's output, this may greatly increase the expense of an evaluation. An evaluation which could use a scoring program instead of human judgements has the advantage that it is easily repeatable.

3. **Varying compression rates**: Summarization involves compression, so it is important to be able to evaluate summaries at different compression rates. This increases the scale and complexity of the evaluation. Creating a gold summary for multiple compression rates or a single gold summary that can be used as a reference to compare summaries generated at varying compression rates makes the evaluation task complex.

4. **User or task specific summaries**: Since summarization involves presenting information in a manner sensitive to a user's or an application's needs, these factors need to be taken into account while evaluating summaries. This in turn complicates the design of an evaluation as each summary needs to be

evaluated by measuring the extent it is useful to a user for completion of their task.

Mani (2001b) distinguishes between the following two evaluation types:

- *Intrinsic Evaluation:* Comparing result summaries against standard answers predefined for evaluation. This involves calculating similarity between system generated summaries and gold summaries generated by humans. Common evaluation measures used are Precision, Recall as described in Chapter 2 and Rouge which is defined below.

  ROUGE: Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004) is a commonly used measure for automatically determining the quality of a summary by comparing it to other (ideal) summaries created by humans. ROUGE counts the number of overlapping units such as n-gram, word sequences and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.

- *Extrinsic Evaluation:* Performing task specific evaluation. Summary output is tested and evaluated in a real application, e.g. when a user compares its performance, utility for completing a search task. Typically two common task based evaluation paradigms are used: 1) Judging relevance of a document, 2) Evaluating comprehension of summaries.

Extrinsic evaluation is expensive and the judgements for a given task and lab-based setup cannot be re-used. Whereas intrinsic evaluation compares multiple automatically generated summaries efficiently, yet a problem lies in development of annotated judgements. Previous studies have shown that typically there is low agreement among users (Radev et al., 2002). Human-generated summaries tend to agree only approximately 60% of the time, measuring sentence content overlap. Thus in general, there are multiple reference summaries that are generated and compared against a system-generated summary to account for subjectivity and variance in the information being captured by different individuals.

We refer readers to Mani (2001a), Das and Martins (2007) and Nenkova and McKeown (2012) for more detail on automatic summarization.

Within the big area of summarization the topic that interests us is *query biased extractive summarization.* In query biased extractive summarization the goal is to generate document summaries by first extracting sentences from a document that matches a user issued query or topic and then combining these sentences to help a user to gauge the usefulness of a document for IR applications. Next, we overview the topic of query biased extractive summarization.

3.2 Query Biased Extractive Summarization (QBES)

In this section, we first review work on *QBES*, and then describe the topics of sentence extraction and sentence combination.

3.2.1 Initial work on QBES

Initially, abstracts of the documents in the form of top sentences were displayed to the user to present clues to judge the relevance of the retrieved documents for their information needs (Rush et al., 1971). As introduced in Chapter 1, these abstracts were static thus irrespective of the query the document snippet generated remained same thus it becomes difficult to judge the usefulness of a document and determine why a particular document was retrieved for a given user query.

Initial work by Tombros and Sanderson (1998) investigated the utility of document summarization in the context of information retrieval, more specifically in the application of query biased (or user directed) summaries. They found in their study that query biased summarization minimise user's need to refer to the full document text, while at the same time provide enough information to support their retrieval decisions in the context of web search. This research led to more focus on understanding and generating effective query biased summaries and became a popular task in SUMMAC and TAC benchmark campaigns. Query biased summaries evaluation in SUMMAC focused on the task of relevance assessment and document categorisation. In the relevance assessment, a subject is presented with either a complete document content and a topic, or a summary of the document and a topic, and is asked to determine the relevance of the document to the topic. In the categorisation task, the subject is asked to categorise a document into one of the predefined categories by either reading the complete document or summaries of a document. The influence of summarization on accuracy and time spent in the task is then studied (Mani, 2001b).

Investigations by White et al. (2003, 2002) showed a positive trend towards using query biased summaries in search engine result pages. A summarization system was developed, and a summary tailored to the user's query is generated automatically for each document retrieved. In general, they found that query biased summarization techniques appear to be more useful and effective in helping users gauge document relevance than the traditional ranked titles/abstracts approach.

The development of search systems in early 2000's and with them a massive collection of user logs led to the study of user search behaviour and interactions during a web search. More studies were conducted to inspect and study user interactions with effective summaries and document representations (Chen and Dumais, 2000; Cutrell and Guan, 2007; Clarke et al., 2007). The whole process of generating QBES has been studied as consisting of two main steps: 1) Sentence extraction, 2) Sentence combination to form coherent summaries which we discuss next.

3.2.2 Sentence Extraction

Sentence extraction[4] is an important part of the overall task of generating effective summaries. Within the context of query biased summarization sentence extraction has received major attention (Allan et al., 2003; Metzler and Kanungo, 2008;

[4] "Sentence extraction" and "Sentence selection" phrases are quite interchangeably used, in the IR and NLP research community, where former is mostly used in the context of Summarization application and the latter for Question Answering application. Both terms should be considered similar for our work unless mentioned otherwise.

Leal Bando et al., 2015). Next we discuss previous work done on query biased sentence extraction for generating summaries to be presented to the users in a web search. We group the earlier work based on the type of approaches used for sentence extraction.

Unsupervised – Most commonly used unsupervised approaches for the task of sentence selection focus on sentence-level retrieval models and query expansion techniques (Allan et al., 2003; Soboroff and Harman, 2005) to capture the topical information related to the user query. Other features which are commonly used for sentence selection involves sentence length, sentence position and presence of significant words typically obtained from the collection using term frequency measures (Edmundson, 1969; Luhn, 1958).

In a recent work, Leal Bando et al. (2015) investigated multiple features such as sentence length, sentence position, significant terms, and query expansion techniques for ranking sentences for the construction of query biased summaries. They found that query expansion significantly improve selection of relevant sentences. After sentence selection, the top sentences are combined based on the order in which they occur in the documents to effectively generate document summaries. Further, they conducted a user-based pairwise evaluation of document summaries generated by their model with and without using query expansion techniques. They found that majority of users preferred summaries generated using query expansion techniques than the ones without using query expansion.

Supervised – The availability of training datasets led to development of machine learning techniques (Cao et al., 2007) for sentence selection. Initial work by Metzler and Kanungo (2008) explored machine learning techniques for sentence selection for QBES generation. They created a set of query dependent and query independent features, comprising of sentence query overlap, LM scores, sentence location, sentence position etc. They found that using Gradient Boost Decision Trees (GBDT's) work well as compared to commonly used support vector regression based learning to rank (L2R) approaches. Recent work by Yulianti et al. (2016); Chen et al.

(2015); Yang et al. (2016) use word embedding features for addressing the problem of sentence selection and answering non-factoid web queries using L2R approaches. Yulianti et al. (2016) generate extractive summaries from each retrieved document using semantic and context-based features for a given topic. Chen et al. (2015) experimented with semantic approaches for finding answer summaries using a L2R retrieval setting. They showed that using semantic representations learned from external resources such as Wikipedia or Google News substantially improve the quality of retrieved answers.

Deep Learning Based Models – Recent work by Chen et al. (2017); Lee (2015); Wang and Nyberg (2015); Severyn and Moschitti (2015) experimented with deep learning based models for sentence selection. Severyn and Moschitti (2015) presented a convolutional neural network architecture for reranking pairs of short texts, where they learn the optimal representation of text pairs and a similarity function to relate them in a supervised way from the available training data. They showed that results using deep learning system on two popular retrieval tasks from TREC: Question Answering and Microblog Retrieval were quite better as compared to previously used support vector machines (SVM) based approaches. Chen et al. (2017) showed that combining a set of query matching, readability, and query focus features into a simple convolutional neural network lead to effective sentence selection performance.

The main challenge of supervised approaches is the need for large training datasets (Mitra and Craswell, 2017). As discussed in Zhang et al. (2016), ML models can very well memorise the data but overfit on smaller datasets. In our work, we explore *unsupervised* models for sentence selection, where we focus on different features comprising of relevance, novelty and readability (as introduced in Chapter 1) to develop effective snippets.

3.2.3 Sentence Combination

Once sentences have been extracted, the next task lies in effectively combining these sentences to generate summaries. Mishra and Berberich (2017) performed pairwise comparison of summaries generated using top sentences from Wikipedia. They study the impact of altering the ordering between sentences in a summary to measure its readability and comprehensibility. They generate four different summaries using the top 10 sentences from a Wikipedia document: i) Original order of sentences, ii) Reverse order of sentences, iii) Randomly shuffled sentences, and iv) Originally consecutive sentences placed as far as possible. They conducted their studies using a crowdsourced platform. They found that sentence ordering had a significant impact on the coherence quality of fixed-length summaries, and the summary generated using original order of sentences was the most coherent one as compared to other alternative summaries explored in their work.

How to effectively combine sentences is an important problem that impacts the coherency and readability of the summaries. In general, combining the sentences based on their original order of occurrence in a document seems to work well (Leal Bando et al., 2015; Mishra and Berberich, 2017). Next, we review the work done on snippet generation and evaluation.

3.3 Snippet Generation and Presentation

In this section, we present an overview of the earlier work done on snippet generation and its evaluation. Generation of effective snippets can be categorised into two main types:

1. **Sentence based snippets**: Complete sentences are used to generate snippets. These snippets are more readable and coherent facilitating quick scanning and reading of the main content (Kanungo and Orr, 2009).

2. **Keyword-in-context (KIC) snippets**: Incomplete sentences, with a goal to combine multiple text fragments and phrases matching query terms are used for generating snippets. In KIC based snippets, these incomplete sentences are combined together poorly which affects the readability of the snippets (Kanungo and Orr, 2009).

Using sentence based snippets can reduce poor readability and coherence of query biased summaries that consist of KIC based snippets (Kanungo and Orr, 2009). Thus in our work, we focus on sentence based snippet generation as sentences have the capability to present single and complete ideas.

Next, we review work done on studying users interactions with snippets presented in a SERP.

Presentation bias in a SERP – Yue et al. (2010); Marcos et al. (2015); Lagun and Agichtein (2011) studied how the presentation of a snippet affects the user behaviour and their search experience. They show substantial evidence of presentation bias in clicks towards results with more attractive titles with bolded query terms in them. These studies found that query terms in the document snippets plays an important role in determining whether to click a snippet or not in determining relevance of a document. Most of these studies just rely on user's click as the only measure which might not be a complete indicative of actual user behaviour and experience. Previous study by Joachims et al. (2005) analysed the user's decision process using eye-tracking and compared implicit feedback against manual relevance judgements, they concluded that clicks are informative but biased. They found that interpretation of clicks as absolute relevance judgements is difficult, though relative preferences derived from clicks are reasonably accurate on average.

Length of snippets – Cutrell and Guan (2007); Maxwell et al. (2017) studied how the length of summaries and snippets affect user behaviour and their search experience. Cutrell and Guan (2007) explored the effects of changes in the presentation of search results and found that adding information to the contextual snippet significantly improved performance for informational tasks but degraded performance for navigational tasks. The queries they used had a definitive answer. Maxwell et al. (2017) examined result summaries of different lengths and selected four conditions where the change in information gain was the greatest: (i) title only; (ii) title plus one snippet; (iii) title plus two snippets; and (iv) title plus four snippets. They found that participants broadly preferred longer result summaries, as they were perceived to be more informative. However, their performance in terms of correctly identifying relevant documents was similar across all four conditions.

Snippet Evaluation – As described in Chapter 1, evaluation of snippets is a complex problem. The notion of a "good snippet" is hard to define and measure. It is often hard to generate informative snippets, having many query terms from different fragments of the original document might generate less readable summaries and longer summaries bearing more information can hinder the overall comprehension of the SERP, and can lead to misleading information. Further, performing user-based IR evaluation can be expensive and difficult to do correctly. Thus complex and diverse task of snippet evaluation is addressed with a range of different methods:

- Automated measures measuring Precision and Rouge (Keikha et al., 2014b; Leal Bando et al., 2015; Yulianti et al., 2016): Snippets quality is measured in terms of the ranking of relevant sentences (using Precision), and how well the summaries scores against human ideal summaries (using ROUGE). This is a system-based evaluation (intrinsic evaluation) and provide initial signals on the quality of different models being explored for snippet generation.

- Manual pairwise comparative evaluation (Leal Bando et al., 2015; Ageev et al., 2013; Kanungo and Orr, 2009; Mishra and Berberich, 2017): Typically two different type of snippets are compared manually in a pairwise setting, where two alternative snippets are placed next to each other. Users are asked to relatively compare snippets based on a pre-defined evaluation measure such as readability, topicality, usefulness etc. This is a user-based evaluation and provide user-based feedback on the quality of snippets being generated but in a pairwise assessment setting.

- User task-based studies (Maxwell et al., 2017; White et al., 2003; Ageev et al., 2011): Snippets effectiveness is measured in terms of their utility in identifying useful and relevant documents by reading the document snippets. This is a user-based evaluation (extrinsic evaluation) and provide user-based feedback on the quality of snippets when presented in a SERP in a task-based setting.

Next, we introduce the approach to snippet generation in this thesis and its relation to earlier work.

3.4 Overview of Our Work

In this thesis, we study how to select and present information for a given information need to users to promote learning and effective engagement by development of enhanced document snippets. We focus on sentence-level relevance, novelty and readability aspects as reviewed in Chapter 2, for generating document snippets to be presented in a SERP.

Our work is different from a single document summarization or a multi-document summarization where the goal is to generate a combined effective answer for a topic of inquiry given a single document or multiple documents respectively as explored previously, as shown in Figure 3.1. We focus on generating effective snippets for each of the ranked web results returned by a search engine. Our model is motivated by and comprises the techniques used for single document and multi-document summarization: our relevance model aims to score sentences within a document for a given query, and thus is similar to a query biased single document summarization, whereas our novelty model compares sentences within, as well as across, the documents similar to the task of duplicate and redundancy detection as in the case of multi-document summarization.

Next, in Chapter 4, we introduce the topic of distributional semantics and semantic similarity. We present a background on the dense vector representation of words and sentences commonly called as *embeddings* which provide a mechanism for semantically motivated comparison of information.

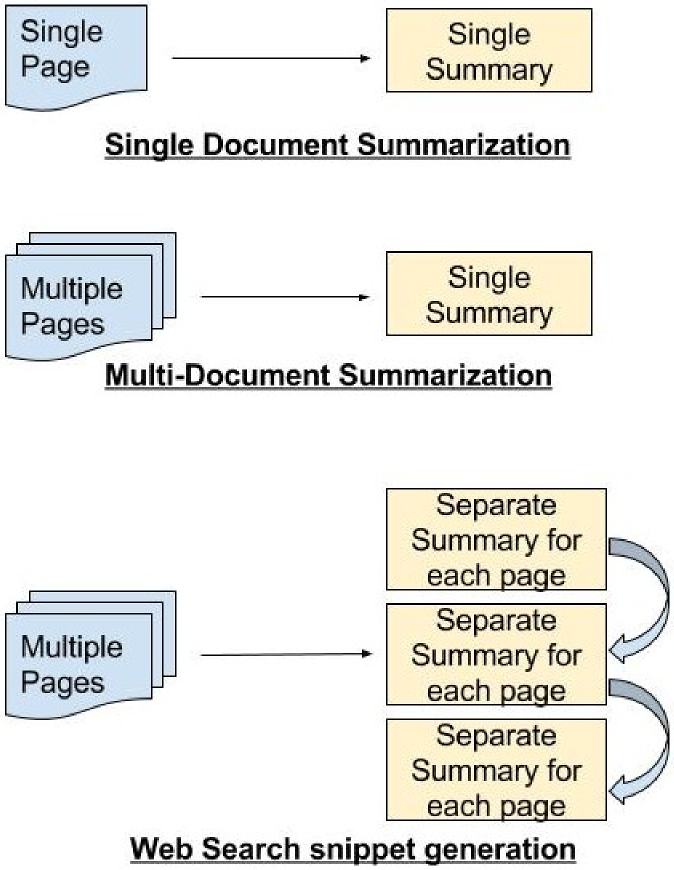

Single Page
Single Summary
Single Document Summarization
Multiple Pages
Single Summary
Multi-Document Summarization
Multiple Pages
Separate Summary for each page
Separate Summary for each page
Separate Summary for each page
Web Search snippet generation

Figure 3.1: Illustration of summarization and snippet generation

Chapter 4

Distributional Semantics and Semantic Similarity

In this chapter, we introduce the main concepts of distributional semantics and vector representation of words and sentences and their application to our thesis work. We present some background on the dense vector representation of words and sentences commonly known as *embeddings* and discuss few models to learn word and sentence embeddings. We then present an overview of application of embeddings, in recent years, in IR and in our work.

4.1 Introduction

Human language is evolved to convey the speaker or writer's meaning. Natural language processing and computational linguistics focuses on construction of computational representations of natural language. These representational models enable applications such as web search, question answering, text classification and natural language understanding.

4.1.1 Vector Representation and Distributional Semantics

Most of NLP tasks involving text, use words as features. Words are commonly represented as vectors, as it facilitate combination of words (sum or average of the words vector) to represent a phrase or a sentence. Vector representation helps to efficiently compare textual information by calculating vector based similarity (e.g. cosine similarity) over the word or the sentence vectors. Under vector representation each word can be represented in two ways:

- *One-hot vector or Local representation*: Under local (or one-hot) representations, every word in a vocabulary is represented by a binary vector $\vec{v} \in \{0, 1\}$, where only one of the values in the vector is one and all the others are set to zero. Each position in the vector $\vec{v}$ corresponds to a term as shown in Figure 4.1, where a black dot represents one and white dot represents zero.

- *Distributed representation*: Under distributed representations every word is represented by a vector $\vec{v} \in R$, where R is a real number. Vector $\vec{v}$ can be a sparse or a dense vector. Vector $\vec{v}$ can be generated using hand-crafted features or by learning the vector representation from the corpus. The individual dimensions may or may not be interpretable in isolation as shown in Figure 4.1.

Under a local or one-hot representation every item is distinct, but when items have distributed or feature based representation, the similarity between two items can be determined based on the similarity between their vector representations. As shown in Figure 4.1, under hot vector representation each term is a unique entity, and the word "notebook" is distinct from the word "laptop". Terms outside of the vocabulary either have no representation, or are denoted by a special "UNK" symbol, under this scheme. In general, local representation does not capture relationship of words, when the meaning of words and phrases are similar. Thus if we search for "Dell Laptop" then we might miss results for "Dell Notebook", as laptop and notebook are distinct vectors using hot vector representation. However, when using

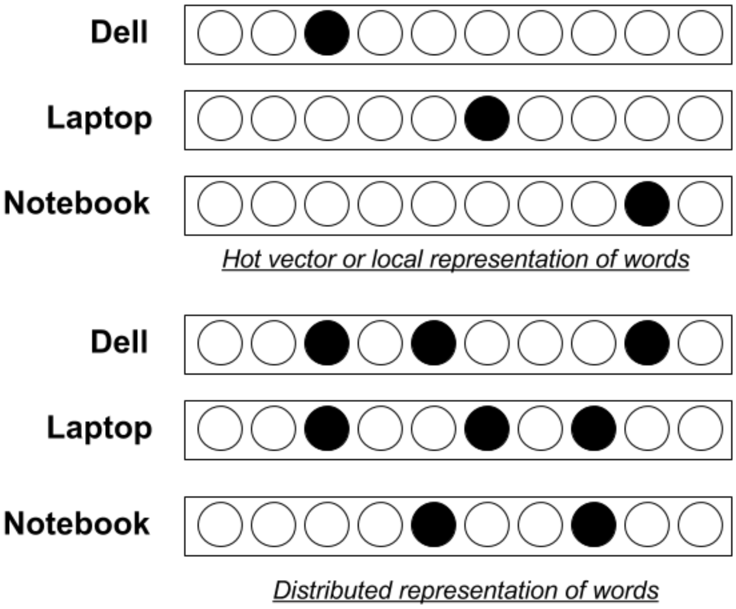

Dell
Laptop
Notebook
Hot vector or local representation of words
Dell
Laptop
Notebook
Distributed representation of words

Figure 4.1: Local and Distributed Representation of words

distributed representation the vector representation of “laptop” and “notebook” capture similarities between these terms and thus if we search for “Dell Laptop” we will find results for “Dell Notebook” as the similarity between “laptop” and “notebook” is non-zero. Thus matching terms which are lexically different but semantically similar.

Several distance metrics can be used to define similarity between terms for distributional semantics, the most common used is the cosine similarity shown in Equation 4.1 (introduced earlier in Chapter 2).

$$cosine\ similarity(A, B) = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4.1}$$

where A_i and B_i are elements of the vectors A and B.

Distributed vector representation helps to compare semantic similarity between words and measure the degree of similarity between them. The main challenge lies in learning effective distributed representations of words which is the focus of study of distributional semantics. Harris (1954) proposed the idea of the “distributional hypothesis", which states that linguistic items with similar distributions have similar meanings. This key idea that "a word is characterised by the company it keeps" was further studied and popularised by Firth (1957). These initial ideas led to the development of the whole research area of distributional semantics, to study how to represent words based on their distributional properties in large samples of language data and collections, to capture semantic similarities. The main idea is to infer the meaning of a word based on its context and neighbouring words. Thus the problem reduces to effectively learning co-occurrence counts of words in their neighbourhood over the whole collection. Using distributed representation for semantic similarity while comparing text, has been widely used in recent years for evaluating text similarity, sentiment analysis, question answering, query-document matching and summarization tasks (Mihalcea et al., 2006; Mikolov et al., 2013a,b; Le and Mikolov, 2014; Agirre et al., 2013; Bogdanova and Foster, 2016). We turn our attention to different types of approaches which have been explored in past for learning distributed representations and discuss same next.

4.1.2 Different types of distributed vector representations

Research in distributional semantics has lead to multiple techniques for learning distributed representation for words. Following is an overview of the principal ones:

- **HAL** proposed by Lund and Burgess (1996) is a method to construct vectors for terms using words co-occurrences in a large corpora of text. Separate counts of the words occurring to the left and right side are maintained while building vector representation of a word. General vectors obtained using this method can have very high dimensionality – typically the size of the vocabulary V.

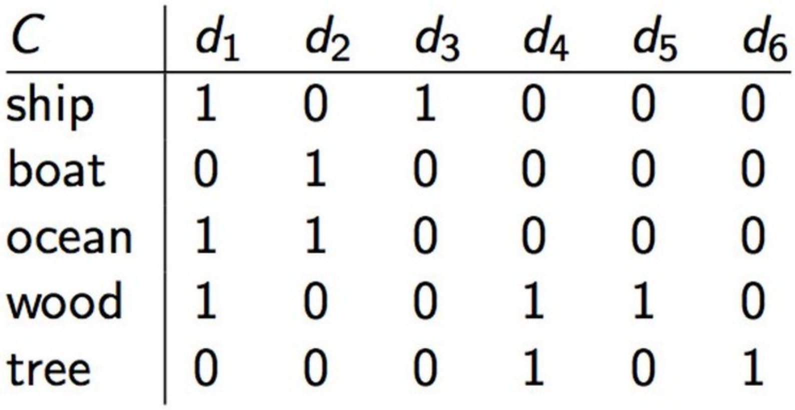
- **Latent Semantic Analysis (LSA)** proposed by Deerwester et al. (1990) is a method in which vector representation of words are learnt using the documents in which the terms occur. A term-document matrix is learnt using the corpus collection. An example is shown in Figure 4.2. Generally the term-document matrix is quite large and sparse, typically the size of the number of documents in the collection.

- **Explicit Vector Representation (EVR)** proposed by Levy and Goldberg (2014) is a method to learn word representation as a sparse vector capturing the contexts in which it occurs. They use a positive pointwise mutual information (PPMI) metric proposed by Church and Hanks (1990) to learn association strength between a word and its context. They show that traditional distributed representation using the sparse vector representation can be used to effectively measure relational similarities and that they perform quite well for word analogy and similarity prediction tasks. The size of vectors formed using EVR are typically of the order of $|V| * |C|$, where V is the vocabulary size and C is the size of the context which is empirically determined, typically C= 5 or 10, is commonly used.

- **Neural language model based representations** proposed by Bengio et al. (2003), and further developed by Mikolov et al. (2013a,b); Pennington et al. (2014), proposed methods to learn distributed word representation using neural network based methods which are trained over large collection of texts. These representations are commonly referred to as *embeddings*, as they embed an entire vocabulary into a relatively low-dimension vector space, where dimensions are real values. In general, the size of the embeddings learnt lies between 50 and 1000 dimensions.

4.1.3 Comparative analysis of dense and sparse representation

While techniques like LSA, HAL and EVR based on distributional features can capture interesting notions of term-term similarity, they have one significant drawback – the resultant vector spaces are highly sparse and high dimensional. The number



C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

Figure 4.2: An example of a term-document matrix representation

of dimensions is generally of the same order as the number of documents or the vocabulary size, which makes their usage for matching similarity and text comparison complex and computationally expensive. High dimension vectors need more space for storage, performing algebraic operations and combining these vectors becomes difficult and inefficient. An alternative is to learn lower dimensional representations of terms from the data that retain similar attributes as the higher dimensional vectors. This desirable feature is very well captured by *embeddings*.

Baroni et al. (2014) empirically demonstrate that embeddings which learn lower dimensional representations, in fact, perform better than explicit counting based models on different word analogies and relational similarities tasks (Turney and Bigham, 2003), possibly due to better generalisation across terms. Although Levy and Goldberg (2014), found that explicit vector representation performs comparably to embeddings for relational similarity tasks. They conclude that the power of embeddings in comparison to EVR lies in effectively learning dense representations by optimally learning the co-occurrence counts of words.

For more detail and a comprehensive overview of background of distributional semantics and embedding the reader is referred to Baroni and Lenci (2010); Turney and Pantel (2010); Mitra and Craswell (2017), which are good surveys of many existing vector representation schemes.

Learning effective distributed representation using neural network based mod-

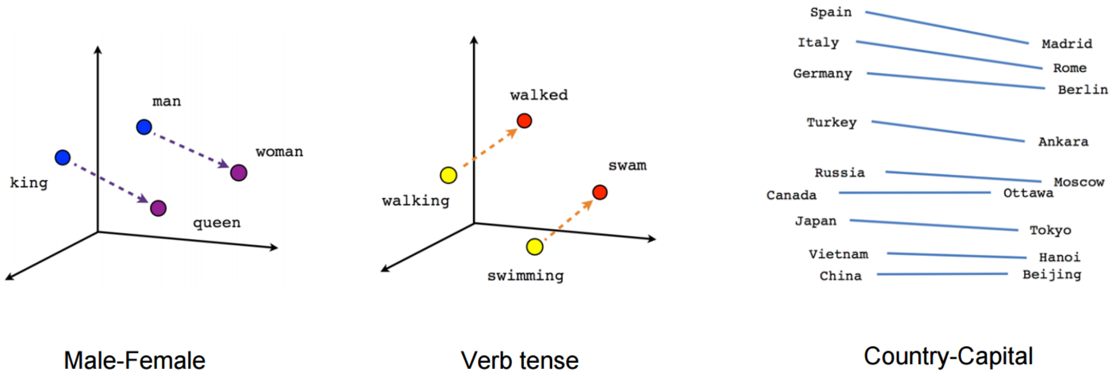


Figure 4.3: Relationships learnt using Word Embeddings

els has become extremely popular, and several new models are being proposed and are emerging on a regular basis. *Word2Vec* is a method to efficiently obtain distributed representation for words that co-occur together in a corpus, where each word representation is learnt in an unsupervised fashion (Mikolov et al., 2013b).

Some examples of the words relationship learnt by word2vec method are shown in Figure 4.3. These models for learning distributed representation of words were further extended and expanded to learn distributed representation for sentences by Le and Mikolov (2014) commonly called as *Paragraph Vectors*. In our work we explore Word2Vec and Paragraph Vectors to capture semantic association between words and sentences to improve retrieval effectiveness and novelty prediction (more details provided in Section 4.3.2). Although we focus only on Word2Vec and Paragraph Vectors approach to effectively learn vector representation of words and sentences in this work, but our experimental investigations using embeddings can be easily extended and replicated using new method being proposed for learning effective word and sentence embeddings using neural network based models such as **Glove** (Pennington et al., 2014) and **FastSen** (Hill et al., 2016) etc. Next, we discuss the working of word2vec and paragraph vector approaches.

4.2 Modelling Distributed Representation

In this section, we discuss methods used for learning distributed representation of words and sentences, that we investigate in this work. All these methods work on the principle of distributional hypothesis as discussed in Section 4.1.1 to infer the meaning of a word based on its context and neighbouring words from the corpus.

4.2.1 Modelling Word Embeddings

We describe in brief two approaches i) continuous bag of words (Cbow) and ii) skip-gram (Cskip) as presented in Figure 4.4, to learn semantic representation of words (Mikolov et al., 2013b,a).

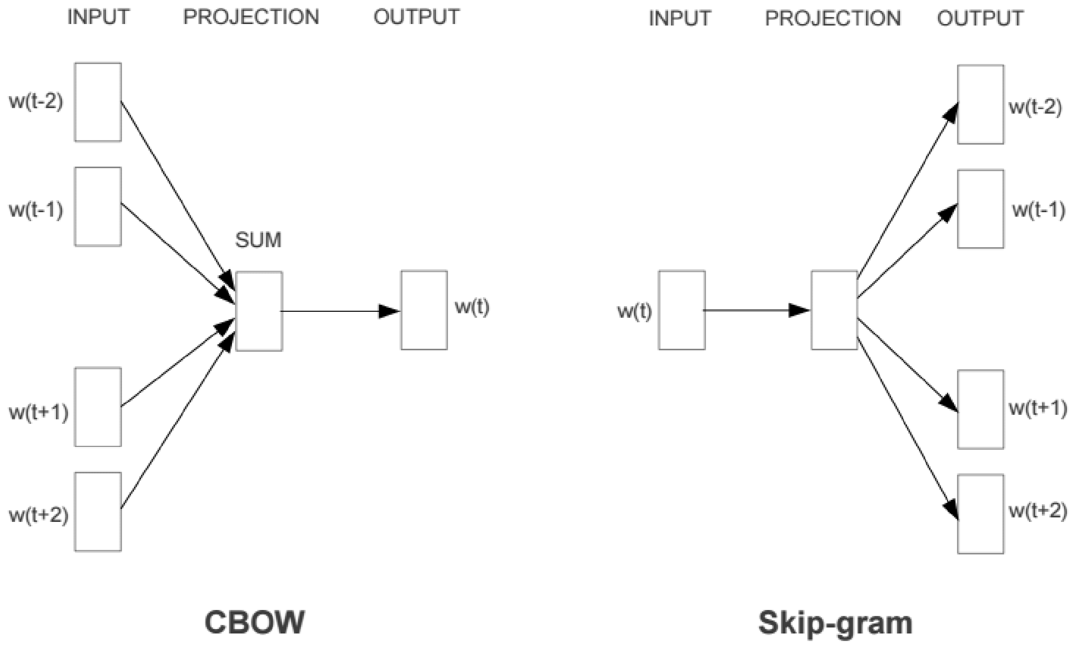


Figure 4.4: Word2Vec CBOW and CSKIP Model

Continuous Bag of Words (CBOW) model

In this model, the current word is predicted based on the context (words in a fixed size window) as shown in Figure 4.4. Each word is represented as a vector of V dimension which is the size of the vocabulary. For each of the word in the corpus a log-linear classifier is learnt where the words from the context are the input, and the training criterion is to correctly classify the current (middle) word. Each word is represented by a vector which is concatenated or averaged with other word vectors in a context, and the resulting vector is used to predict other words in the context. Given a sequence of words w_1, w_2, w_3,..., w_T the objective of a bag of words model is to maximise the average probability, where k is the size of training context, w_t is the centre word as shown in Equation 4.2.

$$\frac{1}{T}\prod_{t=k}^{T-k} p(w_t|w_{t-k}, ...w_{t+k}) \tag{4.2}$$

The prediction task is typically done via a multiclass classifier such as softmax to calculate the probability distribution of predicting the centre word as shown in Equation 4.3.

$$p(w_t|w_{t-k}, ...w_{t+k)} = \frac{e^{y_{w_t}}}{\sum_i^V e^{y_i}} \tag{4.3}$$

We present an example based description of how CBOW method works. For a given sentence “a cat sat on the mat”, we predict the word *sat* using a window of one i.e using the one word to the left (cat), and the one word to the right (on), as shown in Figure 4.5.

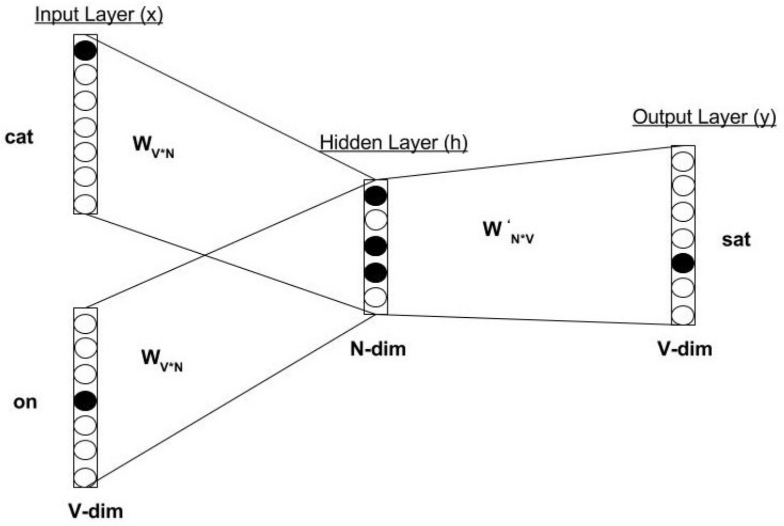


Figure 4.5: CBOW Model

Thus our task is to calculate $p(w_{sat}|w_{cat}, w_{on})$. Given one hot vector representation of words *cat* and *on* say x_{cat} and x_{on}, we form a **h** vector of dimension N by averaging the input vectors, as shown in Equation 4.4.

$$h(x_{cat}, x_{on}) = \frac{1}{2}W^T(x_{cat} + x_{on}) = \frac{1}{2}(v_{cat} + v_{on}) \tag{4.4}$$

where W is a matrix of size $V * N$, where each row indicates an input vector, and v_{cat} and v_{on} are dense input vector representation of words cat and on.

The output dense vector y_{sat} is obtained as shown in Equation 4.5.

$$y_{sat} = b + (W')^T h(x_{cat}, x_{on}) \tag{4.5}$$

where W' is a matrix of size $N * V$ where each column is an output vector, and b is a bias (constant).

Finally, $p(w_{sat}|w_{cat}, w_{on})$ can be calculated using Equation 4.3 and Equation 4.5 as shown in Equation 4.6.

$$p(w_{sat}|w_{cat}, w_{on}) = \frac{e^{y_{sat}}}{\sum_i^V e^{y_i}} \tag{4.6}$$

Thus overall task is to learn the effective weights for the matrix W and W' which represents the dense vector representation of vocabulary terms while training the model over the whole corpus as shown in Equation 4.2 and Equation 4.3.

Skip Gram model

The second model skip-gram is similar to CBOW, but instead of predicting the current word based on the context, it tries to maximise classification of a word based on another word in the context. In this model, the surrounding words are predicted given the current word as shown in Figure 4.4. Each current word is used as an input to a log-linear classifier, and predicts context words within a certain range before and after the current word. Each word is represented as a vector of V dimension.

Given a sequence of words w_1,w_2,w_3,...,w_T the objective of skip-gram model is to maximise the average probability, where c is the size of training context, w_t is the centre word as shown in Equation 4.7.

$$\frac{1}{T}\prod_{t=1}^{T}\prod_{-c\leq j\leq c,j\neq 0} p(w_{t+j}|w_t) \tag{4.7}$$

The basic skip-gram formulation defines $p(w_{t+j} \mid w_t)$ using the softmax function as shown in Equation 4.8.

$$p(w_O|w_I) = \frac{exp(\acute{v}_{w_O}^{T} v_{w_I})}{\sum_{w=1}^{V} exp(\acute{v}_{w}^{T} v_{w_I})} \tag{4.8}$$

where v_{w_I} and $v_{w_O}^{\prime}$ are the input and output vectors, V is the number of words in the vocabulary, v_w represents input vector representation and $\acute{v_w}$ represents output vector representation of word w.

We present an example based description of how the skip-gram method works. For a given sentence “a cat sat on the mat”, we predict the context word *cat* using the word *sat* as shown in Figure 4.6.

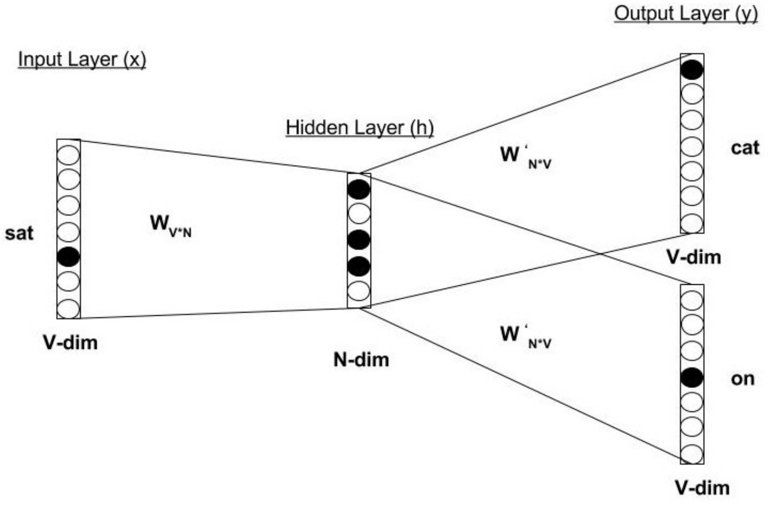


Figure 4.6: Skip-Gram Model

Thus our task is to calculate $p(w_{cat}|w_{sat})$. Given one hot vector representation of word *sat* as x_{sat} we form a **h** vector of dimension *N*, as shown in Equation 4.9.

$$h_{sat} = W^T(x_{sat}) = v_{sat} \tag{4.9}$$

where W is a matrix of size $V * N$, where each row indicates an input vector, and h_{sat} is a dense input vector representation of the word “sat”.

The output dense vector y_{cat} is obtained as shown in Equation 4.10.

$$y_{cat} = h_{sat}^T W' = \acute{v}_{cat} \tag{4.10}$$

where W' is a matrix of size $N * V$ where each column is an output vector.

Finally, $p(w_{cat}|w_{sat})$ can be calculated using Equation 4.8 as shown in Equation 4.11.

$$p(w_{cat}|w_{sat}) = \frac{exp(\acute{v}_{cat}^T v_{sat})}{\sum_{w=1}^{V} exp(\acute{v}_w^T v_{sat})} \tag{4.11}$$

Thus the overall task is to learn the effective weights for the matrices W and W' which represent the dense vector representations of the vocabulary terms, while training the model over the whole corpus as shown in Equation 4.8.

In practice calculating Equation 4.8 is impractical because of the cost of computing p(w_O/w_I) which is proportional to V, which is often large. The novelty of the skip-gram model lies in using an approach called negative sampling (Gutmann and Hyvärinen, 2012). The main idea of negative sampling is to distinguish data from noise by means of logistic regression. Mikolov et al. (2013b) suggests that the unigram distribution raised to the 3/4rd power as shown in Equation 4.12 performs better when drawing negative samples from unigram distribution as compared to unigram and uniform distributions.

$$P_n(w) = U(w)^{3/4}Z \tag{4.12}$$

Using negative sampling the loss function reduces to Equation 4.13. Thus, the task becomes to distinguish the target word w_O drawn from the noise distribution $P_n(w)$ using logistic regression, where k negative samples are drawn for each data sample.

$$\log \sigma(\acute{v}_{w_O}^{T} v_{w_I}) + \sum_{i=1}^{k} E_{w_i \sim P_n(w)}[\log \sigma(-\acute{v}_{w_i}^{T} v_{w_I})] \tag{4.13}$$

The word embeddings model obtained using these models has been shown to perform well on word semantic similarity and word analogies tasks (Mikolov et al., 2013b; Baroni et al., 2014).

4.2.2 Modelling Sentence Embeddings

Based on the idea of learning distributed representation for words, Le and Mikolov (2014) proposed a similar method for learning distributed representation for paragraphs and sentences, which is commonly know as Paragraph Vectors. These sentence representation learnt from the corpus can be effectively used to compare sentences and calculate similarities between them. Next, we describe in brief two models which are commonly used for learning semantic representation of sentences.

Distributed Memory Model (DMM)

In the CBOW model discussed above in Section 4.2.1 the word vectors contribute to a prediction task about the centre word in a sentence. So despite the fact that the word vectors are initialised randomly, they can eventually capture semantics as an indirect result of the prediction task. A similar idea is used in the DMM model where the paragraph vectors along with word vectors predict the centre word given many contexts sampled from the paragraph. In the paragraph vector framework, as shown in Figure 4.7, every paragraph in the corpus is mapped to a unique vector, and every word is also mapped to a unique vector. The paragraph vector and word vectors are averaged (DMM-Mean) or concatenated (DMM-Concat) to predict the next word in a context as shown in Equation 4.14.

$$\begin{aligned} DMM - Mean = h(x_D, x_{cat}, x_{on}) = \frac{1}{3}(v_D + v_{cat} + v_{on}) \\ DMM - Concat = h(x_D, x_{cat}, x_{on}) = v_D + v_{cat} + v_{on} \end{aligned} \tag{4.14}$$

After being trained, similar to CBOW, the paragraph vectors can be used as effective representation of the input sentence.

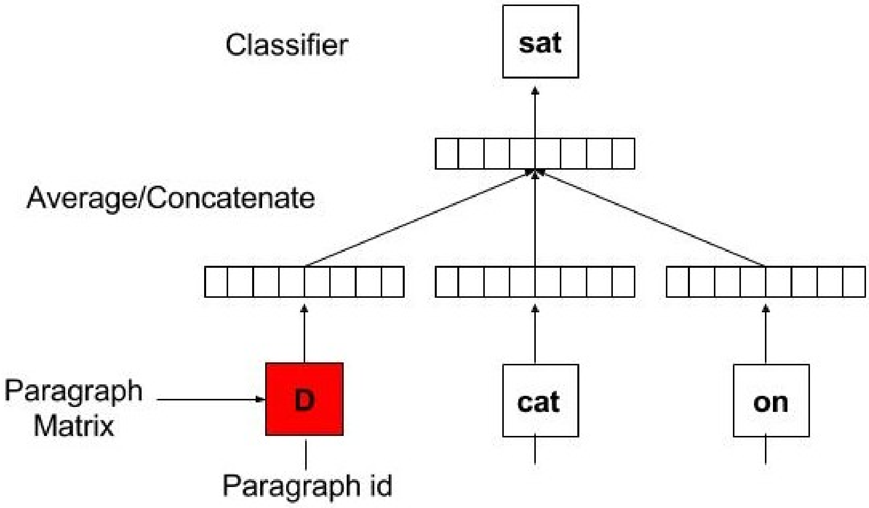


Figure 4.7: DMM model for learning paragraph vectors

Distributed Bag of Words (DBOW) model

The DMM model considers the concatenation of the paragraph vector with the word vectors to predict the next word in a text window. In the DBOW model, the paragraph vector is trained to predict the words in a small window as shown in Figure 4.8. This model is similar to the *skip-gram model* used for learning word vector representation. The context words are ignored in the input, and the model is trained to predict words randomly sampled from the paragraph in the output. While training, a word is randomly sampled from a paragraph D, and the paragraph vector v_D is used to predict context words in the paragraph.

Words and sentence embeddings have shown to perform well for NLP tasks

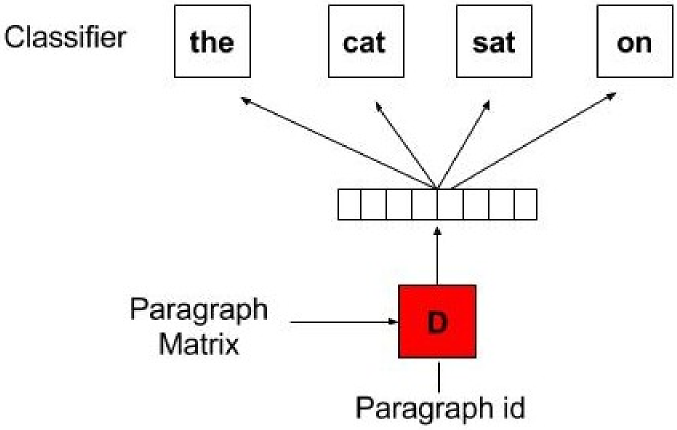


Figure 4.8: DBOW model for learning paragraph vectors

(Mikolov et al., 2013a; Bogdanova and Foster, 2016; Baroni et al., 2014). In recent years, they have also been successfully applied for IR applications which we discuss next.

4.3 Application of Embedding in IR

In this section, first we present an overview of application of word embedding in IR and then present an overview of use of word and sentence embeddings in our work.

4.3.1 Overview of embedding application in IR

Word embedding is of increasing interest among the IR community, in recent years word embedding have been widely and successfully used for IR applications. Most of the previous work using embedding in IR can be categorised in two types:

1. *Incorporating embedding in retrieval model*: Zuccon et al. (2015) and Ganguly et al. (2015) proposed techniques to learn better word probability estimate for Language Model (LM) (reviewed in Chapter 2), using embedding-based approach. To learn effective document frequency and collection frequency counts of query terms they used counts of semantically similar terms obtained using

word2vec, while modelling relevance scores for document retrieval. They found that incorporating embedding-based information improves the effectiveness of document retrieval.

2. *Using embedding for query expansion*: Word embedding has also been explored for learning good expansion terms to retrieve effective documents using the expanded query (Kuzi et al., 2016; Diaz et al., 2016; Roy et al., 2016). Different functions have been proposed for finding potential expansion terms, where the candidate terms are compared to every query term using their vector representations, and then the similarity scores are aggregated, an example is shown in Equation 4.15 for finding potential expansion terms (Diaz et al., 2016; Roy et al., 2016).

$$score(t_c, q) = \frac{1}{q} \sum_{t_q \in q}^{b} cos(\overrightarrow{v}_{t_q}, \overrightarrow{v}_{t_c}) \tag{4.15}$$

   Previous approaches have found that embedding based query expansion on its own performs worse than PRF, but works well in combination with PRF for document retrieval (Zamani and Croft, 2016; Roy et al., 2016).

For more details and a comprehensive overview reader is referred to Mitra and Craswell (2017), which is a good survey of the application of embeddings in IR.

Next, we discuss an overview of application of embeddings in our work.

4.3.2 Application of embeddings in our work

We present an overview of application of word and sentence embeddings in our work for the task of relevance and novelty prediction.

Using word embeddings in our work

Similar to vector space model (VSM) (reviewed in Chapter 2), we use word embeddings, for sentence-level relevance prediction as shown in Table 4.1. We investigate use of word embeddings for sentence retrieval which has not been explored much. We perform query expansion using our proposed embedding-based expansion approaches (described later in Chapter 6). We also use word embedding to compare sentences for finding novel information (described later in Chapter 7).

Task	Similarity Prediction	Method	Type of Embedding
Relevance prediction	Query-Sentence	Word2Vec	In-domain & General embeddings
Query Expansion	Query-Similar words	Word2Vec	In-domain & General embeddings
Novelty prediction	Sentence-Sentence	Word2Vec	In-domain embeddings
Novelty prediction	Sentence-Sentence	Doc2Vec	In-domain embeddings

Table 4.1: Use of embeddings for our work

Types of Embeddings: Embedding of words is learnt based on their neighbourhood and in context words from a large corpus, thus the corpus being used to learn these co-occurrences counts is of vital importance. Previous research (Diaz et al., 2016) and our own investigations (Arora et al., 2017) have shown that learning word embeddings on in-domain and general purpose data such as wikipedia, Google-Ngram corpus captures diverse set of semantically similar terms. Thus we investigate two types of word embeddings in our work which are described below:

- *General domain embeddings*: Embeddings trained on Google news, consisting of about 3 million 300 dimension English word vectors which are released for research, for more details check the link below.[1]

- *In-domain embeddings*: Embeddings learnt using the task specific document collection. We tried both models CBOW and CSKIP, and also varied other parametric settings such as window length, dimension size, window size for our experiments.

More details on the type of embeddings with their hyper-parameters setting and their application in query-sentence similarity, query expansion and sentence-sentence similarity experiments are discussed in our investigation on relevance and novelty predictions described later in Chapters 6 and 7.

[1] `https://github.com/mmihaltz/word2vec-GoogleNews-vectors`

Using sentence embeddings in our work

We use vector representation of sentences and compare sentences within and across documents for novelty prediction. Using Paragraph vectors, semantic similarity between two sentences s_i and s_j is obtained by measuring cosine similarity between their vector representation, as shown in Equation 4.16.

$$similarity(s_i, s_j) = cosine(\overrightarrow{s_i}, \overrightarrow{s_j}) \tag{4.16}$$

In our work, we learn in-domain sentence embeddings. We tried all three models DBOW, DMM-Concat and DMM-Mean, and also varied other parametric settings such as window length, dimension size for our experiments. More details regarding the type of embeddings used for our experiments on novelty prediction is described in Chapter 7.

Next, in Chapter 5, we discuss our initial investigations on user- and system-specific aspects of IR models.

Chapter 5

Initial Investigations

In this work our focus is on the development of richer document snippets and its presentation in a SERP, to improve user search experience and gain of knowledge. We work on system-based experiments for development of effective snippets, and user-based evaluation for measuring snippets utility when presented in a SERP, in a task-based setting. We conducted a couple of initial investigations to learn more about the user and system centric aspects of an IR system. In this chapter, first we describe our system centric preliminary investigations that are done to experiment with different retrieval models, query expansion techniques and tools to be potentially used for generating richer snippets. Next, we describe our user centric preliminary investigations which are done to understand how users engage and interact with web documents and document snippets in a task-based setting. We conclude with the findings and lessons learnt from these investigations, and show how they lead to the main investigations of this thesis presented in subsequent chapters.

5.1 Preliminary Investigations

The experiments carried out for the initial investigations do not directly answer the PhD research questions, but rather lead to the formulation of potential solutions for them. We divide our preliminary studies into two main aspects: *System centric* *investigations* and *User centric investigations* which we describe in this section.

5.1.1 System centric investigations

We conducted three system centric investigations to explore tool and technologies that can potentially be used for snippets generation. To explore different query formulation techniques to address vocabulary mismatch issues we perform initial investigation on the benchmark task of *Cross Lingual Indian News Story Search (CLINSS)* (Gupta et al., 2013), the main motivation was to learn effective techniques that could be used for sentence-level relevance prediction. Comparing sentence-level information to calculate semantic similarity is a complex challenge. We carried out investigations on the benchmark task of *Semantic Textual Similarity (STS)* (Agirre et al., 2015) with the main motivation to investigate methods that could be explored for sentence-level novelty prediction. Further, for the third investigation we focused on *Question Quality prediction using Similar Questions detection (QSQD).* The main aim for conducting QSQD investigation was to explore sentence-level embeddings based question expansion and perform comparative analysis of different retrieval models to be potentially used for relevance and novelty prediction for snippets generation.

5.1.1.1 Cross Lingual Indian News Story Search (CLINSS)

TASK: The CLINSS task is to identify the same news story in different languages where the query is an English news document and retrieved documents are equivalent news documents in the Hindi language.

There are two main challenges of this task:

1) Language barrier: The languages of the source and the target documents are different thus to effectively match English source documents to Hindi target documents we need to process the source and target collection into the same language space.

2) Query formulation: Another key challenge of the CLINSS task is formation of effective search queries from the English news articles, since these are much longer than the queries typically encountered in IR applications. Query formation needs to identify the key elements of the news story, and form an effective query from these capable of identifying articles in the target language describing the same topic.

Our Approach: To address the challenge of the source and target document being in different languages, we investigated and contrasted translation of input queries from English to Hindi using the Google[1] and Bing[2] translation services. Further, to address issues of named entity translation in standard machine translation systems, we also performed transliteration of named entities found in the English queries using Google transliteration[3]. To address the challenge of query formulation we explored the following techniques:

- *Use of query summarization:* We explored a sentence-based summarizer (Kelly et al., 2013) to score and rank the sentences in a document. We hypothesised that selecting the k sentences which are most important to the topic of the document and using these as the basis of our search query can prune noise and divergent content, and hence yield a more effective query. We used summarization over the input queries, and varied the length of the summaries to try to ensure that we do not lose relevant information by removing too many sentences and capture the main aspects of query document effectively.

- *Pseudo relevance feedback (PRF)*: We explored a PRF-based query expansion approach where the initial results returned by a summarised query were used to select potential good expansion terms to improve retrieval performance.

Further, we explored data fusion approaches, which are established technique in IR for merging results from multiple retrieval systems or merging results obtained by varying queries and searching over the same system (Croft, 2002). We combine the ranked system output obtained using:

[1] `https://translate.google.com/`

[2] `https://www.bing.com/translator`

[3] `https://www.google.com/intl/en-GB/inputtools/try/`

- Query translation using different translation services (Google and Bing)
- Summaries (Summ) of different length as input queries (Top 3 sentences, one-third length of the input document)
- PRF-based query expansion performed using the summarised query as an initial query and
- Named entities (NE) transliteration

Experimental Setup: The target document collection had 50,691 news documents in the Hindi language. The training and test dataset had 50 and 25 documents in the English language respectively. For each query the relevance judgements had been carried out by the task developers on a scale of [0-2] where "0" indicates different news event, "1" indicates same news event but different focal event, and "2" indicates same news event and same focal event. As the relevance was done on a scale of [0-2] the evaluation measure NDCG (reviewed in Chapter 2) at rank 1, 5, 10 was used for comparing systems. We used the open source Lucene search engine library[4] to perform indexing of the input documents and searching the queries over the target collection. We used Lucene's default scoring function (a variant of term frequency and inverse document frequency function) for our experiments.[5]

Results and Analysis: The results of our best run shown in Table 5.3 was ranked first among official submissions based on NDCG@5 and NDCG@10 values and second based on NDCG@1 values. Results of different combination approaches explored in this task are reported in Tables 5.1 and 5.2. We compared different combination approaches and performed detail analysis of different explored approaches. The result analysis indicates that:

- For long documents such as news stories, it is advantageous to use a summary of the whole news documents as a query as shown in Table 5.1. The main challenge comes in determining the optimum length of summaries to choose.

[4] `http://lucene.apache.org/core/`

[5] `http://ipl.cs.aueb.gr/stougiannis/default.html`

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Complete Query: Google Translation	0.584	0.523	0.529	0.556
Complete Query: Bing Translation	0.469	0.495	0.508	0.523
Using Google				
Summarized Query	0.622	**0.574**	0.574	0.590
Summarized Query+PRF	0.622	0.550	0.562	0.573
Summarized Query+NE	**0.632**	0.573	**0.579**	**0.591**
Summarized Query+NE+PRF	0.602	0.552	0.560	0.577
Using Bing				
Summarized Query	**0.602**	0.526	0.548	0.558
Summarized Query+PRF	0.571	**0.558**	**0.575**	**0.583**
Summarized Query+NE	0.581	0.545	0.552	0.564
Summarized Query+NE+PRF	0.571	0.548	0.554	0.566

Table 5.1: Fusion Results on training dataset

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Using GTS				
Summarized Query	**0.760**	**0.682**	**0.708**	**0.700**
Summarized Query+PRF	**0.760**	0.669	0.695	0.697
Summarized Query+NE	**0.760**	0.681	0.706	**0.700**
Summarized Query+NE+PRF	0.740	0.680	0.701	**0.700**
Using BTS				
Summarized Query	**0.720**	0.709	0.724	0.732
Summarized Query+PRF	**0.720**	0.709	0.722	0.731
Summarized Query+NE	**0.720**	**0.713**	**0.730**	**0.737**
Summarized Query+NE+PRF	**0.720**	0.702	0.725	0.731

Table 5.2: Fusion Results on test dataset

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Run-1	**0.740**	0.665	0.675	0.684
Run-2	**0.740**	0.670	0.704	0.704
Run-3	**0.740**	**0.680**	**0.726**	**0.724**

Table 5.3: System combinations results on test dataset

Application of summarization may remove important topical terms which are valuable for the search. Generally the summary length has to be determined empirically and can be only selected and learnt if one has a training set, as we had for this task.

- Transliteration of named entities appears to be useful for English Hindi cross language search, with improvements for both training and test queries as shown in Table 5.1 and Table 5.2. Hindi target documents had words which were the translated and transliterated form of input queries. However, automatic transliteration may be incorrect leading to failures to match.
- In general, we found that using PRF has a positive effect if the performance of initial query issued to the system is low. Determining the optimum values for the number of documents and words for query expansion is a complex problem which can be empirically learnt using the training set.
- Our results indicated that using data fusion improves cross language search effectiveness by combining multiple types of information together. Data fusion helped to effectively capture and combine different signals learnt from different models explored in this work.

For more details we direct readers to Arora et al. (2013a,b) which describes our system submission paper and our post-submission extended work done for the CLINSS task.

Following the findings from this work that query formulation (focusing on query summarization and PRF) is an important aspect to improve retrieval effectiveness, we explore on the lines of query expansion for sentence-retrieval experiments to find potential relevant sentences from the documents to be used for generating document snippets. We explore PRF-based approach and propose new techniques of performing query expansion for our main investigation on sentence-level relevance prediction (described later in Chapter 6).

5.1.1.2 Semantic Similarity Task (STS)

TASK: The goal of the STS task was to predict how similar in meaning two sentences $S1$ and $S2$ are by calculating a similarity score between them. The similarity between two sentences was defined on a scale from 0 (no relation) to 5 (semantic equivalence). Thus, given a sentence pair, the aim was to learn a model which outputs a score between 0 and 5 reflecting the semantic similarity between the two sentences. The main challenge was to effectively capture the subtle differences at syntactic and semantic-level while comparing two sentences.

Our Approach: Our system to address this challenge exploited distributional semantics-based embeddings information in combination with tried-and-tested features from previous tasks in order to compute sentence similarity. We used the Word2Vec (W2V) representation as described in Chapter 4, to compute semantic similarity between two words. We then expanded the word-level semantic similarity to incorporate the similarity between two sentences. We combined bag-of-words-based cosine similarity, word embeddings-based sentence similarity, syntactic information (parts-of-speech and dependency relations) based sentence similarity and other features to learn a regression model. We used M5P regression algorithm (Quinlan, 1992) to predict a sentence pair semantic similarity score, which was empirically determined by exploring other regression approaches on the training dataset.

Experimental Setup: The STS task organisers provided participants with training data consisting of pairs of sentences annotated with gold-standard semantic similarity scores. The training data for the task comprised of all the corpora from three years [2012-2014] for which the STS task was conducted previously (Agirre et al., 2012, 2013, 2014). Crowdsourced similarity scores were given on a scale from 0 (no relation) to 5 (semantic equivalence). The test data was taken from five different domains: answers-forums, answers-students, belief, headlines and images. The goal was to predict the semantic similarity values as close to the gold-standard values (human annotation) as possible. The *Pearson coefficient* (described in Chapter 2) was used to measure the correlation between the predicted values and the gold value.

Test Set	**Baseline**	**Run-1**	**Run-2**	**Run-3**	**Top System**	Our Rank
Images	0.604	0.8394	0.835	**0.843**	0.871	19
Headlines	0.531	**0.828**	0.819	0.818	0.842	4
Belief	0.652	0.546	**0.755**	0.698	0.772	2
Answers-students	0.664	**0.660**	0.623	0.611	0.788	47
Answers-forum	0.445	0.556	0.563	**0.653**	0.739	30
Mean		0.720	0.734	**0.737**		26

Table 5.4: Results of our final runs compared to the baseline and the best system for each test set.

Results and Analysis: Our team submitted three runs for each of the five English test sets, the results for our run are shown in Table 5.4. For two of the test sets, belief and headlines, our best system ranked second and fourth out of the 73 submitted systems. Our best submission averaged over all test sets ranked 26 out of the 73 systems. The analysis of the submitted runs and the best features indicated that: different set of features comprising of bag-of-words-based cosine similarity, embeddings-based sentence similarity features, sentence alignment features and syntactic information-based features captured complementary signal and a regression model combining these different set of features perform quite good across all datasets. Our model failed to capture the syntactic relationship effectively between sentences where specific details and entities were being compared for e.g. “Terminals 1 and 4 are connected” and “Terminal 1 is connected to Terminal 2”, where the gold similarity is 1.4 and our system predicted high scores (4.57). Thus our system did not perform well for answer-students and answer-forums test set.

For more details we direct readers to Arora et al. (2015b) which describes our system submission paper for the STS task.

Following the findings from this investigation that a combination of multiple features comprising of bag-of-words (BOW) based cosine similarity, embeddings-based sentence similarity, syntactic-based features work effectively for semantic similarity prediction, we explore these different features for effective sentence comparison to determine sentence-level novelty information.

5.1.1.3 Question Quality prediction using Similar Questions detection (QSQD)

TASK: The aim of this investigation was to perform question classification as “good” or “bad” for StackOverflow[6], a technical community question answering forum. StackOverflow prescribes a comprehensive set of guidelines which a newly asked question should adhere to. A question is classified as “good” if a question conform to the community guidelines, or otherwise it is classified as “bad” if it is a vague, imprecise or a controversial question. This is a cold start problem, where a new question without any community feedback and review, needs to be classified using only the textual features. There are two significant challenges: the relatively short length of the questions as compared to traditional web documents, and the considerable vocabulary overlap that exist between the good and bad questions. Overall, the text of a current question may not have sufficient information to accurately classify it. To alleviate this problem of the lack of sufficient discriminative content in the questions, we propose to make use of other existing questions previously asked in the forum.

Our Approach: We investigated the usefulness of current question expansion techniques for improving question quality prediction. Our approach is somewhat similar to document expansion in information retrieval (IR), where a short document is expanded with the textual content from other documents in order to improve its informativeness and retrievability (Efron et al., 2012).

To perform question expansion we divided the investigation into two parts:

1. *Retrieval models based expansion*: We explored traditional retrieval models for question expansion. Given a question we sought to retrieve similar questions from the collection using LM, BM25 models (reviewed in Chapter 2) and BM25F model which is a variant of BM25 model that allows the flexibility to weigh different fields such as title and body of the question separately

[6] `https://stackoverflow.com/`

Method	Accuracy	F-Measure (Macro average)
Only title content	0.9707	0.503
Title + Body content	0.9735	0.503

Table 5.5: Classification effectiveness for all SO questions, negative and positive samples being 30,163 and 1,315,731, respectively using MNB classifier.

(Robertson et al., 2004). We combined the textual information from these potential similar questions to perform classification.

2. *Question embedding based expansion*: We explored the use of question embeddings for finding similar questions. For each question in the collection we learnt a question embedding using a *Paragraph Vectors* approach (reviewed in Chapter 4). Given a question we found similar questions from the collection using a cosine similarity measure. We linearly combined the similar questions vectors and used the expanded question vector for classification.

For our text-based classification experiments where we had textual features, we used a Multinomial Naive Bayes (MNB) based classifier. For document embedded vector-based experiments, where we had real valued features, we used a Support Vector Machines (SVM) based classifier.

Experimental Setup: We used StackOverflow data dump (released in 2014) from StackExchange platform[7] for our experiments. In total, the number of questions in the collection that were indexed was about 1.35M. Due to the strong class frequency imbalance (1.31M good questions and 31k bad questions), using the whole dataset for training did not produce satisfactory classification effectiveness as shown in Table 5.5. Despite giving high accuracy, the average F-score was close to that of random classification due to the strong class imbalance. Thus we randomly selected 1,000 questions from each class (good and bad) to conduct our experiments. We evaluated classification performance using accuracy and F-score measure. We used Lucene library for IR experiments for indexing and retrieval of questions.

[7] `https://archive.org/details/stackexchange`

k	Neighbourhood	Accuracy	F-Measure
0	N/A	0.713	0.704
3	LM	0.729	0.720
3	BM25	0.719	0.713
3	BM25F	**0.738**	**0.733**

Table 5.6: Multinomial Naive Bayes classification using retrieval models for question expansion, *k* is the number of question used for expansion which were empirically determined.

k	Accuracy	F-Measure
0	0.743	0.743
1	0.740	0.739
3	0.747	0.746
5	0.750	0.749
9	**0.769**	**0.768**
11	0.765	0.764

Table 5.7: SVM classification after performing question embedding based expansion, *k* is the number of question used for expansion which were empirically determined.

Results and Analysis: The best question classification results using different retrieval models based question expansion were obtained by BM25F model. Using BM25F model to find similar questions and performing classification using multinomial naive bayes had an accuracy of 74% and a F-score of 0.73 as shown in Table 5.6. The best results using question embeddings to find similar question and performing classification using support vector machines had an accuracy of 77% and F-score of 0.77 as shown in Table 5.7.

Our analysis revealed that performing expansion for initial questions perform considerably better than using only raw features from the question itself, which in some cases might not be informative and formulated well enough. Consistent trends in improvements of classification results were observed with the expansion applied on both text and question vector embedding.

For more details on our investigation on question quality prediction using similar question detection we direct readers to Arora et al. (2015a, 2016).

Following the experiments using question-based embeddings for similar question detection effectively, we learn sentence-based embeddings for finding novel information in our main investigation (described later in Chapter 7). Further we explore word embeddings-based techniques for effective query expansion for sentence-level relevance prediction, for finding informative sentences to be used for snippet generation.

5.1.2 User centric investigations

We conducted two user centric investigations to study user interaction with snippets and documents in a search task. We sought to determine the fundamental units in a document which are deemed important and useful to satisfying a given information need. Further, we investigate how user knowledge gain varies as users interact with documents snippets in a SERP.

5.1.2.1 Identifying Useful and Important Information within Retrieved Documents

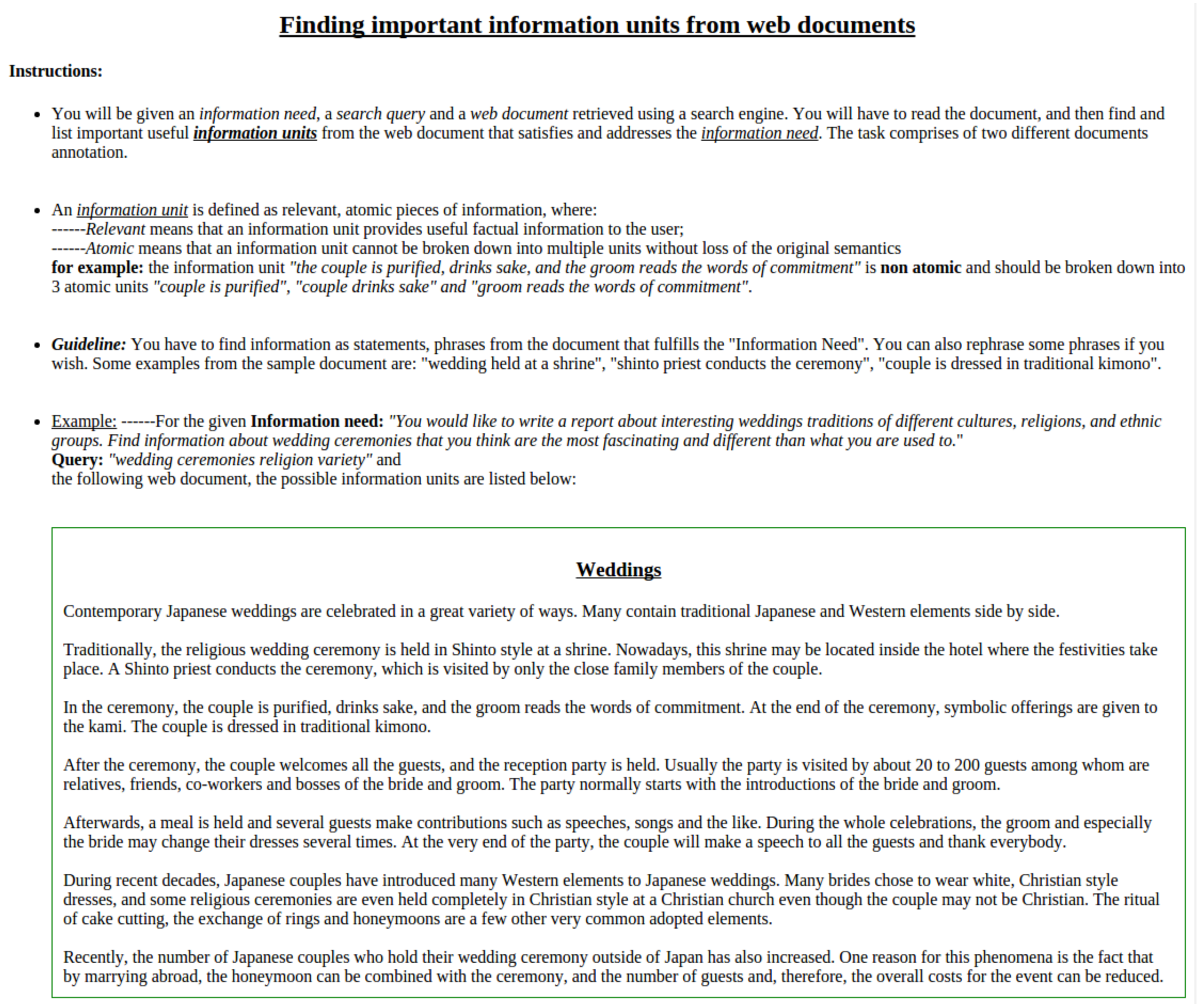
Investigation: We performed an initial study into the identification of important and useful information units within documents retrieved by an information retrieval system, in response to a user query created in response to an underlying information need. We anticipate that understanding what constitute as important and useful textual information can help us to generate effective snippets. We conducted three user studies using a crowdsourcing platform. Participants were first asked to read an *information need* and *contents of a relevant document* and then to perform actions depending on the type of study:

- **Write important information units (WIIU)**: In this study, participants were shown the textual content from a web document and were asked to find textual information units which seemed useful and important to them with respect to a given information need. First they were shown the instructions and then presented with the content of the document. They had to write information units in the text box provided in the interface. Copying and pasting directly from the text document was disabled, to allow participants to read, and engage with the content and then write important and useful units in the space provided. A snapshot of the interface is shown in Figure 5.1.

- **Highlight important information units (HIIU)**: In this study, similar to WIIU study, participants were shown the textual content from a web document. Instead of writing information units participants were asked to highlight textual information units which seemed useful and important to them with respect to a given information need. A snapshot of a sample output (after the completion of highlighting information units) is shown in Figure 5.2.
- **Assess importance of already highlighted information units (AI-HIU)**: In this study, participants were presented with already highlighted information units in a document as shown in Figure 5.2. The annotation of the highlighted information was done by the author, following the definition and guidelines of information units from NTCIR benchmark campaign (Kato et al., 2014). Participants were asked to rate each information units using 4 classes of relevance and importance: i) C1: Highly relevant and important, ii) C2: Fairly relevant and important, iii) C3: Slightly relevant and important, and iv) C4: Neither relevant nor important. They were asked to provide reason for each annotation to capture user's perception of what deemed useful to them.

Our studies focused on the following specific research questions:

- **RQ-1**: Are there consensus for information units between users for the WIIU and HIIU studies?
- **RQ-2**: Can we compare and measure information units identified by the users? (in the WIIU and HIIU studies)



Finding important information units from web documents

Instructions:

- You will be given an *information need*, a *search query* and a *web document* retrieved using a search engine. You will have to read the document, and then find and list important useful ***information units*** from the web document that satisfies and addresses the *information need*. The task comprises of two different documents annotation.

- An *information unit* is defined as relevant, atomic pieces of information, where:
------*Relevant* means that an information unit provides useful factual information to the user;
------*Atomic* means that an information unit cannot be broken down into multiple units without loss of the original semantics
for example: the information unit *"the couple is purified, drinks sake, and the groom reads the words of commitment"* is **non atomic** and should be broken down into 3 atomic units *"couple is purified", "couple drinks sake" and "groom reads the words of commitment".*

- ***Guideline:*** You have to find information as statements, phrases from the document that fulfills the "Information Need". You can also rephrase some phrases if you wish. Some examples from the sample document are: "wedding held at a shrine", "shinto priest conducts the ceremony", "couple is dressed in traditional kimono".

- Example: ------For the given **Information need:** *"You would like to write a report about interesting weddings traditions of different cultures, religions, and ethnic groups. Find information about wedding ceremonies that you think are the most fascinating and different than what you are used to."*
Query: *"wedding ceremonies religion variety"* and
the following web document, the possible information units are listed below:

Weddings

Contemporary Japanese weddings are celebrated in a great variety of ways. Many contain traditional Japanese and Western elements side by side.

Traditionally, the religious wedding ceremony is held in Shinto style at a shrine. Nowadays, this shrine may be located inside the hotel where the festivities take place. A Shinto priest conducts the ceremony, which is visited by only the close family members of the couple.

In the ceremony, the couple is purified, drinks sake, and the groom reads the words of commitment. At the end of the ceremony, symbolic offerings are given to the kami. The couple is dressed in traditional kimono.

After the ceremony, the couple welcomes all the guests, and the reception party is held. Usually the party is visited by about 20 to 200 guests among whom are relatives, friends, co-workers and bosses of the bride and groom. The party normally starts with the introductions of the bride and groom.

Afterwards, a meal is held and several guests make contributions such as speeches, songs and the like. During the whole celebrations, the groom and especially the bride may change their dresses several times. At the very end of the party, the couple will make a speech to all the guests and thank everybody.

During recent decades, Japanese couples have introduced many Western elements to Japanese weddings. Many brides chose to wear white, Christian style dresses, and some religious ceremonies are even held completely in Christian style at a Christian church even though the couple may not be Christian. The ritual of cake cutting, the exchange of rings and honeymoons are a few other very common adopted elements.

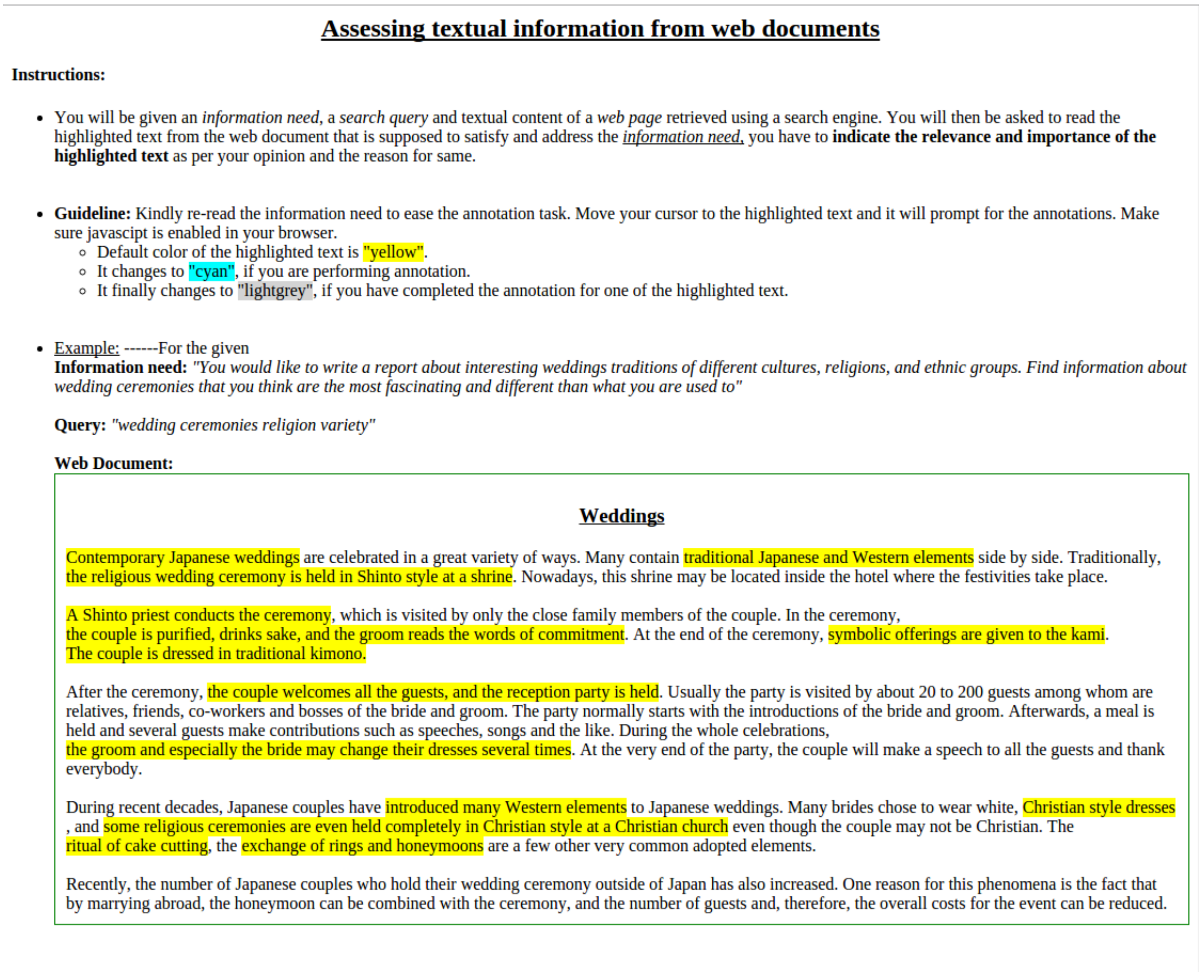
Recently, the number of Japanese couples who hold their wedding ceremony outside of Japan has also increased. One reason for this phenomena is the fact that by marrying abroad, the honeymoon can be combined with the ceremony, and the number of guests and, therefore, the overall costs for the event can be reduced.

Figure 5.1: Finding information units

- **RQ-3**: What is the agreement among users while assessing information units already marked in a document? (for the AIHIU study)

Experimental setup: We used data from the TREC 2012 session track for our study (Kanoulas et al., 2012). We selected 3 information needs (Wedding Traditions, Smoking Cessation, and Junk Food) for this dataset and at random one relevant document from the *qrels* for each of the three information needs.

Since this is a cognitively intensive task for our participants, we opted to concentrate on detailed analysis of a small number of documents for this initial study, with the main goal of analysing important and useful richer units within a document. After conducting a pilot run with 5 volunteers we carried the annotations using the



Assessing textual information from web documents

Instructions:

- You will be given an *information need*, a *search query* and textual content of a *web page* retrieved using a search engine. You will then be asked to read the highlighted text from the web document that is supposed to satisfy and address the information need, you have to **indicate the relevance and importance of the highlighted text** as per your opinion and the reason for same.

- **Guideline:** Kindly re-read the information need to ease the annotation task. Move your cursor to the highlighted text and it will prompt for the annotations. Make sure javascript is enabled in your browser.
  - Default color of the highlighted text is "yellow".
  - It changes to "cyan", if you are performing annotation.
  - It finally changes to "lightgrey", if you have completed the annotation for one of the highlighted text.

- Example: ------For the given
  Information need: *"You would like to write a report about interesting weddings traditions of different cultures, religions, and ethnic groups. Find information about wedding ceremonies that you think are the most fascinating and different than what you are used to"*

  Query: *"wedding ceremonies religion variety"*

  Web Document:

Weddings

Contemporary Japanese weddings are celebrated in a great variety of ways. Many contain traditional Japanese and Western elements side by side. Traditionally, the religious wedding ceremony is held in Shinto style at a shrine. Nowadays, this shrine may be located inside the hotel where the festivities take place.

A Shinto priest conducts the ceremony, which is visited by only the close family members of the couple. In the ceremony, the couple is purified, drinks sake, and the groom reads the words of commitment. At the end of the ceremony, symbolic offerings are given to the kami. The couple is dressed in traditional kimono.

After the ceremony, the couple welcomes all the guests, and the reception party is held. Usually the party is visited by about 20 to 200 guests among whom are relatives, friends, co-workers and bosses of the bride and groom. The party normally starts with the introductions of the bride and groom. Afterwards, a meal is held and several guests make contributions such as speeches, songs and the like. During the whole celebrations, the groom and especially the bride may change their dresses several times. At the very end of the party, the couple will make a speech to all the guests and thank everybody.

During recent decades, Japanese couples have introduced many Western elements to Japanese weddings. Many brides chose to wear white, Christian style dresses , and some religious ceremonies are even held completely in Christian style at a Christian church even though the couple may not be Christian. The ritual of cake cutting, the exchange of rings and honeymoons are a few other very common adopted elements.

Recently, the number of Japanese couples who hold their wedding ceremony outside of Japan has also increased. One reason for this phenomena is the fact that by marrying abroad, the honeymoon can be combined with the ceremony, and the number of guests and, therefore, the overall costs for the event can be reduced.

Figure 5.2: Assessing highlighted information units

Prolific crowdsourcing platform[8].

Results and Analysis: For each study we had seven participants who read documents and label useful and important information (in the WIIU and HIIU studies) and rate already identified and highlighted textual units (for the AIHIU study).

We carried out data analysis for the information units collected for the WIIU and HIIU studies. While writing information units in the WIIU study, participants rephrased the textual information, and in some cases summarised the text and information units in their own words, different from the content words expressed in the documents. While highlighting information units in the HIIU study, participants freely highlighted the textual content i.e the starting and ending points of the highlighted text varied considerably across users. Thus calculating normal agreement

[8] https://www.prolific.ac/

between users was quite challenging. Hence we calculated word overlap and cosine similarity between the information units and the original document to analyse the user's responses. This gave us a rough measure of the consensus between participant annotations. We found average cosine similarity of 0.50 and 0.57 between participant annotations and documents in the WIIU and HIIU studies respectively. In the WIIU and HIIU studies participants reported that it is quite challenging to identify important and useful text when they read documents. Some people encounter new information and find everything useful and important, whereas others who know about the topic can be too critical when judging the importance and usefulness of textual information without strict guidelines.

For AIHIU study, we had 47 textual units categorised by 7 annotators at 4 levels of relevance and importance. We found majority agreement of about 0.489 and pairwise agreement of 0.340 among users annotation in the AIHIU study. Overall, the results and analysis indicate that it would be more practical to work with fixed boundary units such as sentence-level rather than free annotation of textual units for finding useful and important information within documents.

For more details on this investigation we direct readers to Arora and Jones (2017a,b).

Following the findings in this investigation, we work at sentence-level for relevance, novelty and readability prediction for generation of effective snippets.

5.1.2.2 Measuring user knowledge gain

Investigation: We performed a pilot experiment examining learning behaviour when users interact with web documents presented as snippets in a SERP for a given topic. We evaluated gain in user knowledge through measures of self-assessment and reporting, and analysis of pre- and post-study summaries (measuring changes in user concepts and key terms).

This investigation was designed to study the following research question:

- **RQ:** How does user's search experience and knowledge gain vary when they interact with document snippets presented in a SERP in a task-based setting?

Variable	Id	Question	Scale	Source
Topic familiarity	**PR-1**	How familiar are you with this topic?	[0-4], where 4=Very familiar... 0=Not familiar at all	Pre-task
Perceived knowledge	**PR-2**	How will you rate your knowledge on this topic?	[0-4], where 4=Expert... 0=New to the topic	Pre-task
Knowledge summary	**PR-3**	Write a summary about the topic in 5-10 sentences in terms of what you know	N/A	Pre-task
Finding information	**PO-1**	How difficult was it to find information relevant to the topic from the documents?	[0-4], where 4=Very difficult.. 0=Not difficult at all	Post-task
Understanding content	**PO-2**	How difficult was it to understand the content of the documents ?	[0-4], where 4=Very difficult.. 0=Not difficult at all	Post-task
Informativeness	**PO-3**	How informative were the documents with respect to the given topic?	[0-4], where 4=Very informative... 0=Not informative at all	Post-task
Document complexity	**PO-4**	How complex was the language of the documents? language of the documents?	N/A 4=Very complex... 0=Not complex at all	Post-task
Topic familiarity	**PO-5**	How familiar are you with this topic?	[0-4], where 4=Very familiar... 0=Not familiar at all	Post-task
Perceived knowledge	**PO-6**	How will you rate your knowledge on this topic?	[0-4], where 4=Expert 0=New to the topic	Post-task
Perceived learning	**PO-7**	How much do you think you learnt on this topic?	[0-4], where 4=Quite a lot 0=Nothing at all	Post-task
Knowledge summary	**PO-8**	Write a summary about the topic in 5-10 sentences in terms of what you learnt, after interacting with documents.	N/A	Post-task
Task feedback	**PO-9**	Feedback on the whole exercise and if faced any difficulty?	N/A	Post-task

Table 5.8: User study experimental design pre- and post-task questionnaire

Experimental setup: The experiment had three stages as described below:

- Pre-task stage (initial knowledge assessment): This was the first stage of the study, where we seek to assess the prior knowledge of the user on the topic of inquiry. Participants were asked three questions **PR-[1-3]** as shown in Table 5.8.

- Main task stage (interaction with documents): In this stage user's were given a search task statement and were shown snippets of 10 relevant documents as in a general web setting and were asked to read the documents and gather information to accomplish the task of writing a report.

- Post-task stage (after task knowledge assessment): After the main study task users were given a questionnaire where they were asked nine questions **P0-[1-9]** as shown in Table 5.8, to assess their experience and post-task knowledge.

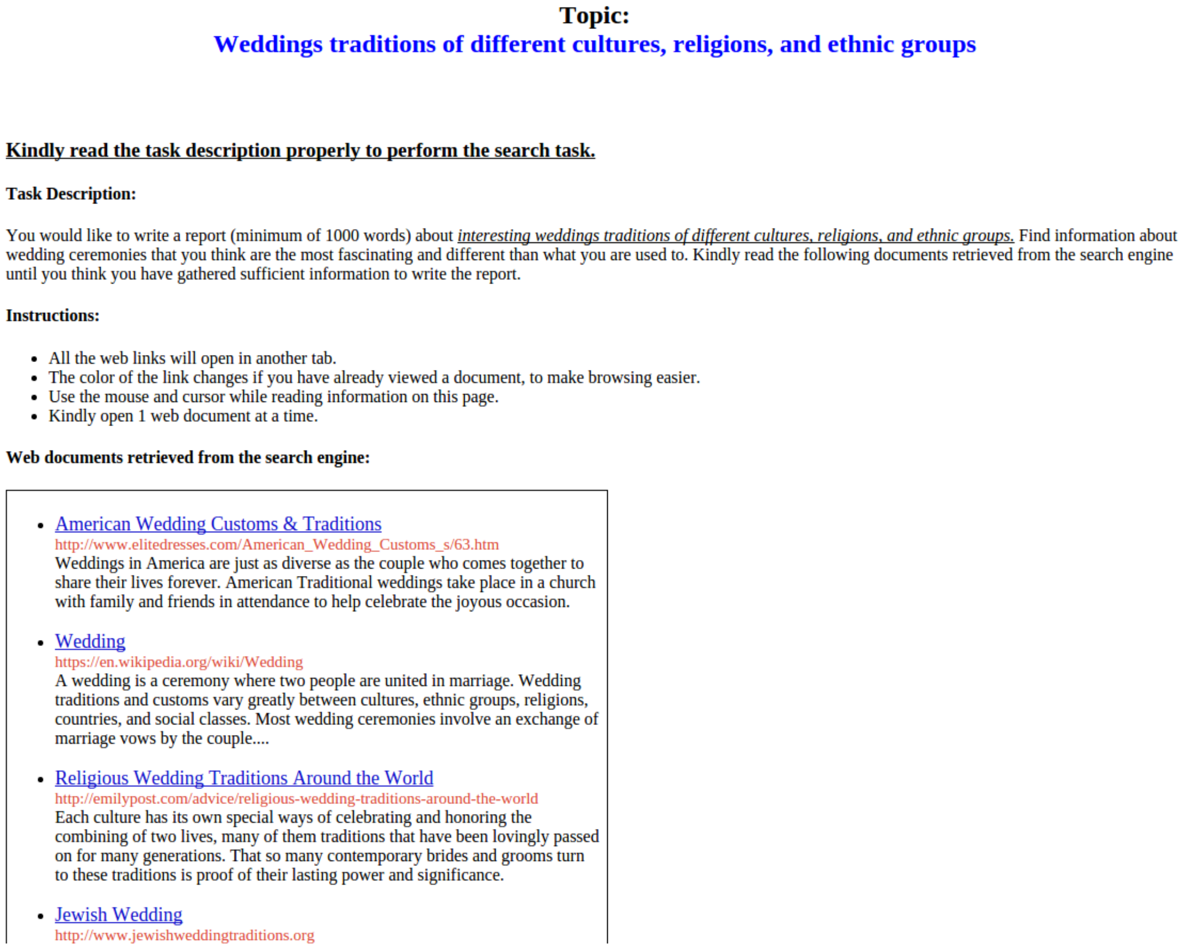
This type of study is quite costly time wise as completing the 3 stages it takes about 25-30 minutes to complete, so we focused only on the topic *Wedding Traditions* in this pilot investigation. We selected 10 relevant documents from the test collection to ensure we capture diversity of results. The 10 documents shown to the users belonged to different topics: i) American weddings, ii) German weddings, iii) Jewish weddings, iv) Religious weddings, v) Wedding in general (wikipedia-article), vi) Religious-Indian (Hindu) weddings, vii) Arab weddings, viii) Bizzare wedding traditions, ix) Bai-ethnic group (chinese wedding) and x) Custom and traditions around globe: Italian, Mexican and Sweden weddings.

The snippets were made using the title, url and document summary consisting of top 2-3 sentences (about 250 characters) from the documents (static summary approach as described in Chapter 1 and Chapter 2). We adopted the same colour coding mechanism as used by Google[9] and the colour of the snippets changed if a user had already clicked a document to mimic the normal search system behaviour. A snapshot of the document interaction stage is shown in Figure 5.3.

We conduced a remote-study which was hosted on our server and shared through an url with the participants. We had 9 people who participated in the study, but 3 people did not complete all the three stages, thus we had data from six participants for analysis. All the participants were native English speakers.

Results and Analysis: We assessed the *pre-* and *post*-study summaries and analysed them based on the topics being covered (key terms and concepts related to different wedding traditions). Further, we analysed user's pre- and post-test sum-

[9] `www.google.com`



Topic:
Weddings traditions of different cultures, religions, and ethnic groups

Kindly read the task description properly to perform the search task.

Task Description:

You would like to write a report (minimum of 1000 words) about *interesting weddings traditions of different cultures, religions, and ethnic groups.* Find information about wedding ceremonies that you think are the most fascinating and different than what you are used to. Kindly read the following documents retrieved from the search engine until you think you have gathered sufficient information to write the report.

Instructions:

- All the web links will open in another tab.
- The color of the link changes if you have already viewed a document, to make browsing easier.
- Use the mouse and cursor while reading information on this page.
- Kindly open 1 web document at a time.

Web documents retrieved from the search engine:

- American Wedding Customs & Traditions
  http://www.elitedresses.com/American_Wedding_Customs_s/63.htm
  Weddings in America are just as diverse as the couple who comes together to share their lives forever. American Traditional weddings take place in a church with family and friends in attendance to help celebrate the joyous occasion.
- Wedding
  https://en.wikipedia.org/wiki/Wedding
  A wedding is a ceremony where two people are united in marriage. Wedding traditions and customs vary greatly between cultures, ethnic groups, religions, countries, and social classes. Most wedding ceremonies involve an exchange of marriage vows by the couple....
- Religious Wedding Traditions Around the World
  http://emilypost.com/advice/religious-wedding-traditions-around-the-world
  Each culture has its own special ways of celebrating and honoring the combining of two lives, many of them traditions that have been lovingly passed on for many generations. That so many contemporary brides and grooms turn to these traditions is proof of their lasting power and significance.
- Jewish Wedding
  http://www.jewishweddingtraditions.org

Figure 5.3: Documents interaction stage

maries based on the count of overlap of topics and the length of the summary (counting the number of sentences).

Study	PR-1	PR-2	PO-5	PO-6	PO-7
User-1	2	1	1	3	3
User-2	2	1	1	1	3
User-3	1	1	1	3	3
User-4	2	3	2	3	3
User-5	1	1	2	2	3
User-6	2	3	2	3	3

Table 5.9: Questionnaire Data Results, PR and PO are pre-test and post-test questions asked in the questionnaire defined in Table 5.8

Tables 5.9, 5.10 and 5.11 show results of user experience, changes in topics covered in pre- and post-task summaries and analysis of pre- and post-task summaries. It was difficult to make definitive conclusion from such a small study. However, we were able to make some interesting observations:

Users	PSS Topics	POS Topics
User-1	Religious, Christian, Indian, Other	Arabic, German, Jewish, American
User-2	Indian, African, Chinese, German	Jewish, Indian, German, Chinese, American
User-3	Irish, Middle-eastern, Judaism	Opinion about different culture marriages
User-4	Western, Japanese, Korean, African	Jewish, Muslim, Chinese, Hindu
User-5	Wedding in general	Arabic, German, American, Jewish
User-6	Wedding in general, Bulgarian wedding	Arabic, Italian, American, German, Hindu

Table 5.10: Data Coding Results, where PSS Topics: topics covered in Pre-Study summary, POS Topics: topics covered in Post-Study summary

Users	PSS len (sen)	POS len (sen)	PSS-DOC ovp	POS-DOC ovp
User-1	8	7	2	4
User-2	5	10	3	5
User-3	3	3	0	0
User-4	5	5	1	4
User-5	2	7	1	4
User-6	8	7	1	5

Table 5.11: Data Coding Results, PSS len: Pre-Study summary length (no. of sentences), POS len: Post-Study summary length, PSS-DOC ovp: no of topics overlap between the Pre-study summary and gold document set i.e. 10, POS-DOC ovp: no. of topics overlap between Post-study summary and gold document set i.e 10

- Change in topical knowledge and learning can be measured in terms of changes in rating from pre- and post-task stage, comparing (PR-2, PO-6, PO-7) as shown in Table 5.9. Stronger signals can be measured in terms of concepts captured in the summaries written before and after the main task as shown in Tables 5.10 and 5.11.

- The difference in the perceived knowledge (PO-6 & PR-2) on the topic increases or remains the same from pre to post-task stage for all users. Although if the user's prior topical knowledge is high, then it is less evident what improvement is made based on ratings. All the users reported that they learn *Fairly*, about different aspects but at shallow level as shown in Table 5.9.

- The topics evolve and get more focused and factual rather than being quite abstract and general in nature from pre- to post-task stage (PR-3 & PO-8) respectively as shown in Table 5.10.

- The familiarity level with the topic decreases or remains the same from pre- to post-task stage (PR-1 & PO-5) for all except one user, which can be explained as people tend to overestimate their familiarity with a topic before encountering the actual information as shown in Table 5.9.

- Snippets play an important part in deciding whether to view a document or not. In the feedback, users stated that they were reading snippets to look for wedding traditions and cultures which were different from what they knew or had read before.

This study helped us to learn: how to design, set up and conduct user studies for measuring search behaviour, and knowledge gain in a user study setting. Following this investigation we follow similar design mechanism for our study on measuring snippets utility, when presented in a SERP (described later in Chapter 9).

5.2 Summary and Conclusions

In this section we summarise the conclusions of our preliminary investigations, and then present an overview of the main investigations to be undertaken in this thesis based on these findings.

5.2.1 Main findings and lessons learnt

To understand the challenges and experiment with the tools and technologies to develop effective snippets and measure user behaviour and interaction effectively, we conducted initial investigations as described in Section 5.1. From these investigation we can make the following conclusions:

- **Query formulation and expansion:** How to correctly find and semantically match words expressed in a query with a document is a complex problem of vocabulary mismatch that needs to be addressed in order to improve retrieval effectiveness. PRF-based approach explored in CLINSS task and question-based embeddings for investigation on question classification using question expansion exhibit that traditional PRF-based approach and embeddings-based techniques can be potentially exploited for improving retrieval effectiveness.

- **Sentence Similarity:** In the STS task we found that for computing the similarity of a sentence pair, a combined representation using a raw bag-of-words-based cosine similarity feature, word embedding-based features and syntactic information-based features form a robust model that tends to perform well in general across different datasets. Thus these features can be explored and investigated to compare sentences effectively to find sentence-level novel information across and within documents.

- **Textual units granularity:** For finding relevant and useful parts of the documents that satisfy user information needs, it is effective to work at sentence-level instead of phrase-level as basic meaningful units.

- **Interactions with text:** The task of highlighting textual units shows more inconsistencies in annotation (overlap across users) as compared to writing textual units as a summary of the important information. We speculate that users are cognitively more engaged when they have to write and summarise points than when they are freely highlighting text. In general, people find it easier to rate textual units rather than writing or highlighting useful and important parts.

- **User Knowledge Gain:** User knowledge gain measured by changes in user's pre- and post-task ratings and comparing pre- and post-task topic summaries vary quite a lot across users. From the small scale of our study it is hard to make conclusions but our results shows that providing information in a richer way can help users to learn about a topic, and improve their search performance.

- **Navigation in a SERP:** As reported by two participants: snippets played an important part in deciding whether to view a document or not. Snippets assisted them to find documents which contained topics which were different from what they knew or had read before.

5.2.2 Overview of our main investigations

Following our initial investigation and the lessons learnt, next we turn our attention to an overview of our main thesis investigations to address the questions introduced in Chapter 1.

1. *Sentence-level relevance prediction*: The main focus is to find **relevant information** from the documents for generating effective snippets, discussed in detail in Chapter 6. We explore how to address the issues of vocabulary mismatch problem. We investigate different traditional retrieval models, propose and experiment approaches which performs query expansion using word embeddings, and combine traditional PRF-based query expansion technique with embeddings-based expansion techniques.

2. *Sentence-level novelty prediction*: The main focus is to find **novel information** for generating effective snippets, discussed in detail in Chapter 7. We perform sentence comparison across and within documents to find new and novel information to be shown to the users. We investigate bag-of-words-based different distance metrics approach, embedding and syntactic information-based sentence comparison approaches for novelty prediction.

3. *Snippet Generation framework*: The main focus is to combine different signals of *relevance*, *novelty* and *readability* for **development of effective snippets**, discussed in detail in Chapter 8. We build a framework which combines relevance, novelty and readability information to form effective snippets to be shown to the users. We study the variations in the quality of snippets gener-

ated by varying the proportion of relevance, novelty and readability information.

4. *User studies to measure effectiveness of snippets generated by our framework*: The main focus is to study and analyse how effective are the snippets generated by our framework as compared to *BM25* relevance model-based snippets when presented in a SERP, discussed in detail in Chapter 9. We evaluate snippets in a task-based setting and investigate how user search behaviour, experience and knowledge gain varies when snippets of different quality are presented in a SERP.

We begin these investigations in the next chapter where we describe our main investigation on sentence-level relevance prediction.

Chapter 6

Sentence-level Relevance Prediction

In this chapter we address the main question: *How to extract topically relevant information from the documents for a given information need expressed as a query to generate snippets which are informative and useful.* We present our investigation on sentence-level retrieval. First we revisit the definition of relevance for our work and then present the challenges in sentence-retrieval. Next, we describe the baseline retrieval models, query expansion techniques applied which seek to improve retrieval effectiveness and a novel method for query expansion using *embeddings* for detecting relevance at sentence-level. Further, we discuss the measures used for evaluating our retrieval models. We then present the experimental results of different approaches investigated in this work and conclude with a detailed analysis on the effectiveness of our sentence-level retrieval work.

6.1 Introduction

We focus on relevance-based sentence selection to build better document snippets. We are interested in capturing the topical aspect of relevance as discussed in Chapter 2, where "relevance" measure the extent to which an information (document or sentence) is related to the information need expressed as a query. For sentence-level retrieval, we retrieve the top sentences for a given query from each document to effectively represent the contents of the documents in relation to the user's information need. Similar work has been done on sentence-level relevance prediction by Habernal et al. (2016); Yulianti et al. (2016); Leal Bando et al. (2015); Losada (2010); Allan et al. (2003) etc.

6.1.1 The Main Challenges of Sentence Retrieval

As described in Chapter 2, sentence-level relevance prediction is more challenging problem as compared to document-level relevance prediction as sentences are typically shorter than the documents, thus there is less textual information to work with to decide whether a sentence is relevant to a user query or not. Next, we describe the main challenges while performing sentence-level relevance prediction:

- **Vocabulary mismatch** – As reviewed in Chapter 2, users do not write queries effectively and often do not use effective words to describe their information need (Salton et al., 1975; Losada, 2010; Leal Bando et al., 2015). Words expressed as a query to describe an information need or a concept differ as compared to their usage in the documents. Some examples of how words are used in a query and their usage in the relevant sentences: *global warming – climate change, heating, greenhouse effect; bombings – explosion, blast etc.* It is a complex problem to handle the vocabulary mismatch between a user's query and content of a document and we aim to address this in our work.

- **Interpreting user's query to retrieve relevant information** – A user's query is a manifestation of their information need. Thus it can be hard to interpret exactly the words expressed by the individual user as the query words depend on user's topical knowledge and can vary a lot in terms of the word usage, and the nature of the user's past search experience. Sometimes query words can be too general so it becomes essential to interpret it correctly given the context. For example in the query *Driving Cell Phone Usage*: The word *usage* has been used in general, where the actual relevant context or the query is more around *benefits, dangers and safety hazards.* So with conventional approaches to query sentence word overlap and matching, it is really hard to capture those sentences that discuss about benefits, dangers and safety hazards while driving which we aim to address in our work.

- **Effective incorporation of contextual and semantic information** – Traditional retrieval models perform exact word matching of the terms in a query and a sentence. Both at the query and sentence-level it is very challenging to incorporate the meaning of words while building relevance models that can maximise the matching of the query with a sentence. Following are some examples of query terms and their semantically related words: i) kenya – nairobi, east africa; ii) global warming – climate change, heating, greenhouse effect; iii) bombings – air strikes, missile strikes, explosion, blast. A model which can capture the words which are semantically similar, thus conveying the same or similar information would be really useful to retrieve relevant information effectively which we investigate in this work.

6.1.2 Research Questions

To answer the research question: *Can we develop effective models to address the vocabulary mismatch issues for sentence-level relevance prediction?*, as described in Chapter 1, we divide our investigation into following sub-questions:

1) *How do different traditional information retrieval algorithms and approaches perform for retrieving relevant sentences for a given query?*

We compare different traditional approaches proposed in previous research for retrieving sentences that satisfy the user's information need effectively. We explore LM, BM25 and VSM-based semantic similarity approaches for sentence retrieval. We conduct a comparative evaluation of different methods in terms of their performance in retrieving relevant sentences.

2) *Can we exploit query expansion techniques to address the query-sentence vocabulary mismatch problem?*

We explore query expansion techniques to address the main challenges of query-sentence vocabulary mismatch. We explore pseudo relevance feedback (PRF) approach and propose three query expansion techniques based on word embeddings approach to capture semantic information.

3) *Can we leverage semantic information effectively to improve sentence retrieval effectiveness?*

We investigate different types of embeddings to incorporate semantic similarity into relevance estimation. We learn different embeddings using in-domain data and also explore general-domain embeddings for our work. We analyse which kind of embeddings work effectively for our task of sentence retrieval.

4) *Can we effectively combine traditional query expansion methods and semantic information-based query expansion methods to improve relevance estimation for sentence-level retrieval?*

We explore methods to combine PRF-based query expansion approach and our proposed semantic-based embeddings approach for query expansion. We investigate a linear interpolation technique to combine expansion terms learnt from PRF and semantic-based query expansion.

6.2 Methodology

In this section, we discuss our experiments that are designed to investigate the research questions discussed in Section 6.1.2.

6.2.1 Baseline Models

Below we describe traditional models (reviewed in Chapter 2) which we investigate in this work for sentence retrieval.

Traditional Retrieval Models: We explore the task of sentence retrieval using *BM25* and *LM* retrieval models as our baseline models.

BM25 model: The BM25 model developed by Robertson et al. (1995) is a probabilistic model that assigns a probability score to each document indicating its relevance to a given query. We use the BM25 retrieval model for the task of sentence retrieval. We calculate sentence score for a given query q using the BM25 model as shown in Equation 6.1.

$$score(s, q) = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i, s).(k_1 + 1)}{f(q_i, s) + k_1.(1 - b + b.\frac{|s|}{avgsl)})} \tag{6.1}$$

In Equation 6.1, f(q_i, s) is the term frequency of q_i in the sentence s, $|s|$ is the length of the sentence s in words, and $avgsl$ is the average sentence length in the text collection from which sentences are drawn, k_1 and b are free parameters to weight term frequency and normalise sentence length variations, and IDF(q_i) is represented in Equation 6.2.

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{6.2}$$

In Equation 6.2, N is the total number of sentences in the collection, and n(q_i) is the number of sentences containing term q_i.

Language Model: We use the LM retrieval model where instead of document (as reviewed in Chapter 2), we work at sentence-level, where we calculate probability of a sentence being generated given a query q, which is represented as $P(s_j|q)$ as shown in Equation 6.3. Using the Bayes theorem, the prior probability of $P(s_j|q)$ can be calculated using the likelihood model, and is reduced to $P(q|s_j)$, as $P(q)$ is constant and sentences are assumed to come from a uniform distribution thus we can ignore $P(s_j)$.

To avoid the problem of zero probability we perform Jelinek-Mercer smoothing and learn weight distributions of a word by combining the query term occurrence in a sentence and collection as shown in Equation 6.4. The main intuition behind using the collection count of a term is to assign a non-zero probability to the unseen words and improve the accuracy of word probability estimation.

$$P(s_j/q) = \frac{P(q/s_j) * P(s_j)}{P(q)} \cong P(q/s_j) \tag{6.3}$$

$$P(q/s_j) = (1-\lambda) * p(q/s_j) + \lambda * p(q/C) \tag{6.4}$$

The BM25 and LM models have shown to exhibit high retrieval effectiveness over a wide range of search applications (Croft et al., 2010) and are quite widely used for IR experiments, so we investigate them in our initial experiments. We explore LM and BM25 retrieval models to effectively compare our results of relevance prediction to select the best model for further experiments and as a baseline for our investigation.

6.2.2 Embedding-based Semantic Similarity (ESS)

Motivated by the VSM model (reviewed in Chapter 2), we experiment with an approach to measure the effectiveness of matching vector representation of words in a query and sentence for relevance prediction. In this approach each word w_i in the query and sentence is represented as a dense vector of p dimensions (embedding) learnt using a neural representation as shown in Equation 6.5 and discussed earlier in Chapter 4. For a given query (q) and sentence (s), we represent all words using their vector representation (embeddings representation) and combine the word vectors to form a query vector and a sentence vector as shown in Equation 6.6 and 6.7. We use cosine similarity to calculate similarity scores between the query and sentences within a document as shown in Equation 6.8. We study how effective is the matching of embeddings-based vector representation of query and sentence for sentence-level retrieval in comparison to the BM25 and LM approaches.

$$\overrightarrow{w_i} = [v_i^1, v_i^2, ..., v_i^p] \tag{6.5}$$

$$\overrightarrow{q_i} = \overrightarrow{w_i^1} + \overrightarrow{w_i^2} + ... + \overrightarrow{w_i^n} \tag{6.6}$$

$$\overrightarrow{s_j} = \overrightarrow{w_j^1} + \overrightarrow{w_j^2} + ... + \overrightarrow{w_j^m} \tag{6.7}$$

$$similarity(q_i, s_j) = cosine(\overrightarrow{q_i}, \overrightarrow{s_j}) \tag{6.8}$$

In the above Section 6.2.1 and 6.2.2, we describe the LM, BM25, and ESS models for the task of sentence retrieval. But as discussed in Section 6.1.1, a query might not be a good representation of user's information need due to lack of background knowledge and user's understanding of the task. Thus we perform query expansion to minimise the vocabulary mismatch problem between a query and a sentence.

6.2.3 Query Expansion

To find potential terms that can be used to effectively represent users query, we investigate the technique of query expansion to address the vocabulary mismatch problem between query and sentences. We explore three different query expansion approaches in this work: i) pseudo relevance feedback (PRF), ii) our proposed semantic expansion-based approach, and iii) our proposed approach combining PRF and semantic expansion-based method.

Pseudo Relevance Feedback (PRF): In PRF, initial query is refined using top ranked documents (reviewed in Chapter 2). In this approach the terms from top retrieved documents which can act as good clues or representative terms to capture a user's query intent are selected to expand the initial query to boost the retrieval effectiveness. We use PRF techniques to expand the input queries using Robertson selection value i.e. rsv scores to select the terms for expansion as shown in Equation 6.9 and 6.10.

$$rsv(i) = r(i) * rw(i) \tag{6.9}$$

where $r(i)$ is number of assumed relevant sentences containing term i, and $rw(i)$ is the standard Robertson/Jones relevance weight

$$rw(i) = log\frac{(r(i)+0.5)(N-n(i)-R+r(i)+0.5)}{(n(i)-r(i)+0.5)(R-r(i)+0.5)} \tag{6.10}$$

where $n(i)$ = total number of sentences containing term i, R = total number of assumed relevant sentences for this query, N = total number of sentences in the collection.

Important challenge while performing PRF is determining the assumed potentially relevant sentences R and the number of terms to be used for query expansion, which are generally explored using a grid search for a given test collection.

Semantic expansion (Using word embeddings): Recent work on the use of embeddings for query expansion for the task of document retrieval have shown to be quite effective (Roy et al., 2016; Kuzi et al., 2016; Diaz et al., 2016; Zamani and Croft, 2016) as reviewed in Chapter 4. In this work, we propose three methods for incorporating embeddings-based expansion for the task of sentence retrieval. The main idea behind these three proposed methods is to find terms which are semantically similar to the initial query and can be used as effective expansion terms to improve the retrieval effectiveness.

We learn vector representation of all the words in the vocabulary, V. Figure 6.1 represents all word vectors in the vocabulary represented in two dimension, where the red dots represent the query word vectors for a given query Q. All non-query terms in the collection are represented as general words vector (possible candidates for expansion). Our goal is to select potential expansion terms that are similar to the query word vectors (the red dots). Next, we propose three methods to perform expansion using the initial query vectors:

1) *QueryWord approach*: In this approach, for a given query Q, instead of using all the general word vectors for expansion we filter the words and form a collection C (potential candidates for expansion), consisting of words which are

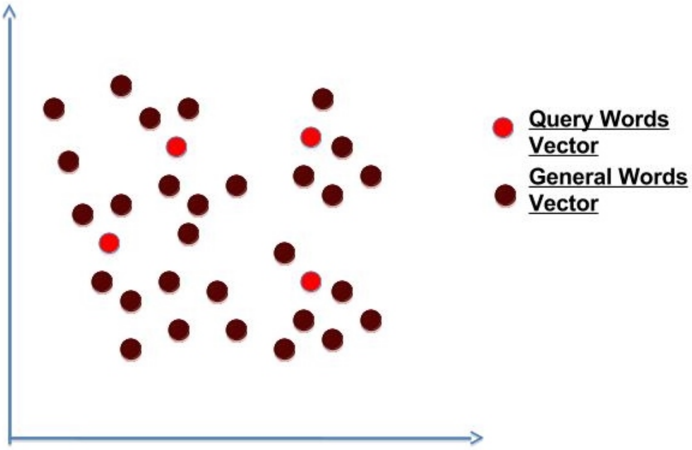


Figure 6.1: Word vectors in 2 dimension

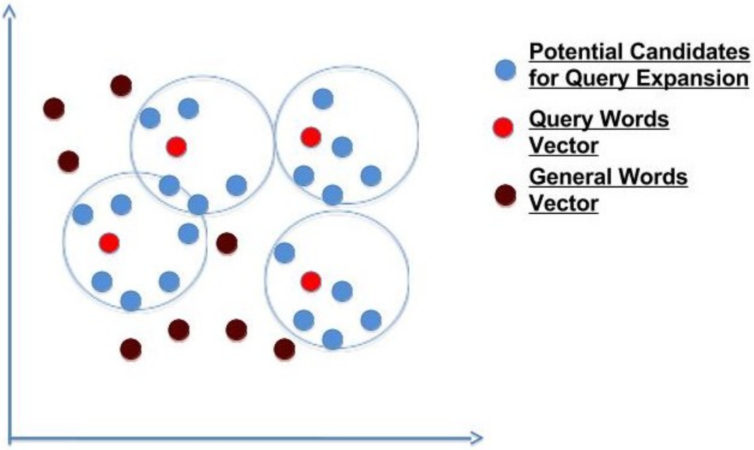


Figure 6.2: Potential candidates selection from complete set of word vectors

semantically similar to individual query words q_i. We obtain potential candidates for query expansion as shown in Figure 6.2. For a query Q consisting of n words say $\{q_1,...,q_n\}$, we generate a pool of potential candidates C, where $C = \{c_1,...,c_n\}$, such that c_i contains top z similar words to q_i, in the embedding space, where similar words are calculated using cosine similarity score between q_i and all the word vectors in the collection.

We sort all the words in C, based on the cosine similarity score between each term t_j and corresponding query word q_i. We select top k terms as effective expanded terms as shown in Figure 6.3. The top words selected are biased towards specific query words rather then being general to the whole query. The main focus is to capture keywords, synonyms and entities which are semantically related to query words.

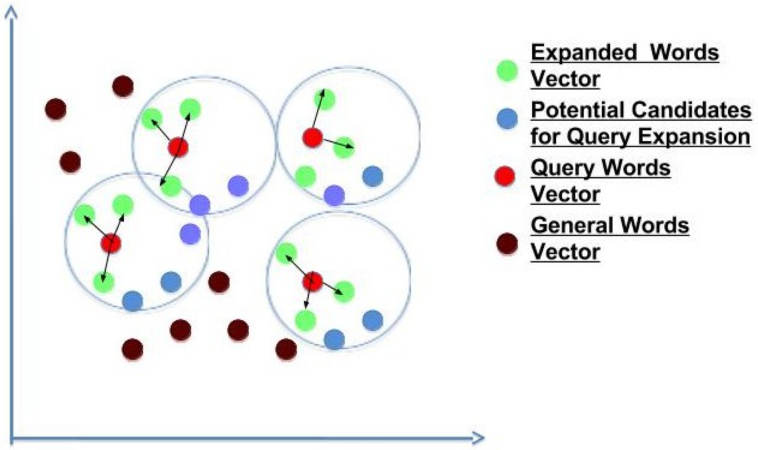


Figure 6.3: Top expansion terms selection using QueryWord approach

2) *Centroid approach*: In this approach, we generate an initial pool of candidates C, similar to QueryWord approach. Instead of using all the general word vectors for expansion we filter the words and form a collection C, consisting of words which are more similar to individual q_{i} as discussed above in QueryWord approach. Next, we form a centroid vector CV, by summing all the query words which are a vector of D dimension i.e $CV = \sum_{i=1}^{n} q_{\mathrm{i}}$.

We sort all the words in C, based on the cosine similarity score between each term t_{j} and the centroid vector CV. We select top k terms as effective expanded terms which are more similar to combined vector consisting of all the query terms as shown in Figure 6.4. The main focus is to retrieve words which are semantically related to all the query words as a single unit or a phrase.

3) *Global Centroid approach*: There is a limitation of the Centroid approach in that it does word vector filtering by creating a collection C as shown in Figure 6.2, to reduce the number of comparisons while finding most similar words to the query centroid vector CV, which we address in the Global Centroid approach. In the Global Centroid approach, we compare centroid vector CV with all word vectors in the collection as shown in Figure 6.5 and investigate whether filtering of information results in significant difference in the performance and identification of

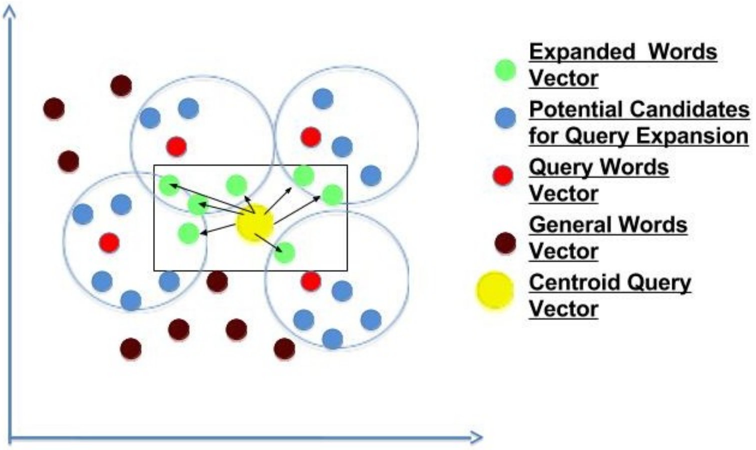


Figure 6.4: Top expansion terms selection using Centroid approach

potential expansion terms.

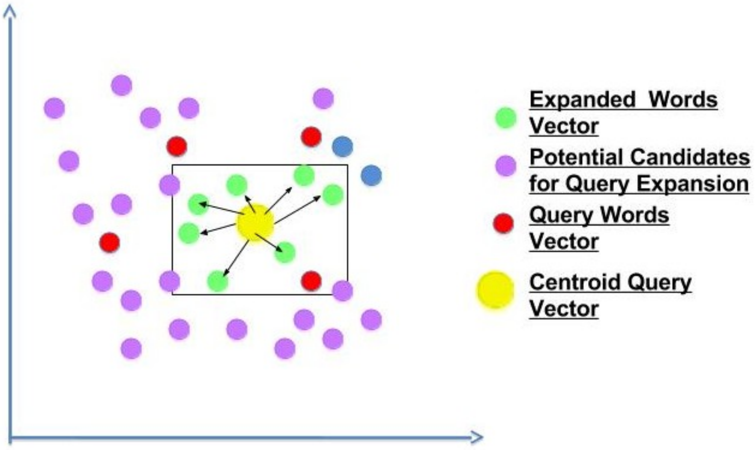


Figure 6.5: Top expansion terms selection using Global Centroid approach

Combined Approach (PRF + Semantic Expansion): Finally, we investigate how can we effectively combine potential expansion terms learnt using traditional and embeddings-based expansion. The embedding-based expansion tries to find potential words that are used in similar context and thus might indicate similar meaning. Pseudo relevance feedback looks for potential candidates in top ranked retrieved sentences which are similar to a query and thus are highly specific, contextual as compared to expansion terms obtained using embeddings techniques which are more general in nature as they capture co-occurrence counts of words from the complete collection. We investigate capturing multiple signals which can be combined together for improving relevant effectiveness. We propose a combined model for sentence retrieval to capture the semantic distribution of words while performing query expansion effectively. We perform query expansion using a linear combination of expansion terms learnt using traditional PRF and word embeddings-based techniques.

$$COW = \alpha * WE_{EQT} + (1 - \alpha) * PRF_{EQT} \tag{6.11}$$

where EQT stands for expanded query terms, WE stands for word embedding approach and PRF stands for pseudo relevance feedback approach, α is empirically calculated by varying it in the range of [0-1], with an increment of 0.1.

6.3 Experimental Setup

In this section we describe the main tools, resources and datasets used for our experiments for the task of sentence relevance estimation.

6.3.1 Datasets

In this section, we describe the dataset that is used to measure the performance of our relevance prediction model. We use the standard TREC Novelty track dataset (Soboroff and Harman, 2005) developed for evaluating sentence-level relevance and novelty detection models. In this thesis, we focus on generation of effective snippets which are topically relevant, novel and readable. Thus we use this collection as there are gold-level annotations available for relevance and novelty measures at sentence-level. There are not many other sentence-level benchmarks collection at both novelty and relevant information available. Hence, we focus on the topics from the TREC 2003 and TREC 2004 Novelty track and the AQUAINT corpus for our experiments. Each of the track has 50 topics (information needs), which are events and opinionated-based topics. This track uses a document collection from the AQUAINT corpus, which consists of newswire text in English, drawn from three sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service. Relevance information annotation is done as a sentence-level binary classification as *Relevant* or *Not Relevant*.

Track	Topics	Documents	Sentences	Relevant Sentences
2003 track	50	1187	39820	15557
2004 track	50	1214	52447	8343

Table 6.1: 2003 and 2004 Novelty track data distribution

Topics Distribution	2003 topics set	2004 topics set
Average Length of Topics (only title)	3.2	3.4
Minimum Length of Topics (only title)	1	2
Maximum Length of Topics (only title)	5	6
Average Length of Topics (only description)	10.5	14.26
Minimum Length of Topics (only description)	4	6
Maximum Length of Topics (only description)	25	32
Sentence Distribution	**2003 documents set**	**2004 documents set**
Average Length of Sentence	17.6	17.35
Minimum Length of Sentence	1	1
Maximum Length of Sentence	96	114
Average Number of Sentence (in a document)	33.5	43.2
Minimum Number of Sentence (in a document)	4	4
Maximum Number of Sentence (in a document)	146	260

Table 6.2: Query and Sentence length distribution for TREC 2003 and TREC 2004 topics set and documents set respectively.

Table 6.1 shows the overall statistics of the topics, documents[1] and sentence distribution in the two data sets, which we use in our experiments. TREC 2003 and 2004 have 39% and 16% of relevant sentences respectively, thus it seems the TREC 2004 documents have more noisy and non-relevant sentences as compared to the TREC 2003 documents. Table 6.2 show details of the topic and sentence length distribution for TREC 2003 and 2004 topics and documents set. The average length of a title field resembles typical web search query of about 3-5 words.

[1] All documents are relevant, where each document is judged by a TREC assessor.

6.3.2 Tools

Below we outline the details of tools and resources used for conducting our experiments.

- Lucene toolkit: We use the lucene toolkit[2] to perform retrieval of sentences for a given query from the documents collection. We performed stemming (Porter, 1980) and stopword removal using the Lucene English Analyser while indexing the collection as well as while searching queries over the collection. We use the LM and BM25 model implementation of lucene and our own implementation of PRF for our experiments. The parameters and other details are provided with the experimental details in Section 6.4.

- Semantic Compositionality: We use gensim[3] implementation of Word2Vec in our work to learn and incorporate *word embeddings* (reviewed in Chapter 4), in our experiments for performing query expansion. We use two types of word embeddings in our work as follows:

  1) *General-domain, Google embeddings*: Embeddings pre-trained on Google news, consisting of about 3 million 300 dimension English word vectors which are released for research, for more details check the link below.[4]

  2) *In-domain, AQUAINT embeddings*: Embeddings learnt using the in-domain AQUAINT document collection. We varied different parameter settings such as training method, dimension size, window size, for our internal experiments and compare the performance and effectiveness of these parameters for semantic similarity and query expansion techniques (details are described later in Section 6.3.4).

  As we perform sentence retrieval for documents from AQUAINT collection, these embeddings learnt using AQUAINT corpus form in-domain embeddings,

[2] `https://lucene.apache.org/core/4_4_0/`

[3] `https://radimrehurek.com/gensim/`

[4] `https://github.com/mmihaltz/word2vec-GoogleNews-vectors`

however Google embeddings which are pre-trained on a big Google news corpora form general-domain embeddings.

6.3.3 Evaluation Metrics

In this section we introduce the evaluation metrics used for our investigation of sentence retrieval. As in standard document based IR we are interested in evaluating how good are our approaches at identifying relevant sentences and ranking them higher than non-relevant sentences, hence we focus on measuring precision (P), as defined in Equation 6.12. Within our use case of snippet generation we are interested in finding potential relevant sentences in top k sentences.

$$Precision = \frac{relevant\ sentences\ retrieved}{retrieved\ sentences} \tag{6.12}$$

Previous work by Leal Bando et al. (2015) used {P@2, P@4, P@6} for the task of sentence retrieval for generating query biased summaries. Thus, for the task of retrieving sentences for a given query and a given retrieved document we use precision at rank two and rank five (P@2, P@5) to compare the performance of our methods, and to study the best parameter settings, as these measures have been commonly used for the task of sentence retrieval.

6.3.4 Data pre-processing

In this section we describe as initial processing done at query and sentence-level.

Query Processing: The TREC 2003 and 2004 track topics have three fields i) Title, ii) Description, and iii) Narrative. The average length of title and description fields is indicated in Table 6.2. We use the title field of each topic since these resemble typical web queries where word length is between 1-5 words with an average of 3.2 words. Similar settings were also used in previous work by Leal Bando et al. (2015). Further, we perform Porter stemming (Porter, 1980) and stopword removal from the title before processing using Lucene.

Sentence Processing: We use Lucene for indexing and searching over sentences within a document. We perform stopword removal and Porter stemming before indexing the document collection. We perform stemming over the queries and sentences since it helps to handle problem of word form variations quite common in matching items in IR.

A small number of queries contain acronyms and abbreviation such as NATO, U.S, ANWAR, NAFTA, JR., and while others have spelling mistakes such as withold, pyonyang. We do not seek to handle these special cases to reflect the general scenario of web search, and aim to address the retrieval issues that these introduces using query expansion techniques.

Learning word embeddings: To learn word embeddings from the AQUAINT document corpus, first we perform stopword removal and Porter stemming over the raw corpus, and then use the processed corpus for training embeddings. Previous research (Diaz et al., 2016; Arora et al., 2017) has shown that it is better to train word embeddings from an in-domain corpus, as well as using embeddings from a big pre-trained model, as these embeddings (in-domain and general-domain) capture and learn quite different information to represent the semantics of the words with a positive impact on retrieval.

We learn different in-domain embeddings using AQUAINT corpus while varying the parameter settings as mentioned below.

- Algorithm: Continuous bag of words (Cbow) and Continuous skip gram (Cskip) model (different models for learning word embeddings, reviewed in Chapter 4)
- Window Size: 5 and 10 (values commonly used for IR and textual similarity experiments), it indicates neighbouring words for learning context information
- Embeddings Size: 100, 200 and 300 dimensions of word vector (values commonly used for IR and textual similarity experiments)

Details of the embedding training algorithm are given in Chapter 4. Overall, we have 12 different in-domain embeddings (2 algorithms * 2 window size * 3 embeddings size), and 1 general-domain embedding.

6.4 Results

In this section we report experimental results for our sentence retrieval experiments.

6.4.1 Traditional retrieval models (baseline approach)

Below we describe the results of sentence retrieval using traditional retrieval models.

LM model: As described in Section 6.2.1, we explored a Jelinek-Mercer smoothing-based language model for sentence retrieval. We tried parametric optimisation by changing λ in the LM retrieval model in the range of [0.05, 1.0] with an increment of 0.05. Figure 6.6 shows our results for sentence retrieval using the language modelling approach for the 2003 and 2004 topic set respectively. λ controls the weight distribution from a document and a collection, the results are more consistent for λ in the range of [0.05, 0.5] across both data sets. As λ varies from [0.5, 0.95] both P@2 and P@5, drop more rapidly for 2004 topic set as compared to 2003 topic set. As λ increases, more weight is given to term counts of collection frequency than sentence frequency thus allowing non-matched query terms to dominate leading to non-relevant sentences being scored higher. In previous work by Zhai and Lafferty (2001), different smoothing techniques were investigated with results indicating that optimal value of λ depends on both the collection and the query, on a general level the optimal value of λ is around 0.1 for title queries for Jelinek-Mercer smoothing-based language model. Similarly we observed high precision values for lower values of λ for both datasets.

BM25 model: For the BM25 model we performed grid search in the range of [0.1, 2.0] and [0.0, 1.0] with an increment of 0.1 for k_1 and b parameters in the BM25 retrieval model. In our experiments we found that changing the k_1 parameter does not have much affect in P@2 and P@5 results for either dataset, but changing the

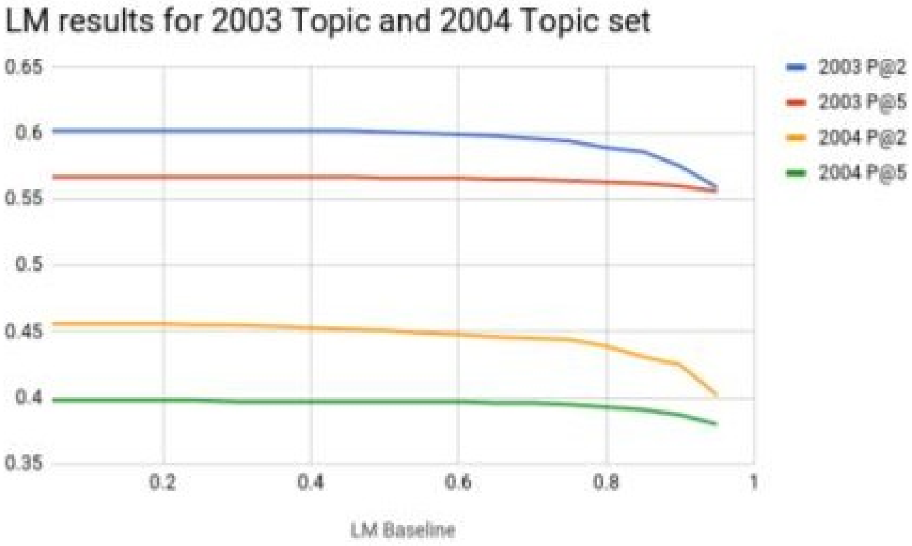
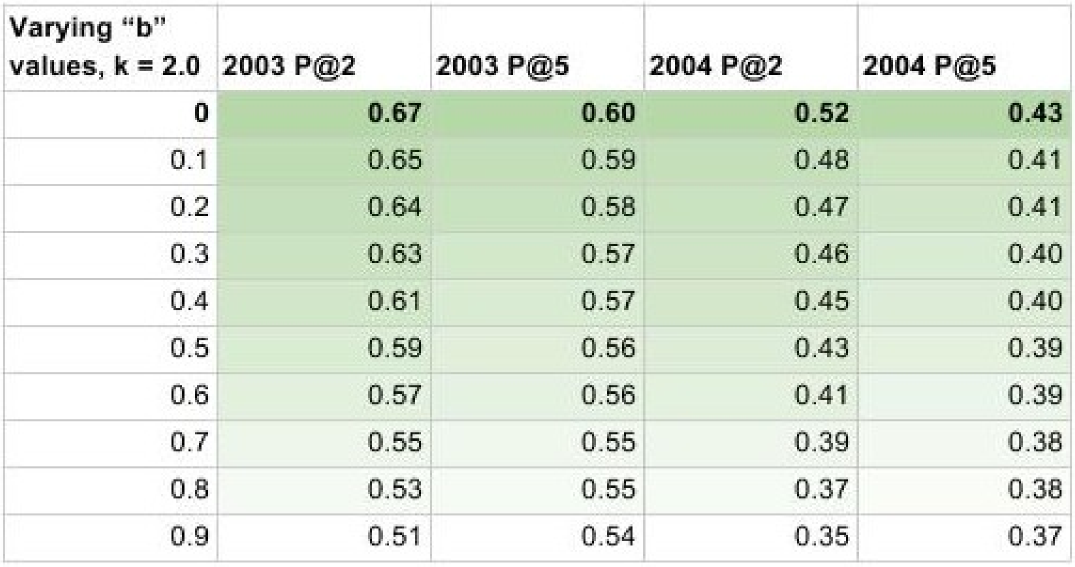


Figure 6.6: P@2, and P@5 scores for LM model for 2003 and 2004 Topic Set

b values result in quite significant changes. The results with the value of $k_1 = 2.0$ were slightly better so we fixed it to this, and varied the b value for our experiments as shown in Figure 6.7. Dark green cells indicate higher values of precision within each column. The results are better for low values of b. Similar observations were made in previous work by Losada (2010), which obtained best results for sentence retrieval when b was set to 0. One possible explanation seems to be that the length distribution across sentences within a document does not vary significantly unlike the case of document retrieval, where documents length could vary a lot in the collection. Thus the hyperparameter b which penalises lengthy documents in the case of document retrieval does not have a positive impact while penalising lengthy sentences for the case of sentence retrieval.

6.4.2 ESS model

We investigated an embeddings-based semantic similarity approach for estimating relevance of each sentence to a query. To represent a query and sentence we simply sum up the individual word vectors and form a combined vector q_i and s_j. The cosine similarity between q_i and s_j then represents the estimated relevance score for



Varying “b” values, k = 2.0	2003 P@2	2003 P@5	2004 P@2	2004 P@5
0	**0.67**	**0.60**	**0.52**	**0.43**
0.1	0.65	0.59	0.48	0.41
0.2	0.64	0.58	0.47	0.41
0.3	0.63	0.57	0.46	0.40
0.4	0.61	0.57	0.45	0.40
0.5	0.59	0.56	0.43	0.39
0.6	0.57	0.56	0.41	0.39
0.7	0.55	0.55	0.39	0.38
0.8	0.53	0.55	0.37	0.38
0.9	0.51	0.54	0.35	0.37

Figure 6.7: P@2, P@5 results in the form of grid search for BM25 retrieval model while varying values of b and fixing k_1=2.0, for 2003 and 2004 Topic Set

a query-sentence pair, as described in Section 6.2.2.

We present the results using 13 different representations (12 AQUAINT embeddings and 1 Google embeddings) in Table 6.3. We find that the AQUAINT embeddings perform far better than Google embeddings for both the data sets. Results also show that the Cbow method of training is slightly better than Cskip for the task of sentence retrieval. Varying the size of embeddings, and context window does not much affect performance as compared to the data source used for training embeddings and the training method used. Slightly better results are obtained using $AQUAINT$ embeddings which are learnt using Cbow model, with window size being set to 10.0 and dimensions being set to 100, so we use this settings of embedding (among 12 different AQUAINT embedding explored for ESS) for semantic-based query expansion approach.

The ESS approach performs poorly compared to the LM or BM25 sentence retrieval as represented in Table 6.4. The possible reasons for this appear to be: i) significant word length variation between query and sentences, ii) combined vector representation of sentences averaging multiple words might capture noise and drift from the main query intent and representation. Previous work by Mitra et al. (2016)

on document retrieval also report that using only embeddings-based models perform more poorly than standard retrieval model such as BM25.

	Track 2003		Track 2004	
Model	P@2	P@5	P@2	P@5
Google News	0.34	0.36	0.19	0.21
AQUAINT_100_5_M1	0.39	0.41	0.24	0.24
AQUAINT_100_10_M1	**0.41**	**0.42**	**0.24**	**0.24**
AQUAINT_200_5_M1	0.40	0.41	0.25	0.24
AQUAINT_200_10_M1	0.40	0.41	0.24	0.24
AQUAINT_300_5_M1	0.40	0.41	0.24	0.24
AQUAINT_300_10_M1	0.41	0.42	0.24	0.24
AQUAINT_100_5_M2	0.40	0.40	0.23	0.23
AQUAINT_100_10_M2	0.40	0.40	0.22	0.23
AQUAINT_200_5_M2	0.40	0.41	0.22	0.23
AQUAINT_200_10_M2	0.40	0.40	0.22	0.23
AQUAINT_300_5_M2	0.40	0.41	0.24	0.24
AQUAINT_300_10_M2	0.39	0.40	0.22	0.22

Table 6.3: ESS model-based results for sentence retrieval, best model scores are in boldface, where M1 indicates Cbow model and M2 indicates Cskip model.

	Track 2003		Track 2004	
Model	P@2	P@5	P@2	P@5
ESS model (AQUAINT_100_10_M1)	0.41	0.42	0.24	0.24
LM (λ=0.2)	0.60^{+}	0.57^{+}	0.46^{+}	0.40^{+}
BM25 (k_1=2.0 & b=0.0)	$\mathbf{0.67}^{*+}$	$\mathbf{0.60}^{*+}$	$\mathbf{0.52}^{*+}$	$\mathbf{0.43}^{*+}$

Table 6.4: Different retrieval model baseline results for sentence retrieval, best scores are in boldface. + indicates that the difference in the results compared to the ESS method (Cbow model) is statistically significant, and $*$ indicates that the difference in the results compared to the LM method is statistically significant with p<0.01, using student's t-test.

Table 6.4 presents the best results using LM, BM25 and ESS. As the BM25 results are statistically significantly better than the LM and ESS, they form the baseline results for our experiments. For further experimental investigation we fix the values of $b = 0.0$ and $k_1 = 2.0$ for the BM25 model.

6.4.3 Query Expansion

To address the vocabulary mismatch problem between queries and sentences for retrieval, we perform query expansion using three methods as described in Section 6.2.3. Below we explain how the QE methods are applied, and the results of applying QE for sentence retrieval.

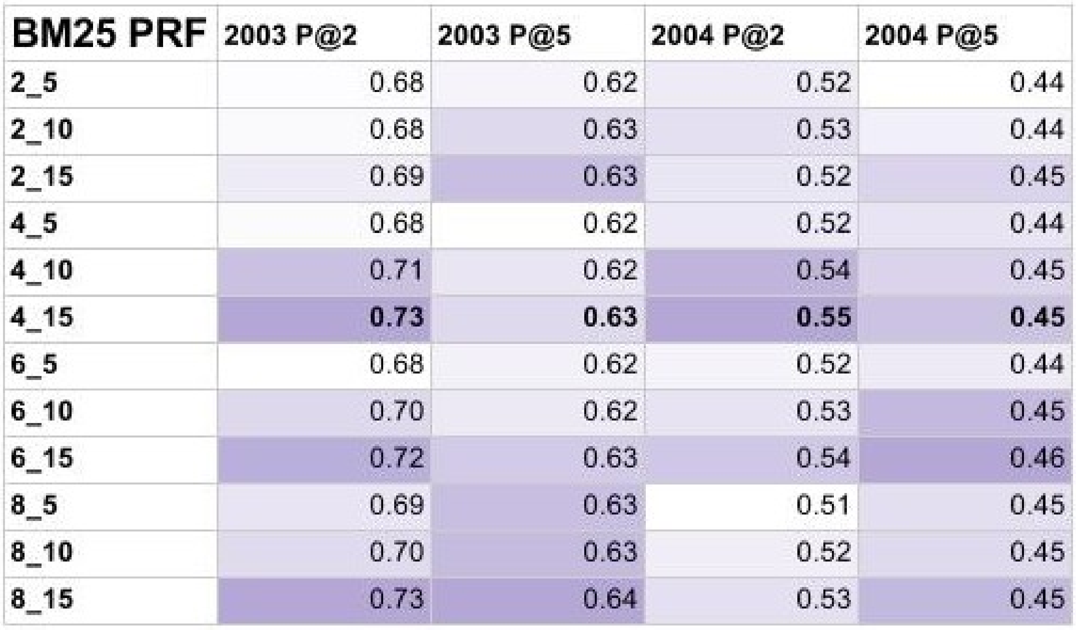
Pseudo Relevance Feedback (PRF): We varied the assumed relevant sentences R in the range of {2, 4, 6, 8}, and expanded query terms (EQT) in the range of {5, 10, 15}. We linearly varied the weight of initial query terms Q and expanded terms EQT as shown in Equation 6.13.

$$Combined\ Query = \beta * Q + 1.0 * EQT \tag{6.13}$$

where β varies in range [1, 2], with an increment of 0.1.

Figure 6.8 shows results of sentence retrieval using PRF-based QE while varying the number of terms, and sentences respectively with β being set to 2.0. Dark shade of purple indicates higher value of precision within each column. We find that results vary significantly depending on the parameters such as the number of expansion terms and assumed relevant sentences, but does not change much while varying weight (β) of the initial query terms.

As more terms are added the P@2 and P@5 scores for both TREC 2003 and 2004 datasets improves. Results with 15 expansion terms are the highest while keeping the number of sentences fixed. While keeping the terms fixed it is not always the case that the result increases with the number of sentences. Across both the datasets, the parameters R=4 and EQT=15 with a weight of initial terms (β) set to 2.0 performs moderately better than alternative combinations of R and EQT explored, and have high values of P@2 and P@5. As reported in Table 6.5, the PRF technique shows statistically significant improvement over the baseline using the BM25 retrieval model for both datasets. Similar improvements using QE techniques for sentence retrieval were reported in earlier work on sentence retrieval (Allan et al.,



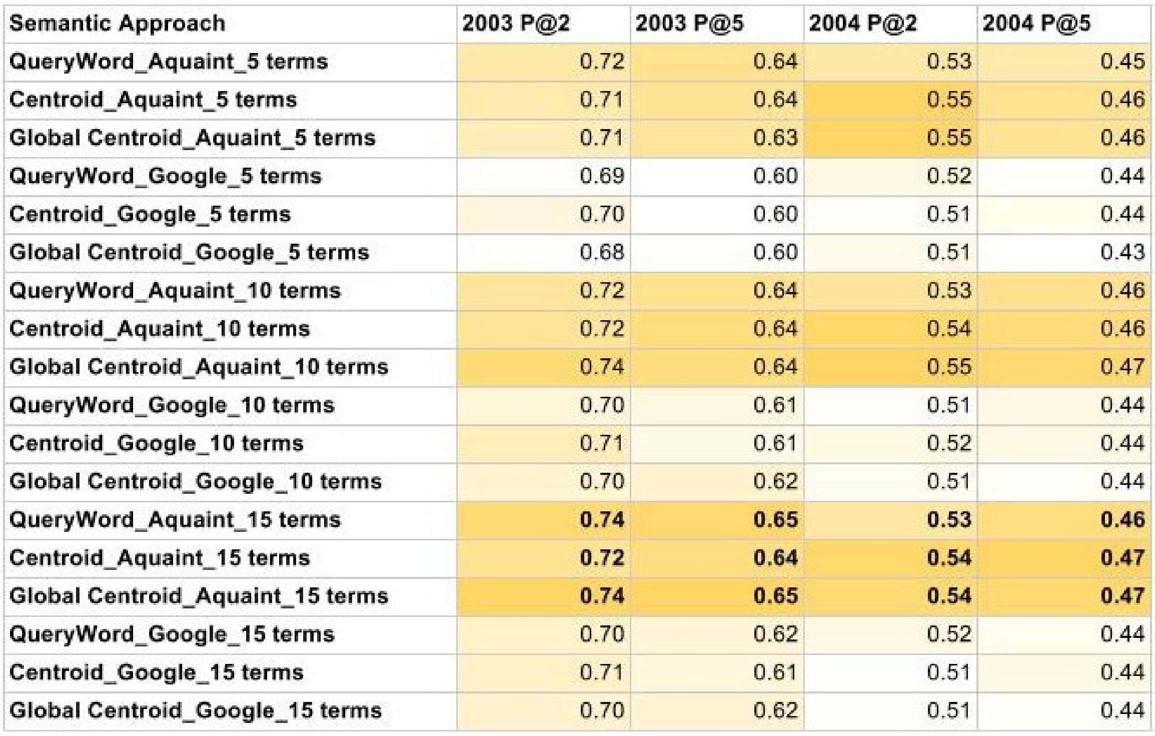
BM25 PRF	2003 P@2	2003 P@5	2004 P@2	2004 P@5
2_5	0.68	0.62	0.52	0.44
2_10	0.68	0.63	0.53	0.44
2_15	0.69	0.63	0.52	0.45
4_5	0.68	0.62	0.52	0.44
4_10	0.71	0.62	0.54	0.45
4_15	**0.73**	**0.63**	**0.55**	**0.45**
6_5	0.68	0.62	0.52	0.44
6_10	0.70	0.62	0.53	0.45
6_15	0.72	0.63	0.54	0.46
8_5	0.69	0.63	0.51	0.45
8_10	0.70	0.63	0.52	0.45
8_15	0.73	0.64	0.53	0.45

Figure 6.8: P@2, P@5 results in the form of grid search for BM25 PRF retrieval model while varying values of sentence_term pair represented vertically, for 2003 and 2004 Topic Set

2003; Losada, 2010; Leal Bando et al., 2015).

Semantic expansion using word embeddings: As an alternative to standard PRF, we also explored the use of semantic expansion using word embeddings. We set the value of $z = 10$, which determines the number of similar words for a query term in the embedding space to be included for obtaining potential candidates C, for QueryWord and Centroid approach as discussed in Section 6.2.3. The number of expanded terms k was varied with the values $\{5, 10, 15\}$. We linearly varied the weight of query terms and expanded terms as shown in Equation 6.13. For each of the embeddings-based expansion methods i) QueryWord approach, ii) Centroid approach, and iii) Global Centroid approach, we experimented with two different embeddings: *Google* embeddings and *AQUAINT* embeddings as described in Section 6.3.2. Figure 6.9 shows results of semantic-based QE. Dark shades of orange indicate higher values of precision within each column.

In general, the performance of AQUAINT embeddings is far better than Google embeddings. The results of the three semantic expansion-based QE methods using



Semantic Approach	2003 P@2	2003 P@5	2004 P@2	2004 P@5
QueryWord_Aquaint_5 terms	0.72	0.64	0.53	0.45
Centroid_Aquaint_5 terms	0.71	0.64	0.55	0.46
Global Centroid_Aquaint_5 terms	0.71	0.63	0.55	0.46
QueryWord_Google_5 terms	0.69	0.60	0.52	0.44
Centroid_Google_5 terms	0.70	0.60	0.51	0.44
Global Centroid_Google_5 terms	0.68	0.60	0.51	0.43
QueryWord_Aquaint_10 terms	0.72	0.64	0.53	0.46
Centroid_Aquaint_10 terms	0.72	0.64	0.54	0.46
Global Centroid_Aquaint_10 terms	0.74	0.64	0.55	0.47
QueryWord_Google_10 terms	0.70	0.61	0.51	0.44
Centroid_Google_10 terms	0.71	0.61	0.52	0.44
Global Centroid_Google_10 terms	0.70	0.62	0.51	0.44
QueryWord_Aquaint_15 terms	**0.74**	**0.65**	**0.53**	**0.46**
Centroid_Aquaint_15 terms	**0.72**	**0.64**	**0.54**	**0.47**
Global Centroid_Aquaint_15 terms	**0.74**	**0.65**	**0.54**	**0.47**
QueryWord_Google_15 terms	0.70	0.62	0.52	0.44
Centroid_Google_15 terms	0.71	0.61	0.51	0.44
Global Centroid_Google_15 terms	0.70	0.62	0.51	0.44

Figure 6.9: P@2, P@5 results in the form of grid search for Semantic Expansion-based BM25 retrieval model while varying types of embedding and terms for expansion, for 2003 and 2004 Topic Set

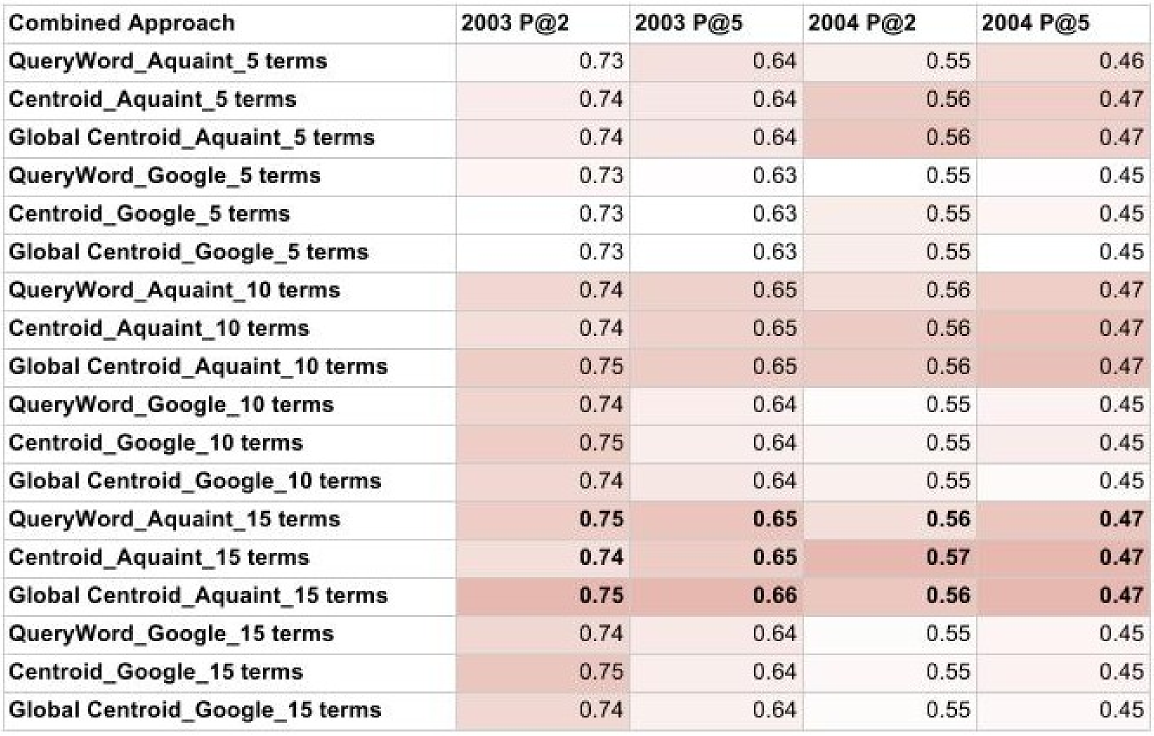
QueryWord, Centroid, and Global Centroid are quite similar. Overall, the Global Centroid approach performs slightly better than the Centroid and the QueryWord approach. As more terms are added the P@2 and P@5 scores increase. The best results are obtained using AQUAINT embeddings with 15 expanded terms and weight of the initial terms being set to 1.0.

The best results corresponding to different embeddings-based expansion technique are presented and compared in Table 6.5. All best results corresponding to the three semantic QE techniques show statistically significant performance over the baseline of BM25 retrieval model for P@2 and P@5 for the two datasets. QE using word embeddings shows statistically significant performance over PRF-based query expansion approach for P@5 only, across both datasets. Word embeddings-based QE appear to identify effective terms to reduce the vocabulary mismatch problem and thus improves retrieval performance considerably. We speculate two reasons for same: i) training embedding algorithm incorporates co-occurrence counts from the corpus effectively, ii) PRF focuses on terms from top R sentences, whereas embeddings looks for potential candidate from the whole vocabulary thus can capture better terms which might not appear in top ranked sentences for the initial query.

Combined Approach (PRF + Semantic Expansion): We present the results for a combined model of QE which integrates semantic distribution of words along with PRF-based QE method. The number of expanded terms k for expansion using word embeddings was varied as $\{5, 10, 15\}$, while for PRF we used the best settings as identified for PRF ($R = 4$ and $EQT = 15$). We varied the weights for the initial query terms, the expanded query using PRF and the expanded query using semantic-based expansion. The best results were obtained with the weight of the initial query being set to 1.2, the PRF-based expanded term set to 0.2 and the embeddings-based term weight set to 0.8. We refer to this configuration of the combination model as **BestRelModel** for further discussion which we use for snippet generation (described later in Chapter 8). Figure 6.10 shows the result of the combined semantic and PRF-based expansion results. Table 6.5 presents and compares the best results using combined expansion approach.

	Track 2003		**Track 2004**	
Approach	P@2	P@5	P@2	P@5
Baseline	0.67	0.60	0.52	0.43
BM25 PRF	0.73*	0.62*	0.55*	0.45*
SE: QueryWord approach	0.74*	0.65$^{*\delta}$	0.53	0.46$^{*\gamma}$
SE: Centroid approach	0.72*	0.64$^{*\delta}$	0.54*	0.47$^{*\delta}$
SE: Global Centroid approach	0.74*	0.65$^{*\delta}$	0.54*	0.46$^{*\delta}$
Com: QueryWord approach	0.74$^{*\delta}$	0.65$^{*\delta}$	0.56*	0.47$^{*\delta}$
Com: Centroid approach	0.74$^{*\gamma}$	0.65$^{*\delta}$	**0.57**$^{*\delta}$	**0.47**$^{*\delta}$
Com: Global Centroid approach	**0.75**$^{*\delta}$	**0.66**$^{*\delta}$	0.56$^{*\gamma}$	**0.47**$^{*\delta}$

Table 6.5: Best Results for query expansion approach for sentence retrieval, the best scores are in boldface. $*$ indicates that the difference in the results compared to the baseline is statistically significant with p<0.01, δ, and γ indicates that the difference in the results compared to the PRF approach is statistically significant with p<0.01, and p<0.05 respectively using student's t-test



Combined Approach	2003 P@2	2003 P@5	2004 P@2	2004 P@5
QueryWord_Aquaint_5 terms	0.73	0.64	0.55	0.46
Centroid_Aquaint_5 terms	0.74	0.64	0.56	0.47
Global Centroid_Aquaint_5 terms	0.74	0.64	0.56	0.47
QueryWord_Google_5 terms	0.73	0.63	0.55	0.45
Centroid_Google_5 terms	0.73	0.63	0.55	0.45
Global Centroid_Google_5 terms	0.73	0.63	0.55	0.45
QueryWord_Aquaint_10 terms	0.74	0.65	0.56	0.47
Centroid_Aquaint_10 terms	0.74	0.65	0.56	0.47
Global Centroid_Aquaint_10 terms	0.75	0.65	0.56	0.47
QueryWord_Google_10 terms	0.74	0.64	0.55	0.45
Centroid_Google_10 terms	0.75	0.64	0.55	0.45
Global Centroid_Google_10 terms	0.74	0.64	0.55	0.45
QueryWord_Aquaint_15 terms	**0.75**	**0.65**	**0.56**	**0.47**
Centroid_Aquaint_15 terms	**0.74**	**0.65**	**0.57**	**0.47**
Global Centroid_Aquaint_15 terms	**0.75**	**0.66**	**0.56**	**0.47**
QueryWord_Google_15 terms	0.74	0.64	0.55	0.45
Centroid_Google_15 terms	0.75	0.64	0.55	0.45
Global Centroid_Google_15 terms	0.74	0.64	0.55	0.45

Figure 6.10: P@2, P@5 results in the form of grid search for Combined Query Expansion Approach while varying type of semantic expansion and values of expansion terms, for 2003 and 2004 Topic Set

As shown in Figure 6.10, the precision scores increase as more terms are added in the combined model. It is worth noting that the scores of combination approach using Google embeddings are also quite competitive and close to the results obtained using the AQUAINT embeddings, contrary to when only semantic expansion without PRF is used as shown in Figure 6.9. Thus adding semantic and PRF expansion terms capture complementary signals which collectively boost the retrieval performance. Across all three semantic expansion approaches the results are quite similar for the combined approach, we speculate that the differences learnt from different approaches and its combination with PRF are not clear as the results have been averaged over all the document in the collection. We inspect in detail how retrieval results, using the best models explored in this work, vary across documents having different amount of gold relevant sentences, which we describe next in Section 6.5.

The combination approach using PRF and embeddings-based expansion perform statistically significantly better than the BM25 baseline and PRF approach for both P@2 and P@5 scores. The best approach for each dataset and metric combination is one which employs a combination of embeddings and PRF expanded terms as shown in Figure 6.10. Our findings although on sentence retrieval are on the similar lines as reported by Roy et al. (2016); Kuzi et al. (2016); Diaz et al. (2016) for document retrieval, that the combination approach performing query expansion using relevance feedback and embeddings-based approach perform significantly better than the relevance feedback-based approach. However, contrary to their findings that individual embeddings based expansion perform inferior to PRF-based approach, in our investigation we find that for sentence retrieval embeddings-based QE perform similar or better than the PRF-based approach. We speculate three reasons for same: i) The nature of the dataset and the collection, ii) our findings are on sentence-retrieval where mismatching problem is more acute (due to short length of the sentences), using embeddings captures potentially better signals and boost the retrieval effectiveness as compared to document retrieval, and iii) the semantic-based approaches explored in our work are more effective than previously used embeddings-based expansion techniques.

6.5 Discussion & Analysis

We perform detailed analysis of our results to see fine distinctions and the effects of different types of semantic-based expansion techniques explored in this work. We study how different models perform when datasets are split into three sets based on the relative number of relevant sentences in a document. Table 6.6 shows the data split and the number of documents within each set. In general, set 1 represents the documents having the least of relevant sentences, set 2 represents documents that have moderate number of relevant sentences and set 3 represents documents that have the greatest number of relevant sentences.

Tables 6.7, 6.8 and 6.9, show results of the best models of different approaches

	2003 Topic Set		2004 Topic Set	
Distribution	Docs	Data split	Docs	Data split
Set 1	420	threshold < 0.28	450	threshold < 0.15
Set 2	415	$0.28 \geq$ threshold < 0.65	495	$0.15 \geq$ threshold < 0.4
Set 3	352	threshold ≥ 0.65	269	threshold ≥ 0.615

Table 6.6: Set-based topic distributions, where threshold is calculated by dividing the number of relevant sentences in a document by the total number of sentences in a document

	2003 Set 1		2004 Set 1	
Method	P@2	P@5	P@2	P@5
BM25 baseline	0.41	0.32	0.29	0.21
BM25 PRF	0.44^{+}	0.33^{+}	0.29	0.21
SE: QueryWord approach	0.45	0.34^{*}	0.31	$0.23^{*\delta}$
SE: Centroid approach	0.44	0.34^{*}	$0.32^{+\gamma}$	$0.23^{+\delta}$
SE: Global Centroid approach	0.45	0.34^{*}	$0.32^{+\gamma}$	$0.23^{+\delta}$
Combined QueryWord approach	0.46^{*}	$0.35^{*\delta}$	$0.32^{*\delta}$	$0.23^{*\delta}$
Combined Centroid approach	0.46^{*}	$0.35^{*\delta}$	**0.33**$^{*\delta}$	**0.24**$^{*\delta}$
Combined Global Centroid approach	**0.48**$^{*\delta}$	**0.36**$^{*\delta}$	**0.33**$^{*\delta}$	$0.23^{*\delta}$

Table 6.7: Set 1 results, the best scores are in boldface. $*$, and $+$ indicates that the difference in the results compared to the baseline is statistically significant with $p<0.01$, and $p<0.05$ respectively, δ, and γ indicates that the difference in the results compared to the PRF approach is statistically significant with $p<0.01$, and $p<0.05$ respectively using student's t-test

explored in this work of sentence retrieval for three different subsets of the 2003 and 2004 dataset. The best approach performing QE using combined PRF and semantic expansion techniques performs statistically significantly better than the baseline models across all three sets and PRF model across all three sets apart from P@2 for 2003 and 2004 set 2 and P@2 for 2004 set 3. The relative improvement of our combined model for the scores of P@2, and P@5 for all different sets, topic collection is shown in Table 6.10. Some key observations from Tables 6.7, 6.8 and 6.9:

- Across all three sets, for 2004 topic set P@2 results using QueryWord approach performs lower than the Centroid and Global Centroid approaches.

	2003 Set 2		2004 Set 2	
Method	P@2	P@5	P@2	P@5
BM25 baseline	0.73	0.64	0.56	0.47
BM25 PRF	0.81*	0.69*	**0.62***	0.50*
SE: QueryWord approach	0.82*	0.71$^{*\gamma}$	0.57*	0.50*
SE: Centroid approach	0.80*	0.70*	0.58*	0.50*
SE: Global Centroid approach	**0.83***	0.71*	0.57*	0.50*
Combined QueryWord approach	**0.83***	**0.72**$^{*\delta}$	0.61*	0.51$^{*\delta}$
Combined Centroid approach	0.82*	0.71$^{*\delta}$	**0.62***	**0.52**$^{*\delta}$
Combined Global Centroid approach	**0.83***	**0.72**$^{*\delta}$	0.61*	**0.52**$^{*\delta}$

Table 6.8: Set 2 results, the best scores are in boldface. $*$ indicates that the difference in the results compared to the baseline is statistically significant with p<0.01, δ, and γ indicates that the difference in the results compared to the PRF approach is statistically significant with p<0.01, and p<0.05 respectively using student's t-test

- The PRF approach does not improve P@2 and P@5 results for set 1 for Topics 2004 as compared to the baseline, while the semantic expansion and combined approaches perform significantly better than the baseline. This indicates that the embeddings approach provides better expanded terms for set 1.

- For set 1 and set 2, which have relatively quite less relevant content the results using the combined expansion are always better than baseline, PRF and only semantic expansion approaches. Thus combining PRF and semantic expansion captures different signals and boost the retrieval effectiveness.

- For set 3 which has most amount of relevant information, the results using the semantic expansion and the combined expansion approach are significantly better as compared to the baseline and PRF results. However, the performance of both semantic expansion and combined approach is quite similar indicating that adding PRF-based terms to the embeddings approach does not help when the initial results using just embeddings-based expansion are quite high.

Overall our best model performs better than the BM25 baseline model by about 12%, and 10% for P@2 and P@5 respectively for the complete 2003 document set and 10% and 9% for P@2 and P@5 for the complete 2004 document set.

	2003 Set 3		2004 Set 3	
Method	P@2	P@5	P@2	P@5
BM25 baseline	0.91	0.87	0.82	0.73
BM25 PRF	0.96^{*}	0.90^{*}	0.85^{+}	0.76^{*}
SE: QueryWord approach	**0.99**$^{*\delta}$	$0.94^{*\delta}$	0.83	0.77^{*}
SE: Centroid approach	0.98^{*}	$0.94^{*\delta}$	0.86^{+}	**0.79**$^{*\delta}$
SE: Global Centroid approach	**0.99**$^{*\delta}$	**0.95**$^{*\delta}$	0.86^{+}	0.78^{*}
Combined QueryWord approach	$0.98^{*\delta}$	$0.93^{*\delta}$	0.85^{+}	$0.78^{*\gamma}$
Combined Centroid approach	$0.98^{*\delta}$	$0.93^{*\delta}$	**0.87***	**0.79**$^{*\delta}$
Combined Global Centroid approach	**0.99**$^{*\delta}$	$0.93^{*\delta}$	**0.87***	**0.79**$^{*\delta}$

Table 6.9: Set 3 results, the best scores are in boldface. $*$, and $+$ indicates that the difference in the results compared to the baseline is statistically significant with $p<0.01$, and $p<0.05$ respectively, δ, and γ indicates that the difference in the results compared to the PRF approach is statistically significant with $p<0.01$, and $p<0.05$ respectively using student's t-test

	2003 Topic Set		2004 Topic Set	
Method	P@2	P@5	P@2	P@5
Set-1	17.0%	12.5%	14.0%	9.5%
Set-2	14.0%	12.5%	9.0%	11.0%
Set-3	9.0%	7.0%	6.0%	8.0%

Table 6.10: Set-based relative improvements of our best model as compared to baseline results

We performed manual analysis of alternative QE outputs to explain the effects of different QE techniques with the objective of improving the task of sentence retrieval. Table 6.11 shows examples of expansion terms obtained from different query expansion techniques.

Some key observations from our analysis of results of the alternative QE techniques:

- In general, QE techniques capture words which are semantically related along with some noisy words as shown in Table 6.11.

- Google embeddings help to capture spelling variation effectively, for example for a query word *pyonyang*, which was misspelt, different expansion terms learnt are: pyongyang, pyongang, pyeongyang.

Topic – Query Words: Expansion Words
PRF-based expansion
India Pakistan Nuclear Tests: treaty, weapon, condemned, tension
China Spaceflight Program: spacecraft, satellite, rocket, launch, astronaut
Global Warming threat: catastrophe, antarctica, weather, snowfall, extinct
Driving Cell Phone Usage: handset, highway, accident, safety, cellular
AQUAINT embeddings-based expansion
India Pakistan Nuclear Tests: hindu, kashmir, sharif, rivalry, restraint, treaty
China Spaceflight Program: spacecraft, voyage, aerospace, chinese
Global Warming threat: climate, ice, antarctic, temperature, greenhouse, melting
Driving Cell Phone Usage: car, collision, telephones, cellular, mobile, distraction
Google embeddings-based expansion
India Pakistan Nuclear Tests: pakistani, islamabad, delhi, bangladesh, subcontinent, kashmir
China Spaceflight Program: chinese, beijing, shanghai, shenzhen, payloads, nasa
Global Warming threat: danger, worldwide, melting, warmed, cyberthreat, melting
Driving Cell Phone Usage: telephone, cells, cellphone, speeding, phones, cellular

Table 6.11: Example of Query Expansion learnt using different query expansion techniques explored in our work

- The QueryWord approach helps to capture synonyms, for example for the word: *gun*, expansion terms are: handgun, guns, pistol, firearm, firearms, handguns and rifle, for word: *phone*, expansion terms are: telephone, phones, cellphone, landlines, etc.

- The QueryWord approach helps to capture alternative variants of same words, for example for the word: *ban*, expansion terms are: bans, banned, banning, and for the word: *launched*, expansion terms are: launches, launching, re-launched.

- Along with good semantically related words, the expansion technique also add many noisy words which can sometimes lead to the problem of query drift. For the topic: “Atlanta Olympics bombing”, the expansion terms learnt using the AQUAINT centroid approach are “tanzanian, centennial, nairobi, injured, blast, dead, sympathy, kenyan, blasts” which are quite misleading and are not on the topic, thus hampering the retrieval performance.

Table 6.12 shows some examples of the top sentences from a document returned by the baseline model and our best model for different topics. A basic analysis reveals that the top sentence returned by our best model of combined approach using PRF and semantic-based expansion (**BestRelModel**), is more informative, as seen from Summaries 1, 2, 3 and 4, but in some cases it can also return sentences with too much information and drift from the main focus as seen for Summary 5.

We discuss few limitations of this work and present important direction for the future work.

1) Combination approach: We propose a linear combination approach for combining expansions terms learnt using traditional and semantic composition method and demonstrate that they work significantly better for the task of sentence retrieval. Developing better techniques of combining information learnt from semantic distribution and PRF-based approach is an area worth pursuing.

2) Combination of Embeddings: Our experimental investigation shows that the embeddings learnt with different parameter settings, and in-domain and general-domain corpora captures different information. How to effectively select expansion terms and train effective embeddings for the dedicated task needs to be further explored and investigated and opens a vast area for future research. We didn't explore how can we effectively combine terms from different in-domain and general-domain embeddings such as Google and AQUAINT embeddings but is worth pursuing in future.

6.6 Summary

We investigated and explored different models for topical relevance-based sentence selection for generating effective snippets. We found that the BM25 retrieval results are significantly better than the LM and query-sentence embedding-based matching approach (ESS). We perform query expansion using PRF and our proposed approach for semantic-based expansion. Our proposed approach using embeddings for query expansion perform statistically significantly better than the baseline, and comparatively similar or better than PRF-based approaches. In our experiments we found that varying the size of embeddings and context window does not affect the performance as compared to the data source (training data for embeddings) and the method used for training the embeddings (Cbow and Cskip).

We also proposed a linear interpolated mechanism of merging expansion terms obtained from embeddings-based approach and traditional PRF-based approach. The performance of combined approach out perform individual expansion approaches and PRF-based expansion. The linearly combined expansion terms perform significantly better than the baseline retrieval model and PRF-based query expansion approach. Our analysis shows that combining expansion terms learnt from both embeddings and PRF-based expansion terms provide complimentary signals and thus helps to improve the retrieval performance significantly. In this chapter, we proposed novel models for addressing the challenges of vocabulary mismatch for the task of relevance prediction. We experimentally show that our proposed models are effective and better than commonly used BM25-based retrieval model and PRF-based query expansion approach for sentence-level relevance prediction.

Table 6.12 shows some examples of the top sentences from a document returned by the baseline and our best model for different topics. At present, these top sentences are independently extracted from the documents. When these sentences are used to generate summaries to be represented in a SERP, based on the ranked order of the document relevance, some of these summaries are alike and express similar information as these summaries are independently generated. To generate summaries which are topically relevant, as well as to do not have repetitive or redundant information when presented in a SERP, we focus on the task of novel sentence selection which is discussed next in Chapter 7.

Summary-1: Topic – Japan Nuclear Accident
Baseline model: That's the early consensus among authorities on nuclear power as Japanese emergency officials struggled to contain the country's worst nuclear power accident ever, and President Clinton offered to do "whatever we possibly can" to help the people affected by the accident 70 miles northeast of Tokyo.
Best model: The deadly accident at a Japanese uranium processing plant Thursday can't compare with the disastrous explosion at the Chernobyl nuclear reactor in 1986, but it's probably more serious than the 1979 meltdown at Three Mile Island that crippled the US nuclear industry.
Summary-2: Topic – Driving Cell Phone Usage
Baseline model: But the experience was a powerful lesson, and she no longer uses the cell phone while driving.
Best model: While cell phone users were busy dialing, conversing, answering the phone or hanging up, their attention simply was not on the road.
Summary-3: Topic – human genome decoded
Baseline model: If the underestimate with Drosophila and other species is also true of the human genome, then its size "may have to be readjusted to as much as 4.0 billion base pairs," the company said, adding that the larger estimate had been built into its timetable for completing work on the human genome.
Best model: The public consortium had originally planned to complete the human genome sequence by 2005, but the project became a race when the Celera Corp., founded last year, announced that it would sequence the genome by the end of 2001.
Summary-4: Topic – Atlanta Olympics bombing
Baseline model: Greece Condemns Bomb Attack in Atlanta
Best model: ATHENS, July 27 (Xinhua) – The Greek Government today strongly condemned the bomb attack in Atlanta early Saturday which left two dead and over 100 injured.
Summary-5: Topic – First Human Hand Transplant
Baseline model: Biology, not doctors, ultimately will determine whether Matthew Scott's hand transplant is effective, one of his surgeons says.
Best model: Breidenbach said Scott was being treated in two ways – as a transplant patient, with an anti-rejection regimen like that of a kidney transplant patient, and as a limb-reattachment patient.

Table 6.12: Examples of top relevant sentence for a document returned by the baseline and our best model

Chapter 7

Novelty Detection

In this chapter we address the main question: *How to find new and novel information as users go through ranked documents from top to bottom in a SERP to avoid repetitive and redundant information and thus improve the user experience.* We predict sentence level novelty scores for a given set of relevant ranked documents for a given topic of inquiry (information need). We use the sentence level novelty scores to generate snippets for web documents to be presented in a SERP. In this chapter, first we provide our working definition of novelty. Next, we describe the baseline model, bag-of-words (bow) based distance metrics approach and our proposed models i) using word and sentence embeddings, and ii) using syntactic information for novelty detection. Then we discuss an approach that combines the output of bow-based distance metrics, embeddings and syntactic information based sentence comparison approaches for novelty detection. We go on to discuss the measures used for evaluating our novelty prediction models. We then present and analyse the experimental results.

7.1 Introduction

We follow the definition of *novelty* as defined in the TREC Novelty task (Soboroff and Harman, 2005), where the definition of **new** is relevant information that has not appeared previously in a set of documents on a given topic (as reviewed in Chapter 2). Novel sentences are determined by identifying which relevant sentences add new information as users read sentences from top to bottom in a linear fashion.

In the task of novelty detection, we are given an ordered list of relevant documents and a list of sentences from each document in the order of their occurrence in the ranked documents. Each sentence in the ordered list is classified as novel, if it contains new information which has not appeared previously, otherwise it is classified as not novel. The following is an example:

> **Example:** *Topic* – Egyptian Air disaster 990.
>
> *Sentence 1:* BOSTON (AP) – A Boeing 767 plane with 197 passengers aboard disappeared over the ocean about 60 miles south of Nantucket after taking off from New York's Kennedy International Airport, officials said Sunday. **(Novel)**
>
> *Sentence 2:* EgyptAir Flight 990 was headed to Cairo, Egypt, Coast Guard Lt. Rob Halsey said. **(Novel)**
>
> *Sentence 3:* There were 197 passengers on the flight, an EgyptAir official said. **(Not Novel)**

Sentence 1 is *novel* as it is the first sentence and provide novel information on the topic, and Sentence 2 is also *novel* as it provides new information as compared to sentence 1 (occurring higher up the order). But, Sentence 3 contains information (197 passengers on the flight) which is already present in Sentence 1, thus it is classified as *not novel*.

7.1.1 The Main Challenges of Novelty Prediction

Next, we discuss the main challenges associated with sentence-level novelty prediction.

- **Relevance-based filtering:** The task of novelty prediction is to find new information on the topic of inquiry. Without considering the topic of inquiry most of the information in the relevant documents seems to be novel, as each sentence provide some potential new information unless it is a complete duplicate. Thus, it becomes essential to avoid noise, and misleading sentences which are non-relevant and get wrongly classified as novel. The performance of novelty models is very sensitive to the presence of non-relevant sentences (Allan et al., 2003; Soboroff and Harman, 2005). Filtering of non-relevant sentences from the documents is a main challenge which affects the performance of novelty prediction models which we address in our work.

- **Handling partial duplicates:** Identifying complete duplicates or near duplicates where sentences are syntactically and lexically identical appears to be an easier task, and most of the general sentence comparison approaches (e.g. Cosine similarity) perform quite well (Tsai et al., 2010). The main challenge lies in identifying novel information from partial duplicates i.e. sentences which cover different sub-topics and discusses multiple aspects, where some aspects overlap with the previously occurring sentences and some aspects are new. The following is an example:

  Example: *Topic* – Egyptian Air disaster 990.

  Sentence 1 : BOSTON (AP) – A Boeing 767 plane with 197 passengers aboard disappeared over the ocean about 60 miles south of Nantucket after taking off from New York's Kennedy International Airport, officials said Sunday. (**Novel**)

  Sentence 2 : EgyptAir Flight 990 was headed to Cairo, Egypt, Coast Guard Lt. Rob Halsey said. (**Novel**)

  Sentence 3 : EgyptAir Flight 990, bound for Cairo, took off from New York's Kennedy International Airport early Sunday and went down in the ocean roughly 60 miles south of the Massachusetts island of Nantucket. (**Not Novel**)

Sentences 1 & 2 are novel as they both provide new information, while sentence 3 is not novel as it has: i) partial information which is overlapping with sentence 1 (took off from New York's Kennedy International Airport, and went down in the ocean roughly 60 miles south of the Massachusetts island of Nantucket.), and ii) partial information which is overlapping with sentence 2 (EgyptAir Flight 990, bound for Cairo).

A single sentence combining multiple novel aspects occurring separately in previous sentences is *not novel* as described in an example above. It is a challenging problem to automatically measure the extent of new or repetitive information among partial duplicates. We investigate this problem in this work.

- **Handling paraphrases and incorporating semantic similarity:** Similar information can be described and written in different ways commonly known as *paraphrasing*. A paraphrase is an alternative for expressing the same meaning with different words, in the same language (Ştefănescu et al., 2014). Effectively capturing and comparing paraphrases is another challenge while finding novel information among sentences. For example: "JFK international airport in New York" is same as "New York's Kennedy International Airport". And "plane with 197 passengers aboard" is similar to "197 passengers on the flight". We attempt to address this challenge of handling paraphrases by incorporating semantic similarity while comparing sentences.

- **Handling sentence length variations**: Another challenge in sentence-level novelty predictions lies with handling sentences of varying length. Longer sentences tend to discuss multiple aspects of the topic whereas shorter sentences tend to focus on one aspect of the topic. It is a complex problem to have a general model which can compare and score sentences of varying length effectively, for e.g. i) long sentence vs long sentence, ii) long sentence vs short sentence, and iii) short sentence vs short sentence. The following is an example:

  Example: *Topic* – Swissair crash Nova Scotia.

Sentence 1: So far, the massive search-and-rescue effort has found no survivors, Canadian police told a news briefing at Peggy's Cove, Nova Scotia. (**Novel**)

Sentence 2: Earlier, Swissair confirmed that there were no survivors from the crash. (**Not Novel**)

Sentence 3: We have no survivors. (**Not Novel**)

The count of unique words in sentence 1, 2 and 3 is 21, 11 and 4, respectively. Word overlap between sentence 1 & 2 (*no, survivors*) is 2, and word overlap between sentence 1 & 3 (*no, survivors*) is also 2. Average word overlap between sentence 1 & 2 is 2/11 = 0.18, and average word overlap between sentence 1 & 3 is 2/4 = 0.5 . Cosine similarity between sentence 1 & 2 is 0.0125, and cosine similarity between sentence 1 & 3 is 0.0625.

Example using two different sentence comparison approaches i) word overlap and ii) cosine similarity, show quite varying sentence similarity results. Sentence 2 & 3 both are not novel but an average word overlap scores (0.18 for Sentence 1 & 2, 0.5 for Sentence 1 & 3) and cosine similarity scores (0.0125 for Sentence 1 & 2, 0.0625 for Sentence 1 & 3) seems to show quite diverse similarity results. Thus it is challenging to perform effective sentence comparison when sentences are of varying length. We investigate this problem in our work.

7.1.2 Research Questions

To answer the research question: *Can we find novel information by comparing information within and across documents effectively?*, as discussed in Chapter 1, we divide our investigation into following sub-questions:

1: *How do different BOW-based distance metrics for sentence comparison perform for novelty prediction?*

We explore various BOW-based distance metrics for measuring sentence similarity. We experiment with Overlap similarity, Jaccard coefficient, Dice coefficient and Cosine similarity.

2: *How do different techniques for representing the semantics of words and sentences work for novelty prediction?*
We investigate word and sentence-based embedding techniques for detecting novel sentences. We learn different embeddings using our in-domain data and compare what kind of embedding method works well for novelty prediction.

3: *How does comparing information across sentences using syntactic information perform in comparison to complete sentence based comparison?*
Typically, a short sentence captures one aspect of the topic of inquiry, but comparing complete sentences can penalise shorter sentence when compared with a longer sentence capturing multiple aspects as discussed in Section 7.1.1. Thus we explore syntactic information based sentence comparison for novelty detection.

4: *How can we develop effective models that combine different approaches explored in question 1, 2 and 3 for novelty prediction?*
After individually exploring various methods, we investigate an approach that combines the different techniques explored for novelty prediction in questions 1, 2 and 3.

7.2 Methodology

In this work, we focus on unsupervised approaches to perform novelty prediction. To predict whether a sentence is novel or not, given previously occurring sentences, we compare each sentence in a document with all the sentences occurring above it and across all the sentences occurring in the document higher up the order. We measure degree of similarity between two sentences (S_i and S_j) and assign a novelty score to each sentence S_i as shown in Equation 7.1 where i indicates the sentence position in the ordered list of sentences given a ranked order of documents, and $1 \leq j \leq i-1$.

$$novelty\_score(S_i) = 1 - max(similarity(S_i, S_j)) \tag{7.1}$$

The motivation of Equation 7.1 is to calculate the highest degree of similarity which a sentence S_i shares with another sentence S_j occurring higher up the order. Thus 1 - max(similarity(S_i, S_j)), measures the extent of new information in Sentence S_i. A complete duplicate will have a similarity score of 1 and a novelty score of 0. Similar modelling of novelty scores has been successfully explored in earlier work by Allan et al. (2003); Zhang et al. (2003); Abdul-jaleel et al. (2004); Tsai et al. (2010); Tang et al. (2010).

Once each sentence is assigned a score in the range of [0-1], the next task lies in determining the threshold θ, as shown in Equation 7.2 to select novel sentences. All the sentences which score higher than the threshold θ, are classified as *novel* else they are classified as *not novel* as shown in Equation 7.2. Similar threshold based pruning is commonly applied in previous work on novelty prediction (Zhang et al., 2003; Abdul-jaleel et al., 2004; Tsai et al., 2010; Tang et al., 2010). The value of θ is empirically calculated by varying θ in the range of [0, 1], for the test collection.

$$\begin{aligned} S_i \geq \theta \qquad & \{Novel\} \\ S_i < \theta \qquad & \{Not\ Novel\} \end{aligned} \tag{7.2}$$

Our novelty prediction model makes the following assumptions:
1) We perform novelty prediction given a fixed ordering of documents for each topic. Documents can be ordered chronologically or based on scores of a relevance model. In this work, the dataset contains an initial ordering of the documents based on the published date of the articles (Soboroff and Harman, 2005).
2) Following Cutrell and Guan (2007) and Joachims et al. (2005) who found that users read from top to bottom while interacting with a SERP, our novelty prediction models assume that users read from top to bottom.

Next, we present our various methods investigated for predicting sentence-level novelty scores.

7.2.1 Baseline model

As discussed in Section 7.1.1, filtering non-relevant information is one of the main challenge of novelty prediction. We explore filtering of information using different threshold of relevance scores using our *best relevance prediction model (BestRelModel)* as described in Chapter 6. The *BestRelModel* is a BM25 model, where query expansion approach is performed using traditional pseudo relevance feedback (PRF) and global-centroid-based semantic expansion. For a given topic we score each sentence in a list of ranked documents using the BestRelModel. All the sentences that have a relevance score greater than the threshold ϕ, are included in the filtered collection.

We linearly varied the relevance threshold ϕ, in the range of [0-0.3] with an increment of 0.05 to filter out non-relevant content from a collection. For the baseline novelty model, all the sentences in the collection are considered as novel. Thus the baseline model has the highest recall of novel sentences as compared to all other methods and variations. We vary different relevance thresholds to prune the whole collection and compare their recall scores, to select the optimum value of ϕ, that we fix for further investigation on novelty prediction.

7.2.2 Bag-of-words (BOW) based distance metrics approach

Most of the earlier work on novelty detection investigated BOW based complete sentence comparison approaches. In the BOW approach, each sentence is represented as a bag of independent words disregarding the structure, and the word order of the sentence. Different distance metrics such as cosine similarity and jaccard coefficient, have been investigated and are commonly used for comparing sentences for novelty prediction (Tsai et al., 2010; Tang et al., 2010). We explore and analyse how these different distance metrics compare with each other and perform for the task of novelty detection using the filtered collection.

Given two sentences S1 and S2, after performing basic data processing such as stopword removal and stemming, S1 is represented as set of words A, and S2 as set of words B. The similarity between S1 and S2 is calculated using different distance metrics which are discussed next, where $|A|$ represents the number of words in set A, $|B|$ represents the number of words in set B, $|A \cap B|$ represents the number of common words in set A and B.

- **Jaccard coefficient** (Jaccard, 1901): Given two sets A and B jaccard coefficient is calculated as shown in Equation 7.3:

$$\frac{\mid A \cap B \mid}{\mid A \mid + \mid B \mid - \mid A \cap B \mid} \tag{7.3}$$

- **Dice coefficient** (Sørensen, 1948; Dice, 1945): Given two sets A and B dice coefficient is calculated as shown in Equation 7.4:

$$\frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \tag{7.4}$$

- **Sentence overlap:** Given two sets A and B the sentence overlap is calculated as shown in Equation 7.5:

$$\frac{\mid A \cap B \mid}{\mid A \mid} \tag{7.5}$$

- **Cosine similarity:** Given two sets A and B or sentence vectors $\vec{A}$ and $\vec{B}$, cosine similarity is calculated as shown in Equation 7.6 and 7.7, where i indicates a term in a sentence vector of dimension n. In our work we use set-based cosine similarity using BOW approach and vector-based cosine similarity for embeddings approach (described in the next section).

$$\frac{\mid A \cap B \mid}{\mid A \mid * \mid B \mid} \tag{7.6}$$

which is calculated as

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{7.7}$$

We present an example walk-through of these distance metrics.

Sentence 1: EgyptAir Flight 990 was headed to Cairo, Egypt, Coast Guard Lt. Rob Halsey said. **(NOVEL)**

Sentence 2: It originated in Los Angeles, according to EgyptAir officials at Cairo International Airport. **(NOVEL)**

After stopword removal and stemming, we represent sentences as a set of words.

Set 1: [“egyptair”,“flight”,“990”,“head”,“cairo”,“egypt”,
“coast”,“guard”,“lt” “rob”,“halsey”,“said”]

Set 2: [“origin”,“lo”,“angel”,“accord”,“egyptair”,“offici”,
“cairo”,“intern”,“airport”]

Intersection of set 1 and set 2: [“egyptair”,“cairo”]

Cosine level similarity score: 0.018

Jaccard coefficient score: 0.105

Dice coefficient score : 0.190

Sentence overlap score : 0.222

Using Equation 7.1, novelty score of sentence 2 as compared to sentence 1 is 0.982 using cosine similarity, is 0.895 using jaccard coefficient, 0.81 using dice coefficient and 0.778 using sentence overlap. All these different metrics varies in terms of how they perform sentence length normalisation (denominator of the Equations 7.3, 7.4, 7.5, 7.6, 7.7), while comparing two sentences.

7.2.3 Sentence comparison using embeddings

To address the challenge of handling paraphrases and incorporating semantic similarity for capturing similar information as discussed in Section 7.1.1, we use word and sentence level embeddings to detect novel information. Using embeddings for calculating textual similarities has been commonly used in recent years and our previous investigations using embeddings for semantic similarity have shown good results (Arora et al., 2015b) as discussed in Chapter 5. On similar lines we explore embeddings for finding sentence similarity which we discuss next.

Using word embeddings (WE)

In this approach, each word in a sentence is represented as a vector of D dimensions also know as embeddings as discussed in Chapter 4. For each word (w_k) in a sentence (S_i) we combine their vector representation to form a centroid vector i.e $\overrightarrow{CV_i}$ $= \sum_{k=1}^{n} \overrightarrow{w_{\mathrm{k}}}$, where n represents the number of words in S_i. To calculate similarity between two sentences S_i and S_j as shown in Equation 7.1, we compare the centroid vector $\overrightarrow{CV_i}$ with $\overrightarrow{CV_j}$. We calculate cosine similarity between centroid vectors to find the amount of overlapping information across sentences. We explore continuous bag of words (Cbow) and continuous skip-gram (Cskip) based approaches as discussed in Chapter 4 for learning word embeddings for novelty prediction.

Using sentence embeddings (SE)

In this approach, instead of combining each word embedding to represent a centroid vector for a sentence we learn sentence embeddings directly trained from the corpus in an unsupervised way. Each sentence S_i is represented as a vector of D dimension, $\overrightarrow{S_i}$. We explore distributed bag of words (DBOW) and distributed memory model (DMM) based approaches as discussed in Chapter 4 for learning sentence embeddings. To calculate similarity between two sentences S_i and S_j as shown in Equation 7.1, we compare cosine similarity between the sentence vectors $\overrightarrow{S_i}$ and $\overrightarrow{S_j}$. We hypothesise that the sentence embedding learnt from the corpus capture sentence similarities in a more effective manner than word embeddings approach. Since sentence embedding are trained using the corpus rather than averaging individual words vector representation, we anticipate that they would incorporate the contextual information effectively and perform better than simply averaging the word-vectors in an ad hoc manner.

7.2.4 Syntactic Information

As discussed in Section 7.1.1, one of the main challenges of novelty detection is finding new information among sentences which have a partial overlap of information. Approaches focusing on comparing complete sentences using BOW and embeddings might not be that effective when there is a partial overlap of information. Thus, we perform syntactic processing of sentences and use NLP cues and markers (parts of speech, phrases) to compare sentences.

The meaning of a sentence is made up of not only the meanings of its individual words, but also the structural way the words are combined (Oliva et al., 2011). For comparing semantic similarity of sentences previous work have explored syntax based features, focusing on comparing *nouns*, *verbs*, *noun phrases*, *dependency relations* between two sentences and have obtained good results (Ştefănescu et al., 2014; Oliva et al., 2011) as compared to using cosine similarity approach with bags-of-words. Earlier work on sentence level novelty detection explored named entities and part of speech (POS) information (Schiffman and McKeown, 2004; Abdul-jaleel et al., 2004; Li and Croft, 2005) as discussed in Chapter 2.

We parse each sentence in the document collection using bllip parser (details described later in Section 7.3.2) and extract three types of syntactic features: parts-of-speech information, noun phrases (NP) information, sentence segments (Segm) information which we describe next. An illustration of these features extraction is provided in Example 7.2.4.

We explore following three syntactic features for novelty prediction in this work.

1) Using parts-of-speech: For calculating similarities between two sentences S_i and S_j, instead of comparing complete sentences we extract all the nouns and verbs from the two sentences and represent them as sets, then we compare similarity between the set of all nouns from S_i and S_j, and the set of all verbs from S_i and S_j. We explore BOW and embedding models for comparing similarity between nouns ($Nouns_i$ and $Nouns_j$) and verbs ($Verbs_i$ and $Verbs_j$) for sentence S_i and S_j. Equation 7.8, describes how we calculate the similarity score between sentence S_i

and S_j.

$$similarity(S_i, S_j) = \frac{(similarity(Nouns_i, Nouns_j)) + (similarity(Verbs_i, Verbs_j))}{2} \tag{7.8}$$

2) Using noun phrases: Instead of comparing complete sentences S_i and S_j, we compare similarity between the noun phrases (NP) extracted from the sentences S_i and S_j. Equation 7.9, describes how we calculate the similarity score between sentence S_i and S_j, where m and n represents the number of NP phrases in S_i and S_j, respectively. We explore BOW and embedding models while comparing NP similarities between sentence, where the latter capture phrases which might have different lexical items but are semantically similar.

$$similarity(S_i, S_j) = \frac{\sum_{z=1}^{m} max \sum_{t=1}^{n} (similarity(NP_z, NP_t))}{m} \tag{7.9}$$

As shown in Equation 7.9, each NP phrase in S_i (NP_{S_i}) is compared with all the NP phrases in S_j (NP_{S_j}), and the highest similarity score between each NP_{S_i} and all NP_{S_j} is added to calculate the overall sentence similarity score between S_i and S_j.

3) Using sentence segments: Sentence segmentation is the process of dividing a sentence into elementary units, which may be clauses or phrases from which a sentence tree is constructed (Tofiloski et al., 2009). These elementary units (Segm) can be compared for finding sentence similarity which we investigate in our work. We used the parser output to perform sentence segmentation using the following sentence markers: *S*, *Sbar*, *SQ*, *SInv*, *SBARQ*, *SInvQ*. These sentence markers are used as a baseline system for sentence segmentation (Tofiloski et al., 2009). For each parsed sentence we split the sentence into multiple segments depending on the presence of these sentence markers. Instead of comparing complete sentences, we compare segments across two sentences to find similar information effectively. As before we explore BOW and embedding models. Equation 7.10, describes how we calculate the similarity score between sentence S_i and S_j, where m and n represents the number of segments in S_i and S_j, respectively.

$$similarity(S_i, S_j) = \frac{\sum_{z=1}^{m} max \sum_{t=1}^{n} (similarity(Segm_z, Segm_t))}{m} \tag{7.10}$$

As shown in Equation 7.10, similar to the noun phrases approach, each $Segm$ in S_i ($Segm_{S_i}$) is compared with all the $Segm$ in S_j ($Segm_{S_j}$), and the highest similarity score between each segments $Segm_{S_i}$ and all $Segm_{S_j}$ is added to calculate the overall sentence similarity score between S_i and S_j.

Next, we present an example walk-through of our different methods using syntactic information for sentence comparison.

Sentence 1 : EgyptAir Flight 990 was headed to Cairo, Egypt, Coast Guard Lt. Rob Halsey said.

Sentence 2 : There were 197 passengers on the flight, an EgyptAir official said.

Figure 7.1 and Figure 7.2 represents the syntactic parse tree for Sentence 1 and Sentence 2. After performing sentence parsing we extract different information for comparing sentences, which is discussed next.

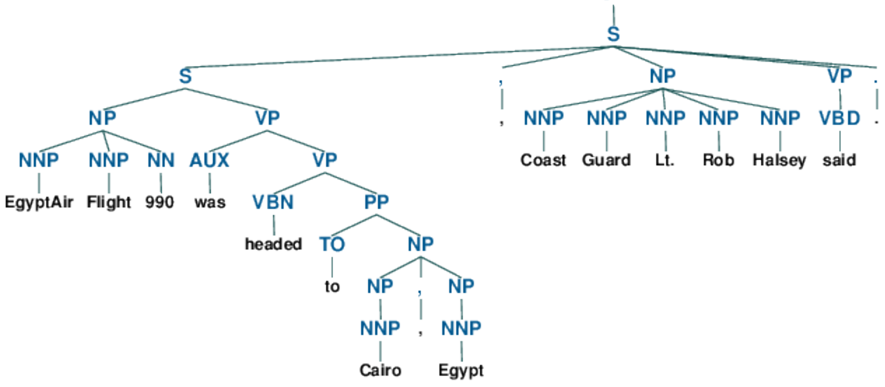


Figure 7.1: Syntactic parsed tree for sentence 1

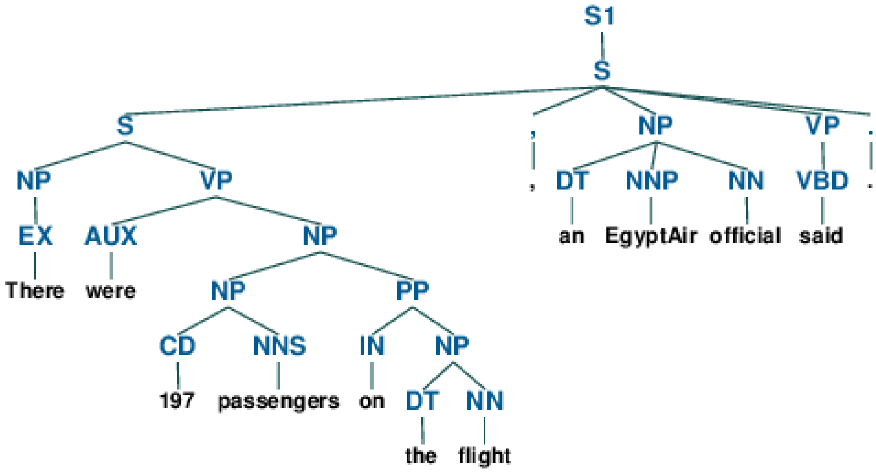


Figure 7.2: Syntactic parsed tree for sentence 2

Parts of speech example

Set of nouns for sentence 1: {"EgyptAir","Flight","990","Cairo", "Egypt","Coast","Guard","Lt.","Rob","Halsey"}

Set of verbs for sentence 1: {"said","headed"}.

Set of nouns for sentence 2: {"passengers","flight","EgyptAir","official"}

Set of verbs for sentence 2: {"said"}

As described earlier, we use Equation 7.8 for calculating similarity between the set of nouns and verbs from sentence 1 and sentence 2 respectively, using BOW based distance metric and word embeddings approach.

Noun phrases example

Noun Phrases for Sentence 1: {"EgyptAir Flight 990","Cairo Egypt","Coast Guard Lt. Rob Halsey"}

Noun Phrases for Sentence 2: {"There","197 passengers on the flight","an EgyptAir official"}

As described earlier, we use Equation 7.9 for calculating similarity between the noun phrases for sentence 1 and sentence 2, using BOW based distance metric and word embeddings based approach. In some cases there are bigger noun phrases which have multiple smaller noun phrases within it, as in Sentence 2 in Example 7.2.4. The noun phrase "197 passengers on the flight" has two smaller noun phrases within it, "197 passengers" and "the flight", in such cases we use the bigger NP phrase while performing sentence comparison. We explored using the smaller noun phrase units but the results were poor, we speculate that using the smaller noun phrases reduces the sentence to a bag-of-words representation consisting of only Nouns, and hence does not use the syntactic information from the sentences effectively.

Sentence segments example

Segments for Sentence 1: {"EgyptAir Flight 990 was headed to Cairo Egypt", "Coast Guard Lt Rob Halsey said."}

Segments for Sentence 2:{"There were 197 passengers on the flight", "an EgyptAir official said."}

As described earlier, we use Equation 7.10 for calculating similarity between the segments for sentence 1 and sentence 2, using BOW based distance metric and embeddings based approach.

7.2.5 Combination Model

Previous studies (Tang et al., 2010; Tsai et al., 2010) combined different distance metrics (cosine similarity, jaccard similarity) and showed that a combined method seems to be more robust. They found that combined models perform better in general for different sub-collections, which were obtained by dividing topics based on the threshold of the novelty content. In our work, we hypothesise that our methods capture different complementary signals and a combined model focusing on the combination of the best of each method will perform better for the task of novelty detection. We explore combining the output of the best of each model discussed in Section 7.2.2 (BOW), 7.2.3 (Embeddings) and 7.2.4 (Syntactic filtering). The output of a novelty model is a set of novel sentences and we calculate the intersection of each model output to select the combined set of novel sentences.[1] In a combination approach a sentence will be classified as *novel* if a sentence is classified *novel* by each of the individual approaches being combined together.

7.3 Experimental Setup

In this section we discuss the main tools, resources, data sets used for our experiments and evaluation metrics for novelty prediction.

7.3.1 Datasets

We use the standard TREC Novelty track dataset (Soboroff and Harman, 2005) for building and evaluating novelty detection models because it contains sentence level novel annotation (**New** or **Not New**) for a set of sentences from a set of ranked relevant documents on a given topic. We focus on the topics from the TREC 2003 and 2004 Novelty track. The topics from these tracks consist of events and opinionated topics. This track uses collections from the AQUAINT corpus, which consists of newswire text data in English, drawn from three sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service.

The TREC novelty track had two separate tasks for novelty prediction.

- **Task1 – Using Gold collection:** Given all the relevant sentences from the collection for the topic of inquiry, predict the set of novel sentences for each topic.

- **Task2 – Using Complete collection:** Given all the sentences from the collection, predict the set of novel sentences for each topic. In this task, a system has to first determine whether a sentence is topically relevant and then predict if it is novel or not.

[1] We investigated the union of each model output but the results were poor, thus we only explore in detail the intersection of each model output.

Table 7.1 details the sentence distribution across the 2003 and 2004 document collection. In the collection the amount of novel sentences is about 25.7% and 6.6% for 2003 and 2004 document set respectively. In the *gold collection* comprising of only the relevant sentences the amount of novel sentence is about 65.7% and 41.3% for 2003 and 2004 document set respectively. For both 2003 and 2004 collection the percentage distribution of novel sentences varies considerably across *complete* and *gold* collection. The 2004 document collection has less novel content than the 2003 collection.

Track	Topics	Documents	Sentences	Relevant Sentences	Novel Sentences
2003 track	50	1187	39820	15557	10226
2004 track	50	1214	52447	8343	3454

Table 7.1: 2003 and 2004 Novelty track data distribution

Using *gold collection* for novelty detection is artificial in nature, as it assumes that we know all the relevant sentences for a given topic, which in general does not happen. Thus in our work we investigate the complete collection in which documents comprise of sentences which are relevant as well as non-relevant. For the task of snippet generation, we are more interested in the complete collection, where we do not have the relevant sentence information a priori, and given all the information from a document have to generate snippets to represent in a SERP.

7.3.2 Tools

Next, we describe different tools and resources used for conducting our experiments.

- **Sentence Similarity**: We compute BOW based sentence similarity using our own implementation of different distance metrics as discussed in Section 7.2.2.
- **Embeddings**: We use the gensim (Rehurek and Sojka, 2011) implementation of Word2Vec and Paragraph vectors in our work to learn *word* and *sentence*

embeddings as discussed in Chapter 4 and Section 7.2.3. We use in-domain embeddings in our experiments, which are learnt using the AQUAINT document collection. We varied different parameter settings such as training method, dimension size, window size, for our experiments and compare the performance and effectiveness of these parameters for novelty detection.

- **Sentence Parser**: We use bllip parser (Charniak and Johnson, 2005) to parse sentences to extract different syntactic information such as POS, NP and sentence markers (S, SBAR, SINV) for performing sentence segmentation, as discussed in Section 7.2.4.

7.3.3 Evaluation Metrics

Novel sentences are returned as an unranked set in the novelty track. The number of novel sentences varies across topics and so precision, recall and F-score averaged across all topics were used as standard evaluation measures in TREC Novelty detection task (Soboroff and Harman, 2005). Similarly we measure precision, recall and f-measure for Novelty detection, as shown in Equation 7.11.

In general, precision and recall for a topic has an exponential decay relationship as shown in Figure 7.3. High values of precision is obtained for lower values of recall and high values of recall is obtained for lower values of precision. Thus it becomes hard to analyse results and compare either of these values. Thus, we select the best parameter settings and models based on the F-score. For models and settings which have similar values of F-score we compare them using precision values as we want the best model to have high F-scores and also have high precision scores to avoid mis-classification of not novel sentences as novel, i.e avoid false positives, similar to Allan et al. (2003). We report all the scores at 3 decimal points, which are averaged over all the topics in the dataset.

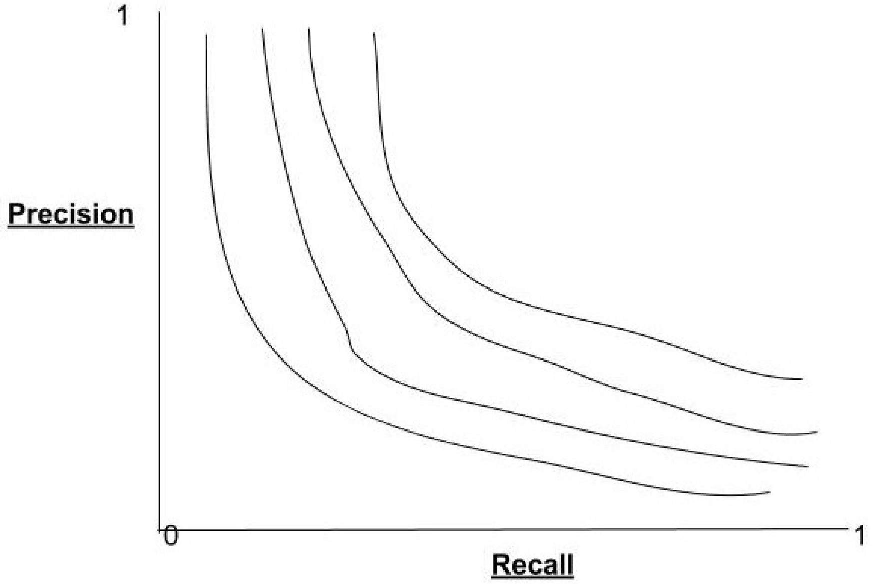


Figure 7.3: Precision and Recall variation

$$
\begin{aligned}
Precision &= \frac{novel\ sentences\ retrieved}{retrieved\ sentences}\\
Recall &= \frac{novel\ sentences\ retrieved}{novel\ sentences}\\
F\text{-}score &= \frac{(1+\beta^2) * (Precision * Recall)}{(\beta^2 * Precision) + Recall}\\
F\text{-}score &= \frac{2 * Precision * Recall}{Precision + Recall} \quad (\beta = 1)
\end{aligned}
\tag{7.11}
$$

7.3.4 Data pre-processing

Next, we discuss an initial processing done at sentence level.

Stopword Removal and Stemming: We perform stopword removal and stemming using the NLTK toolkit (Bird and Loper, 2004).

Word Embeddings: Similar to our investigation for sentence retrieval using embeddings, we learn different in-domain embeddings using the AQUAINT corpus after performing stopword removal and stemming. We vary embeddings training algorithm (CBOW and CSKIP), embeddings size (100, 200, 300) and window size (5, 10), these values of embeddings and window size are commonly used for IR and textual similarity experiments.

Details regarding the embedding training algorithm are discussed in Chapter 4. Overall, we have 12 different in-domain embeddings (2 algorithms * 2 window size * 3 embedding size).

Sentence Embeddings: Similar to word embeddings, we learn different in-domain embeddings using AQUAINT corpus after performing stopword removal and stemming. We vary embeddings training algorithm (DBOW, DMM Mean and DMM Concat), embeddings size (100, 200, 300) and window size (5, 10) for learning sentence embedding in this work, these values of embeddings and window size are commonly used for IR and textual similarity experiments. Details regarding the embedding training algorithm are discussed in Chapter 4. Overall, we have 18 different in-domain embeddings (3 algorithms * 2 window size * 3 embeddings size) that we explore in our work.

Sentence Parsing: Parsing of sentences is done on the raw corpus. We extract POS and NP based information from the parsed sentence. We perform sentence segmentation on the parser output to extract sentence segments. While comparing segments, noun phrases, nouns and verbs information between sentences we perform stemming and stopwords removal.

7.4 Results

All methods for novelty prediction discussed in Section 7.2, assign a novelty score to each sentence. The main challenge lies in determining the threshold θ for novelty prediction to determine whether a sentence is novel or not. We linearly vary the novelty threshold θ in the range of [0-1] with an increment of 0.05 to determine the optimum threshold that performs best across both document collection. For word embeddings we vary the novelty threshold θ in the range of [0-0.10] with an increment of 0.01 .

7.4.1 Baseline Model

As discussed in Section 7.2.1, we perform filtering of complete collection by varying different values of relevance threshold ϕ. We select the best relevance model (*BestRelModel*) from Chapter 6. Table 7.2 presents the result of varying relevance threshold for selecting the baseline model. We see a big increment in precision and F-score when the collection is pruned using a relevance threshold as compared to using a raw corpus. We select the relevance threshold $\phi = 0.25$ for filtering information, and removing non-relevant content from the collection as our **baseline model**. After $\phi = 0.25$, recall decreases below 0.8 for both the collection, so to ensure that we do not miss much novel content (i.e. have a reasonable good recall) and comparatively strong F-score as compared to scores using the raw corpus, we use pruned corpus at ϕ=0.25, as our baseline model. For all following experiments we use this pruned collection for comparing the performance of different novelty models discussed in Section 7.2.

	Track 2003			Track 2004		
Relevance Pruning	Precision	Recall	F score	Precision	Recall	F score
Raw Corpus	0.270	1.00	0.394	0.081	1.00	0.144
Pruned Corpus, $\phi = 0.10$	0.366	0.866	0.470	0.142	0.900	0.233
Pruned Corpus, $\phi = 0.15$	0.366	0.865	0.470	0.142	0.900	0.233
Pruned Corpus, $\phi = 0.20$	0.368	0.855	0.470	0.143	0.884	0.233
Pruned Corpus, ϕ **= 0.25**	**0.373**	**0.831**	**0.470**	**0.145**	**0.846**	**0.234**
Pruned Corpus, $\phi = 0.30$	0.379	0.775	0.463	0.148	0.794	0.235

Table 7.2: Results of novelty model for relevance based pruning. Baseline model scores are in boldface.

7.4.2 Bag-of-words (BOW) based distance metrics approach

In this section, we discuss the results obtained by BOW-based distance metrics for novelty prediction as discussed in Section 7.2.2. Table 7.3 present results for both 2003 and 2004 data collections. The different distance metrics performance varies in the order of Jaccard coefficient > Dice Coefficient > Cosine Similarity > Sentence

Overlap. Cosine similarity model has the highest recall, and jaccard coefficient model has the highest precision across both collections. Similar findings were reported in earlier work by Tsai et al. (2010) though only for the gold collection. They found that cosine similarity leads to high-recall systems and jaccard similarity leads to high-precision systems. We select *Jaccard coefficient* as the best BOW model for further comparison. We also use the *Jaccard coefficient* in the syntax-based approach.

	Track 2003			Track 2004		
Bag-of-words approach	Precision	Recall	F score	Precision	Recall	F score
Baseline (relevance only)	0.373	0.831	0.470	0.145	0.846	0.234
Overlap, $\theta = 0.20$	0.417	0.767	0.489	0.156	0.754	0.242
Jaccard, $\theta = 0.65$	**0.437**	**0.781**	**0.505**	**0.165**	**0.767**	**0.252**
Dice, $\theta = 0.45$	0.434	0.792	0.505	0.163	0.783	0.252
Cosine, $\theta = 0.40$	0.429	0.795	0.504	0.161	0.783	0.249

Table 7.3: Results of bag-of-words based different distance metrics for novelty prediction. Best model is in boldface.

7.4.3 Embedding Results

Next, we present results of our experiments using embedding based sentence comparison approaches as discussed in Section 7.2.3.

Word embeddings

The word embedding results are shown in Table 7.4. We tried all 12 embedding combinations, and found that in general the Cbow approach works slightly better than the Cskip model. However, similar results are obtained for each algorithm irrespective of varying embeddings size, window size. We speculate two reasons for similar results: i) small size of the corpus, ii) the effect of subtle differences between embeddings is diminished because all the words are averaged in a sentence to represent a centroid vector.

Using word embeddings for sentence comparison leads to high recall scores across both the datasets but the precision and F-scores are relatively lower than the BOW models. The best results using word embeddings are obtained using *Cbow* model with an embedding size of 200 and window size of 10. We use this embedding configuration, *Cbow-200-10*, in our syntax-based models.

	Track 2003			Track 2004		
Embedding Name	Precision	Recall	F score	Precision	Recall	F score
Baseline	0.373	0.831	0.470	0.145	0.846	0.234
Cbow-200-10	**0.421**	**0.804**	**0.501**	**0.155**	**0.829**	**0.246**
Cbow-200-5	0.424	0.784	0.499	0.155	0.828	0.246
Cbow-100-10	0.417	0.818	0.501	0.155	0.828	0.246
Cbow-100-5	0.418	0.818	0.501	0.155	0.828	0.246
Cbow-300-10	0.418	0.818	0.501	0.155	0.829	0.246
Cbow-300-5	0.418	0.817	0.502	0.155	0.828	0.246
Cskip-200-10	0.417	0.818	0.501	0.155	0.828	0.246
Cskip-200-5	0.418	0.818	0.501	0.156	0.827	0.246
Cskip-100-10	0.417	0.818	0.501	0.157	0.823	0.247
Cskip-100-5	0.417	0.818	0.501	0.157	0.823	0.247
Cskip-300-10	0.417	0.818	0.501	0.155	0.829	0.246
Cskip-300-5	0.418	0.817	0.501	0.155	0.828	0.246

Table 7.4: Results of word embedding based sentence comparison approach for novelty prediction. Best model is in boldface. $\theta = 0.02$, for all different types of embeddings. Embedding Name: Algorithm-Dimension-WindowSize

Sentence embeddings

Table 7.5 present results of using sentence embeddings. We tried all 18 embedding combinations, and found that in general DBOW approach works quite better than both the DMM Mean model and the DMM Concat model. Overall, varying the algorithm leads to quite varying results as compared to varying the embeddings size and window size. The best results using sentence embeddings are obtained using the *Dbow* model with an embedding size of 100 and window size of 10.

Using sentence embeddings for sentence comparison leads to overall high precision, recall and F-scores across both the datasets. Results are better than using word embeddings, and are relatively better or similar to the BOW based Jaccard coefficient model. As we hypothesised sentence embedding learnt from the corpus seems to capture sentence similarities in a more effective manner than word embeddings approach and shows relatively quite better results for novelty prediction.

	Track 2003			Track 2004		
Embedding Name	Precision	Recall	F score	Precision	Recall	F score
Baseline	0.373	0.831	0.470	0.145	0.846	0.234
Dbow-200-10	0.434	0.807	0.510	0.162	0.809	0.252
Dbow-200-5	0.433	0.811	0.510	0.161	0.813	0.252
Dbow-100-10	**0.437**	**0.803**	**0.512**	**0.163**	**0.802**	**0.252**
Dbow-100-5	0.435	0.806	0.511	0.162	0.808	0.253
Dbow-300-10	0.434	0.810	0.511	0.161	0.811	0.251
Dbow-300-5	0.432	0.811	0.510	0.161	0.813	0.252
DMM Concat-200-10	0.373	0.831	0.470	0.145	0.846	0.234
DMM Concat-200-5	0.373	0.830	0.470	0.145	0.845	0.234
DMM Concat-100-10	0.373	0.831	0.470	0.145	0.846	0.234
DMM Concat-100-5	0.374	0.830	0.470	0.145	0.841	0.234
DMM Concat-300-10	0.373	0.831	0.470	0.145	0.846	0.234
DMM Concat-300-5	0.373	0.830	0.470	0.145	0.844	0.234
DMM Mean-200-10	0.429	0.800	0.505	0.161	0.801	0.251
DMM Mean-200-5	0.426	0.808	0.505	0.160	0.813	0.250
DMM Mean-100-10	0.426	0.804	0.504	0.160	0.810	0.250
DMM Mean-100-5	0.426	0.811	0.506	0.159	0.815	0.249
DMM Mean-300-10	0.430	0.797	0.505	0.162	0.801	0.252
DMM Mean-300-5	0.427	0.805	0.505	0.160	0.811	0.250

Table 7.5: Results of sentence embedding based sentence comparison approach for novelty prediction. Best model is in boldface. θ = 0.2, for Dbow embeddings, θ = 0.25, for DMM Mean embeddings, and θ = 0.05, for DMM Concat embeddings. Embedding Name: Algorithm-Dimension-WindowSize

7.4.4 Syntactic Information Results

Table 7.6 show results for syntactic information as discussed in Section 7.2.4 for both TREC 2003 and 2004 datasets. The POS-based method comparison using only nouns and verbs across sentences for novelty prediction perform well and similar to the BOW based Jaccard coefficient approach for complete sentence comparison as shown in Table 7.7. For the POS-based approach, comparing using BOW works better than comparing word embeddings. For the POS-based approach using BOW based Jaccard coefficient, the precision values slightly decrease and the recall values slightly increase as compared to the BOW-based complete sentence comparison.

Comparing sentences using only NP information does not seem to work well as compared to BOW-based Jaccard coefficient approach. We speculate that too much important information (e.g. verbs) might be filtered out using this approach.

Performing sentence segmentation using markers like S, SBAR etc as discussed in Section 7.2.4, performs best while comparing sentences for novelty prediction. Using BOW comparison seems to outperform embeddings comparison. Comparing sentence segments performs better than comparing noun phrases, nouns and verbs and better than the BOW-based Jaccard coefficient approach.

	Track 2003			Track 2004		
Method Used	Precision	Recall	F score	Precision	Recall	F score
Baseline	0.373	0.831	0.470	0.145	0.846	0.234
BOW POS, $\theta = 0.6$	**0.434**	**0.786**	**0.505**	**0.164**	**0.779**	**0.252**
Cbow POS, $\theta = 0.05$	0.422	0.607	0.447	0.159	0.797	0.249
BOW NP, $\theta = 0.3$	0.426	0.792	0.501	0.162	0.789	0.251
Cbow NP, $\theta = 0.05$	0.425	0.785	0.500	0.155	0.822	0.246
BOW Segments, $\theta = 0.65$	**0.439**	**0.787**	**0.509**	**0.164**	**0.791**	**0.254**
Cbow Segments, $\theta = 0.35$	0.402	0.803	0.488	0.152	0.832	0.242

Table 7.6: Results of syntax-based sentence comparison approaches for novelty prediction. Best models using parts of speech (POS) and segmentation approach (Segments) are in boldface. NP indicates results for noun phrases based approach. BOW indicates the Jaccard coefficient model, Cbow indicate the word embedding based model using the Cbow-200-10 configuration.

	Track 2003			Track 2004		
Filtered Task-1	Precision	Recall	F score	Precision	Recall	F score
Baseline	0.373	0.831	0.470	0.145	0.846	0.234
Jaccard, $\theta = 0.65$	0.437*	0.781	0.505*	**0.165***	0.767	0.252*
Dbow-100-10, $\theta = 0.2$	0.437*	0.803$^{\delta}$	**0.512**$^{*\delta}$	0.163*	0.802$^{\delta}$	0.252*
BOW POS, $\theta = 0.6$	0.434*	0.786$^{\delta}$	0.505*	0.164*	0.779$^{\delta}$	0.252*
BOW Segments, θ =0.65	**0.439***	0.787	0.509$^{*\gamma}$	0.164*	0.791$^{\delta}$	**0.254***

Table 7.7: Best results for different models investigated for novelty prediction, the best scores are in boldface. $*$ indicates that the difference in the results compared to the baseline is statistically significant with $p<0.01$, δ, and γ indicates that the difference in the results compared to the Jaccard approach is statistically significant with $p<0.01$, and $p<0.05$ respectively using student's t-test. BOW POS, BOW Segments indicate results for BOW based parts-of-speech and segments based comparison approaches respectively.

7.4.5 Combination approach

Sentence embedding and syntax-based sentence comparison results are statistically significantly better than the BOW-based Jaccard coefficient approach for complete sentence comparison as shown in Table 7.7. Next, we used the best settings from each method to combine the rich information captured by each model.

Table 7.8 shows the result for different combination methods explored. Best results are obtained using a combination approach comprising *sentence embeddings* and comparing *sentence segments* using BOW based Jaccard coefficient. This best model will be referred as **BestNovelModel** for further discussion. All the combination approaches as shown in Table 7.8 perform better than the baseline scores and using the individual approaches. All the combination approach results are statistically significantly better than the baseline and the BOW based Jaccard coefficient approach as shown in Table 7.8. As we hypothesised it seems all these approaches capture different signals for detecting novel sentences thus a combined model taking the intersection of each model seems to perform well.

7.5 Discussion

We explored different types of sentence comparison techniques for determining sentence level novelty prediction. Similar to prior results by Tsai et al. (2010), we found that cosine similarity lead to high recall and jaccard coefficient lead to high precision values. Though Tsai et al. (2010) only explored the gold set comprising all relevant sentences, it seems a similar trend of cosine similarity leading to high recall and jaccard coefficient leading to high precision is observed for novelty prediction over the complete collection.

Further, we investigated and hypothesised that instead of comparing complete sentences it seems more apt to compare sentence segments, noun phrases, nouns and verbs information across sentences. The POS-based comparison seems to perform similar to the BOW-based comparison. We found that segmenting sentences

	Track 2003			Track 2004		
Filtered Task-1	Precision	Recall	F score	Precision	Recall	F score
Baseline	0.373	0.831	0.470	0.145	0.846	0.234
Jaccard, $\theta = 0.65$	0.437^{*}	0.781	0.505^{*}	0.165^{*}	0.767	0.252^{*}
All combined	$0.457^{*\delta}$	0.749	$0.510^{*\gamma}$	$0.171^{*\delta}$	0.730	$0.256^{*\gamma}$
BOW + BOW POS + Dbow	$0.446^{*\delta}$	0.770	$0.508^{*\gamma}$	$0.168^{*\delta}$	0.754	$0.255^{*\delta}$
BOW + BOW Segments + Dbow	$0.455^{*\delta}$	0.756	$0.511^{*\delta}$	$0.170^{*\delta}$	0.737	$0.255^{*\gamma}$
BOW + Dbow	$0.444^{*\delta}$	0.777	$0.509^{*\delta}$	$0.167^{*\delta}$	0.762	$0.254^{*\delta}$
BOW POS + Dbow	$0.444^{*\delta}$	0.781	$0.510^{*\delta}$	$0.167^{*\delta}$	0.774	$0.256^{*\delta}$
BOW Segments + Dbow	$\mathbf{0.451}^{*\delta}$	**0.774**	$\mathbf{0.514}^{*\delta}$	$\mathbf{0.168}^{*\delta}$	**0.772**	$\mathbf{0.256}^{*\gamma}$
BOW Segments + BOW POS	$0.451^{*\delta}$	0.762	$0.510^{*\gamma}$	$0.169^{*\delta}$	0.752	$0.256^{*\gamma}$
BOW Segments + BOW	$0.451^{*\delta}$	0.759	$0.509^{*\delta}$	$0.168^{*\delta}$	0.741	0.254^{*}
BOW + BOW POS	$0.440^{*\delta}$	0.773	0.505^{*}	$0.166^{*\delta}$	0.757	$0.254^{*\gamma}$

Table 7.8: Results for combination of best models investigated for novelty prediction, the best combination model is in boldface. $*$ indicates that the difference in the results compared to the baseline is statistically significant with $p<0.01$, δ, and γ indicates that the difference in the results compared to the Jaccard approach is statistically significant with $p<0.01$, and $p<0.05$ respectively using student's t-test. BOW indicates Jaccard coefficient and Dbow indicates sentence embedding based complete sentence comparison approaches for novelty prediction. BOW POS, BOW Segments indicate Jaccard coefficient based parts-of-speech and segments based sentence comparison models respectively. All Combined model = BOW + BOW POS + BOW Segments + Dbow

using sentence markers such as *S, SBAR, SQ* and comparing segments between sentences seems to work reasonably well. Using sentence segments seems to work better than using BOW-based complete sentence comparison. Using BOW-based segment matching seems to work better than using embedding based segment matching.

We discuss limitations of this work and present important directions for the future work.

1) Sentence comparison: We perform individual sentence based comparison for determining novelty prediction. We do not handle the cases where information is spread across multiple sentences occurring higher up the order. We revisit our previous example discussed in Example 7.1.1.

Example: *Topic* – Egyptian Air disaster 990.

Sentence 1 : BOSTON (AP) – A Boeing 767 plane with 197 passengers aboard disappeared over the ocean about 60 miles south of Nantucket after taking off from New York's Kennedy International Airport, officials said

Sunday. (**Novel**)

Sentence 2 : EgyptAir Flight 990 was headed to Cairo, Egypt, Coast Guard Lt. Rob Halsey said. (**Novel**)

Sentence 3 : EgyptAir Flight 990, bound for Cairo, took off from New York's Kennedy International Airport early Sunday and went down in the ocean roughly 60 miles south of the Massachusetts island of Nantucket. (**Not Novel**)

Sentence 3, consist of information already covered in Sentence 1 and Sentence 2. Thus comparing sentences which cover different sub-topics and discusses multiple aspects, where some aspects overlap with the previously occurring sentences and some aspects are new is a complex challenge. We anticipate that approaches which goes beyond sentence-level comparison and combine information from all the sentences occurring previously can further improve novelty performance. However, how to combine information from previously occurring sentences is a research challenge and is worth pursuing.

2) Combination approach: We combined the intersection of the output list of different models. An intersection of the output of different models seems to be effective and show consistently positive results for both collections as shown in Table 7.8. We do not explore other alternative approaches for combining multiple signals captured using different models but is worth pursuing in future.

7.6 Summary and Conclusion

We explored different distance metrics similar to the earlier work by Allan et al. (2003); Tsai et al. (2010), and novel techniques of using embeddings and syntactic information in sentence comparison for novelty prediction. Our method show using syntactic cues and embeddings techniques for sentence comparison improve novelty performance. We explored a combination of different models which captures complementary signals for novelty prediction. Combination approach performs quite well and showed statistically significantly better results than the baseline and distance metrics-based approach for both 2003 and 2004 document collection. In this chapter, we proposed novel methods to perform sentence-level novelty detection. Our proposed models perform better and are effective than commonly used bag-of-words approach for novelty detection. Our best results are obtained using a combination approach (**BestNovelModel**) comprising of *sentence embedding* comparison and comparing *sentence segments* using a BOW based Jaccard coefficient approach. We use this **BestNovelModel** for generating document snippets to be presented in a SERP.

Next, in Chapter 8, we discuss the task of sentence-level readability prediction, and the task of snippet generation combining relevance, novelty and readability output to generate snippets to be shown to the users in a web search.

Chapter 8

Snippet Generation

Document snippets presented in a SERP are intended to assist users to identify retrieved information which may be useful to them in satisfying their information need. Readability of snippets is a key factor that influences the user experience and their search behaviour. In this work we aim to generate snippets which are easy to read. Thus we explore sentence-level readability scores prediction for snippet generation in this chapter. First, we describe the readability prediction model that we use in this work. Next we discuss different combination approaches explored to combine the output of *relevance*, *novelty* and *readability* model to form snippets to be shown to the users. Then we present the measures used for evaluating different document snippets. We report the results of our evaluation to select the best combination approaches for representing snippets in a SERP. Finally we conclude with the main findings of our snippet generation framework.

8.1 Readability Prediction

The task of readability prediction investigates the ease of reading textual content. This is typically based on analysis of features such as combination of counts of words, syllables, characters and sentences in a piece of a text as reviewed in Chapter 2. Next, we discuss unsupervised and supervised approaches which have been commonly used for calculating the readability scores.

8.1.1 Unsupervised models

Unsupervised approaches rely on simple features such as count of words, syllables, characters in a document or the textual content for readability prediction. We discuss four different unsupervised approaches.

FOG index: The Fog Index (Gunning, 1952) estimates the years of formal education a person needs to understand the text on the first reading. The mathematical formula for calculating fog index is shown in Equation 8.1.

$$Grade\ Level = 0.4 * (ASL + PHW) \tag{8.1}$$

where ASL = Average Sentence Length (i.e., number of words divided by the number of sentences) and PHW = Percentage of Hard Words, where hard words are calculated by counting number of words which have more than 3 syllables. This formula was developed using empirical investigation from which the weighting factor 0.4 was also selected (Gunning, 1952). The FOG index is generally used for scoring a textual paragraph or a document which has more than 100 words.

Flesch Reading Ease Sores (FRES) and Flesch-Kincaid Grade Level (FKGL): FRES and FKGL are readability tests which were designed to indicate how difficult a passage in English is to understand (Kincaid et al., 1975). FRES output is a number ranging from 0 to 100. A higher FRES score indicates that the text is easier to read. FKGL was developed to map the output of the reading scores to US grade level. FRES and FKGL are calculated using the formulas shown in Equation 8.2.

$$\begin{gathered} FRES = 206.835 - (1.015 * ASL) - (84.6 * ASW) \\ FKGL = 0.39 * (ASL) + 11.8 * (ASW) - 15.59 \end{gathered} \tag{8.2}$$

where ASL = Average Sentence Length and ASW = Average number of syllables per word (i.e., the number of syllables divided by the number of words). The weights used in the FRES and FKGL formulas were determined by empirical investigations (Kincaid et al., 1975).

SMOG Readability Formula: SMOG index estimates the years of education a person needs to understand a piece of writing (Mc Laughlin, 1969).

$$SMOG\ Grade = 3 + \sqrt{Polysyllable\ Count} \tag{8.3}$$

Polysyllable Count is calculated as the sum of words with three or more syllables in three groups of sentences (Group 1, 2 and 3), even if the same word appears more than once. Group 1, 2 and 3 consist of 10 sentences in a row near the beginning, 10 sentences in the middle, and 10 sentences in the end respectively of a document. Thus SMOG grade works better when the document or textual information has more than 30 sentences.

Automated readability index: The automated readability index (ARI) is a readability test for English texts, designed to measure the understandability of a text (Senter and Smith, 1967). The formula used to calculate the automatic readability score is shown in Equation 8.4.

$$ARI = 4.71 * (ACW) + 0.5 * (ASL) - 21.43 \tag{8.4}$$

where ASL = Average Sentence Length and ACW = Average number of characters per word (i.e., the number of characters divided by the number of words). A higher ARI score indicates that the text is difficult to read.

8.1.2 Supervised models

Work on manually curated datasets for sentence and paragraph level readability scores led to exploration of more machine learning (ML) feature based approaches to readability prediction. As reviewed in Chapter 2, initial work by Kanungo and Orr (2009) explored supervised models for predicting readability scores for document snippets. They collected user judgements for 5000 document snippets and trained ML models using features comprising of different readability metrics such as the FOG index, FRES, FKGL and other textual features such as punctuation, capital words. Their ML based model comprising of multiple features showed substantially better correlation with user judgements as measured by Pearson's correlation coefficient, than unsupervised approaches (FOG index, FRES) commonly used for readability prediction.

8.1.3 Our Investigation

The limitation of supervised models lies in the need for development of readability datasets depending on the task and application. Creation of readability datasets involves collection of user judgements on the readability of a sentence or passage level text corpus, similar to the one developed by Kanungo and Orr (2009). The non-availability of an annotated corpus at readability level for document snippets and the reasonable performance of unsupervised approaches to attain readability measurements of the textual content (Kanungo and Orr, 2009) motivated us to use unsupervised models for sentence-level readability prediction in this work.

We performed a comparative manual analysis of different unsupervised models. Table 8.1 presents a number of example sentences with corresponding readability scores calculated using different unsupervised models. In general most of the unsupervised approaches work well for larger textual content and document-level information. Without the availability of any gold data it was hard to compare alternative models to determine the one that worked best. Based on a manual analysis of readability of sentences for two topics and about 50 sentences from each topic using alternative unsupervised models as shown in Table 8.1, we selected *FRES* model to explore for our work on snippet generation. We selected FRES model for our exploration because of two reasons: i) FRES model output is between [0-100] thus provide a more broader range to compare and differentiate sentences which are easy to read than the ones which are complex and difficult. ii) FRES has been quite popularly used for measuring the readability of web documents and snippets (Kanungo and Orr, 2009; Collins-Thompson et al., 2011)

In our work, we calculate the readability score for each sentence in the ranked set of documents using the FRES model which we refer as **BestReadModel** for further discussion. We use this **BestReadModel** for snippets generation in our work.

Sentence	ARI	**FRES**	FKGL	FOG	SMOG
Venter said in May that he would start sequencing the human genome next year and complete it by 2001	8.40	**89.60**	5.48	9.70	8.48
Celera's sequencing strategy is quite different from the safe and methodical approach of its rival.	12.45	**39.33**	11.50	19.34	15.25
The genome also contains a wealth of information about human evolutionary history and early migration patterns.	14.50	**10.82**	15.72	16.40	13.95
The DNA sequencing method won him his second Nobel Prize in 1980.	5.38	**88.90**	3.84	8.13	8.50

Table 8.1: Readability score output for the sentences related to the Topic: Human Genome Decoded. Bold values indicates the best readability model.

Next, we discuss our snippet generation method which combines relevance, novelty and readability scores of retrieved sentences to generate effective snippets.

8.2 Snippet Generation

To answer the research question: *How to combine sentence-level relevance, novelty and readability features to generate effective snippets?*, introduced in Chapter 1, we divide our investigation into following question:

RQ: *What combination of relevance, novelty and readability can be used to form the most effective document snippet for predicting the usefulness of a document for a given topic?*

8.2.1 Methodology

Combining different features comprising of relevance, novelty and readability scores to generate snippets and evaluating their effectiveness manually in a user-based setting is a complex task. Thus we divided our investigation of the development and evaluation of snippets into three stages: i) Development stage, ii) Pilot stage, and iii) User study stage. Evaluating all snippet combination approaches in a SERP would be very expensive (time-wise and cost-wise). Thus in our development phase our focus was on independent evaluation of document snippets generated using different combination approaches to identify the best snippet generation methods. The best two snippet combination approaches identified in this initial phase were used to generate snippets for a SERP. We then examined changes in the user behaviour, experience and knowledge gain when interacting with these SERPs. In this chapter we discuss only the development stage of our study, the other stages of the investigation are described in Chapter 9.

To perform manual and user-based evaluation of snippets generated using our framework, we selected six topics from the TREC 2003 and 2004 collection (Soboroff and Harman, 2005) as shown in Table 8.2. This same TREC collection was used for the sentence-level relevance and novelty experiments described in Chapter 6 and 7 respectively. As user studies are difficult to operate, and are cognitively intensive where interacting with each topic can take about 20-30 minutes, thus we focused on small number of topics (commonly done for IIR studies) to investigate and capture the variation in user behaviour and knowledge gain effectively. All the topics are exploratory and investigative in nature ensuring that they are interesting, simple and engaging for the users. Out of the six topics, we selected two topics (D1 and D2) for the development stage, two topics (L1 and L2) for the pilot stage and two topics (C1 and C2) for the final user study stage. We used separate topics for each stage to demonstrate that the snippet generation method being developed is not biased towards specific topics and can be easily adapted for other topics.

Experimental Phase	Title	Description
Development (D1)	Human Genome Decoded	Human genome decoded at NIH
Development (D2)	Snowmobiles Banned National Parks	Identify documents that express an opinion either for or against banning snowmobiles in National Parks.
Pilot study (L1)	First Human Hand Transplant	The first human hand transplant in the United States was performed on Matthew Scott on January 25, 1999.
Pilot study (L2)	Microsoft Antitrust Charges	What are opinions on Microsoft's guilt or innocence on charges of antitrust?
User study (C1)	Clone Dolly Sheep	Cloning of the sheep Dolly
User study (C2)	Nobel Peace Prize	1998 Nobel peace prize

Table 8.2: Topics used for Snippet evaluation

Sentence selection and Combination

As discussed in Chapter 3, two main aspects of snippet generation are *Sentence Selection* and *Sentence Combination.* We manually explored different weighting options for combining sentences based on Topical relevance (Rel), Novelty (Nov), and Readability (Read) models as shown in Equation 8.5. We select the top k scoring sentences based on the combination of features as the document snippets. As discussed in Chapter 3, combining sentences in the initial order works well as studied by Mishra and Berberich (2017); Leal Bando et al. (2015). The top k scoring sentences are presented in their sequence from the source document.

$$F(X) = W_{rel} * Rel(X) + W_{read} * Read(X) + W_{nov} * Nov(X) \tag{8.5}$$

Length of snippets: Exploring different length of snippets has been studied quite extensively previously (Maxwell et al., 2017; Cutrell and Guan, 2007; Yulianti et al., 2016) as discussed in Chapter 3. A recent work on query biased summaries found that summaries comprising of sentence lengths = 3 are quite effective, in a user based pairwise setting (Leal Bando et al., 2015). Another work reported that the average length of answer summaries is 2.67 sentences (Yulianti et al., 2016). Thus we fix length of snippets to 3 sentences and instead focus on different sentence selection approaches for snippet generation.

Varying quality of snippets: For each of the two topics in the development phase we used the best relevance (*BestRelModel*) model, the best novelty (*BestNovelModel*) model, and the *BestReadModel* to score sentences within a document. The BestRelModel is the BM25 model with pseudo relevance feedback and semantic-based global centroid approach for query expansion as described in Chapter 6. The BestNovelModel is the combined model integrating syntactic-based and embedding-based sentence similarity approaches as described in Chapter 7. The BestReadModel is the FRES readability model described earlier in Section 8.1.1. We normalise the scores of BestRelModel, BestNovelModel and BestReadModel in the range of [0-1] and combine them using Equation 8.5.

The novelty model operates by comparing sentences within and across the documents occurring higher up the retrieval ranked list. Thus measuring document snippets independently using the *Novelty* model and its combination may not truly reflect the nature of Novelty model. Their true effect may not reflect in an individual assessment independent of analysis of other snippets.

We investigate 7 different types of snippet combination by varying the relative weights of the relevance (W_{rel}), novelty (W_{nov}) and readability (W_{read}) scores as shown in Equation 8.5. The following combinations were examined:

a) **Only Relevance:** Top 3 sentences selected using only the output of the *BestRelModel* model ($W_{rel} = 1.0$, $W_{read} = 0.0$ and $W_{nov} = 0.0$).

b) **Relevance + Readability:** Top 3 sentences selected using an average of the output of the *BestRelModel* and the *BestReadModel* ($W_{rel} = 0.5$, $W_{read} = 0.5$ and $W_{nov} = 0.0$).

c) **Only Novelty:** Top 3 sentences are selected using only the output of the *BestNovelModel* model ($W_{rel} = 0.0$, $W_{read} = 0.0$ and $W_{nov} = 1.0$).

d) **Novelty + Readability:** Top 3 sentences selected using an average of the output of the *BestNovelModel* and the *BestReadModel* ($W_{rel} = 0.0$, $W_{read} = 0.5$ and

$W_{nov} = 0.5$).

e) **Relevance + Novelty:** Top 3 sentences selected using an average of the output of the *BestNovelModel* and the *BestRelModel* ($W_{rel} = 0.5$, $W_{read} = 0.0$ and $W_{nov} = 0.5$).

f) **Combined (Relevance + Novelty + Readability):** Top 3 sentences selected using an average of the output of the *BestRelModel*, the *BestNovelModel* and the *BestReadModel* ($W_{rel} = 0.33$, $W_{read} = 0.33$ and $W_{nov} = 0.33$).

g) **Combined (Relevance + Novelty {pruned by Readability}):** Instead of combining readability scores with novelty and relevance output, we use readability scores to include only those sentences which have a readability score greater than 0.5. Then we combine the output of the *BestNovelModel* and the *BestRelModel* (W_{rel} $= 0.5$ and $W_{nov} = 0.5$).

8.2.2 Evaluation Measure

Pairwise evaluation of snippets is commonly used for comparing snippet generation methods (Leal Bando et al., 2015; Ageev et al., 2013). Since we have seven snippets for each document, conducting a pairwise approach would result in 21 pairs of snippet per document which would be very expensive to compare and evaluate (time-wise and cost-wise). To overcome this challenge a general mechanism that is commonly used for evaluating document snippets focuses on scoring the snippets based on different notions such as coherence, readability, usefulness, grammatical correctness. In line with this approach we follow the evaluation mechanism and its definition proposed in the DUC benchmark campaign (Harman and Over, 2002), and score each snippet on a scale of 1-5 using measures of *grammatical correctness, clarity and coherence*. We add two more evaluation measures of *topicality and usefulness*, since we are interested in measuring how effective the snippets are in providing useful and relevant on-topic information to satisfy a user's information need. Different evaluation measures, with their scale and description are presented in Table 8.3. Each of the snippets generated using 7 different methods are scored by the author using the evaluation criteria defined in Table 8.3. As the goal was to compare alternative approaches to find the top approaches, it seems reasonable to have only 1 annotator, following the typical TREC evaluation paradigm of one primary assessor for the relevance judgements where the goal is to compare system rankings (Voorhees, 2000)

Overall we manually evaluated 2 (topics) * 20 (documents per topic) * 7 (different combination approaches) = 280 snippets, to select the best snippet combination approaches for SERP representation.

Evaluation Measure	Range	Description
Grammatical Correctness	1-5	Snippets have no spelling mistakes, meaningless words, meaningless sentences
Clarity	1-5	Snippets have no pronoun errors and/or hard to understand words
Coherence	1-5	Check for the flow of the sentences, semantic closeness of the information, whether information is in a sequential order
Topicality	1-5	Snippets are related to the topic of inquiry
Usefulness	1-5	Snippets contribution to understanding the topic or addressing the information need

Table 8.3: Measures used for snippet evaluation, where 1 indicates lower degree and 5 indicates higher degree of the evaluation measure.

8.2.3 Results & Analysis

In this section we describe the results of the evaluation of our snippet creation methods.

Tables 8.4, 8.5 and 8.6 show results of our snippet creation methods for topics $D1$ and $D2$. The average of the grammatical correctness, clarity, coherency, topicality and usefulness measured over 20 documents is shown in Table 8.4 and 8.5. Table 8.6 shows the average results across the two topics. In general all the approaches score highly for grammatical correctness and topicality. Clarity and coherency results vary across the different approaches. Our main focus is on the average scores

Model	Grammatical	Clarity	Coherent	Topicality	Usefulness	Length
Only Rel	**5.00**	**3.95**	**4.05**	**4.90**	**3.90**	**108**
Rel + Read	4.90	3.33	3.29	4.76	3.38	77
Only Nov	4.95	3.95	3.48	4.62	2.71	89
Nov + Read	4.81	3.62	2.95	4.14	2.86	65
Combined	5.00	4.05	3.71	4.86	3.48	81
Rel + Nov	**5.00**	**4.43**	**4.14**	**5.00**	**4.05**	**105**
Combined Pruned	4.86	4.09	4.00	4.86	3.62	85

Table 8.4: Snippet Generation output for Topic D1. Rel, Nov and Read indicates Relevance, Novelty and Readability model respectively, Combined indicates model combining output of relevance, novelty and readability models, Combined Pruned indicates model combining output of novelty and relevance models where each sentence has been pruned by readability threshold.

Model	Grammatical	Clarity	Coherent	Topicality	Usefulness	Length
Only Rel	**4.92**	**4.40**	**3.72**	**4.68**	**2.80**	**91**
Rel + Read	4.76	3.92	3.44	4.68	2.16	63
Only Nov	4.88	4.00	3.68	4.04	2.40	73
Nov + Read	4.76	3.60	3.40	4.20	2.40	58
Combined	4.84	4.12	3.92	4.80	2.56	65
Rel + Nov	**5.00**	**4.28**	**4.04**	**4.76**	**2.96**	**92**
Combined Pruned	4.76	4.12	3.56	4.16	2.76	68

Table 8.5: Snippet Generation output for Topic D2. Rel, Nov and Read indicates Relevance, Novelty and Readability model respectively, Combined indicates model combining output of relevance, novelty and readability models, Combined Pruned indicates model combining output of novelty and relevance models where each sentence has been pruned by readability threshold.

of *usefulness* for different approaches, because we are more interested in how useful and effective the snippets are in helping the user to perform a search task and satisfy their information needs. For both topic $D1$ and $D2$, the *Relevance + Novelty*, and *Only Relevance* approaches perform far better than the other snippet combination approaches for the usefulness measure. However, these snippets were examined and evaluated independently, in this thesis we are more interested in how user behaviour, experience and knowledge gain varies when our snippets are presented in a SERP. Thus we select 1) **Relevance + Novelty** and 2) **Only Relevance** as two alternative snippet generation approaches for examination in a task-based user study to measure their utility when presented in a SERP, this study is presented in Chapter

Model	Grammatical	Clarity	Coherent	Topicality	Usefulness	Length
Only Rel	**4.96**	**4.18**	**3.88**	**4.79**	**3.35**	**99**
Rel + Read	4.83	3.63	3.36	4.72	2.77	70
Only Nov	4.92	3.98	3.58	4.33	2.56	81
Nov + Read	4.78	3.61	3.18	4.17	2.63	61
Combined	4.92	4.08	3.82	4.83	3.02	73
Rel + Nov	**5.00**	**4.35**	**4.09**	**4.88**	**3.50**	**98**
Combined Pruned	4.81	4.11	3.78	4.51	3.19	76

Table 8.6: Combined Score for Topic D1 and D2. Rel, Nov and Read indicates Relevance, Novelty and Readability model respectively, Combined indicates model combining output of relevance, novelty and readability models, Combined Pruned indicates model combining output of novelty and relevance models where each sentence has been pruned by readability threshold.

9.

All combination approaches produce different results and re-rank the sentences based on the scores of relevance, novelty and readability or a combination of these scores. We select the top three candidates and combine them in their order of occurrence in the document. Thus if 2 methods of combination produce exactly 3 similar candidates but with different ranking the snippets generated are actually identical. Table 8.7 presents snippet output using the alternative methods for a relevant document for topic D1.

Next, we give some observations from our analysis of results for our alternative snippets combination approaches:

a) **Only Relevance:** Snippets generally have longer sentences, meaning that they are generally more informative.

b) **Only Novelty:** Most of the snippets are quite poor, they contain new and diverse information, but are badly joined together, less coherent and less useful overall.

c) **Relevance + Readability:** Snippets have shorter sentences, do not capture the topic and its underlying aspects effectively.

d) **Novelty + Readability:** As the readability model favours shorter sentences, the coherence and usefulness is lower than when using only the novelty model. This model performed worst among all the seven combination approaches explored.

e) **Relevance + Novelty:** Snippets are quite good compared to all other approaches, more coherent, useful and clearer sentences. The main drawback is the larger size of the snippets.

f) **Combined (Relevance + Novelty + Readability):** Snippets miss some topical information and this impacts on their usefulness, but they are shorter in length compared to snippets created using the *Only Relevance* and *Relevance + Novelty* model.

g) **Combined (Relevance + Novelty pruned by Readability):** Snippets are informative, they are shorter in length compared to snippets created using the *Only Relevance* model. Most snippets are similar to either the *Relevance + Novelty* or Combined (*Relevance + Novelty + Readability*) approach.

Contrary to our expectation that the true effect of *Novelty model* might not be reflected in individual assessment without other snippets in context, we found that the combination approach using *Relevance + Novelty* perform best compared to all the other approaches in terms of generating snippets which are more useful and provide information which is relevant to the topic of inquiry. We speculate this is due to the effectiveness of novelty model which compares sentences within and across the documents. In *Relevance + Novelty* model the sentences which are selected are relevant as well capture information which is new thus overall capturing multiple aspects (potential more useful information) in top 3 sentences as compared to using *Relevance* model where top sentences may have repetitive aspects, thus slightly less useful than the former.

8.3 Summary and Conclusion

In this chapter we have introduced the model used for sentence-level readability prediction for generating effective snippets in this thesis. We described a snippet generation approach. We explored seven alternative combinations of relevance, novelty and readability information to generate effective snippets. We performed manual evaluation of snippets generated by combination of sentence-level relevance, novelty and readability scores using measures like *grammatical correctness, clarity, coherence, topicality and usefulness* on the scale of [**1-5**]. **Relevance + Novelty** and **Only Relevance** snippet combination approaches score best in terms of *usefulness* as compared to other different snippet generation approaches explored in this work. We study and investigate these best snippet generation settings for our SERP representation in the pilot study and the crowdsource-based study, which we discuss next in Chapter 9.

Relevance: Celera Corp. said Thursday that it has sequenced the human genome, in effect discovering the sequence of DNA molecules in which information on human heredity is inscribed. Craig Venter, Celera's chairman and chief scientific officer , told Reuters Thursday that "Now that we have completed the sequencing of one human being's genome, we will turn our computational power to the task of ordering the human genome." Celera, whose operation employs 300 sequencing machines and the largest known civilian computer, last month reported assembling the sequencing of the genome of a fruit fly.
Relevance Pruned: Discovering the sequence is an essential step toward mapping the entire human genome, though the real challenge will be to put the sequence together properly. The company next plans to assemble the final human genome sequence from the genes of five anonymous people. The fragments, they said, provided enough DNA to cover the genome 14 times over.
Novelty: The company, which was established barely two years ago, reported that it would now attempt to assemble the genetic fragments in their proper order. The difficulty in comprehending the formula has been that genomes in living creatures contain numerous repeats, which are copies of stretches of DNA that follow nearly identical sequences, confusing the assembly process. Writing in the magazine Science, two researchers, Eugene W. Myers and Edward Winstead, explained that the assembly began with 3.1 million fragments of the genome, which they described as "random bits of fly DNA that have been converted into characters that a computer can read".
Novelty Pruned: Celera's announcement Thursday drove the stock up nearly 25 percent, to $143 a share, by noon. Celera's scientists said earlier this year that they had developed techniques for bridging the repeat stretches to unscramble at least some of them. This would allow the researchers to compare the donated genes to detect the minute changes that make one human being different from another.
Relevance + Novelty: Celera Corp. said Thursday that it has sequenced the human genome, in effect discovering the sequence of DNA molecules in which information on human heredity is inscribed. Craig Venter, Celera's chairman and chief scientific officer, told Reuters Thursday that "Now that we have completed the sequencing of one human being's genome, we will turn our computational power to the task of ordering the human genome." Celera, whose operation employs 300 sequencing machines and the largest known civilian computer, last month reported assembling the sequencing of the genome of a fruit fly.
Relevance + Novelty + Readability: Discovering the sequence is an essential step toward mapping the entire human genome, though the real challenge will be to put the sequence together properly. The company next plans to assemble the final human genome sequence from the genes of five anonymous people. The fragments, they said, provided enough DNA to cover the genome 14 times over.
Relevance + Novelty (Readability Pruned): Celera's announcement Thursday drove the stock up nearly 25 percent, to $143 a share, by noon. Discovering the sequence is an essential step toward mapping the entire human genome, though the real challenge will be to put the sequence together properly. The company next plans to assemble the final human genome sequence from the genes of five anonymous people.

Table 8.7: Examples of snippet generation using alternative feature combinations

Chapter 9

Snippet Evaluation

The main focus of our work is to generate document snippets and measure its utility when presented in a SERP in a task-based setting. We evaluate how effective are our snippets in helping the participants to perform information seeking and gathering tasks and gain knowledge during their engagement with the SERP. In this chapter, we present our work done on the user-centred task-based evaluation of snippets. First, we introduce the three different snippet models explored in this work. Then we describe our user-based evaluation which is conducted in two steps: i) pilot study and ii) crowdsource-based study. Pilot-study was a lab-based study to see the effectiveness of two best models of snippet generation described in Chapter 8, to get user-feedback on the study-design and interface developed for measuring SERP interactions. Crowdsource-based study was conducted to evaluate the effectiveness of two best models developed by our framework and a baseline model of snippet creation, when presented in a SERP. We describe the results and analysis of our pilot and crowdsource-based investigations. Finally, we discuss our main findings, and present the conclusions of this chapter.

9.1 Introduction

User-based search evaluation is challenging in nature. Based on the results of the snippet generation study described in Chapter 8, we select the two best models in the study described in this chapter for snippet evaluation. We compare these snippets to a baseline model for snippet creation. The three different snippets models explored for SERP representation are:

- *Baseline model*: Document snippets are generated using the output score of the BM25-based sentence-level relevance model as described in Chapter 6.
- *Novelty model*: Document snippets are generated using an average of the output score of the sentence-level relevance model (*BestRelModel*) and novelty model (*BestNovelModel*) as described in detail in Chapter 8.
- *Relevance model*: Document snippets are generated using the output score of the sentence-level relevance model (*BestRelModel*) proposed in this work (described in Chapter 8).

For each model, the top 3 ranked sentences are combined in their original order of occurrence in a document to represent a snippet.

Interacting with documents in a SERP to gather information and learn about a topic is a comprehensive and cognitive intensive task for the participants. Thus we concentrate on detailed analysis of a small number of topics for this study, with the main goal to capture richer and stronger signals for measuring changes in user search behaviour, experience and knowledge gain aspects. We used separate topics for the pilot and crowdsource-based study.

Next, we describe our pilot investigation where *Novelty* and *Relevance* models were used to present snippets in a SERP.

9.2 Pilot Experimental Investigation

In this section, we describe our pilot study investigation. Two main goals of this study were:

1. To measure the changes in the user experience, interactions and knowledge gain aspects in a simulated work task setup for the SERP generated using *Novelty* and *Relevance* models.

2. To gather feedback on the interface design and experimental setup before conducting crowdsource-based user study.

9.2.1 Experimental Design

To design the experimental protocol for measuring users interactions with SERP, we follow the procedure used in our initial investigation on measuring user knowledge gain where participants interacted with document snippets in a web search (as discussed in Chapter 5), which is described next.

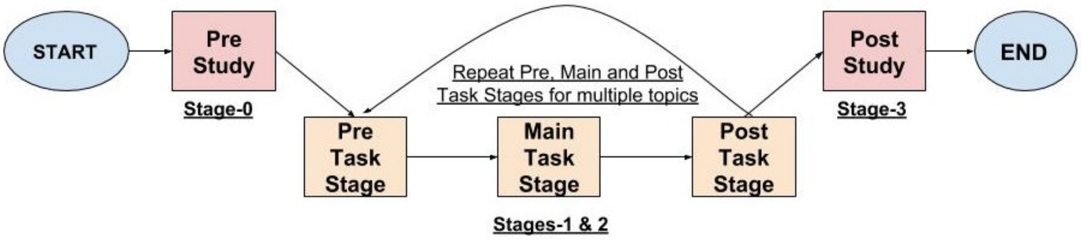


Figure 9.1: Flow diagram for snippet evaluation

Figure 9.1 shows the data flow of the stages of our study. The study consisted of four main stages which are described below. Snapshots[1] of different stages of our study are shown in Figures 9.2, 9.3, 9.4, 9.5 and 9.6.

- *Stage-0*: Participants entered the first stage of the study, they were provided with general information about the study indicating different stages as shown

[1] Pilot and Crowdsource-based studies had the same experimental design consisting of 4 stages, however the topics used in both studies were different and the exact questions asked within each stage were also slightly different.

in Figure 9.2. Then they completed an initial questionnaire consisting of three questions ES-[1-3].

- *Stage-1*: Participants were given their first simulated work-task, in which they were given scenarios that simulate real life information needs where they were asked to write a report for a college project on the first topic. This stage consisted of three steps.

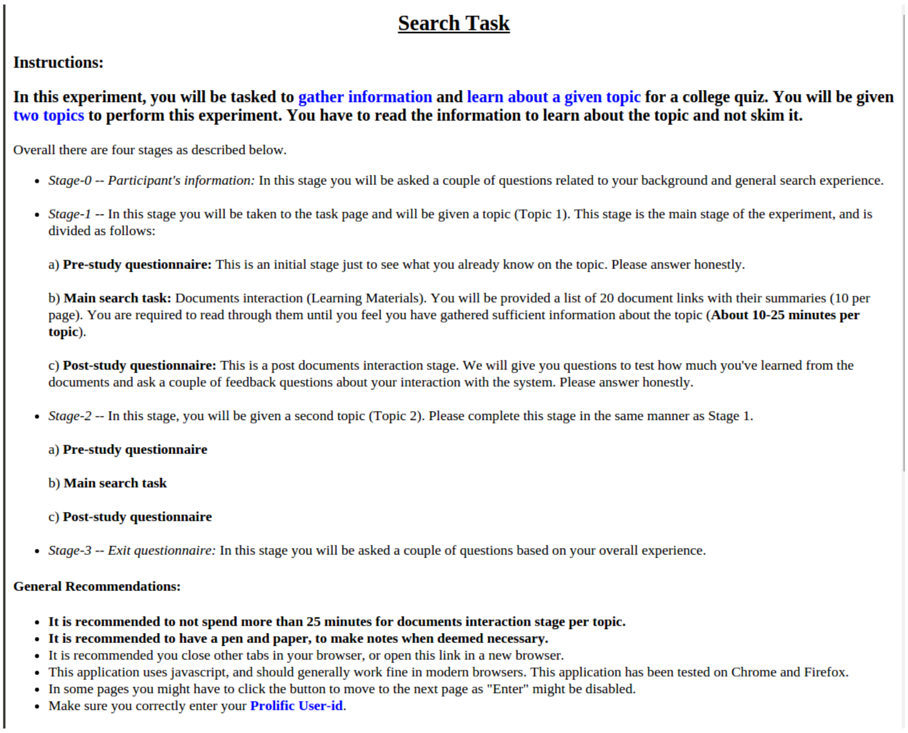
  Pre-task stage (initial knowledge assessment): This was the first step of the stage-1, where we sought to assess the prior knowledge of the users on a given topic. Subjects were asked three questions PR-[1-3], as shown in Table 9.2.

  Main task stage (interaction with documents): In this stage users were given a work-task and shown top 20 relevant documents on the given topic and were asked to read the documents and to gather information which would enable them to accomplish the task of writing a report on the topic. There were two SERP pages where 10 snippets were presented per page, as is commonly used in a general web search systems (e.g. Google).

  Post-task stage (after task knowledge assessment): After the document interaction stage, users were given a post-task questionnaire in which they were asked nine questions PO-[1-9], as shown in Table 9.2.

- *Stage-2*: Participants were given a second simulated work-task and were asked to repeat the pre-task, the main task and the post-task stages on this second topic.

- *Stage-3*: Participants entered the last stage of the study and completed an exit questionnaire in which they were asked five question EX-[1-5], as shown in Table 9.1.



Search Task

Instructions:

In this experiment, you will be tasked to gather information and learn about a given topic for a college quiz. You will be given two topics to perform this experiment. You have to read the information to learn about the topic and not skim it.

Overall there are four stages as described below.

- *Stage-0 -- Participant's information:* In this stage you will be asked a couple of questions related to your background and general search experience.
- *Stage-1* -- In this stage you will be taken to the task page and will be given a topic (Topic 1). This stage is the main stage of the experiment, and is divided as follows:

  a) **Pre-study questionnaire:** This is an initial stage just to see what you already know on the topic. Please answer honestly.

  b) **Main search task:** Documents interaction (Learning Materials). You will be provided a list of 20 document links with their summaries (10 per page). You are required to read through them until you feel you have gathered sufficient information about the topic (**About 10-25 minutes per topic**).

  c) **Post-study questionnaire:** This is a post documents interaction stage. We will give you questions to test how much you've learned from the documents and ask a couple of feedback questions about your interaction with the system. Please answer honestly.
- *Stage-2* -- In this stage, you will be given a second topic (Topic 2). Please complete this stage in the same manner as Stage 1.

  a) **Pre-study questionnaire**

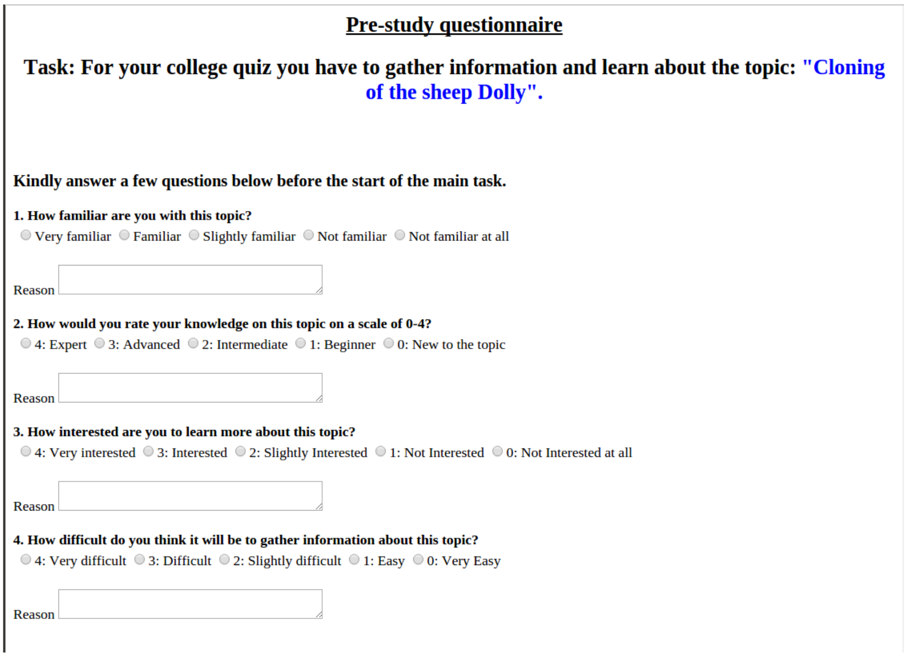
  b) **Main search task**

  c) **Post-study questionnaire**
- *Stage-3 -- Exit questionnaire:* In this stage you will be asked a couple of questions based on your overall experience.

General Recommendations:

- **It is recommended to not spend more than 25 minutes for documents interaction stage per topic.**
- **It is recommended to have a pen and paper, to make notes when deemed necessary.**
- It is recommended you close other tabs in your browser, or open this link in a new browser.
- This application uses javascript, and should generally work fine in modern browsers. This application has been tested on Chrome and Firefox.
- In some pages you might have to click the button to move to the next page as "Enter" might be disabled.
- Make sure you correctly enter your **Prolific User-id**.

Figure 9.2: Stage 0: Entry stage, Topic:C1 used in crowdsource-based study



Pre-study questionnaire

Task: For your college quiz you have to gather information and learn about the topic: "Cloning of the sheep Dolly".

Kindly answer a few questions below before the start of the main task.

1. How familiar are you with this topic?
○ Very familiar ○ Familiar ○ Slightly familiar ○ Not familiar ○ Not familiar at all

Reason

2. How would you rate your knowledge on this topic on a scale of 0-4?
○ 4: Expert ○ 3: Advanced ○ 2: Intermediate ○ 1: Beginner ○ 0: New to the topic

Reason

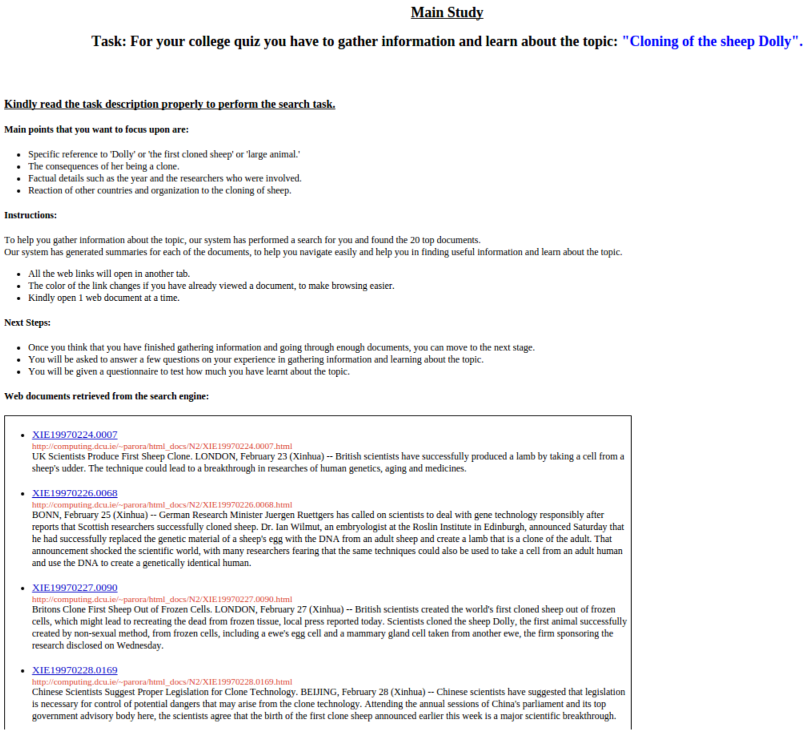
3. How interested are you to learn more about this topic?
○ 4: Very interested ○ 3: Interested ○ 2: Slightly Interested ○ 1: Not Interested ○ 0: Not Interested at all

Reason

4. How difficult do you think it will be to gather information about this topic?
○ 4: Very difficult ○ 3: Difficult ○ 2: Slightly difficult ○ 1: Easy ○ 0: Very Easy

Reason

Figure 9.3: Stage 1: Step-1 Pre-task questionnaire, Topic:C1 used in crowdsource-based study



Main Study

Task: For your college quiz you have to gather information and learn about the topic: "Cloning of the sheep Dolly".

Kindly read the task description properly to perform the search task.

Main points that you want to focus upon are:

- Specific reference to 'Dolly' or 'the first cloned sheep' or 'large animal.'
- The consequences of her being a clone.
- Factual details such as the year and the researchers who were involved.
- Reaction of other countries and organization to the cloning of sheep.

Instructions:

To help you gather information about the topic, our system has performed a search for you and found the 20 top documents.
Our system has generated summaries for each of the documents, to help you navigate easily and help you in finding useful information and learn about the topic.

- All the web links will open in another tab.
- The color of the link changes if you have already viewed a document, to make browsing easier.
- Kindly open 1 web document at a time.

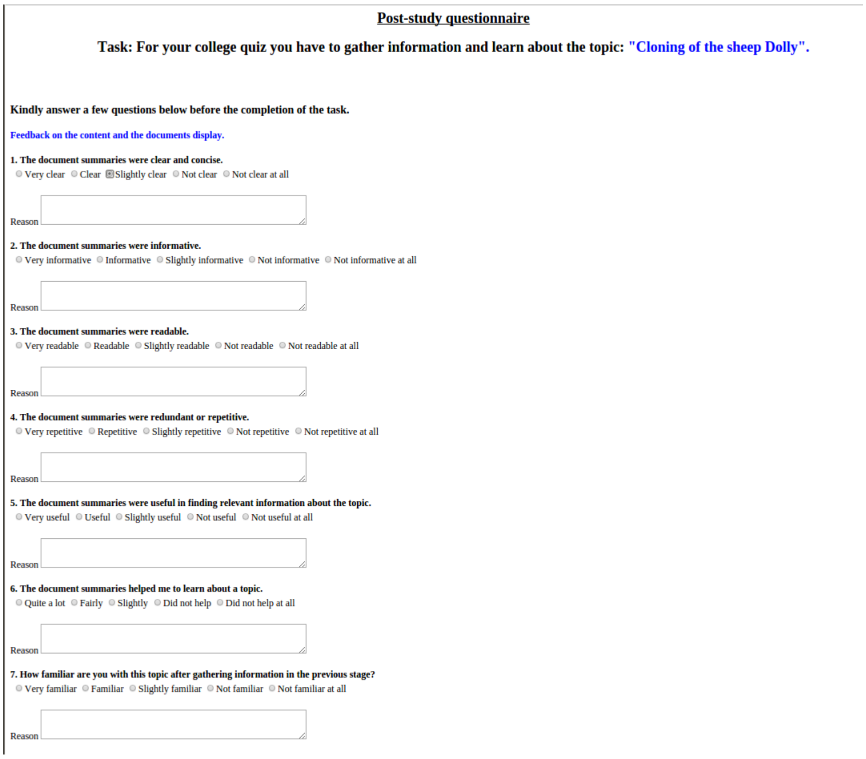
Next Steps:

- Once you think that you have finished gathering information and going through enough documents, you can move to the next stage.
- You will be asked to answer a few questions on your experience in gathering information and learning about the topic.
- You will be given a questionnaire to test how much you have learnt about the topic.

Web documents retrieved from the search engine:

- XIE19970224.0007
  http://computing.dcu.ie/~parora/html_docs/N2/XIE19970224.0007.html
  UK Scientists Produce First Sheep Clone. LONDON, February 23 (Xinhua) -- British scientists have successfully produced a lamb by taking a cell from a sheep's udder. The technique could lead to a breakthrough in researches of human genetics, aging and medicines.
- XIE19970226.0068
  http://computing.dcu.ie/~parora/html_docs/N2/XIE19970226.0068.html
  BONN, February 25 (Xinhua) -- German Research Minister Juergen Ruettgers has called on scientists to deal with gene technology responsibly after reports that Scottish researchers successfully cloned sheep. Dr. Ian Wilmut, an embryologist at the Roslin Institute in Edinburgh, announced Saturday that he had successfully replaced the genetic material of a sheep's egg with the DNA from an adult sheep and create a lamb that is a clone of the adult. That announcement shocked the scientific world, with many researchers fearing that the same techniques could also be used to take a cell from an adult human and use the DNA to create a genetically identical human.
- XIE19970227.0090
  http://computing.dcu.ie/~parora/html_docs/N2/XIE19970227.0090.html
  Britons Clone First Sheep Out of Frozen Cells. LONDON, February 27 (Xinhua) -- British scientists created the world's first cloned sheep out of frozen cells, which might lead to recreating the dead from frozen tissue, local press reported today. Scientists cloned the sheep Dolly, the first animal successfully created by non-sexual method, from frozen cells, including a ewe's egg cell and a mammary gland cell taken from another ewe, the firm sponsoring the research disclosed on Wednesday.
- XIE19970228.0169
  http://computing.dcu.ie/~parora/html_docs/N2/XIE19970228.0169.html
  Chinese Scientists Suggest Proper Legislation for Clone Technology. BEIJING, February 28 (Xinhua) -- Chinese scientists have suggested that legislation is necessary for control of potential dangers that may arise from the clone technology. Attending the annual sessions of China's parliament and its top government advisory body here, the scientists agree that the birth of the first clone sheep announced earlier this week is a major scientific breakthrough.

Figure 9.4: Stage 1: Step-2 Documents Interactions, Topic:C1 used in crowdsource-based study



Post-study questionnaire

Task: For your college quiz you have to gather information and learn about the topic: "Cloning of the sheep Dolly".

Kindly answer a few questions below before the completion of the task.

Feedback on the content and the documents display.

1. The document summaries were clear and concise.
○ Very clear ○ Clear ● Slightly clear ○ Not clear ○ Not clear at all

Reason

2. The document summaries were informative.
○ Very informative ○ Informative ○ Slightly informative ○ Not informative ○ Not informative at all

Reason

3. The document summaries were readable.
○ Very readable ○ Readable ○ Slightly readable ○ Not readable ○ Not readable at all

Reason

4. The document summaries were redundant or repetitive.
○ Very repetitive ○ Repetitive ○ Slightly repetitive ○ Not repetitive ○ Not repetitive at all

Reason

5. The document summaries were useful in finding relevant information about the topic.
○ Very useful ○ Useful ○ Slightly useful ○ Not useful ○ Not useful at all

Reason

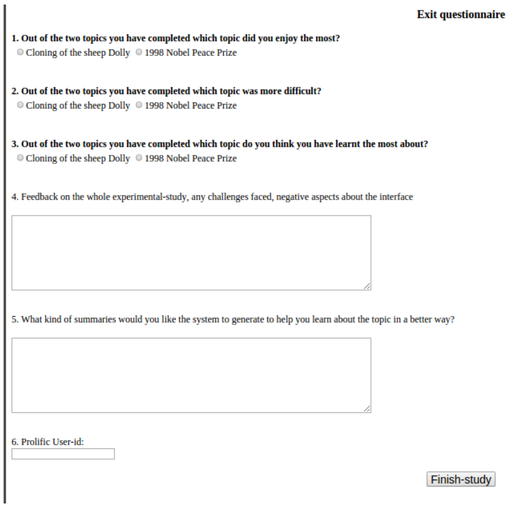
6. The document summaries helped me to learn about a topic.
○ Quite a lot ○ Fairly ○ Slightly ○ Did not help ○ Did not help at all

Reason

7. How familiar are you with this topic after gathering information in the previous stage?
○ Very familiar ○ Familiar ○ Slightly familiar ○ Not familiar ○ Not familiar at all

Reason

Figure 9.5: Stage 1: Step-3 Post-task questionnaire, Topic:C1 used in crowdsource-based study



Exit questionnaire

1. Out of the two topics you have completed which topic did you enjoy the most?
○ Cloning of the sheep Dolly ○ 1998 Nobel Peace Prize

2. Out of the two topics you have completed which topic was more difficult?
○ Cloning of the sheep Dolly ○ 1998 Nobel Peace Prize

3. Out of the two topics you have completed which topic do you think you have learnt the most about?
○ Cloning of the sheep Dolly ○ 1998 Nobel Peace Prize

4. Feedback on the whole experimental-study, any challenges faced, negative aspects about the interface

5. What kind of summaries would you like the system to generate to help you learn about the topic in a better way?

6. Prolific User-id:

Figure 9.6: Stage 3: Exit questionnaire, Topic:C1 used in crowdsource-based study

Variable	Id	Question	Scale	Source
Past search experience	ES-1	How long have you been using search engines like Google, Bing etc?	0=[0-1] years 1=[1-3] years 2= >3 years	Entry Stage
Search frequency	ES-2	How often do you use search engines like Google, Bing in a day?	0=once 1=5-10 times 2=>10 times	Entry Stage
Language	ES-3	Is English your first/native language?	0=Yes 1=No	Entry Stage
Task enjoyed	EX-1	Which task did you liked and enjoyed most?	0=Task L1 1=Task L2	Exit Stage
Task deemed difficult	EX-2	Which task seemed more difficult?	0=Task L1 1=Task L2	Exit Stage
Task deemed having useful documents	EX-3	Which task you think had more useful documents?	0=Task L1 1=Task L2	Exit Stage
Task perceived learnt	EX-4	Which task you think you have learnt more about?	0=Task L1 1=Task L2	Exit Stage
Study feedback	EX-5	Overall feedback on the interface and if faced any challenges	N/A	Exit Stage

Table 9.1: Pilot study experimental design entry and exit stages questionnaire

9.2.2 Tasks and System

We used two tasks for the pilot study. The simulated work-task statement that was shown to the users is shown below.

- *Task 1:* For your college project you have to write a report on the Topic (L1): "Microsoft's guilt or innocence on charges of antitrust".

Variable	Id	Question	Scale	Source
Pre-task familiarity	**PR-1**	How familiar are you with this topic?	[0-4], where 4=Very familiar... 0=Not familiar at all	Pre-Task
Pre-task perceived knowledge	**PR-2**	How will you rate your knowledge on this topic?	[0-4], where 4=Expert... 0=New to the topic	Pre-Task
Pre-task knowledge summary	**PR-3**	Write a summary about the topic (about 5 sentences) in terms of what you know.	N/A	Pre-Task
Find useful information	**PO-1**	How difficult was it to find information relevant to the topic from the documents?	[0-4], where 4=Very difficult... 0=Very easy	Post-Task
Understand content	**PO-2**	How difficult was it to understand the content of the documents?	[0-4], where 4=Very difficult... 0=Very easy	Post-Task
Useful	**PO-3**	How useful were the document summaries and snippets?	[0-4], where 4=Very useful... 0=Not useful at all	Post-Task
Readable	**PO-4**	How readable were the document summaries and snippets?	[0-4], where 4=Very readable... 0=Not readable at all	Post-Task
Post-task familiarity	**PO-5**	How familiar are you with this topic after interacting with documents?	[0-4], where 4=Very familiar... 0=Not familiar at all	Post-Task
Post-task perceived knowledge	**PO-6**	How will you rate your knowledge on this topic after interacting with documents	[0-4], where 4=Expert... 0=New to the topic	Post-Task
Perceived learning	**PO-7**	How much do you think you learnt on this topic? interacting with documents?	[0-4], where 4=Quite a lot... 0=Nothing at all	Post-Task
Post-task knowledge summary	**PO-8**	Write a small abstract for your report (minimum 5 sentences) based on the gathered information.	N/A	Post-Task
Task feedback	**PO-9**	Feedback on the whole exercise.	N/A	Post-Task

Table 9.2: Pilot study experimental design pre- and post-task questionnaire

- *Task 2:* For your science project you have to write a report on the Topic (L2): “First human hand transplant in the United States” which was performed on Matthew Scott on January 25, 1999.

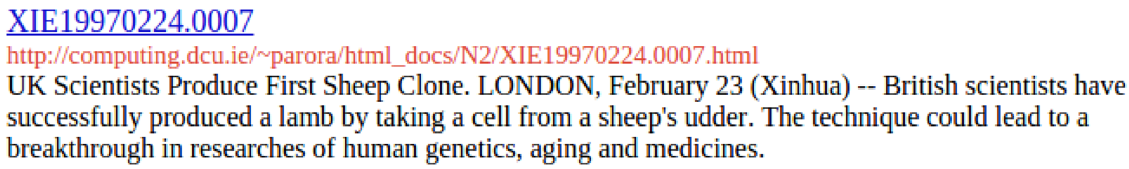
After the task statement participants were shown the instructions.

Instructions: To help you gather information about the topic, our system has performed search for you and found 20 top documents. Our system has generated snippets for each of the documents, to help you navigate easily and help you in finding useful information to complete the task of writing the report. Once you think that you have finished gathering information and read enough documents, you can move to the next stage. You will be asked to answer a few questions on your experience in finding information, and to provide feedback on the interface. You will be asked to write a small abstract (about 5 sentences) for the report based on the gathered information.

Interface Setup: Next, we outline the system developed to conduct our pilot study.

- *SERP layout and design*: Two types of SERP pages are generated comprising of snippets generated using *Novelty* and *Relevance* models. We adopted the same colour coding mechanism as used by Google[2], document-id is represented in blue, url of the document is represented in red, and the document summary generated by our snippet models is represented in black as shown in Figure 9.7. The colour of the document-id changes if a user has already clicked a document to imitate the web search systems behaviour which we anticipate users are familiar with to help them interact with the documents efficiently.

- *System development*: The system is built using javascript and html. We used nodjs for making the server and back end for running the application. Images of our interface representing four different stages of our system are presented in Figures 9.2, 9.3, 9.4, 9.5 and 9.6.



XIE19970224.0007
http://computing.dcu.ie/~parora/html_docs/N2/XIE19970224.0007.html
UK Scientists Produce First Sheep Clone. LONDON, February 23 (Xinhua) -- British scientists have successfully produced a lamb by taking a cell from a sheep's udder. The technique could lead to a breakthrough in researches of human genetics, aging and medicines.

Figure 9.7: Example of a document snippet

[2] www.google.com

9.2.3 Participants and Setup

This was a lab-based study, participants were recruited through e-mail on a voluntary basis. The study was hosted on our server and shared through an url. Eight participants performed the task. The demographics of the users were 3 Female and 5 Male. 3 participants were native English speakers. Average age of the participants was 28 years.

Tasks Distribution: We had two types of SERP pages (M1 and M2) consisting of snippets from *Novelty* and *Relevance* models respectively. We had two topics (L1 and L2) for which we performed the search task. We followed a factorial design mechanism for our pilot study as shown in Table 9.3. Each group performed a task using both type of SERP pages (M1 and M2) and for both topics (L1 and L2). Due to the limited number of participants we did not alter the order of topics. We had 4 participants who were randomly allocated to group-1 and 4 participants who were randomly allocated to group-2.

Group	Stage-1	Stage-2
Group-1	Topic:L1, Model-1 (L1M1)	Topic:L2, Model-2 (L2M2)
Group-2	Topic:L1, Model-2 (L1M2)	Topic:L2, Model-1 (L2M1)

Table 9.3: Pilot study tasks distribution

Data collection

We describe the ways in which we captured data in the pilot study, we then describe the method used for scoring summaries for the assessment of user's topical knowledge.

- **Questionnaire:** As shown in Tables 9.1 and 9.2, we collected data from the participants at several stages as described in Section 9.2.1. For questions ES-[1-3], EX-[1-4] for the overall study and PR-[1-2], PO-[1-7] for each task we asked users to rate their search experience, perceived topic familiarity and knowledge gain on a likert scale. We also asked users to write open feedback for each task (PO-9) and overall feedback on the complete study (EX-5).

- **Logging of users interactions:** All the user interactions with the system in terms of the documents clicked, mouse movements and the time spent at different stages were recorded.

- **Summaries:** To measure users prior-topical knowledge (PR-3) we asked users to write a summary on what they know about the topic. To assess what they learned while interacting with documents in the simulated work-task, we asked them to write an abstract for their college report (PO-8).

Scoring of summaries (data coding): It is challenging to compare user summaries effectively to measure changes in user learning and gain of knowledge in a task-based setting. Thus, we used the techniques proposed by Wilson and Wilson (2013) for comparing summaries to measure knowledge gain of the participants in our experimental setup by measuring two aspects: *Number of unique facts*, and *Quality of facts* on the given topic. We counted the number of unique facts in a summary and assigned it a score out of 5 (minimum sentences users are asked to write). For each unique fact in the summary a participant was assigned a score of 1 (max score=5). The description of the ratings used for measuring the quality of users summary is presented in Table 9.4.

Rating	**Description**
0	Most of the sentences and information is irrelevant.
1	Some sentences and facts are irrelevant.
2	Sentences and facts are generalised to the overall subject matter. Holds little useful information or advice.
3	Most of the sentence and facts fulfil the required information need and are useful.
4	A level of detail is given via at least one key aspect on the topic, along with providing facts which are useful.
5	Exhibits a level of analysis, comparison, opinion and insights on the subject.

Table 9.4: Measuring quality of summaries

9.2.4 Results & Analysis

Tables 9.5, 9.6 and 9.7 show the results of the user experience, knowledge gain and user interactions for our pilot study. Results are quite mixed, it is difficult to see the difference between different SERP pages (M1 & M2) and topics (L1 & L2) because of the limited number of participants (4 in each group).

	Group-1		Group-2	
Question	L1M1	L2M2	L1M2	L2M1
Find useful information	1.00	1.00	1.00	1.50
Understand content	1.75	0.75	2.00	1.50
Useful	2.75	2.75	3.25	2.75
Readable	1.50	1.00	1.75	2.00
Post-Pre task familiarity	0.75	1.75	1.5	1.75
Post-Pre task perceived knowledge	1.00	1.00	1.5	1.25
Perceived learning	3.75	3.25	3.25	1.75

Table 9.5: Participants pre- and post-task user experience results

User Experience: As shown in Table 9.5, for Topic:L1, scores of *finding useful information* are similar across both the SERP models M1 and M2, for *understanding content* model-M1 is slightly better than model-M2 (lower scores indicate the ease of understating content), average scores for *useful* and *readable* snippets for Topic:L1 using model-M2 is higher as compared to using model-M1. The difference in *post- and pre-task familiarity* and *post- and pre-task perceived knowledge* scores are also higher using model-M2 than using model-M1. For the *perceived learning* average scores using model-M1 are higher as compared to model-M2. However, when we examine the results for Topic:L2, where we change the SERP models (M1 and M2) for both the groups, we observe opposite behaviour for *finding useful information* and *understanding content* where results score higher for model-M2 as compared to model-M1, average scores for *readable* summaries is higher in model-M1 as compared to model-M2. However, the *perceived learning* for Topic:L2 is more in Group-1 using model-M2 as compared to Group-1 using model-M1.

Knowledge Gain: For the pre-task knowledge summary (PR-3), 7 out of 8 participants reported that they knew nothing about the subject and were totally new to the topic, 1 participant wrote general information on "Microsoft", where most of the information was not useful and relevant to the topic on "Microsoft guilt or antitrust charges". Thus we consider that the users have no prior knowledge on the topics and score "0" for the quality of facts and number of facts measure as described in Section 9.2.3.

	Group-1		Group-2	
Knowledge Measure	L1M1	L2M2	L1M2	L2M1
Number of Facts	4.75	4.25	4.00	3.50
Quality of Facts	4.50	4.00	3.75	3.25

Table 9.6: User knowledge gain results

Table 9.6 shows the results of our analysis of the post-task knowledge summary (PO-8). For both Topic:L1 & L2 Group-1 scores are higher as compared to Group-2 irrespective of the topic and SERP system being used. The high scores of the knowledge measure are also reflected in the terms of high rating for the perceived learning reported by the users in Group-1 as compared to Group-2 for both topics as shown in Table 9.5. Thus we speculate that there seems to be some relation in the perceived learning (measured using users rating) and actual learning (measured using evaluating users post-task summary), which we explore in detail in our later crowdsource-based study with a larger number of participants.

User Interactions: When we analysed the user behaviour and interaction results with the interface as shown in Table 9.7, we find that irrespective of the topic and the SERP model participants in Group-1 spent more time viewing the documents on SERP, clicked more documents, spent more time on the task overall as compared to the participants in Group-2. We speculate there is some relation between the user behaviour measures and the knowledge gain where more time spent viewing the documents and the number of documents being clicked and viewed also relate to high knowledge gain reflected as high scores of summaries as reported in Table 9.6. We also explore this relation between user behaviour and knowledge gain in our later crowdsource-based study.

	Group-1		Group-2	
	L1M1	L2M2	L1M2	L2M1
Time Spent (documents interaction)	34 mins	25 mins	17 mins	14 mins
Documents Clicked	7	9	5	5
Overall Time Spent	50 mins	38 mins	32 mins	23 mins

Table 9.7: User interaction results

Task Feedback: Most of the participants reported that the exercise was interesting and they learnt a lot on both topics. The topic on "Microsoft anti-trust charges (L1)" was on government and legal aspects, some participants reported that it was difficult at the start, but after a while it became easy and interesting. Participants found the other topic (L2) on "Scott human hand transplant" quite easy to understand content-wise, two participants reported that some documents seemed non-relevant and confused them a bit while interacting with documents. Three participants reported that snippets were larger than those which they are used to, but they found them useful providing more on-topic information. An overview of the study feedback provided by the participants is summarised in Table 9.8.

Positive aspects	**Negative aspects**
Study was easy to follow.	Documents were too long to read.
Very impressed by the summarization tool.	Documents were quite old.
Interface was user friendly & easy to access.	Snippets were a bit lengthy.
Exercises and the interface was pretty handy.	Reading the documents took time, felt like doing an English reading test.

Table 9.8: Study feedback

We had a one-to-one post-study interview with the participants, where we asked for detailed feedback to learn more about their experience. Participants answered the following questions, which are accompanied by the most notable responses.

Following are the questions asked and the main responses we received from the participants:

- Do we need to keep time limit for the document interactions stage?
  All participants preferred not to have a time limit as it will impact their natural behaviour to interact with the system without time pressure constraints. A few users who spent more than 30 minutes on reading documents, suggested including a general note in the instructions regarding the maximum time to be spent on reading the documents.

- Aspects about the interface they did not like?
  One participant suggested increasing the size of the text-boxes used for collecting user responses and feedback.

- Were they able to capture the information learnt through the documents interaction stage in their report effectively?
  A few people reported that they were not sure what to include in the report as there were multiple aspects on the topic which they read, thus they presented a general overview about the topic based on what they learnt. When asked participants liked the idea of having specific questions in the questionnaire, to test their knowledge on the topic, as if in a test.

Pilot study helped us to test the system developed for measuring document snippets utility when presented in a SERP and capture user feedback, comments and suggestions effectively. We incorporated the feedback from the pilot lab-based study into a modified version of the system for our crowdsource-based study. We refined some questions to capture user experience effectively. Instead of asking users to write a report or a long summary which can be hard to evaluate and assess, we designed topic specific questions which were used to measure changes in the user's knowledge in a search task, similar to the work reported in Collins-Thompson et al. (2016) for measuring learning outcome. We developed topic specific questions based on Bloom's taxonomy paradigm, which is described in detail in the next section, to effectively measure changes in the user's knowledge, when they interact with documents in a search task. Next we present our investigation using a crowdsource-based study.

9.3 Crowdsource-based Investigation

In this section we describe our investigation to evaluate the effectiveness of snippets generated using: *Baseline*, *Relevance* and *Novelty* models.

9.3.1 Experimental Design

We divide our investigation into two main questions:

- How does the user search experience, interactions and gain in topical knowledge vary for snippets generated by our framework (*Relevance* and *Novelty*) as compared to the baseline model?
- Are there correlations between the user search experience, document interactions variables with the knowledge gain aspects in a search task?

For the crowdsource-study we follow the design protocol as used for the pilot study as discussed in Section 9.2.1. Complete details of the specific questions asked within each stage are given in Table 9.9 and 9.10. We discuss the four main stages of the user study.

- *Stage-0*: Participants entered the first stage of the study and completed an initial questionnaire consisting of three questions ES-[1-3], as shown in Table 9.9. A snapshot of this stage is shown in Figure 9.2.
- *Stage-1*: Participants were given a first simulated work-task to gather information and learn about a topic. This stage consisted of three steps.

  Pre-task stage (initial knowledge assessment): This is the first step of the Stage-1. Subjects were asked seven questions PR-[1-7], as shown in Table 9.10. A snapshot of this stage is shown in Figure 9.3.

Main task stage (interaction with documents): In this stage users were given a search task and were shown 20 relevant documents on the given topic and were asked to read the documents to learn about the topic. A snapshot of this stage is shown in Figure 9.4.

Post-task stage (after task knowledge assessment): After the document interaction stage users were given a post-task questionnaire where they were asked twenty questions PO-[1-20], as shown in Table 9.10. A snapshot of this stage is shown in Figure 9.5.

- *Stage-2*: Participants were given a second simulated work-task and repeated the pre-task, the main task and the post-task stages for the second topic.
- *Stage-3*: Participants entered the last stage of the study and completed an exit questionnaire where they were asked five questions EX-[1-5], as shown in Table 9.9. A snapshot of this stage is shown in Figure 9.6.

Variable	Id	Question	Scale	Source
Past search experience	ES-1	How long have you been using search engines like Google, Bing etc?	0=[0-1] years 1=[1-3] years 2= >3 years	Entry Stage
Search frequency	ES-2	How often do you use search engines like Google, Bing in a day?	0=once 1=5-10 times 2=>10 times	Entry Stage
Language	ES-3	Is English your first/native language?	0=Yes 1=No	Entry Stage
Task enjoyed	EX-1	Which task did you liked and enjoyed most?	0=Task C1 1=Task C2	Exit Stage
Task deemed difficult	EX-2	Which task seemed more difficult?	0=Task C1 1=Task C2	Exit Stage
Task perceived learnt	EX-3	Which task you think you have learnt more about?	0=Task C1 1=Task C2	Exit Stage
Study feedback	EX-4	Overall feedback on the interface and if faced any challenges	N/A	Exit Stage
Suggestions	EX-5	What kind of document snippets and summaries would you like the system to generate to help you learn about the topic in a better way?	N/A	Exit Stage

Table 9.9: User study experimental design entry and exit stages questionnaire

Variable	Id	Question	Scale	Source
Pre-task familiarity	**PR-1**	How familiar are you with this topic?	[0-4], where 4=Very familiar... 0=Not familiar at all	Pre-Task
Pre-task perceived knowledge	**PR-2**	How will you rate your knowledge on this topic?	[0-4], where 4=Expert... 0=New to the topic	Pre-Task
Interest	**PR-3**	How interested are you to learn about this topic?	[0-4], where 4=Very interested 0=Not interested at all	Pre-Task
Perceived difficulty	**PR-4**	How difficult do you think it will be to gather information about this topic	[0-4], where 4=Very difficult.. 0=Very easy	Pre-Task
Pre-task knowledge test	**PR-[5-7]**	Questionnaire for measuring pre-task topical knowledge Refer Tables 9.12 & 9.13	N/A	Pre-Task
Clarity	**PO-1**	The document snippets were clear and concise.	[0-4], where 4=Very clear... 0=Not clear at all	Post-Task
Informative	**PO-2**	The document snippets were informative.	[0-4], where 4=Very informative... 0=Not informative at all	Post-Task
Readable	**PO-3**	The document snippets were readable.	[0-4], where 4=Very readable... 0=Not readable at all	Post-Task
Useful snippets	**PO-4**	The document snippets were useful in finding relevant information about the topic.	[0-4], where 4=Very useful... 0=Not useful at all	Post-Task
Helped to learn	**PO-5**	The document snippets helped me to learn about a topic.	[0-4], where 4=Quite a lot.. 0=Did not help at all	Post-Task
Post-task familiarity	**PO-6**	How familiar are you with this topic after gathering information on the topic?	[0-4], where 4=Very familiar... 0=Not familiar at all	Post-Task
Post-task perceived knowledge	**PO-7**	How would you rate your knowledge on this topic?	[0-4], where 4=Expert 0=New to the topic	Post-Task
Perceived learning	**PO-8**	How much do you think you learnt on this topic?	[0-4], where 4=Quite a lot 0=Nothing at all	Post-Task
Duplicate documents	**PO-9**	How many documents do you think were duplicates?	[0-4], where 4=Quite a lot 0=No duplicates at all	Post-Task
Useful documents	**PO-10**	How many documents were useful to help you learn about the topic?	[0-4], where 4=Quite a lot 0=Not useful at all	Post-Task
Post-task knowledge test	**PO-[11-19]**	Questionnaire for measuring post-task topical knowledge Refer Tables 9.12 & 9.13	N/A	Post-Task
Task Feedback	**PO-20**	Feedback on the whole exercise and any difficulties (if faced)?	N/A	Post-Task

Table 9.10: User study experimental design pre- and post task questionnaires

9.3.2 Tasks and System

The simulated work-task statement that was shown to the users is presented below.

- *Task statement:* For your college quiz you have to gather information and learn about the Topic (C1):“Cloning of the sheep Dolly”.
- *Task statement:* For your college quiz you have to gather information and learn about the Topic (C2):“1998 Nobel peace prize”.

Next we describe the interface setup and the instructions that were given to the participants:

- *Interface Setup:* Similar to the pilot study we selected top 20 relevant documents for each topic. As described earlier, three types of SERPs were generated per topic using *Baseline*, *Relevance* and *Novelty* snippet generation models.
- *Instructions:* After the task statement participants were shown following instructions: “To help you gather information about the topic, our system has performed a search for you and found the 20 top documents. Our system has generated snippets for each of the documents, to help you navigate easily and help you in finding useful information and learn about the topic. Once you think that you have finished gathering information and have read enough documents, you can move to the next stage. You will be asked to answer a few questions on your experience in gathering information and learning about the topic. You will be given a questionnaire to test how much you have learnt about the topic.”

9.3.3 Participants and Setup

We used the Prolific crowdsourcing platform[3] to conduct the user study. The study was hosted on our server. We performed pre-screening of the participants to select

[3] `https://prolific.ac/`

those who we anticipated would interact with the documents as required by the task rather than skipping the interactions to complete the questionnaires. We pre-screened participants based on the following criteria provided by the platform: first language being English, country of residence and birth being USA and UK, prior approval rate of 90% on the platform, and minimum prior accepted task-submissions on the system >=50. Each participant was paid about 7.5 euro on the completion of the study.

Overall 39 participants completed the study out of which 5 participants only completed the first topic and thus were not included for the data analysis. To make sure that users had read documents and not just skimmed them, we removed four users who spent less than 4 minutes in the documents interaction stage on both the topics, similar to the screening done by Syed and Collins-Thompson (2017) for measuring user knowledge gain in a search task. Data analysis was done on the data collected from the 30 participants who completed the task correctly, the demographics of the users were 17 Female and 13 Male. The average age of the participants was 31.5, with minimum age of 22 and maximum age of 47.

Task Distribution: We wanted to keep the whole experiment around 1 hour to avoid participants feeling fatigue and getting disinterested. Each task took on average about 30 minutes, as indicated by the pilot study thus each participant interacted with two topics in the experiment. We followed a “between-within” group design mechanism for our user-study as presented in Table 9.11, similar to the work done by Collins-Thompson et al. (2016) for measuring learning outcomes in a web search. Study was “between group” since for the same topic we vary the SERP model and “within group” keeping the SERP model the same, we vary the topic. We rotated the order of topics within each group to avoid the learning effect, which balances the effect of the user gaining familiarity with the interface and the experiment, after performing Task-1 which can impact on the results for Task-2. Thus within each Group-1,2,3 and the corresponding model-1,2,3 respectively, half of the participants performed Topic:C1 first and Topic:C2 second, and the other half of the participants performed Topic:C2 first and Topic:C1 second. Overall, we perform analysis of the data captured from 30 participants, 10 per group.

Group	Stage-1	Stage-2
Group-1	SERP model-1, Task-1	SERP model-1, Task-2
Group-2	SERP model-2, Task-1	SERP model-2, Task-2
Group-3	SERP model–3, Task-1	SERP model-3, Task-2

Table 9.11: Tasks distribution

9.3.4 Data collection

Similar to the pilot study we captured data in our user study using Questionnaire and Logging users interactions. To measure knowledge gain we designed topic specific pre- and post-tests.

- Questionnaire: As shown in Table 9.9 and 9.10, we collected data from the participants at several stages as described in Section 9.3.1. For questions ES-[1-3], EX-[1-3] for the overall study and PR-[1-4], PO-[1-10] for each task, we required user responses on a likert scale to measure user experience, perceived topic familiarity and knowledge gain. We also asked the users to provide open feedback (PO-20) for each task, and overall feedback (EX-4) and suggestions (EX-5) on the completion of the study.

- Logging of users interactions: All the user interactions with the system in terms of the document clicked, mouse movements and time spent at different stages were recorded.

Measuring knowledge and learning using Bloom's Taxonomy

We used the bloom's model (Bloom, 1956) modified and developed by Anderson et al. (2001) in our work. This model is commonly used for classifying educational learning objectives into multiple categories: *Remember, Understand, Apply, Analyse, Evaluate and Create.* To measure the changes in user knowledge level and assess how much participants have learnt in a search task, we designed topic specific questions to ask at pre-task and post-task stages. Following the Bloom's taxonomy paradigm, we focused on 3 types of questions for measuring knowledge gain in our work.

- *Factual one-word answer questions* (**FOWA**): We designed questions and measure how well the participants remembered the basic and general information on the topic. The main focus was to measure the *remembering* skill associated with the lower cognitive level in the Bloom's taxonomy.

- *Single answer multi-choice questions* (**SAMC**): We designed questions to measure how well the participants remembered and recognised the general information from the documents. The main focus was to measure the *remembering and understanding* skills associated with the lower cognitive level in the Bloom's taxonomy.

- *Open-ended questions* (**OE**): We designed questions to analyse the user's understanding of the topic and its related aspects, to measure how well user had understood the topic. The main focus was to measure the *remembering, understanding, analysing and creating* skills associated with the higher cognitive level in the Bloom's taxonomy.

For each Topic C1 & C2, we designed nine questions, three question for each type of category: FOWA, SAMC and OE as shown in Table 9.12 and 9.13. People have different prior knowledge on a topic, and their learning can be affected by their existing knowledge of the topic. So instead of comparing participant's average post-test scores across three groups we compare the average of the difference of their post and pre-test scores. Thus for each topic we selected three questions, one question from each category of FOWA, SAMC and OE and ask these in the pre-test and all nine questions in the post-test. Each question was scored by the author based on the answer key shown in Tables 9.12 and 9.13. For each question in FOWA and SAMC, participants were assigned a score of "1" for each correct answer thus giving a maximum score of “3” for both FOWA and SAMC category. For each question in OE we scored the answers out of “3” (minimum number of sentences participants are asked to write on each question), for each aspect that matches and is related to the aspects mentioned in the answer key, participants were assigned a score of “1” with a maximum score of “3”. Thus the overall score for OE category for each topic was “9”. The overall maximum score for each topic was “15 (3+3+9)”.

$$\begin{aligned}
Knowledge\ gain\text{-}1 &= \frac{Post\ test\ score(FOCA) - Pre\ test\ score(FOCA)}{Max\ score(3)}\\
Knowledge\ gain\text{-}2 &= \frac{Post\ test\ score(SAMC) - Pre\ test\ score(SAMC)}{Max\ score(3)}\\
Knowledge\ gain\text{-}3 &= \frac{Post\ test\ score(OE) - Pre\ test\ score(OE)}{Max\ score(9)}\\
Overall\ Knowledge\ gain &= \frac{Post\ test\ score(Combined) - Pre\ test\ score(Combined)}{Max\ score(15)}
\end{aligned} \tag{9.1}$$

Following the learning gain formula used in earlier work by Syed and Collins-Thompson (2017) and Pirolli and Kairam (2013), we measure the knowledge gain using Equation 9.1. To assess user’s knowledge gain, we measured their overall change in the post- and pre-test scores, and individual category based post- and pre-test scores as shown in Equation 9.1. We measure four types of knowledge gain for each topic. Knowledge gain-1,2,3 measures the difference in post- and pre-test scores for FOWA, SAMC and OE question types. Overall knowledge gain is measured by the overall difference in the post- and pre-test scores (combining FOWA, SAMC and OE test scores) in a search task.

9.3.5 Results & Analysis

Tables 9.14, 9.15 and 9.16 show results of the user experience, knowledge gain and user behaviour for the crowdsource-based study for the two topics C1 & C2. Table 9.17 presents the average results for both topics C1 & C2. We conducted a pairwise student’s t-test (independent) across all three systems to measure if the difference between two systems is statistically significant.

ID	Factual one-word answer questions (FOWA)
PO-11	When was the first cloning done (year or time period)?
Answer-11	1996 / 1997 / late 1990's (either of them is acceptable)
PO-12	Where was the first cloning done (country or continent name)?
Answer-12	UK / Scotland / Edinburgh (either of them is acceptable)
PO-13	Can the cloned animal conceive babies?
Answer-13	Yes, they can.
ID	Single answer multi-choice questions (SAMC)
PO-14	Which counties opposed or called for stricter actions regarding genetic cloning?
Options	**i)** Kuwait, Israel, China; **ii)** USA, UK; **iii)** Germany, China; **iv)** None of the above
Answer-14	**i)** Kuwait, Israel, China
PO-15	How was the cloning of the sheep done? (Process)
Options	**i)** Taking a cell from a sheep's egg cell and a sheep's mammary gland ; **ii)**Frozen cells stored in chemicals; **iii)** Both A and B; **iv)**None
Answer-15	**iii)** Both A and B;
PO-16	Which species have been successfully cloned?
Options	**i)** Humans, Cows, Mice, Sheep; **ii)**Cows, Mice, Sheep; **iii)**Only Sheep; **iv)**All of the above are true
Answer-16	**ii)**Cows, Mice, Sheep;
ID	Open-ended questions (OE)
PO-17	What were some dangers to the cloned sheep "Dolly"? (atleast 3 points)
Answer-17	**Different aspects:** Premature death; aging cells; short life expectancy; illness; problems with the offspring.
PO-18	What are the benefits and applications of performing animals cloning? (atleast 3 points)
Answer-18	**Different aspects:** Foodstock & livestock growth; better health care and medical facilities; agriculture benefits; understanding and fighting diseases; growth of best breed of animals
PO-19	What are some benefits and dangers of Human Cloning? (atleast 3 points)
Answer-19	**Different aspects:** Cure genetic diseases; replace organs; better offspring with desired characteristics; ethical issues with respect to human cloning; risks associated with human cloning.
ID	Pre-test questions description
PR-5	Same as PO-12
PR-6	Same as PO-15
PR-7	Same as PO-18

Table 9.12: Questions designed for the topic C1: "Cloning of the sheep Dolly"

User Experience: As shown in Tables 9.14 and 9.17, for different user experience variables measured using user's rating on likert scale, average scores are quite similar. Results for the *novelty* model as compared to the *baseline* and *relevance* models, are much higher for the readability of snippets measure, and slightly better for the usefulness of snippets and perceived learning on the topic measures, but the differences in the results are not significant. Results for the *relevance* model as

ID	Factual one-word answer questions (FOWA)
PO-11	Who established the Nobel Prize (person name) ?
Answer-11	Alfred Nobel
PO-12	The winners of the Nobel Peace Prize 1998 came from which country?
Answer-12	Northern Ireland
PO-13	If there are more winners do everybody gets the same cash prize or is the cash prize divided between the winners?
Answer-13	Cash Prize is divided between the winners.
ID	Single answer multi-choice questions (SAMC)
PO-14	Who were the 1998 Nobel Peace Prize winners ?
Options	**i)** John Hume and David Trimble; **ii)**David Trimble; **iii)**John Hume; **iv)** None of the above
Answer-14	**i)** John Hume and David Trimble
PO-15	Which area/discipline does not have a nobel prize associated with it ?
Options	**i)** Mathematics; **ii)**Literature; **iii)**Physics; **iv)**Peace
Answer-15	**i)** Mathematics
PO-16	Where does the Nobel Prize ceremony takes place (country name) ?
Options	**i)**Norway; **ii)**Sweden; **iii)**Norway for peace and Sweden for rest; **iv)**Sweden for peace and Norway for rest
Answer-16	**iii)**Norway for peace and Sweden for rest
ID	Open-ended questions (OE)
PO-17	What did the 1998 Nobel Peace Prize winners do? Why were they given the prize, their contributions? (atleast 3 points)
Answer-17	**Different aspects:** Peace effort between Norther Ireland and Ireland; signing of Good Friday agreement in 1998; ending guerrilla war by Irish Republican Army (IRA); improving relationship between Ireland and Norther Ireland to end warfare going from couple of decades.
PO-18	How did the other political leaders reacted and responded to the winners declaration?
Answer-18	**Different aspects:** Tony Blair praised the winners for their effort and work; the news was received with appreciation from most politician leaders; people also praised efforts of Gerry Adams in signing of the peace agreement; expected that the prize will further strengthen the peace efforts and bring further stability in the area.
PO-19	Background on Nobel Prize selection process and committee?
Answer-19	**Different aspects:** Selection process is secretive; in 1998 there were about 139 nominations; person has to be nominated by someone; winners are announced on Oct 16 and ceremony happens on Dec 10; 5-6 committee members; 6 categories of Nobel prizes.
ID	Pre-test questions description
PR-5	Same as PO-11
PR-6	Same as PO-16
PR-7	Same as PO-19

Table 9.13: Questions designed for the topic C2: “1998 Nobel Peace Prize”

	Topic:C1			Topic:C2		
	Baseline	Relevance	Novelty	Baseline	Relevance	Novelty
Clarity	2.60	**3.10**	2.70	2.20	**2.60**	2.40
Informative	**3.20**	**3.20**	2.90	**2.90**	**2.90**	**2.90**
Readable	3.10	2.90	**3.40**	2.60	2.70	**3.20**
Useful	**3.00**	2.70	2.90	2.70	**3.00**	2.90
HelpToLearn	**3.10**	3.00	3.00	2.60	**3.00**	2.90
Perceived Learnt	2.80	**2.90**	2.80	2.30	2.50	**2.90**
Duplicates documents	**1.90**	1.80	1.70	**2.40**	2.00	2.10
Useful documents	3.00	**3.10**	**3.10**	**2.80**	**2.80**	**2.80**
Post - Pre Know	**1.40**	1.10	1.00	1.10	0.90	**1.50***
Post - Pre Familiarity	**1.90**$^{\gamma}$	0.40	1.30^{+}	1.80	1.70	**2.20**

Table 9.14: User Experience: Topic C1 & C2, best scores are in boldface. $+$ and $*$ indicates that the difference in the results between the relevance and novelty system is statistically significant with $p<0.05$ and $p<0.1$ respectively, γ indicates that the difference in the results between the relevance and the baseline system is statistically significant with $p<0.05$ respectively using student's t-test.

	Topic:C1			Topic:C2		
	Baseline	Relevance	Novelty	Baseline	Relevance	Novelty
Knowledge gain-1	0.77	0.73	**0.82**	0.75	0.63	**0.77**
Knowledge gain-2	**0.47**	0.40	0.43	0.63	0.70	**0.87**$^{\alpha}$
Knowledge gain-3	0.42	0.41	**0.56***	0.46	0.49	**0.65**$^{*\beta}$
Overall knowledge gain	0.50	0.47	**0.58**	0.55	0.56	**0.72**$^{+\beta}$

Table 9.15: Knowledge Gain: Topic C1 & C2, best scores are in boldface. $+$ and $*$ indicates that the difference in the results between the relevance and novelty system is statistically significant with $p<0.05$ and $p<0.1$ respectively, β and α indicates that the difference in the results between the novelty and the baseline system is statistically significant with $p<0.05$ and $p<0.1$ respectively using student's t-test.

	Topic:C1			Topic:C2		
	Baseline	Relevance	Novelty	Baseline	Relevance	Novelty
Documents Clicked	**15.00**	13.60	12.20	10.80	**12.30**	12.20
Time Spent Viewing	**13.70**	9.50	13.20	11.50	**13.00**	12.45
Overall Time Spent	**28.90**	28.40	28.70	20.10	**31.70**$^{\omega}$	27.10

Table 9.16: User Interactions: Topic C1 & C2. ω indicates that the difference in the results between the relevance and the baseline system is statistically significant with $p<0.1$ respectively using student's t-test.

compared to the *baseline* and *novelty* models are higher for the clarity measure and slightly better for helping participants to learn from the snippets, but the differences in the results are again not significant. Participants reported finding more duplicate documents in the *baseline* model as compared to the *relevance* and *novelty* models

	Topic C1 + C2		
	Baseline	Relevance	Novelty
Clarity	2.40	**2.85**	2.55
Informative	**3.05**	**3.05**	2.90
Readable	2.85	2.80	**3.30**
Useful	2.85	2.85	**2.90**
HelpToLearn	2.85	**3.00**	2.95
Perceived Learnt	2.55	2.70	**2.85**
Duplicates documents	**2.15**	1.90	1.90
Useful documents	2.90	**2.95**	**2.95**
Post - Pre Know	**1.25**	1.00	**1.25**
Post - Pre Familiarity	**1.85**	1.05	1.75
Knowledge gain-1	0.76	0.68	**0.80**
Knowledge gain-2	0.55	0.55	**0.65**
Knowledge gain-3	0.44	0.45	**0.61**
Overall knowledge gain	0.53	0.52	**0.65**
Documents Clicked	12.90	**12.95**	12.20
Time Spent Viewing	12.60	11.25	**12.82**
Overall Time Spent	24.50	**30.05**	27.90

Table 9.17: Average scores for user experience, knowledge gain and user behaviour aspects

but the differences are not significant. The average scores of the difference of the post- and pre-topic familiarly are higher with the *baseline* model as compared to the *relevance* and *novelty* models. We got mixed results for the difference of post- and pre-knowledge scores for *relevance* and *novelty* models where for Topic:C1 the *relevance* model is slightly better than the *novelty* model, but the difference is not significant. However for Topic:C2, the *novelty* model is significantly better than the *relevance* model. Overall, apart from the difference between post and pre-familiarity, among all other nine observed variables, for two variables (snippet informativeness and the difference between post- and pre-task knowledge) the SERP generated by the *novelty* and *relevance* models score is similar to the *baseline* model, and for all other seven variables (clarity, readable, useful, helpful to learn, perceived learning, duplicate document, useful documents) the SERP generated using the *relevance* and *novelty* models scores is better than the *baseline* model.

Knowledge Gain: As shown in Tables 9.15 and 9.17, the *novelty* model shows better knowledge gain for both topics: C1 & C2. Using the *novelty* model the results for the Topic:C1 are significantly better for knowledge gain-3 as compared to the *relevance* model. The results of the *novelty* model for the Topic:C2 are significantly better than both the *baseline* and *relevance* model for the knowledge gain-3 and the overall knowledge gain measures. A major difference is obtained for knowledge gain-3 as compared to the knowledge gain-1 and knowledge gain-2 measures. Knowledge gain-3 is associated with assessing higher cognitive level skills, thus the scores indicate that the SERP generated using *novelty* model helped users to learn about a topic and improve their topical knowledge effectively, as compared to the *baseline* and *relevance* models.

Behaviour and interactions: As shown in Table 9.16, there is no statistical difference across the three models in terms of the time spent interacting with the documents and number of documents clicked. There is considerable difference in the overall time spent for Topic:C2 across three systems, participants interacting with *baseline* model spent more time as compared to the *relevance* and *novelty* models. The difference in results of *relevance* and *baseline* models is statistically significant. Participants clicked more documents for Topic:C1 as compared to Topic:C2 across all three models.

Correlation: We calculated Pearson correlation to investigate how the user perceived knowledge and the user interactions variables correlate with the overall knowledge gain measure. Results indicate that the difference between the perceived post- and pre-task familiarity is positively correlated with the overall knowledge gain (r=0.47, $p < 0.01$), where r indicates the degree and strength of correlation and p indicates the statistical significance of the correlation. The difference between the perceived post- and pre-topical knowledge also shows positive correlation with the overall knowledge gain (r=0.26, $p < 0.05$). Time spent viewing the documents shows positive correlation with the overall knowledge gain (r=0.34, $p < 0.01$) and the number of documents clicked also shows a positive correlation with the overall knowledge gain (r=0.28, $p < 0.05$). For other user experience variables the correlation values are not significant. Overall, the results indicate that user interactions measured using documents clicked and time spent viewing the documents positively correlates with the knowledge gain, and that the user perceived difference in knowledge gain and familiarity is also related to the actual learning measured through the difference of post- and pre-test questionnaires. Similar findings regarding the correlation of i) actual and perceived learning, and ii) time spent interacting with documents and actual learning were observed in an earlier work on measuring learning outcomes in web search (Collins-Thompson et al., 2016).

9.3.6 Discussion

In this section, we present our analysis of the feedback provided by the participants in the user study. Then we discuss the participant suggestions to develop effective snippets and summaries to help them learn about the topic effectively.

Feedback Analysis: The main points from the feedback provided by the participants are shown in Table 9.18. Most of the participants expressed they learnt a lot. However, they found that some information was quite repetitive such as key names, dates in the snippets. As we used the top 20 relevant documents (manually judged for relevance by the TREC assessors) for generating snippets and the documents were judged independently thus they contained overlapping information. In future, we will like to address this problem and try to provide non-repetitive documents to the users.

Positive aspects	**Negative aspects**
Easy to read.	Quite some information to take in.
Easy for a person new to the topic, covers diverse information on the topic.	Hard to remember, memorise key names, dates.
Expressed they learnt a lot	Reported snippets were lengthy.
Can talk and discuss in public about these topics.	Found similar or repetitive information e.g. names of Nobel prize winners.
Enjoyed the exercise and felt more informed on the topics.	Three users felt the topics boring and indicated preference for other topics.

Table 9.18: User study feedback

Participant Suggestions: We asked participants to suggest what kind of document snippets and summaries they would like the system to generate to help them learn effectively. Following are the key points from our analysis of the participant's responses.

- Many participants suggested including audio and visual content such as photos and YouTube videos on the topic in general. They expressed the desire that using images of the Nobel prize winners and the process of cloning would make it easy to memorise and learn about the topic.
- Key names and dates could be presented differently (e.g. font or colour wise), or maybe with highlighting of the key events and facts within the text.
- The interface should present summaries which are easy to read and provide a general overview of the topic and go into detail as participants explore and get into the topic more deeply.
- Participants suggested generating different topic-based summaries (similar to a multi-document summary) rather than traditional individual document specific summaries to help them learn effectively, for e.g time-based summaries for the "Cloning" topic indicating when it happened, and the progress in the latter years related to the topic; separate summaries for the background information on the "Nobel prize", and specific event-based summary for the "1998 Nobel peace prize winners".

In our investigation on measuring snippets utility when presented in a SERP, the results analysis showed some interesting observations and findings, as described in Section 9.3.5. However the sample size is only 30, so the deductions and findings from the study should be explored with larger number of participants and more topics to see if the trends reported in this work are consistent.

9.4 Conclusion

In this work, we investigated how user search experience, interactions and knowledge gain changes when participants interact with snippets generated by three different models (*Baseline*, *Relevance* and *Novelty*) presented in a SERP. We conducted a lab-based study on 8 participants to get detailed feedback on the interface and our experimental setup. We then conducted a crowdsource-based study where we analysed data for 30 participants, where each participant interacted with either of the three different SERPs for two topics. Across the three models of snippet generation, results of the user experience were quite mixed. Overall the results of the user experience in the SERP generated by our framework were quite better than those for the *baseline* model. The difference between the results using different snippet models for the user experience variables were not significant.

Results of the knowledge gain measured using the difference of post- and pre-test scores were statistically significantly better for the *novelty* model as compared to the *baseline* and *relevance* models. The difference was more evident and significant for the open questions category (measuring user understanding of the topic) and the overall test scores, which indicates that the users learnt more in the SERP generated using the *novelty*-based snippet generation model.

User interactions measured in terms of time spent viewing documents and the number of documents clicked did not vary significantly across the three snippet creation models. However, time spent viewing the documents and the number of documents clicked in a SERP showed positive correlation with the overall knowledge gain scores measured using Pearson correlation. There is also a positive correlation between the perceived learning and the actual learning measured using the difference of post- and pre-test scores. Overall, the experimental investigation shows that the SERP generated using *novelty* model improved the user experience and helped participants to learn about the topic effectively as compared to the *baseline* and *relevance* model.

Next, we present the summary of this thesis work.

Chapter 10

Conclusion

In this chapter, we present the summary of our work. We revisit the research questions investigated in this thesis and describe our findings. Finally, we describe some of the limitations of our investigations and directions for future work.

10.1 Summary of our Work

Traditionally retrieved documents are returned in a search engine result page (SERP) where each document is represented as a snippet. A snippet seeks to represent the potential relevant information from a document to assess a user to gauge the usefulness of the document to satisfy their information need. In this thesis, we extend this goal of snippet generation (judging the usefulness of a document) to providing relevant, novel and easy to read information to improve user's topical knowledge and engagement, which we investigated in this work.

We developed a framework for generating document snippets for search engine results presentation in a web search. We focused on three main aspects namely: *relevance, novelty and readability* while generating effective snippets to be shown to the users. To generate effective snippets, we explored distributional semantic techniques (embeddings) along with traditional relevance models and query expansion approaches for relevance prediction. Further we explored bag-of-words, embeddings and syntax-based information for novelty prediction. We evaluated the performance of relevance and novelty models in terms of standard measures like precision and F-score. We used the Flesch reading ease score computational model for sentence-level readability prediction. We combined the relevance, novelty and readability features to generate alternative document snippets. We explored seven combinations of snippets by varying the weights of relevance, novelty and readability scores. We evaluated snippets generated by seven combination methods manually by scoring them using five measures: grammatical correctness, clarity, coherence, topicality and usefulness on a scale of [0-5]. We selected the best two snippet combination approaches for SERP presentation for a user-centred task-based study. Finally, we compared 3 different types of snippets: two generated by our framework (*Novelty model and Relevance model*) and a *Baseline model* (snippet generated by BM25-based relevance model) to measure snippets utility when presented in a SERP.

In our user-based study, we found that effective and richer snippets can help to improve user experience and learning. The average user experience scores for snippets developed using our framework (the *novelty and relevance*) models were higher than the *baseline* model. However, the difference was not significant. Participants found the *novelty*-based snippets to be more readable and useful, and their perceived learning was also higher in *novelty*-based snippets. Further, participants knowledge gain scores measured using the difference between post- and pre-test scores were significantly better for the *novelty* model than the *baseline* and *relevance* model. There was no statistical difference across the three models in terms of the user interactions measuring the time spent interacting with the documents, documents clicked and the overall time spent on a topic. However, time spent viewing the documents and the number of documents clicked in a SERP, showed positive correlation with the overall knowledge gain scores measured using Pearson correlation. There was also a positive correlation between the perceived learning and the actual learning measured using the difference of post- and pre-test scores. Overall, the user study showed that the SERP generated using *novelty* model improved the user experience and helped participants to learn about the topic effectively as compared to the *baseline* and *relevance* model. The participants who interacted with *novelty* model reported to have better user experience and learn effectively than the ones interacting with the *baseline* and *relevance* model. Thus we successfully investigated and addressed our overarching goal to create better informed and more satisfied searchers, with the development of effective snippets capturing sentence-level relevance and novelty information in this work.

Next, we revisit the research questions investigated in this work.

10.2 Research Questions Revisited

In this section we revisit the research questions outlined in Chapter 1.

- **RQ-1:** *Can we develop effective models to address the vocabulary mismatch issues of sentence-level relevance prediction?*

  The investigation for this question is described in Chapter 6. We explored distributional semantics for addressing query-sentence vocabulary mismatch issues. Our experimental results show that incorporating word embeddings information is effective and addresses the main challenges of word mismatch issues. We investigated and proposed different approaches to learn better query expansion terms to aid in understanding user's information need to improve sentence-level retrieval performance.

  A combined proposed model comprising of traditional pseudo relevance feedback and embedding-based expansion shows significant improvements for the task of sentence retrieval. There has not been much work exploring embeddings based QE and its combination with PRF for sentence retrieval experiments. Contrary to previous findings that embeddings-based expansion perform poorly in comparison to PRF-based expansion for document retrieval, our experiments suggest that for sentence retrieval using embedding-based expansion perform similar or better than PRF. We propose three reasons for these contradictory findings: i) the nature of data collections, ii) the semantic-based approaches explored in our work are more effective than previously used embeddings-based expansion techniques, and that iii) the problem of mismatch is more acute at the sentence level than at the document level because of the short nature of the sentences. Thus using embeddings seems to capture the words relationships effectively and find potential good expansion terms to boost the sentence-level retrieval performance.

- **RQ-2:** *Can we find novel information by comparing information within and across documents effectively?*

  The investigation for this question is described in Chapter 7. We explored sentence-based comparison methods for novelty detection. We investigated: i) bag-of-words-based distance metrics approach, ii) word and sentence embeddings-based approach and iii) novel syntax-based approach for novelty detection. A combination model capturing multiple features explored in (i), (ii) and (iii) seems to be quite effective and shows significant improvements for the novelty prediction for both 2003 and 2004 document collection.

- **RQ-3:** *How to combine sentence-level relevance, novelty and readability features to generate effective snippets?*

  The investigation for this question is described in Chapter 8. We developed a novel model, which generates topically relevant summaries and also compares the information across documents to avoid redundant and repetitive information in order to find new information to be presented in a SERP. We explored seven alternative snippet combination approaches by varying threshold of relevance, novelty and readability scores. These snippets were manually scored using different measures such as *grammatical correctness, clarity, coherence, topicality and usefulness* on the scale of [**1-5**]. Two best snippet combination approaches which scored best in terms of *usefulness* as compared to other alternative approaches explored were used for measuring snippets utility when presented in a SERP in a task-based setting.

- **RQ-4:** *How does user search behaviour and gain of topical knowledge vary using snippets generated by our framework?*
  The investigation for this question is described in Chapter 9. We evaluated three snippets generation approaches (*Novelty, Relevance, Baseline*) in a user-based study to measure snippets "utility". The difference between the results using different snippet models for the ten user experience variables were not significant, however there were some noticeable differences as reported in Chapter 9. Results of the knowledge gain measured using the difference of post- and pre-test scores were statistically significantly better for the *novelty* model as compared to the *baseline* and *relevance* models. The difference was more evident and significant for the open questions category (measuring user understanding of the topic) and the overall test scores, which indicates that the users learn more in the SERP generated using the *novelty*-based snippet generation model. User interactions measured in terms of time spent viewing documents and the number of documents clicked did not vary significantly across the three snippet creation models.

10.3 Limitations and Directions for Future Work

In this section we describe limitations of our work that could be addressed and explored in future work:

- **Limitation of using only relevant documents** – In general, web search results are a mixture of varying levels of relevant and non-relevant documents. Whereas in our study all the documents considered are relevant, thus ensuring that people are spending more time engaging and interacting with on-topic information. We speculate that if there were non-relevant results in our study, then participant's time would be divided between finding useful and relevant documents from among the ranked documents and interacting with these useful documents once identified, thus the gain in knowledge would be affected by introducing non-relevant results which we will like to explore in our future work.

- **Supervised techniques** – In this work we focused on unsupervised approaches for snippet generation. In future, we would like to explore more on the lines of deep-learning-based supervised techniques for summaries generation (Li et al., 2015).

- **Assessing snippet combination approaches** – For manual comparison of snippet creation and evaluation approaches described in Chapters 8 and 9, only one person (the author) did the annotation, based on the guidelines and rating schemes that were devised. As the goal was to compare alternative approaches to find the top approaches, it seemed reasonable to have only one annotator, following the typical TREC evaluation paradigm of one primary assessor for the relevance judgements where the goal is to compare system rankings (Voorhees, 2000). However, in future, we would like to repeat this exercise with multiple annotators and analyse how the annotation varies across users and how do it impacts the ranking of different combination approaches.

- **User study scale** – As described in Chapter 2, user-based evaluation is complex and challenging. We conducted a preliminary pilot experiment before the user study with a larger number of participants. We evaluated the effectiveness of three snippet generation models for two topics. We tried to control the learning effects by rotating the topics within each group. Although this study produced interesting findings, the sample size is only 30, so the study should be repeated with a larger number of participants and more topics to see if the trends reported in this work are consistent.

- **Studying knowledge gain** – We measure the changes in user knowledge by measuring their pre- and post-test scores. There may be other information which a participants might have learnt that we did not capture in our tests. Another factor which we did not consider in our experiment is the learning ability of an individual. A few participants mentioned they were not able to remember some names and dates, thus we would like to explore more on how learning ability impacts the user knowledge gain with input from education researchers.

- **Using only textual information** – In this work, we just focused on textual information. As reported in our study feedback, many participants suggested to include audio and visual content such as photos and YouTube videos on the topic in general. They expressed the desire to include images related to the topic (e.g winners of Nobel prize for our study) that would make it easy to memorise and learn about the topic. Thus in future we would like to explore multi-media elements focusing on text, images, videos to design better document representations.

Bibliography

Nasreen Abdul-jaleel, James Allan, W Bruce Croft, O Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. Umass at trec 2004: Novelty and hard. In *Proceedings of TREC-13*. Citeseer, 2004.

Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 345–354, 2011.

Mikhail Ageev, Dmitry Lagun, and Eugene Agichtein. Improving search result summaries by using searcher behavior data. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 13–22. ACM, 2013.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393, Montréal, Canada, 2012.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, 2013.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, 2014.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, 2015.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, et al. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, volume 1998, pages 194–218. Citeseer, 1998.

James Allan, Victor Lavrenko, and Hubert Jin. First story detection in tdt is hard. In *Proceedings of the 9th ACM International Conference on Information and Knowledge Management*, pages 374–381. ACM, 2000.

James Allan, Courtney Wade, and Alvaro Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 314–321. ACM, 2003.

James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the Second Strategic Workshop on Information Retrieval in Lorne. *SIGIR Forum*, 46(1):2–32, 2012.

John R Anderson. Memory, language, and thought, 1976.

Lorin W Anderson, David R Krathwohl, and Benjamin Samuel Bloom. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon, 2001.

Piyush Arora and Gareth JF Jones. Identifying useful and important information within retrieved documents. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 365–368. ACM, 2017a.

Piyush Arora and Gareth JF Jones. How do users perceive information: Analyzing user feedback while annotating textual units. In *Supporting Complex Search Task Workshop at Conference on Conference Human Information Interaction and Retrieval*, pages 7–10, 2017b.

Piyush Arora, Jennifer Foster, and Gareth J. F. Jones. Applying query formulation and fusion techniques for cross language news story search. In *Post-Proceedings* *of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation*, FIRE '12 & '13, pages 10:1–10:9, 2013a.

Piyush Arora, Jennifer Foster, and Gareth JF Jones. Dcu at fire 2013: Cross-language! ndian news story search. In *Forum for Information Retrieval Evaluation (FIRE 2013), New Delhi, India*, 2013b.

Piyush Arora, Debasis Ganguly, and Gareth JF Jones. The good, the bad and their kins: Identifying questions with negative scores in stackoverflow. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1232–1239. ACM, 2015a.

Piyush Arora, Chris Hokamp, Jennifer Foster, and Gareth Jones. Dcu: Using distributional semantics and domain adaptation for the semantic textual similarity semeval-2015 task 2. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 143–147, 2015b.

Piyush Arora, Debasis Ganguly, and Gareth JF Jones. Nearest neighbour based transformation functions for text classification: A case study with stackoverflow. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 299–302. ACM, 2016.

Piyush Arora, Jennifer Foster, and Gareth J. F. Jones. *Query Expansion for Sentence Retrieval Using Pseudo Relevance Feedback and Word Embedding*, pages 97–103. Springer International Publishing, 2017.

Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. Towards a science of user engagement (position paper). In *WSDM workshop on user modelling for Web applications*, pages 9–12, 2011.

Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247, 2014.

Nicholas Belkin, Toine Bogers, Jaap Kamps, Diane Kelly, Marijn Koolen, and Emine Yilmaz. Second workshop on supporting complex search tasks. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 433–435. ACM, 2017.

Nicholas J Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5(1):133–143, 1980.

Nicholas J Belkin. On the evaluation of interactive information retrieval systems. *The Janus Faced Scholar*, 2010.

Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. Ask for information retrieval: Part i. background and theory. *Journal of Documentation*, 38(2):61–71, 1982a.

Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. Ask for information retrieval: Part ii. results of a design study. *Journal of Documentation*, 38(3): 145–164, 1982b.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb): 1137–1155, 2003.

Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the Association for Computational Linguistics 2004, on Interactive poster and demonstration sessions*, page 31, 2004.

Benjamin Samuel Bloom. Taxonomy of educational objectives: The classification of educational goals . 1956.

Dasha Bogdanova and Jennifer Foster. This is how we do it: Answer reranking for open-domain how questions with paragraph vectors and minimal feature engineering. In *Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1290–1295, 2016.

Pia Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research. An International Electronic Journal*, 8(3), 2003.

Katriina Byström and Kalervo Järvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, 1995.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine learning*, pages 129–136. ACM, 2007.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM, 1998.

Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics, 2005.

Hao Chen and Susan Dumais. Bringing order to the web: Automatically categorizing search results. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 145–152. ACM, 2000.

Ruey-Cheng Chen, Damiano Spina, W Bruce Croft, Mark Sanderson, and Falk Scholer. Harnessing semantics for answer sentence retrieval. In *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 21–27. ACM, 2015.

Ruey-Cheng Chen, Evi Yulianti, Mark Sanderson, and W. Bruce Croft. On the benefit of incorporating external features in a neural architecture for answer sentence selection. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1017–1020. ACM, 2017.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Charles LA Clarke, Eugene Agichtein, Susan Dumais, and Ryen W White. The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 135–142. ACM, 2007.

Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 403–412. ACM, 2011. ISBN 978-1-4503-0717-8.

Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 163–172. ACM, 2016.

Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Impacts of time constraints and system delays on user experience. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 141–150. ACM, 2016.

W Bruce Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pages 1–36. Springer, 2002.

W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 283. Addison-Wesley Reading, 2010.

Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–416. ACM, 2007.

Dipanjan Das and André FT Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195, 2007.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, 1990.

Brenda Dervin. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. *Qualitative research in information management*, 9: 61–84, 1992.

Fernando Diaz, Bhaskar Mitra, and Nick Craswell. Query expansion with locally-trained word embeddings. *arXiv preprint, arXiv:1605.07891*, 2016.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Taoufiq Dkaki, Josiane Mothe, and Jérôme Augé. Novelty track at irit-sig. In *Proceedings of TREC-11*, 2002.

Samuel Dodson, Luanne Freund, and Rick Kopak. Do highlights affect comprehension?: Lessons from a user study. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 381–384. ACM, 2017.

Harold P Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285, 1969.

Miles Efron, Peter Organisciak, and Katrina Fenlon. Improving retrieval of short texts through document expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 911–920, 2012. ISBN 978-1-4503-1472-5.

Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pages 223–232. ACM, 2014.

David Ellis. A behavioural approach to information retrieval system design. *Journal of Documentation*, 45(3):171–212, 1989.

J. R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.

Luanne Freund, Jiyin He, Jacek Gwizdka, Noriko Kando, Preben Hansen, and Soo Young Rieh. Searching as learning (sal) workshop 2014. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 7–7, 2014a.

Luanne Freund, Heather O'Brien, and Rick Kopak. Getting the big picture: supporting comprehension and learning in search. *Proceedings of Search As Learning (SAL) workshop, IIiX 2014*, 2014b.

Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth J.F. Jones. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 795–798. ACM, 2015. ISBN 978-1-4503-3621-5.

Pierre-Etienne Genest and Guy Lapalme. Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 354–358. Association for Computational Linguistics, 2012.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*, pages 40–48. Association for Computational Linguistics, 2000.

Robert Gunning. The technique of clear writing. *McGraw-Hill, New York*, 1952.

Parth Gupta, Paul Clough, Paolo Rosso, Mark Stevenson, and Rafael E Banchs. Pan@ fire 2013: Overview of the cross-language indian news story search (clinss) track. 2013.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb):307–361, 2012.

Jacek Gwizdka, Preben Hansen, Claudia Hauff, Jiyin He, and Noriko Kando. Search as learning (sal) workshop 2016. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1249–1250, 2016.

Ivan Habernal, Maria Sukhareva, Fiana Raiber, Anna Shtok, Oren Kurland, Hadar Ronen, Judit Bar-Ilan, and Iryna Gurevych. New collection announcement: Focused retrieval over the web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 701–704. ACM, 2016.

Preben Hansen and Soo Young Rieh. Recent advances on searching as learning: An introduction to the special issue. *Journal of information science*, 42(1):3–6, 2016.

Donna Harman. Overview of the trec 2002 novelty track. In *Proceedings of TREC-11*, 2002.

Donna Harman and Paul Over. The duc summarization evaluations. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 44–51. Morgan Kaufmann Publishers Inc., 2002.

Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

Ahmed Hassan Awadallah, Ryen W White, Patrick Pantel, Susan T Dumais, and Yi-Min Wang. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 829–838. ACM, 2014.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1367–1377, 2016.

David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338. ACM, 1993.

Darwin P Hunt. The concept of knowledge and how to measure it. *Journal of Intellectual Capital*, 4(1):100–113, 2003.

Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. 37:241–72, 01 1901.

Bernard J Jansen, Danielle Booth, and Brian Smith. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, 45(6):643–663, 2009.

Jiepu Jiang, Daqing He, and James Allan. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 607–616. ACM, 2014.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161. ACM, 2005.

Evangelos Kanoulas, Ben Carterette, Mark Hall, Paul Clough, and Mark Sanderson. Overview of the TREC 2012 session track. 2012.

Tapas Kanungo and David Orr. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 202–211. ACM, 2009.

Makoto P Kato, Matthew Ekstrand-Abueg, Virgil Pavlu, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. Overview of the ntcir-11 mobileclick task. In *NTCIR*, 2014.

Mostafa Keikha, Jae Hyun Park, and W Bruce Croft. Evaluating answer passages using summarization measures. In *Proceedings of the 37th International ACM SIGIR Conference on Research & development in information retrieval*, pages 963–966. ACM, 2014a.

Mostafa Keikha, Jae Hyun Park, W Bruce Croft, and Mark Sanderson. Retrieving passages and finding answers. In *Proceedings of the 2014 Australasian Document Computing Symposium*, page 81. ACM, 2014b.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1—2):1–224, 2009.

Diane Kelly and Leif Azzopardi. How many results per page?: A study of serp size, search behavior and user experience. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 183–192. ACM, 2015.

Diane Kelly and Karl Gyllstrom. An examination of two delivery modes for interactive search system experiments: remote and laboratory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1531–1540. ACM, 2011.

Diane Kelly, David J Harper, and Brian Landau. Questionnaire mode effects in interactive information retrieval experiments. *Information Processing & Management*, 44(1):122–141, 2008.

Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*, pages 101–110. ACM, 2015.

Liadh Kelly, Johannes Leveling, Shane McQuillan, Sascha Kriewel, Lorraine Goeuriot, and GJF Jones. Report on summarization techniques. *Khresmoi project deliverable D*, 4:4, 2013.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

Birgit Kopainsky, Stephen M Alessi, and Pål I Davidsen. Measuring knowledge acquisition in dynamic decision making tasks. In *Proceedings of 29th International Conference of the System Dynamics Society, Washington, DC, USA*, pages 24–28, 2011.

David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

Carol C Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991.

Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference* *on Research and Development in Information Retrieval*, SIGIR '95, pages 68–73, 1995. ISBN 0-89791-714-6.

Saar Kuzi, Anna Shtok, and Oren Kurland. Query expansion using word embeddings. In *Proceedings of Conference on Information and Knowledge Management 2016*, pages 1929–1932, 2016.

Dmitry Lagun and Eugene Agichtein. Viewser: Enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th international ACM SIGIR Conference on Research and development in Information Retrieval*, pages 365–374. ACM, 2011.

Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4):1–132, 2014.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

Lorena Leal Bando, Falk Scholer, and Andrew Turpin. Query-biased summary generation assisted by query expansion. *Journal of the Association for Information Science and Technology*, 66(5):961–979, 2015.

Sungjin Lee. Online sentence novelty scoring for topical document streams. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 567–572, 2015.

David M Levy. I read the news today, oh boy: reading and attention in digital libraries. In *Proceedings of the second ACM International Conference on Digital libraries*, pages 202–211. ACM, 1997.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, 2014.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015.

Xiaoyan Li and W. Bruce Croft. Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 744–751. ACM, 2005.

Yuelin Li and Nicholas J Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, 2008.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

David E Losada. Statistical query expansion for sentence retrieval and its effects on weak and strong queries. *Information retrieval*, 13(5):485–506, 2010.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28 (2):203–208, 1996.

Inderjeet Mani. *Automatic summarization*, volume 3. John Benjamins Publishing, 2001a.

Inderjeet Mani. Summarization evaluation: An overview. 2001b.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. The tipster summac text summarization evaluation. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 77–85. Association for Computational Linguistics, 1999.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

Gary Marchionini. *Information seeking in electronic environments.* Number 9. Cambridge university press, 1997.

Mari-Carmen Marcos, Ferran Gavin, and Ioannis Arapakis. Effect of snippets on user experience in web search. In *Proceedings of the XVI International Conference on Human Computer Interaction*, pages 47:1–47:8, 2015.

David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings* *of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 135–144. ACM, 2017.

G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

Donald Metzler and Tapas Kanungo. Machine learned sentence selection strategies for query-biased summarization. In *SIGIR Learning to Rank Workshop*, pages 40–47, 2008.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, pages 775–780, Boston, Massachusetts, 2006.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint, arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013b.

Arunav Mishra and Klaus Berberich. How do order and proximity impact the readability of event summaries? In *European Conference on Information Retrieval*, pages 212–225. Springer, 2017.

Bhaskar Mitra and Nick Craswell. Neural models for information retrieval. *arXiv preprint, arXiv:1705.01509*, 2017.

Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. A dual embedding space model for document ranking. *arXiv preprint, arXiv:1602.01137*, 2016.

Diane Napolitano, Kathleen Sheehan, and Robert Mundkowsky. Online readability and text complexity analysis with textevaluator. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 96–100, 2015.

Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. *Mining text data*, pages 43–76, 2012.

Heather L O'Brien and Elaine G Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the Association for Information Science and Technology*, 59(6):938–955, 2008.

Jesús Oliva, José Ignacio Serrano, María Dolores del Castillo, and Ángel Iglesias. Symss: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering*, 70(4):390–405, 2011.

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

Mikko Pennanen and Pertti Vakkari. Students' conceptual structure, search process, and outcome while preparing a research proposal: a longitudinal case study. *Journal of the American Society for Information Science and Technology*, 54(8): 759–770, 2003.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Willett Peter, Barbara Wildemuth, Luanne Freund, and Elaine G. Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*, 70(6):1118–1140, 2014.

Peter Pirolli and Sanjay Kairam. A knowledge-tracing model of learning from a social tagging system. *User Modeling and User-Adapted Interaction*, 23(2-3):139–168, 2013.

Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM, 1998.

Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Ross J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore, 1992. World Scientific.

Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408, 2002.

R Rehurek and P Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

Rebecca B Reynolds. Relationships among tasks, collaborative inquiry processes, inquiry resolutions, and knowledge outcomes in adolescents during guided discovery-based game design in school. *Journal of Information Science*, 42(1):35–58, 2016.

Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, 42(1):19–34, 2016.

S Robertson, S Walker, S Jones, MM Hancock-Beaulieu, and M Gatford. Okapi at trec-3. *NIST special publication*, (500225):109–123, 1995.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, pages 42–49. ACM, 2004.

Stephen E Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304, 1977.

Stephen E Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.

Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608*, 2016.

Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. Interactive intent modeling: Information discovery beyond search. *Communications of the ACM*, 58(1):86–92, 2015.

Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint, arXiv:1509.00685*, 2015.

J. E. Rush, R. Salvador, and A. Zamora. Automatic abstracting and indexing. ii. production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, 22(4):260–274, 1971.

Ian Ruthven. Interactive information retrieval. *Annual Review of Information Science and Technology*, 42(1):43–91, 2008.

Ian Ruthven and Diane Kelly. *Interactive information seeking, behaviour and retrieval*. Facet Publishing, 2011.

Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the Association for Information Science and Technology*, 26(6):321–343, 1975.

Tefko Saracevic. The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the Annual Meeting-American Society for Information Science*, volume 34, pages 313–327, 1997.

Denis Savenkov, Pavel Braslavski, and Mikhail Lebedev. Search snippet evaluation at yandex: lessons learned and future directions. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 14–25. Springer, 2011.

Barry Schiffman and Kathleen McKeown. Columbia university in the novelty track at trec 2004. In *Proceedings of TREC-13*, 2004.

RJ Senter and Edgar A Smith. Automated readability index. Technical report, Cincinnati University Ohio, 1967.

Aliaksei Severyn and Alessandro Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 373–382. ACM, 2015.

Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184–209, 2014.

Ian Soboroff and Donna Harman. Novelty detection: the trec experience. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112. Association for Computational Linguistics, 2005.

Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34, 1948.

Dan Ştefănescu, Rajendra Banjade, and Vasile Rus. A sentence similarity method based on chunking and information content. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 442–453. Springer, 2014.

Rohail Syed and Kevyn Collins-Thompson. Optimizing search results for human learning goals. *Information Retrieval Journal*, 20(5):506–523, 2017.

Wenyin Tang, Flora S Tsai, and Lihui Chen. Blended metrics for novel sentence mining. *Expert Systems with Applications*, 37(7):5172–5177, 2010.

Milan Tofiloski, Julian Brooke, and Maite Taboada. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 77–80. Association for Computational Linguistics, 2009.

Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10. ACM, 1998.

Elaine G Toms. Task-based information searching and retrieval. *Interactive Information Seeking, Behaviour and Retrieval. Facet Publishing*, pages 43–59, 2011.

Matthew Trappett, Shlomo Geva, Andrew Trotman, Falk Scholer, and Mark Sanderson. Overview of the inex 2011 snippet retrieval track. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 283–294. Springer, 2011.

Matthew Trappett, Shlomo Geva, Andrew Trotman, Falk Scholer, and Mark Sanderson. Overview of the INEX 2013 snippet retrieval track. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*, 2013.

Andrew Trotman and Shlomo Geva. Passage retrieval and other xml-retrieval tasks. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, pages 43–50, 2006.

Flora S Tsai, Wenyin Tang, and Kap Luk Chan. Evaluation of novelty metrics for sentence-level novelty mining. *Information Sciences*, 180(12):2359–2374, 2010.

Yohannes Tsegay, Simon J Puglisi, Andrew Turpin, and Justin Zobel. Document compaction for efficient query biased snippet generation. In *European Conference on Information Retrieval*, pages 509–520. Springer, 2009.

Peter D Turney and Jeffrey Bigham. Combining independent modules to solve multiple-choice synonym and analogy. In *Problems". Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03*. Citeseer, 2003.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

Andrew Turpin, Falk Scholer, Kalvero Jarvelin, Mingfang Wu, and J. Shane Culpepper. Including summaries in system evaluation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 508–515, 2009.

Pertti Vakkari. Relevance and contributing information types of searched documents in task performance. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–9. ACM, 2000.

Pertti Vakkari. Task-based information searching. *Annual Review of Information Science and Technology*, 37(1):413–464, 2003.

Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.

Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

Ellen M Voorhees. The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 355–370. Springer, 2001.

Di Wang and Eric Nyberg. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 707–712, 2015.

Ryen W White, Ian Ruthven, and Joemon M Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 57–64. ACM, 2002.

Ryen W White, Joemon M Jose, and Ian Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing & Management*, 39(5):707–733, 2003.

Mathew J Wilson and Max L Wilson. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology*, 64(2):291–306, 2013.

Tom D Wilson. Models in information behaviour research. *Journal of Documentation*, 55(3):249–270, 1999.

Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W Bruce Croft, Jiafeng Guo, and Falk Scholer. Beyond factoid qa: effective methods for non-factoid answer sentence retrieval. In *European Conference on Information Retrieval*, pages 115–128. Springer, 2016.

Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th International Conference on World wide web*, pages 1011–1018. ACM, 2010.

Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, and Mark Sanderson. Using semantic and context features for answer summary extraction. In *Proceedings of the 21st Australasian Document Computing Symposium*, ADCS '16, pages 81–84, 2016.

Hamed Zamani and W. Bruce Croft. Embedding-based query language models. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 147–156. ACM, 2016.

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, 2001.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint, arXiv:1611.03530*, 2016.

Min Zhang, Chuan Lin, Yiqun Liu, Leo Zhao, and Shaoping Ma. Thuir at trec 2003: Novelty, robust and web. In *Proceedings of TREC-12*, pages 556–567, 2003.

Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium*, page 12. ACM, 2015.

Our Publications

The parts of Chapter 1 and Chapter 3 discussing the overview of our thesis, ideas, background work on snippet generation framework to support user engagement and learning has been published in:

- P. Arora, (2015). *Promoting User Engagement and Learning in Amorphous Search Tasks*. SIGIR 2015 Doctoral Consortium paper.

- P. Arora, and G. J. F. Jones, (2016). *Position Paper: Promoting User Engagement and Learning in Search Tasks By Effective Document Representation.* Search As Learning (SAL) Workshop at SIGIR, 2016.

The parts of Chapter 6, exploring query expansion techniques using Pseudo Relevance Feedback and Word Embedding for sentence retrieval has been published in:

- P. Arora, J. Foster, and G. J. F. Jones, (2017). *Query Expansion for Sentence Retrieval Using Pseudo Relevance Feedback and Word Embedding.* In Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017. Lecture Notes in Computer Science.

The parts of Chapter 5, initial investigations on exploring user search behaviour and user's interactions with information on the web has been published in:

- P. Arora, and G. J. F. Jones, (2017). *How do Users Perceive Information: Analyzing User Feedback while Annotating Textual Units.* Supporting Complex Search Task (SCST) Workshop at CHIIR, 2017.

- P. Arora, and G. J. F. Jones, (2017). *Identifying Useful and Important Information within Retrieved Documents.* In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17).

The parts of Chapter 7, exploring similarities of sentences for novelty detection, is adapted from our initial work and system submission for the task of semantic similarity of sentences published in:

- C. Hokamp, and P. Arora, (2016). *DCU-SEManiacs at SemEval-2016 Task 1: Synthetic Paragram Embeddings for Semantic Textual Similarity.* SemEval@ NAACL-HLT. 2016.

- P. Arora, C. Hokamp, J. Foster, and G. J. F. Jones, (2015). *DCU: Using Distributional Semantics and Domain Adaptation for the Semantic Textual Similarity SemEval-2015 Task 2.* In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 143-147).

The parts of Chapter 5, initial investigations on exploring and experimenting with traditional IR models, query expansion techniques, sentence embeddings for similar questions detection has been published in:

- P. Arora, D. Ganguly, and G. J. F. Jones, (2016). *Nearest Neighbour based Transformation Functions for Text Classification: A Case Study with StackOverflow.* In Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16).

- P. Arora, D. Ganguly, and G. J. F. Jones. (2015), *The Good, the Bad and their Kins: Identifying Questions with Negative Scores in StackOverflow.* In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15).

- P. Arora, J. Foster, and G. J. F. Jones, (2013).*Applying Query Formulation and Fusion Techniques For Cross Language News Story Search.* In Proceedings of Fifth Workshop of the Forum for Information Retrieval (FIRE 2013).

- P. Arora, J. Foster, and G. J. F. Jones, (2013). *DCU at FIRE 2013: Cross-Language !ndian News Story Search.* Working Notes of CL!NSS task at the Forum for Information Retrieval (FIRE 2013).

Other publications during PhD which are not directly related to this Thesis are:

- A.H. Vahid, P. Arora, Q. Liu, and G. J. F. Jones, (2015). *A Comparative Study of Online Translation Services for Cross Language Information Retrieval.* In Proceedings of the 24th International Conference on World Wide Web, MWA 2015, (pp. 859-864).

- J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, , and L. Tounsi, (2014). *DCU: Aspect-based polarity classification for semeval task 4.* In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 223-229).

- X. Wu, R. Haque, T. Okita, P. Arora, A. Way, and Q. Liu, (2014). *DCU-Lingo24 Participation in WMT 2014 Hindi-English Translation task.* ACL 2014, (pp. 215-220).