Utilizing sub-topical structure of documents for Information Retrieval

Debasis Ganguly Johannes Leveling Gareth J. F. Jones
Centre for Next Generation Localisation
School of Computing, Dublin City University, Dublin 9, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

ABSTRACT

Text segmentation in natural language processing typically refers to the process of decomposing a document into constituent subtopics. Our work centers on the application of text segmentation techniques within information retrieval (IR) tasks. For example, for scoring a document by combining the retrieval scores of its constituent segments, exploiting the proximity of query terms in documents for ad-hoc search, and for question answering (QA), where retrieved passages from multiple documents are aggregated and presented as a single document to a searcher. Feedback in ad hoc IR task is shown to benefit from the use of extracted sentences instead of terms from the pseudo relevant documents for query expansion. Retrieval effectiveness for patent prior art search task is enhanced by applying text segmentation to the patent queries. Another aspect of our work involves augmenting text segmentation techniques to produce segments which are more readable with less unresolved anaphora. This is particularly useful for QA and snippet generation tasks where the objective is to aggregate relevant and novel information from multiple documents satisfying user information need on one hand, and ensuring that the automatically generated content presented to the user is easily readable without reference to the original source document.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Query formulation, Relevance Feedback*; H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Abstracting methods*

General Terms

Experimentation, Performance, Measurement

Keywords

Document Segmentation, Query Segmentation

PIKM'11, October 28, 2011, Glasgow, Scotland, UK.
.

1. INTRODUCTION

Recent years have witnessed an upsurge in the quantity of news, encyclopedic articles, blogs, forum and social networking posts etc. on the web and elsewhere. Some of these, such as news and Wikipedia articles are carefully authored, edited and quality controlled, while others such as blogs and social networking posts are not. A document in the former category is often explicitly decomposed into paragraphs or sections to convey a specific aspect of the overall information in a more focused way. The sub-topic organizational pattern is more implicit in a document of the later category, due to the absence of explicitly demarcated paragraphs. The effective management and exploitation of this content often requires us to gain insights into its sub-topical structure of documents access aspects of the information it contains. The process of extracting this sub-topical structure is referred to as *text segmentation*. One popular application of text segmentation is in topic detection and tracking systems, which involve discovering the topical structure in unsegmented streams of news reports appearing across multiple media.

In our investigations we are exploring mechanisms by which application of text segmentation can improve IR effectiveness. Our work to date has demonstrated that text segmentation can play a significant role in areas including pseudo relevance feedback (PRF), automated pseudo query session generation and patent prior art search. Another potential application of text segmentation is in the long answer-type QA task where segmented portions of documents can be aggregated to form a single relevant and readable answer which forms the basis of our future research. In this paper we provide a survey of our existing work on the first three application areas followed by a proposal of extending existing text segmentation approaches to produce more readable segments for increasing the effectiveness of long-answer type QA systems.

2. RELATED WORK

One of the best known studies in text segmentation for natural language applications is the work of Hearst [8], which introduced the TextTiling algorithm designed to automatically segment documents into coherent sub-topics, by selecting the valleys in the smoothed plot of cosine similarities between adjacent blocks of sentences as potential topic shift points. A problem of this approach as pointed out by Reynar is that it is essentially local in nature ignoring long range similarities [19]. He pointed out that the text segmentation problem has an optimal substructure property and proposed a dynamic programming approach for solving the problem. His proposed method involves iteratively selecting a position as topic shift point based on the minimization of inter segment similarity density or maximization of intra segment similarity density of the similarity matrix, and then solving the subproblems

in an identical manner on the two sub parts thus formed at each step. Choi advocated using a sharpened similarity matrix obtained by using a rank filter for better topic shift predictions, and an efficient pre-computation technique of net similarities to gain speed, over Reynar's algorithm [2].

Some early work in IR involved text segmentation to decompose the documents of a collection into automatically constructed segments representing self-coherent sub-topics, and deriving the final retrieval score for a document by summing up the scores of individual segments. Hearst and Plaunt [7] reported that using author marked paragraphs as segments, yields better retrieval effectiveness than automatically constructing sub-topical segments spanning multiple paragraphs. Moffat et al. [16] report an improvement in retrieval effectiveness by the use of fixed size multi-paragraph units as segments. Subsequent research in IR focused on using fixed length textual units, i.e. fixed length passages or word windows. Xu and Croft [24] proposed Local Context Analysis (LCA) which involves decomposing the feedback documents into fixed length word windows so as to overcome the problem of choosing terms from unrelated portions of a long document, and then ranking the terms by a scoring function which depends on the co-occurrence of a word with the query term, the co-occurrence being computed within the fixed word length windows. Mitra et al. [15] use local term correlation weighted idf scores summed over fixed length windows to rerank a subset of the top ranked documents and then assume the reranked set as pseudo relevant. They report an improved mean average precision (MAP) and a decreased query drift with this approach. Lam-Adesina and Jones [10] report successful results for a query expansion based on query related sentences extracted retrieved documents.

3. RESULTS OF OUR STUDY TO DATE

3.1 Document Segmentation

This section summarizes our work so far on utilization of document segmentation for ad hoc IR tasks.

3.1.1 Improving PRF of ad-hoc retrieval

In query expansion for pseudo relevance feedback (PRF) it is wrongly assumed that a pseudo relevant document as a whole is relevant, which is generally not true as shown in [23]. Segmentation of the top ranked documents and restricting the choice of feedback terms only to the most similar segments may be helpful to improve retrieval [**?**].

The conventional feedback strategy in ad-hoc IR is to score terms from the pseudo-relevant set of documents by a function $f : t \rightarrow \mathbb{R}$ and then add the top scoring terms to the original query. The two limitations of this conventional feedback strategy are: i) a document as a whole is assumed to be relevant although only a few sentences of it may actually be relevant; ii) each document from the pseudo-relevant set of documents is assumed to be equally relevant, although it may be the case that some documents are more relevant than the others.

In our proposed feedback method which we call *SBQE* (Sentence Based Query Expansion), we address the first limitation by segmenting each document into sentences and adding the most similar sentences to the query thus restricting the choice of feedback terms at sub-document level. The second limitation is overcome by adding the most number of sentences from the top ranked document and decreasing the number of sentences to add for the subsequent documents of the pseudo-relevant set.

Our experiments on the TREC topics show that *SBQE* significantly outperforms the standard LM term based query expansion [18] and Relevance Based Language Model (RLM) [11]. For more details the reader is referred to [3].

3.1.2 Simulating query sessions

Retrieval over an entire query session is an emerging area of IR where the focus is to retrieve more relevant documents by taking hints from the query reformulation patterns within a session. In navigational browsing, it is observed that a query reformulation is based on the contents of documents viewed by a searcher in response to the previous queries [1]. Text segmentation can be useful in simulating the reading event of a segment of a document, analogous to the real-life event of a searcher reading a particular portion of a document, and using the information read to reformulate the initial query. Thus segmentation may be applied to automatically generate a test corpus of query sessions.

Observation of the fact that in a document specific terms are densely concentrated in a few sentences, whereas general terms are more uniformly distributed [8], led us to devise an algorithm to automatically generate more specific reformulations (by adding specific terms from top ranked documents) and more general reformulations (by removing or substituting more specific terms with general ones) from a given query. Experiments with TREC 8 topic titles as initial queries involving manual assessments to judge the quality of the reformulations, show that both specific and general queries can be generated with reasonable accuracy, with a greater inter-assessor agreement for general reformulations [4].

3.2 Query Segmentation

Queries in ad hoc and web search are typically very short comprising of a few keywords, thus excluding these queries from any potential for the application of text segmentation. However, in associative document search, full documents are used as queries to retrieve related documents from a collection. A particular example of this is patent prior art search, where an entire patent claim is used as a query to retrieve prior articles related to the claimed invention. The long queries being ambiguous are not suitable for high precision retrieval. Segmentation of a patent query may be useful to gain more insights into the intended information need being sought.

Our work centers around two methods of making patent queries less ambiguous and more specific to an information need. Both the methods are compared against a baseline of using a simple frequency cut-off filter to remove unit frequency terms from patent queries, shown to work well in previous work [13].

3.2.1 Using PRF to Reduce Queries

This method attempts to reduce queries by removing sentences or fixed word length windows which are most dissimilar to the pseudo-relevant documents [6]. This can be viewed as a process opposite to that of *SBQE*. Removal of sentences most dissimilar to the query from pseudo relevant documents ensures that the query starts *looking* more like the pseudo-relevant documents with the noisy terms being removed. Experiments with the CLEF-IP 2010 dataset show that the word window based removal of sentences (using a window size of 20) improves MAP by 7.28%.

3.2.2 Retrieval with Sub-queries

This method aims at utilizing the sub-topics of a patent query, which are more specific in nature discussing a particular aspect of the invention, as compared to the whole patent claim. The method involves segmenting the patent queries into sub-topics by TextTiling [8] and using each sub-topic as a separate query to retrieve and finally merge the results [5]. Extracting sub-topics from a full patent description has the effect of making the information need

expressed in each sub-query more focused, and the final merging step of interleaving documents retrieved for each sub-query has the effect of addressing each aspect of the claimed invention. PRF for patent prior art search tasks shows a degradation in MAP [9, 21, 12] primarily because the massive queries lack a specific focus towards a relevance criterion, initial retrieval precision is low, and added terms tend to make the query more ambiguous. However, PRF on the focused query segments expressing a precise information need is expected to benefit retrieval effectiveness. Our experiments on CLEF-IP 2010 data using 50 patent queries show that query segmentation alone increases PRES by 12.14%, whereas segmentation with PRF on each sub-query increases PRES by 14.05%.

4. LONG ANSWER TYPE QA

We now turn our attention to the last application area of our work and provide a brief background followed by our proposal.

4.1 Background

The QA tasks at TREC [22] primarily focused on providing exact answers to factoid questions such as "*Who invented the paper clip*?". List type questions required collecting short answers from multiple documents sources and presenting the final answer as a list, e.g. "*Which countries were visited by first lady Hillary Clinton?*". Additionally in TREC 2003 there was a separate sub-track dealing with definition questions such as "*What is mold?*".

Most participating systems in TREC used a shallow parsing approach on the documents obtained from an initial retrieval step to return desired entities such as persons for *who* questions and time for *when* questions. For the descriptive questions, the common approach was standard passage retrieval. The answers for definition questions are more difficult to judge than the short answer questions, where it is a simple matter of manually assessing whether the desired entity is present in the returned answer text or not.

For judging the answers to the definition questions, the approach undertaken by the NIST assessors were as follows. A pool of answer strings was presented to an assessor who was then asked to create a list of "information nuggets" about the target question. A vital information nugget was defined as a fact which must appear in an answer string to mark it relevant. Computation of nugget recall for an answer string is straightforward. Computing precision is trickier because it is very difficult to quantify the unique number of information nuggets in an answer string. Hence precision was estimated by the length of the answer strings. Final evaluation was done by combining precision and recall with the F-score measure.

Thus evaluation of the TREC QA task did not take into consideration the readability of an answer string. This is clarified with an example answer to the question "*What is a golden parachute?*" as shown below.

```
But if he quits or is dismissed during the two
years after the merger, he will be paid $24.4
million, with DaimlerChrysler paying the "golden
parachute" tax for him and the taxes on the
compensation paid to cover the tax.
```

The answer contains an unresolved pronoun and the clause before the starting *but* is also unknown. Thus, collecting most similar sentences to a query and concatenating them together can score high on relevance since precision will be fairly high due to the shorter length of sentences as compared to passages, but readability is likely to suffer.

The QA task at INEX attempts to strike a balance in the evaluation of long answers by taking into consideration readability as well [17]. In the long answer type QA task, answers have to be generated by aggregation of several passages from different documents on the Wikipedia. For example in response to the question "*Who is Mahatma Gandhi?*" a QA system should not only return "*Indian freedom fighter*" but should also return a short biographical sketch and an outline of major events in his life with dates. The maximum length of the answers is preset to a pre-defined limit.

These long answers are evaluated by measuring the readability and informative content. For measuring readability, assessors are asked to mark a position in each answer snippet where he thinks that the answer becomes unreadable due to inconsistent grammatical structures, unresolved anaphora or redundant information. These are called the "last point of interest" marks. Readability is then measured as a fraction of the position of the "last point of interest" mark with respect to the total length of the answer string. In addition for measuring relevance, the assessors are also asked to mark relevant passages from a pool of retrieved Wikipedia articles. Now given a set R of relevant passages and an answer text T, a distance between the word distributions is computed by the KL divergence. Thus, lesser the KL divergence between the word probability distributions R and T, higher is the relevance of T.

The general observations made from the evaluation of submitted runs in INEX 2010 QA task are as follows [20]:

1. The baseline system which returned sentences having the highest *LexRank* values extracted from the top 20 ranked passages scores highest on relevance and lowest on readability.
2. A system which returns long sentences scores the highest in readability and lowest in relevance.

The reason is that short sentences tend to have a higher probability of relevant word occurrence due to the short length of the answer text string which is used as the denominator in maximum likelihood probability estimate, and hence a less KL divergence. Long sentences on the other hand are typically devoid of anaphora and hence are more readable. However, this approach scores low on relevance due to the higher denominator value used in the maximum likelihood probability estimates.

4.2 Our Proposal

Submitted runs in the INEX QA task did not use text segmentation to define the retrievable units. However, it is reasonable to hypothesize that sub-topics obtained from the application of text segmentation should be self-coherent and hence readable units of text. In fact the INEX QA task in 2011 will define XML elements (sections and passages as authored in the Wikipedia articles) as retrievable units. This is a reasonable approach for informative and carefully written articles such as Wikipedia documents. But considering the general long answer type QA task where the target collection may well be informally written articles such blogs, forum posts etc., it can be postulated that arbitrary sentences aggregated from different documents may not suffice to constitute readable text.

Existing text segmentation methods involve a general methodology requiring the notion of a similarity function $sim(s_i, s_j)$ defined between every sentence pair s_i and s_j [19, 2, 14], or only between consecutive sentences [8]. Similarity is commonly measured cosine similarity between the sentence vectors or other measures such as the number of introductory words or the length of lexical chains etc. None of the existing techniques take into account anaphora resolution and hence can end up generating segments with unresolved anaphora thus hampering readability.

We propose two solutions based on a very simple anaphora resolution technique to augment existing text segmentation methods. The simple method of anaphora resolution involves defining a list of words comprising of pronouns (e.g. *he*), wh-adverbs (e.g. *however*), and co-ordinationg conjunctions (e.g. *but*). One can then use

this list as a post processing step on the output of any text segmentation algorithm to merge a segment, having a presence of one or more these continuation words in its first sentence, with the segment preceding it. Another solution is to use a vector of similarity values in the computational steps of a text segmentation algorithm itself. Instead of using one dimensional cosine similarity values, we propose to use a two dimensional vector to define the similarity $sim(s_i, s_j) = (sim_{cos}, sim_{lnk})$, whose first component sim_{cos} is the cosine similarity between s_i and s_j, and the second component sim_{lnk} is a measure of the amount of linkage depenedency computed as a function of position of the first ocurrence of a continuation word in a sentence s, denoted by $p(s)$ as follows.

$$sim_{lnk}(s_i, s_{i+1}) = \begin{cases} 0 & \text{if } p(s_{i+1}) = 0 \\ \frac{len(s_{i+1}) - p(s_{i+1})}{len(s_{i+1}} & \text{if } p(s_{i+1}) > 0 \end{cases}$$

$$sim_{lnk}(s_i, s_j) = \prod_{k=i}^{j-1} sim_{lnk}(s_k, s_{k+1}) \quad (1)$$

Equation 1 is used to propagate the linkage similarities in a chain. The product becomes zero if any sentence in the chain does not have a continuation word, since for such a sentence s, $p(s) = 0$ and hence the chain breaks.

The research questions are as follows:

1. Can an approach to long answer type QA with text segmentation output sub-topics as retrievable units, help in finding a sweet spot between relevance and readability?
2. Can readability be further improved by augmenting an existing text segmentation algorithm either by merging consecutive segments based on the occurrence of continuation words or extending a text segmentation algorithm itself by augmenting sentence similarities with the notion of anaphora dependencies?

We plan to implement the two approaches of producing readable segments. To test our approach we will use the INEX 2011 QA long answer type question collection. The retrievable unit for INEX 2011 QA task is an XML element in contrast to an arbitrary passage as in 2009 and 2010. The XML elements comprising of Wikipedia paragraphs and sections can be treated as the gold standard readable set. We expect runs with automatically inferred sub-topics as retrievable units, to perform in the middle range between those of arbitrary passages and XML elements. The method can then be used to automatically construct answers from blogs and forum posts which do not possess gold standard paragraph markers.

5. CONCLUSIONS

The paper provided a brief survey of our work to date on applying text segmentation in a diverse range of IR applications. At the time of writing , we have only formulated the research questions pertaining to the long answer type QA task. Section 4 is thus devoid of any experimental results to support our claims.

To summarize, we state that text segmentation has offered benefits to many standard IR methodologies and techniques. We show that it can increase retrieval effectiveness of ad hoc IR and patent prior art search, and can be used to automatically simulate query sessions and thus potentially build create huge query test collections for session IR research. We also propose a framework to extend existing segmentation techniques to produce readable segments which we believe can improve QA applications.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL).

6. REFERENCES

[1] Bates and M. J. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, 1989.

[2] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the NAACL 2000*, pages 26–33, 2000.

[3] D. Ganguly, J. Leveling, and G. J. F. Jones. Query expansion for language modeling using sentence similarities. In *Proceedings of the IRFC 2011*, pages 62–77, 2011.

[4] D. Ganguly, J. Leveling, and G. J. F. Jones. Simulation of within-session query variations using a text segmentation approach. In *Proceedings of the CLEF 2011*. (To appear). Springer, 2011.

[5] D. Ganguly, J. Leveling, and G. J. F. Jones. United we fall, divided we stand: A study of query segmentation and PRF for patent prior art search. In *Proceedings of the 4th International Workshop on Patent Information Retrieval, PAIR'11*. ACM, 2011.

[6] D. Ganguly, J. Leveling, W. Magdy, and G. J. F. Jones. Patent query reduction using pseudo relevance feedback. In *Proceedings of CIKM 2011*. ACM, 2011.

[7] M. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *SIGIR '93*, pages 59–68. ACM, 1993.

[8] M. A. Hearst. Multi-paragraph segmentation of expository text. In *ACL*, ACL '94, pages 9–16, Stroudsburg, PA, USA, 1994. ACM.

[9] K. Kishida. Experiment on pseudo relevance feedback method using taylor formula at NTCIR-3 patent retrieval task. In *NTCIR-3*, 2003.

[10] A. M. Lam-Adesina and G. J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of SIGIR 2001*, pages 1–9. ACM, 2001.

[11] V. Lavrenko and B. W. Croft. Relevance based language models. In *SIGIR 2001*, pages 120–127. ACM, 2001.

[12] W. Magdy, J. Leveling, and G. J. F. Jones. Exploring structured documents and query formulation techniques for patent retrieval. In *10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009*, pages 410–417, 2010.

[13] W. Magdy, P. Lopez, and G. J. F. Jones. Simple vs. sophisticated approaches for patent prior-art search. In *ECIR*, pages 725–728, 2011.

[14] I. Malioutov and R. Barzilay. Minimum cut model for spoken lecture segmentation. In *In Proceedings of the COLING-ACL 2006*, pages 25–32, 2006.

[15] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *SIGIR 1998*, pages 206–214. ACM, 1998.

[16] A. Moffat, R. Sacks-Davis, R. Wilkinson, and J. Zobel. Retrieval of partial documents. In *TREC*, pages 181–190, 1993.

[17] V. Moriceau, E. SanJuan, X. Tannier, and P. Bellot. Overview of the 2009 QA track: Towards a common task for QA, focused IR and automatic summarization systems. In *Focused Retrieval and Evaluation, INEX-2009*, pages 355–365, 2009.

[18] J. M. Ponte. *A language modeling approach to information retrieval*. PhD thesis, University of Massachusetts, 1998.

[19] J. C. Reynar. Statistical models for topic segmentation. In *Proceedings of the ACL-99*, 1999.

[20] E. SanJuan, V. Moriceau, and X. Tannier. Overview of the INEX 2010 question answering track (QA@INEX). In *Comparative Evaluation of Focused Retrieval, INEX 2010*, 2010, (To appear).

[21] H. Takuechi, N. Uramoto, and K. Takeda. Experiments on patent retrieval at NTCIR-5 workshop. In *NTCIR-5*, 2005.

[22] E. M. Voorhees. Overview of the TREC 2003 question answering track. pages 54–68, 2003.

[23] R. Wilkinson, J. Zobel, and R. Sacks-Davis. Similarity measures for short queries. In *In Fourth Text REtrieval Conference (TREC-4)*, pages 277–285, 1995.

[24] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR 1996*, pages 4–11. ACM, 1996.