Toward Higher Effectiveness for Recall-Oriented Information Retrieval:

A Patent Retrieval Case Study

Walid Magdy

BSc., MSc.

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

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

School of Computing

Supervisor:

Gareth J. F. Jones

January 2012

Declaration

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 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.: 58117563

Date:

Abstract

Research in information retrieval (IR) has largely been directed towards tasks requiring high precision. Recently, other IR applications which can be described as recall-oriented IR tasks have received increased attention in the IR research domain. Prominent among these IR applications are patent search and legal search, where users are typically ready to check hundreds or possibly thousands of documents in order to find any possible relevant document. The main concerns in this kind of application are very different from those in standard precision-oriented IR tasks, where users tend to be focused on finding an answer to their information need that can typically be addressed by one or two relevant documents. For precision-oriented tasks, mean average precision continues to be used as the primary evaluation metric for almost all IR applications. For recall-oriented IR applications the nature of the search task, including objectives, users, queries, and document collections, is different from that of standard precision-oriented search tasks. In this research study, two dimensions in IR are explored for the recall-oriented patent search task. The study includes IR system evaluation and multilingual IR for patent search. In each of these dimensions, current IR techniques are studied and novel techniques developed especially for this kind of recall-oriented IR application are proposed and investigated experimentally in the context of patent retrieval. The techniques developed in this thesis provide a significant contribution toward evaluating the effectiveness of recall-oriented IR in general and particularly patent search, and improving the efficiency of multilingual search for this kind of task.

Dedication

To my Lord (Allah): “The One who created me, and guides me. The One who feeds me and quenches my thirst. And when I get sick, He cures me. The One Who puts me to death, then brings me back to life. The One Who hopefully forgive my sins on the Day of Judgement. My Lord, grant me wisdom, and include me among the righteous”

- Prophet Abraham peace be upon him (Quran: Chapter 26, Verses 78-83)

Acknowledgments

First of all, I am deeply grateful to Allah, my Lord, who gave me the power and confidence to pursue this work and complete my PhD studies.

I would like to thank the Centre for Next Generation Localisation (CNGL) for funding the research of my PhD study, and giving me the support for living and travel to conferences.

I would like to express my sincere gratitude to my supervisor Gareth Jones for his constant support and encouragement. He gave limitless portions of his time that made me feel that I am his only PhD student. He always guided me with his advice while giving me the freedom and flexibility to pursue my research in various directions. Thanks to you Gareth.

I would like to thank Johannes Leveling for fruitful discussions on my work and his valuable feedback on my thesis, which was a great support while writing it.

I am very grateful to my colleagues in DCU and CNGL, who provided me with help whenever needed. I would like particularly to thank Rami Ghorab from Trinity College, whom I consider a true brother, for his unconditional support and help when I am not in Dublin.

Last but definitely not least, I would like to thank the person who this thesis would never have been possible without her support, understanding and patience. I would like to express my sincere gratitude to my soul-mate, my beloved wife Asmaa, who always has been my source of power and motivation. Thanks to you Smsm. I love you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim in information retrieval (IR) is to find documents relevant to a user's information need as expressed by some form of search query while minimizing the number of returned non-relevant documents. Research in IR focuses on developing or extending methods which will achieve higher user satisfaction through improving the retrieval effectiveness. The primary concern of most current IR systems has been to achieve high precision through retrieving relevant documents at high ranks. However, high precision at high rank positions is not always the sole requirement to achieve high user satisfaction when using an IR system.

The second concern of the IR systems is recall, where the concern is the proportion of available relevant material which has been retrieved. While recall has often been a lesser consideration compared to precision in IR research, some IR applications can be best described as recall-oriented. In such applications, the key objective of the search task is to retrieve all or at least most of relevant documents, and ideally to do this within maximum precision to enable the user to identify these relevant documents with the minimum overall search effort. There are

several examples of this type of IR application; the most popular of these that are currently receiving attention in the research community are patent retrieval and legal search. These types of IR application are primarily concerned with recall, where missing one relevant document can lead to the granting of a patent for a non-novel invention described in a patent application, or changing the decision of a judge in a legal case. The documents searched in these applications are typically complex and of variable types and formats. Also, these search tasks often involve a multilingual dimension since they can encompass search across international content sources. Due to the complexity of the data and the relatively recent introduction of these tasks to research in IR, much work is still required to fully understand and quantify the nature and challenges of these tasks, and to develop techniques that achieve higher retrieval effectiveness for these search applications.

In the research study presented in this thesis, recall-oriented IR is studied in general with a specific focus on patent search as a key application area emphasizing the challenges of recall-oriented search. The task in patent retrieval is mainly to find relevant patents related to an idea of an invention described in a document (patent application). The objective is to find all possible relevant patent documents for a patent application. This search process acts as a novelty check to the invention described, where missing one relevant patent document can lead to a wrong decision about the novelty of an invention. Since patent search is an important task, often with significant commercial implications, it is performed by professionals (patent examiners) at professional dedicated organisations (patent offices) and represents a multi-million euro business across the world. The majority of existing research investigations in patent search has focused on query formulation, where the objective is to find the best terms to represent a patent application as a query to achieve high retrieval effectiveness by retrieving all possible relevant documents at high ranks. Much less focus has been directed to other important (from our point of view) themes

in patent search, including the evaluation of patent search tasks from a recall-oriented point of view and the efficiency and effectiveness of patent search subtasks such as cross-language patent search. These themes are essential components for a full understanding and interpretation of the behaviour of a patent search system and its application within a patent office.

1.1 Focus of the Thesis

In this thesis, the special nature of recall-orientated IR and specifically patent search is explored and studied to understand the nature and challenges that accompany this kind of IR task. Understanding the nature and challenges of this task leads to investigating possible effective and efficient solutions for two main challenges in patent search and recall-oriented IR in general: retrieval effectiveness evaluation and multilingual search. The importance of these two themes with respect to achieving enhanced patent search effectiveness is explained and novel techniques for each one are introduced and extensively evaluated.

This section introduces these themes, the research questions to be covered, and the specific hypotheses examined in this thesis.

1.1.1 Research questions

The research questions to be addressed in this thesis can be divided into three sets. The first set relates to the nature of recall-oriented IR and patent search. The other two sets relate to the two main themes studied in the thesis, which are the evaluation of recall-oriented IR and multilingual search for such applications.

1. **The special nature of patent search and recall-oriented IR**
    - What are the differences between recall-oriented search tasks and other standard IR tasks, and how can these be demonstrated?

- What are the properties and configuration of a retrieval system to achieve high retrieval effectiveness for patent search?
- Is applying standard retrieval techniques for patent search as effective as for general IR tasks?

2. **Evaluation in Recall-Oriented Information Retrieval Tasks:**
    - What features of performance should be measured for a recall-oriented IR application such as patent search?
    - Are the current evaluation methods and metrics, which are currently being used for patent search and recall-oriented IR applications, appropriate to evaluate this type of application?
    - Can a more meaningful metric be developed for these tasks?
    - How can the suitability and reliability of such a novel metric be established to ensure that it reflects real system performance and ensure that it is robust to different system variables?
3. **Multilingual patent search:**
    - Patent search is often multilingual by nature; what are the differences between cross-language patent search and cross-language information retrieval for standard ad-hoc search tasks?
    - Can the translation process in cross-language patent search be optimized to be more efficient with the long queries typically used in patent search?
    - How can the retrieval effectiveness be maintained or even improved while increasing the efficiency of the cross-language patent retrieval system?

1.1.2 Hypotheses

The two main hypotheses of this thesis can be summarised as follows:

H1. A deeper understanding of the nature and the requirements of patent search and recall-oriented IR tasks in general can lead to an improved understanding of the utility of existing evaluation metrics for these tasks, and the proposal and implementation of a novel metric specifically developed for patent search that addresses these requirements.

H2. The special nature of patent topics, which are considerably long and domain-specific, leads to high cost requirements for effective cross-language search. It is hypothesized that machine translation methods used in cross-language patent search can be adjusted and optimized to significantly improve the efficiency and effectiveness for the high cost cross-language search for such task.

1.2 Structure of the Thesis

The remainder of the thesis is organized as follows:

Chapter 2 examines the differences between recall-oriented IR and more common precision-oriented IR applications, and provides an introduction to patents and patent retrieval. It further explores the different search tasks within patent retrieval and highlights the main research directions for these tasks. Finally, it compares the patent search task to the legal search task in order to highlight the commons and differences between different recall-oriented IR tasks.

Chapter 3 reviews key existing work for different patent search tasks and datasets. It also gives an introduction to the two main IR evaluation campaigns that have included patent search tasks which have encouraged research in this area, namely the NTCIR patent retrieval track and the CLEF-IP track. The description of each of these tasks and the datasets provided are described

with specific focus on the CLEF-IP task since it is the more recent data set and has been used in the reported state-of-the-art work on patent search. Also, it is the dataset used for the studies described in this thesis. In addition, this chapter summarises the methods developed by participants in these evaluation tracks which achieved the best retrieval effectiveness for these patent search tasks.

Chapter 4 reports our work and contribution to the CLEF-IP patent search task in the process of exploring and understanding the nature and the challenges of the recall-oriented tasks represented in patent search. This chapter describes the algorithms we developed in our participation in the CLEF-IP track for two successive years (2009 and 2010). It shows how our work compares to the best run submitted for this task. In addition, it reports some of our attempts to apply standard IR query expansion techniques to improve the retrieval effectiveness of patent search, and describes a novel method for query expansion of patent queries based on automatically generated synonyms set from the patent text.

Chapter 5 presents our first main contribution in the thesis. It examines the effectiveness of using the current IR evaluation metrics for recall-oriented IR applications such as patent search, and gives real and synthetic examples of when these metrics can succeed or fail. It then introduces a novel score metric that shows greater suitability for the evaluation of this type of application. Extensive experimentation is performed to demonstrate its suitability for recall-oriented IR applications and to test its robustness with incomplete relevance judgements.

Chapter 6 presents our second main contribution in the thesis. It reviews the work in multilingual and cross-lingual information retrieval, and shows the challenges of applying these techniques to the patent search task. Following this, a novel approach for using machine translation (MT) systems for translation in multilingual IR is introduced. This chapter applies the techniques for multilingual search in English, French, and German languages for the CLEF-IP

cross-language patent search task. Both query translation and document translation techniques are tested with the new translation approach. Additional work is presented in the chapter that examines the issue of word compounding in the German language that leads to low quality translations with negative impact on retrieval effectiveness.

Chapter 7 concludes the thesis. It highlights the achievements of the thesis and provides directions for possible future work aimed at further improving the effectiveness and efficiency of patent search and recall-oriented IR in general.

The work described in this thesis has been published in 11 publications in top international venues. In addition, a patent was filed on the translation system described in Chapter 6. Appendix C contains the full list of publications.

Chapter 2

Recall-Oriented Retrieval and Patent Search

This chapter provides an overview of the subject of recall-oriented information retrieval (IR) in general and patent search specifically. The first section introduces recall-oriented IR and highlights the differences between this and the more popular topic of precision-oriented IR. Subsequent sections introduce the topic of patents and patent retrieval covering details of the form and structure of patents and the challenges of patent search. A summary of the different patent search tasks and different research directions for these tasks is provided. The chapter concludes with a comparative analysis of patent search and the related recall-oriented legal search task.

2.1 Recall-Oriented IR vs. Precision-Oriented IR

The primary focus of the majority of research in the development of IR algorithms and the evaluation of retrieval accuracy has been on the precision of the retrieved results. This work has centred on tasks where one or a few of the relevant documents are sufficient to satisfy the user

information need. In this common scenario, it is assumed that the user wants to find the answer to his request as rapidly as possible, and usually will not wish to make the effort to check more than a small number of results before either moving on from the search session or revising the query since it has been unsuccessful at locating relevant results at the high rank (Jansen and Spink, 2005). This is why IR systems for the precision-oriented tasks seek to retrieve at least one or a few relevant items at the highest possible rank, without regard for the retrieval or the rank of the other relevant items.

Recall-oriented IR has a different objective, which is to find as many as possible, if not all, of the documents relevant to the user's information need. In this scenario, the user is usually willing to go deep in the ranked results list to search for any retrieved relevant document. A clear example of this situation is patent search in which a patent examiner typically checks hundreds of documents in the results list to locate any possible relevant document (Azzopardi et al., 2010). Of course finding relevant documents at the top of the list remains a desirable feature in recall-oriented IR, since this reduces the search effort. However, the key objective is to retrieve and then locate all the relevant documents. This was shown in (Bonino et al., 2010), where the authors explain that both recall and precision are highly important in patent search. Similar analysis applies to legal search, where there are sometimes thousands of relevant documents for a topic (query); therefore, finding all relevant documents with a minimum number of documents to be checked is a primary objective (Tomlinson et al., 2007).

Another feature that usually characterises recall-oriented IR applications such as patent search and legal search is the length of topic (query). Since the objective for these applications is the recall, the queries are typically relatively long when compared to standard IR applications in order to have a higher chance of matching different variations of terms representing the same topic. For example, in patent search the topic can take the form of a full patent application that

runs to tens of pages, which is unlike the precision-oriented tasks, where the users' queries are typically short and consist of small number of words (Spink et al., 2002).

Additionally, one of the concerns in some recall-oriented IR applications is multilingual search. Since it is important to find everything relevant, then in some situations, there is a high possibility that relevant information exists in languages other than that of the topic. Cross-language information retrieval (CLIR) in standard precision-oriented search has been the topic of much research (Oard, 1998; Oard and Diekema, 1998; Parton et al., 2008). In this situation, the language barrier between queries and documents is typically crossed by translating the query to the document language using standard machine translation (MT) tools, since it is more efficient to translate a short query of two or three words rather than translating a large collection of documents (Oard, 1998). However for recall-oriented search tasks such as patent search, the very long topic statements mean that current standard MT tools can require significantly more time to translate the topic compared to the short queries which are generally used in other IR applications that can be translated much faster. In addition, the special nature and vocabulary of the documents in patent search applications requires the translation system to be trained on data of the same domain and linguistic form. This creates a challenge to locate a sufficient amount of this data for the training process of some language pairs.

There is a strong commercial interest in recall-oriented search tasks, where the patent search and legal search are multi-million euro businesses that cost large amount of time, effort, and money for professional organisations, since critical decisions are taken based on the search results. For example, applying for a patent for an invention in the case of patent search and giving a decision in a legal case in the court in case of legal search. This is one of the main reasons for the increasing interest in developing more effective and efficient retrieval systems for these tasks.

Table 2-1 highlights some of the main differences between precision-oriented and recall-oriented IR with illustrative examples.

Table 2-1: Simple comparison between precision-oriented and recall-oriented IR applications

	Precision-Oriented IR	**Recall-Oriented IR**
Objective	Find one or a few relevant documents at the top of the ranked results list that satisfy the user's information need.	Find all possible relevant documents within a results list.
Users	Keen to find the answer to their information need as quickly as possible and within the top ranked results, otherwise, they reformulate the query	More patient to work through the results list to find any possible relevant document before they stop to reformulate the query. Usually, users are professionals who are expert in the field of search
Multilingual Search	Queries are typically short, therefore automatic translation is fast	Queries are typically very long, which makes translation time significant
Evaluation	Focuses more on precision	Focuses more on recall
Examples	• Web search • News search	• Patent search • Legal search

2.2 Patents

Since patent retrieval forms the core recall-oriented application examined in this research study, this section provides an overview of the nature of patents. This overview is intended to make clear to the reader why the topic of recall-oriented IR, represented by patent search, is an important and challenging area of study at this time. In this context, the introduction of the patent features enables the reader to understand the specific technical details of patents that raise challenges in patent search and which can be exploited and utilized in creative ways to improve

the effectiveness of the patent search. The following section provides a patent document sample to further illustrate all the concepts highlighted in this section.

2.2.1 Patents and the patent application process

A patent is an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem (WIPO, 2010).

The procedure for filing patent applications and the applicant rights vary widely between countries according to national laws and international agreements. The exclusive right granted to the patent applicant in most countries is the right to prevent others from producing, using, selling, or distributing the patented invention without permission (WIPO, 2010).

Patent applications are submitted to a patent office. This is a governmental or intergovernmental organization that is responsible for granting patents in a given country or region. Patent offices may grant or reject the patent application based on whether or not the application fulfils the requirements for patentability (European Commission, 2008).

There are many patent offices across the world, for example the United States patent and trademark office[1] (USPTO), the European patent office[2] (EPO), and the Japan patent office[3] (JPO).

The main role of the patent office is to check the novelty of a patent application through an extensive search process for any relevant prior work. It is this patent search process that is the subject of the recall-oriented IR studied in this thesis. Patent examiners are responsible for

[1] http://www.uspto.gov/

[2] http://www.epo.org/

[3] http://www.jpo.go.jp/

checking for the novelty of a patent application. Patent examiners employed by a patent office are experts in the field of the invention. Actually, some of the patent examiners carry a post graduate degree (masters or doctoral) in the field of the patents they examine (Azzopardi et al, 2010).

During the novelty checking process in the patent office, the patent document passes through different versions. The first version of a patent is the patent application, which is the initial invention disclosure submitted by the patent applicant asking for the granting of a patent for an invention. This patent application is then checked for novelty by the patent examiners. They can subsequently ask the applicants to provide several updates to the application itself, until the patent is granted or rejected as a new invention. The published version of the granted patent is the most accurate one that describes the novel parts of the invention, and is usually the one used to refer to the invention; however, patent applications are still used for referencing in the period before a decision is taken about the patent. The decision about the novelty of a patent usually takes a few years before granting or rejecting a patent application.

2.2.2 Patent structure

A patent is a structured document, which consists of several sections, such as title, abstract, description, citations, inventors … etc. The text of a patent document is usually saved electronically in the patent office as an XML file with specific fields corresponding to each section or subsection in the patent document and some additional metadata about the patent document itself. However, what the normal users get access to is a structured document (not an XML document) as shown in Figure 2.1. The purpose of the XML files is to facilitate the search process within specific fields of text in the patent office.

The structure of patents varies according to the patent office to which the invention is filed. However, there are common sections that are found in most patent documents. The most important one is the "claims" section. Each patent should contain at least one claim. The claim defines the scope of protection granted by the patent and the specific novel aspects of the invention that need to be protected. Other sections such as the "title" of the invention, "classification", "description", "abstract", "summary of invention" may or may not exist according to the patent office. For example, it is very common to find patents filed to the USPTO containing the "abstract" field, but it is not very common in the EPO. Another example of inconsistent use of fields between different patent offices is the presence of explicit fields in USPTO patents called "summary of the invention" and "field of the invention". These two fields are not common in the other patent offices. A detailed example of a patent document structure is provided in Section 2.3 for further explanation.

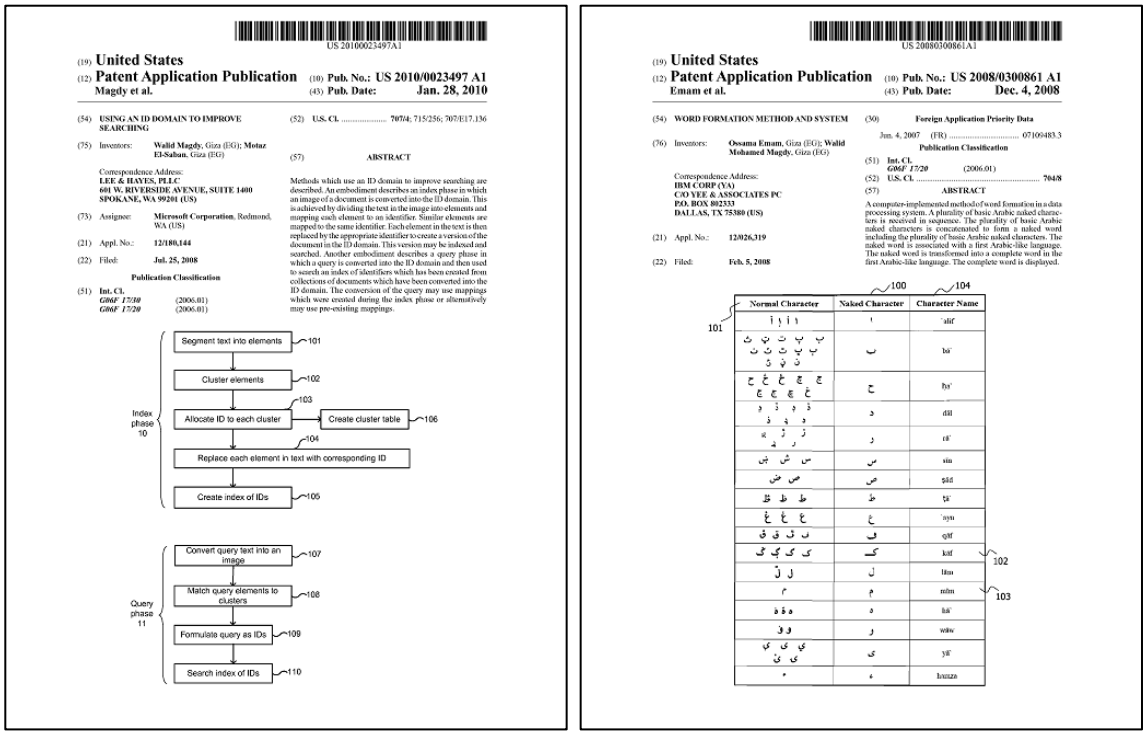


Figure 2.1: Examples of the first page of two USPTO patents

2.2.3 Characteristics of patent text

The language of patents is characterized by the complexity and ambiguity of its terms, where the contribution of a patent is usually intended to be unclear within the text (Krier and Zacc, 2002). Unlike scientific publications or technical reports, the writers of a patent try to generalize the coverage of the invention and focus on emphasizing the novelty of the ideas disclosed, not on helping the reader to understand their technique. The objective is usually to maximize the scope of what is protected by the patent to maximize the potential gain to the inventors arising from owning it. In addition, the writer tries to make finding any relevant prior work by the patent examiner a hard job to avoid invalidation of any of the claims. This leads to the usage of unusual expressions that makes understanding or even finding a patent a difficult job.

A recent comparison between patent text and general English text was reported by Verberne et al. (2010). The British National Corpus (BNC) was used as a general English corpus and 400,000 patents from four different sources (EPO, USPTO, JPO, and WIPO) were used as a patent corpus. This research showed that the length of the sentences in the corpora differs significantly. The sentences in patents were found to be much longer than sentences in general English. It was also found that the median sentence length of the general English corpus is 10 words, but for the patents it was 22 words. Surprisingly, it was found that the terms used in both corpora are nearly the same, where 96% of the unique terms used in the patents exist in the general English corpus. However, the frequency distribution of these terms differs significantly, and the part of speech tagging is very different as well. For example, the term "said" in general English is usually a verb, but in patents it is an adjective (e.g. "*A said claim …*"). Furthermore, word combinations in patents are uncommon in the general text, where the noun phrases in patents, which consist of two or more words, are very uncommon in the general English. This finding is consistent with a study reported in (D'hondt, 2009), which showed that the multi-word

terms in patents are often invented and defined by patent writers, which means that they are not found in any dictionary or lexicon. These findings show how challenging the language of patents is, where the same words are used as in standard texts, but with different meanings and in different ways. This contributes to the challenges of reading patent documents and searching patent document collections.

Another challenge in the text of some patents, especially older ones, is the presence of errors in the text due to the optical character recognition (OCR) process used to create and index digitized copies of patents originally created as paper documents (Adams, 2011; Johannes et al., 2011). The majority of the patents that were filed in the period before the 2000's were in printed form only. The OCR process usually introduces errors to the recognized text, especially for older patents, where the average quality of OCR systems was very poor.

2.2.4 Citations and multilinguality

Different from other technical publications, it is very common in patents to find citations and references to patents from different patent offices and in different languages. This fact means that when a patent is checked for its novelty, the idea should not have been disclosed before in any language. Furthermore, in some patent offices, patents can be filed in more than one language, such as the EPO, where any patent should have certain sections translated in three languages (English, German, and French).

The citations in a patent pass through several stages. Initially, the patent applicant provides some citations to other patents or prior work in general. These are mainly intended to show that the submitted invention is novel. Usually the first citations list is changed by the patent examiners through their search while checking the novelty of the patent application. They remove some citations which are irrelevant or have no value, and add other citations that relate to

the invention. The final citations list comes from the patent examiners' search report. This final citations list in a granted patent document, which is the main outcome in the search report, includes the citations accepted by examiners from the initial patent application, or those assigned by the patent examiners through the patent search process. The final citations are the most important ones since they take large amount of effort and time to be searched for and checked for relevance by the patent examiners. This final citation list from the search report of the patent examiner is taken to be the set of existing patent documents relevant to the filed patent application.

A common behaviour when filing patents is that the patent application can be filed in more than one patent office and in different languages in order to protect an invention in different countries. Patents that refer to the same invention are called a *patent family*. The patents of a patent family are those filed by the same inventors on the same date to describe the same invention but in different regions and can be in different languages and formats. The requirements of patent filings in the different territories can be mean that the details filed are not exact translations. When a citation is used in one of the patent applications to an invention, usually using one of the patents in a patent family is sufficient.

2.2.5 Patent classification

Each patent publication should be assigned to at least one classification, which describes the main field and specific fields of the invention. The classifications are assigned by the patent examiners to a patent application at the most detailed level which is applicable to its contents. There are several standard patent classification schemes such as: International Patent Classification (IPC) which is agreed internationally, the United States Patent Classification (USPC) that is used in the USPTO, and the European Classification (ECLA) which is based on

the IPC, but adapted by the EPO. These classification schemes are a hierarchical patent classification system that classifies the topic of an invention of a patent into a tree of classes representing different fields of technologies. For example, the IPC is created under an agreement administrated by the world intellectual property organization[4] (WIPO) and updated on a regular basis by a committee of experts (WIPO, 2011). The IPC divides technology into eight major sections (e.g. human necessities, fixed constructions, physics, and electricity) with approximately 70,000 subdivisions or classes. Each subdivision has a symbol consisting of Arabic numerals and letters of the Latin alphabet. For example, the classification "A01B 1/00" in Figure 2.2 represents "hand tools". The first letter is the "section symbol" consisting of a letter "A" that stands for "human necessities". By adding the following two digits number "A01", the symbol class represents "agriculture; forestry; animal husbandry; trapping; fishing". The final letter makes up the subclass "A01B", which represents "soil working in agriculture or forestry; parts, details, or accessories of agricultural machines or implements, in general". The subclass is then followed by a 1 to 3 digit number, an oblique stroke, and a number of at least two digits representing "main group" and "subgroup", see Figure 2.2.

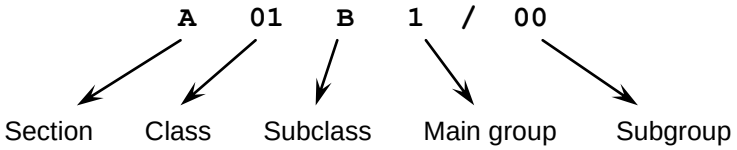


Figure 2.2: IPC class code (A01B 1/00)

The patent classification is indispensable for the retrieval of patent documents in the search for prior art. Such retrieval is needed by patent-issuing authorities, potential inventors, research

[4] http://www.wipo.int/

and development units, and others concerned with the application or development of a technology.

2.3 Sample of a Patent Document

In this section we present a sample patent document from the European patent office (EPO) to further explain the nature and the structure of patents. Figure 2.3 shows some of the fields of the patent document with the structured features of the document highlighted. The patent document example is selected from the EPO since the data collection used for experimentation in this thesis comes from the EPO. The patent document is presented in the standard XML format assigned by WIPO (ST.36[5]) that describes the different fields of a patent document. This example is for a granted patent, which has passed through several versions to reach the current presented version. This is different from a patent application which is initially submitted by the patent applicant and does not contain some of the information contained in the final version, such as translations of claims and some of the citations as described earlier. In addition, the text of the patent itself and the claims can be modified from the initial patent application until the final version of the granted patent has been agreed. The patent example presented in Figure 2.3 shows only some of the sections and subsections of the patent document, which can be considered as the basic sections of a patent document.

The following subsections describe each of the patent fields (sections) presented in Figure 2.3. In addition, further important patent sections not presented in Figure 2.3 are also described briefly.

[5] http://www.wipo.int/standards/en/pdf/03-36-01.pdf

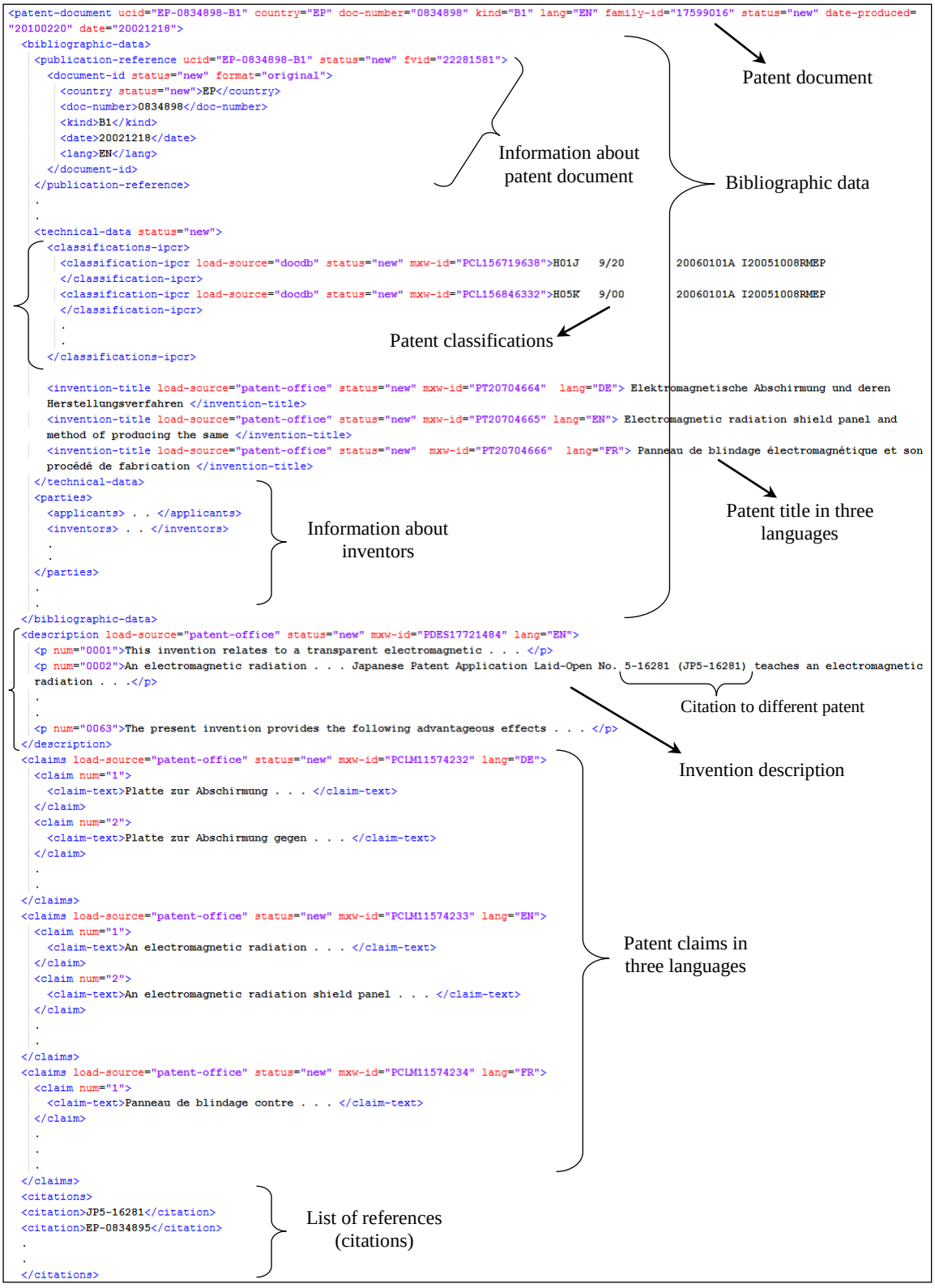


Figure 2.3: A sample XML file for a patent document from the EPO

2.3.1 Bibliographic data

This section of the patent document contains some legal information about the patent document not the invention itself. It is composed of a number of subsections that contain information about the patent document, technical information about the patent, and information about the parties involved in the patent document. The publication reference subsection contains information about the document such as patent ID, document number, the version (called "kind" in Figure 2.3) of the document which indicates whether it is a patent application or a granted patent, the date of the publication, the language of the document, and the country applied to. The technical data subsection contains two important pieces of information about the patent: the patent classifications and the title of the patent in three languages (this is a feature in the EPO patents, where the title is stated in English, French, and German). The parties subsection contains information about the applicant institute, the inventors, the assignees, and the patent agent or agency that wrote and filed the invention. This information includes the names and the addresses of each of the parties.

There are additional subsections that contain more data about the patent such as the application reference which has information about the initial patent application, dates of when the publication became available to the public, and other similar kinds of information.

2.3.2 Description

This section of the patent document represents the core of the invention, since it contains all the technical details of the invention. It consists of a set of paragraphs that describe all the aspects of the invention in detail and the differences between it and the state-of-the-art. The description section can contain tables, experimentation on the performance of the invention, and description

of figures relating to the invention. The first paragraph of the description section usually contains information about the topical field of the invention.

Very important information within the description text is the references to other patent documents as shown in Figure 2.3. These references are part of the citations that a patent examiner would be interested to examine in order to measure the contribution of the invention against prior art.

2.3.3 Claims

The claims section of the patent document lists what aspects of the invention that the patent is going to protect. A successful patent does not have to have all its claims accepted, but at least one of them must be. The examination can lead to dropping some of the claims by showing that they are not novel. This usually happens because patent applicants try to generalize their invention as much as possible, which can lead to the novelty of some of the very general claims being found to be invalid.

The claims section in EPO patents contains the list of claims in three languages (English, French, and German). However, this is not the situation for the initial patent application, where the claims are submitted in one language only, which is the language of the document. The claims translations are only provided for the granted version of the patent.

2.3.4 Citations

This section of the patent document contains the list of older patents that are related to the invention by describing the relevant parts of the prior-art of the invention, or these citations can be for patents that have been located by the patent examiners and were found to invalidate parts

of the invention in the initially submitted patent application, where the final version of the patent get these parts modified or removed.

2.3.5 Additional parts of a patent document

There are several additional sections of a patent document that are not shown in the example in Figure 2.3. One of the sections is the "abstract" section, which is a short paragraph that contains a summary of the invention. This section is not always present in EPO patents since it is an optional section. Another basic section in most patent documents is the "figures" section. These figures are very common in most of patent applications and describe the invention graphically. Figures can include, for example, flowcharts of the workflow of an invention, technical drawings, or structure diagrams. The "figures" section is not included in our study since our focus is only on text retrieval methods for patent search.

2.4 Patent Retrieval

The patent search task is one of the main activities carried out in a patent office to check the novelty of patent applications for filed inventions. This section introduces the patent retrieval task and discusses its nature. We firstly present how this task is handled in real-life in the patent office, and then discuss how this task has been introduced in the IR research domain. The next section highlights the main research directions that are investigated in the research community.

2.4.1 Patent search in the patent office

For many years, Boolean search has remained the common approach used for finding relevant patents and documents that can invalidate the novelty of a patent application (Connett-Porceddu et al., 2005; Adams, 2011). The patent examiner constructs a Boolean query (e.g. "(Term1 OR

Term2) AND (Term3 OR Term4)") that represents the invention described in the patent application and uses it for searching the patent documents database. Patent search performed in the patent office is an interactive search task, where it is performed through multiple search sessions to reach the required information need. The patent examiner prunes the query several times till he/she gets a reasonable number of results to be checked (usually hundreds of results), then he/she sorts the results by different means such as publication data and checks them one by one.

Although this method of search is very exhaustive and time consuming, interactive Boolean search remains the preferred search technique by patent examiners since it is reproducible (Adams, 2011). Reproducibility of search means that the retrieval system will give the same results for the same query each time. This is not the case for probabilistic search, where adding new documents to the collection changes the system variables about the importance of each term, which leads to a different list of results each time the collection is updated. Having reproducible search results is an essential matter for patent examiners to defend their decisions about the novelty of patents whenever required (Adams, 2011).

In recent years, the patent search task has been introduced to the IR research community with the objective of finding more efficient and effective methods for finding relevant documents and sorting the search results using relevance to the topics (Fujii et al., 2004; Roda et al., 2009). These research initiatives have studied the patent retrieval task as a one shot search task, where results are retrieved in a single session. To the best of our knowledge, no work has modelled the interactive nature of the patent search task to date in the manner that is has been modelled and investigated for other search tasks (Hersh and Over, 1999; 2000). Nevertheless, these initiatives have promoted research investigation into the exploration of the nature of patent search, and to finding novel solutions to many of its challenges and issues as described later in the thesis.

2.4.2 History of research in patent retrieval

An initial trial examining patent retrieval was introduced using a very small collection of patent documents consisting of 6,000 US patents in (Osborn and Strzalkowski, 1997). In their work, they conducted what can be viewed as a prior-art patent search task. The search topics were full patents, and they used the citations list within each topic patent as the list of relevant documents. They utilized a combination of a series of shallow natural language processing (NLP) techniques to match between the topics and the documents. The SMART system based on the vector space model (Buckley et al., 1996) was used in their experiments. They showed positive results over the baseline that did not use any NLP processing, but the scalability of their techniques is questionable, since the size of the collection used is far smaller than that of real-life patent collections that extend to millions or tens of millions of patent documents. Nevertheless, their research constitutes the first attempt to study patent search.

At the ACM SIGIR 2000 conference, the first workshop on patent retrieval was held (Kand and Leong, 2000). This workshop was one of the initial and main efforts to promote research in patent retrieval within the IR community. It discussed the nature, challenges, requirements, and existing systems for patent search. The participants in this workshop were researchers and practitioners from communities relevant to the patent domain, and the objective was to share their ideas, approaches, perspectives, and experiences from their work.

An example of the patent systems presented in this workshop was the PATOLIS-e patent search system (Saito, 2000), which is used in Japan Patent Information Organisation (JAPIO), one of the patent information service institutions in Japan. PATOLIS is a commercial patent search system of Japanese patents, utility models, designs, and trademarks. PATOLIS-e is an information service that uses the PATOLIS database in English using a Japanese-English concordance table. Saito (2000) overviewed the characteristics of patent search and the

challenges that face a real patent search system. He also showed that lexical resources are usually required to improve the results in patent search.

Other reports at this workshop from the EPO and JAPIO focused on presenting the challenges of patent search in general and cross-language patent search in particular between different languages, for languages such as English, German, Spanish, and Japanese (Sarasua and Corremans, 2000; Fukui et al., 2000). This work highlighted the characteristics of patent search that can be listed as: (1) the detail of the search, where most of the returned results are carefully checked, (2) the need for deep analysis of the expression of patent examiners, (3) the heavy use of generic terms and vague expressions, (4) the extensive use of acronyms and new words, and (5) sentences are grammatically correct but use limited syntactic structure (Sarasua and Corremans, 2000). One interesting result they showed for experiments for English-Japanese CLIR was the weighting scheme, where they mentioned that they do not use *term frequency*[6] (*tf*) in the weighting of results, since important concepts are often hidden in the text with general or vague expression in the patent documents. Further, the *inverse document frequency*[7] (idf) coefficient is calculated only per technical classes, so that the same term will have different idf's in different fields. Sarasua and Corremans (2000) mentioned that this greatly improves the results.

The rest of the workshop discussed the challenges of patents and patent retrieval in general, and the techniques used in the patent offices to overcome these challenges, especially the vocabulary of patents. The workshop concluded with the importance of having research on patent retrieval to create more effective and efficient patent search systems (Kando, 2000). This

[6] Number of times a term appears in a given document

[7] The inverse of the number of documents that contain a given term

outcome of the workshop was the main motivation to include the patent search tasks in international IR evaluation campaigns.

In the following year, the evaluation of patent retrieval was proposed in one of the major IR evaluation campaigns (NTCIR-2) in 2001 (Leong, 2001). Since then, patent retrieval has featured as a track in each of the NTCIR[8] campaigns for several years. More recently, patent retrieval has been introduced at the CLEF[9] evaluation campaign in 2009, as the CLEF-IP (CLEF Intellectual Property) track (Roda et al., 2009). Patent retrieval is of interest in IR research since it is of commercial importance and also because it is a challenging IR task with different characteristics to existing popular IR research tasks (Leong, 2001; Roda el al., 2009).

2.4.3 Patent search tasks

Various tasks have been created around patents; some are retrieval tasks, others include tasks such as patent mining and patent classification (Larkey, 2001; Fuji et al., 2007). The IR tasks at NTCIR and CLEF related to patent retrieval can be listed as follows:

- **Ad-hoc search**

In the ad-hoc search task, a number of topics are used to search a patent collection with the objective of retrieving a ranked list of patents that are relevant to this topic (Iwayama et al., 2003). This task is also known as the "technology survey" task, where the assumed scenario is that a business organisation or institute is interested in an idea that can lead to an invention and wants to make a technology survey through finding relevant patents to this idea to find what novel aspects exist in the idea to focus on them when writing a patent application.

[8] http://www.nii.ac.jp/

[9] http://www.clef-campaign.org/

The topics are designed as a “clipping”, which is few words describing an idea of an invention. This acts as a title of a topic in standard ad-hoc search tasks, which is in the form of a short query of few terms; this is the main topic of interest to be searched for. In addition to the “clipping”, there is a “memorandum”, which describes the exact information need of the topic (Iwayama et al., 2003). Both the “clipping” and “memorandum” can be utilized in the search.

Ad-hoc search was introduced as one of the initial tasks in patent retrieval, particularly at NTCIR-3 in 2003. The task can be considered a good starting point to explore the patent search task and the challenges that accompany it. However, the ad-hoc patent search task does not ideally model the real life problems in the patent search, where the main scenario in patent search is the one performed in the patent offices to check the novelty of a patent application, which acts as the topic. The real task in the patent office is an exhaustive and challenging one, since formulating the query from the patent application takes much effort, and the search process requires being very thorough in finding all possible relevant patents that can invalidate the novelty of a patent application or at least a part of the patent application.

- **Invalidity search**

This task models a part of the real life patent search tasks which are performed in a patent office. The claims of a patent are taken as the topics for an IR search, where each claim is used as a separate topic. The objective is to search for all relevant patents to a given claim to find out whether this claim is novel or not (Fujii et al., 2004). All relevant documents are needed, since missing only one document can lead to later invalidation of the claim or the patent itself. It is normal that some of the claims in a patent application are invalidated without invalidating the patent itself, since there can be other claims in the patent application which are novel (usually, this should include the first claim, which is the main claim of an invention).

- **Passage search**

This is similar to invalidity search, but because patents are usually long, the task focuses on locating the important fragments in the relevant documents (Fujii et al., 2005). Hence, the claims of a patent application, which are the topics, are used to search within the retrieved patents for the specific relevant passages or paragraphs that can invalidate the claim. The passage retrieval is performed by sorting the passages in the retrieved documents from the patent invalidity search task according to their relevance to the claim topic. This is very similar to the passage retrieval in many other IR tasks, where the objective is to locate the text segments that are the most relevant in a relevant document (Salton et al., 1993).

- **Prior-art search**

This is the main search task carried out in patent offices. The objective of this task is to find all patents relevant to a patent application which can invalidate the novelty of the patent application or at least describe prior-art work in the area of the patent application.

The prior-art search task is investigated in the evaluation campaigns by taking the full patent application as a topic, with the objective of finding all relevant patents (Roda et al., 2009; Piroi, 2010). In this kind of task, the list of citations in the granted version of the patent is taken as the list of relevant documents, and the objective is to find these citations automatically (Graf and Azzopardi, 2008). These relevant patent documents can be classified into three types (Roda et al., 2009): type “A”, which is a relevant patent that contains relevant work considered as prior art but does not invalidate the patent application; type “X”, which is a relevant patent that carries a very similar idea as the patent application and invalidates the full invention; and type “Y”, which is a set of relevant patents that together invalidate the patent application. Prior-art search in patent retrieval focuses on finding any kind of patent relevant to the patent application in hand;

this is different to invalidity search which focuses on finding any type of document that proves that a given claim in a patent application is not novel.

2.5 Research Directions in Patent Search

This section overviews the common research directions for patent search. A brief description of these directions is provided. Relevant prior work is discussed in detail in the next chapters.

2.5.1 Query formulation

A patent topic is much longer than in other standard IR tasks. It can take the form of a long paragraph representing a claim in patent invalidity search task or even a very long document in the form of a patent application running to tens of pages in the case of prior-art patent search. This has led to query formulation becoming one of the major focuses of research in patent search. The objective is to convert the very long patent topic into a suitable query to search the patent collection. Some research has focused on identifying the best parts in the patent document to extract the query terms. Experiments of this type were reported by most of the participants in the patent retrieval tracks in NTCIR and CLEF (Fujii et al, 2004; Roda et al., 2009; Piroi, 2010) and by some individual research such as (Xue and Croft, 2009a; Mahdabi et al, 2011). Other research has focused on extracting features from the patent topics and documents while carrying out the retrieval process, such as utilizing patent citations (Fujii, 2007a; Fujii, 2007b), patent classifications (Xue and Croft, 2009b; Patrice and Lopez, 2010), and extracting keywords and term concepts from patents (Patrice and Lopez, 2010). Other research has divided the retrieval process into multiple stages by applying subtopic retrieval then merging the results as in (Takiki et al., 2004), or applying re-ranking to the top retrieved documents from an initial retrieval stage based on additional patent features as in (Mase et al., 2005; Patrice and Lopez, 2010).

All this research has focused mainly on how to build an effective query from the long patent document in many different ways. This research is discussed in detail in the next chapter where prior work on patent search is analysed and compared.

2.5.2 Query expansion

Query expansion is generally effective in ad-hoc search (Billerbeck and Zobel, 2004; Rocchio, 1971), which motivated researchers to explore its application to patent search. The main objective was to expand the query with additional terms in order to increase the possibility of matching additional relevant documents and consequently improving the retrieval effectiveness. Several attempts to apply different relevance feedback techniques on patent search have been reported in the literature (Kishida, 2003; Itoh, 2004; Takeuchi et al., 2005; Konishi, 2005). Unfortunately, none of these trials have succeeded in achieving a stable significant improvement in the retrieval effectiveness. This topic is discussed in detail in the next chapter.

2.5.3 Multilingual search

Some of the reported existing work has included cross-language patent search, where the patent topics were in one language and the task was to find relevant patents in different languages (Fujii et al., 2004; Roda et al., 2009; Piroi, 2010). Although the patent topics are different from those in standard cross-language search tasks, the basic technique adopted by the majority of the reported work in cross-language patent search is similar to that for standard CLIR. The basic technique was to translate the patent topic into the document language, and then to use this translated topic for formulating a query to search the collection (Piroi, 2010). The dominant approach was to use standard machine translation systems to translate the patent topic into the collection language to enable direct search of the collection (Roda et al., 2009; Piroi, 2010). In addition, some attempts

at using dictionary-based methods for the translation process has been reported (Jochim et al, 2010), however the results reported for this method of translation were significantly lower than when using machine translation systems. A detailed discussion of the work in cross-language patent search is presented in Chapter 6 of this thesis.

2.6 Patent Search vs. Legal Search

Similar to patent search, legal search is another recall-oriented IR application. The main objective is to find all possible relevant documents related to a given legal case. However, legal search differs in nature from patent search in many regards. Legal search is concerned with finding all possible types of documents that can act as evidence in a legal case. The kind of documents searched includes, but is not limited to letters, memos, budgets, reports, agendas, minutes, plans, transcripts, scientific articles, and emails (Tomlinson et al., 2007). The search topics in this case are usually a "complaint" that is filed in court. This complaint outlines the theory of a case, including factual assertions and causes representing the legal theories of the case (Schmidt et al., 2002). It can be expected that with this amount of variation in the types of documents searched and nature of the search request, that the number of relevant documents can vary significantly for each request. In fact, as an evidence of this, the number of relevant documents per query at the Legal Track at TREC 2007 ranged from 18 to 77,467, with an average of 16,904 (Tomlinson et al., 2007). This demonstrates the amount of effort and time needed to find all relevant documents for some of the topics in order to complete the search task. The main users of legal search applications are lawyers and judges, who are characterized with patience and for whom search for relevant documents is part of their paid work, which is similar to patent examiners.

Similar to the current techniques used for patent search in the patent office, Boolean search remains the major technique used for finding these documents in legal search (The Sedano Conference, 2007). The typical aim of research in the area of legal search or patent search is to assess the ability of IR technology to meet the needs of these communities as tools to help them to do a better job with less effort. This mainly comprises: introducing ranking for the retrieved documents according to their relevance to topics, developing effective and efficient retrieval systems for these tasks, and establishing meaningful evaluation methodologies for the retrieval effectiveness of these IR systems.

While legal search has some clear similarity to patent search, the retrieval challenges of the very diverse documents sets encountered in legal search are very different to those of highly structured patents whose form of writing is very particular to the patent domain. We do not consider legal search further in this thesis. However, the issue of evaluation methods for improved retrieval effectiveness and multiliguality developed in this thesis can also be applied to legal search.

2.7 Summary

This chapter has introduced recall-oriented IR applications and compared them to more common precision-oriented IR applications. The objective of recall-oriented IR is to find as many of the available relevant documents as possible. This type of application is generally used in professional environments where users are sufficiently patient to go deep in the retrieved results list to find any possible relevant documents. Typically, these users are experts in the field of their search, such as patent examiners and lawyers. Recall-oriented IR applications can require multilingual IR, since relevant documents can be in multiple languages, meaning that they must appropriately address the challenges of multilingual IR.

This chapter also overviewed the subject of patents and patent search, since they form the core of the work presented in this thesis. A patent is a title of protection for an invention that is filed into a patent office, which is responsible for checking the novelty of a patent application. The text of patents is characterized by its ambiguity since it is written in a different way to standard text in order to cover all aspects of an invention without focusing on helping the reader to understand it. The citations of a patent contain a list of patents related to the current patent application that describe part of the prior-art. These citations are typically from multiple patent offices in different languages. This fact highlights the importance of multilingual search in patent retrieval.

The main patent retrieval tasks were listed and defined. These include the two tasks that are close to the real life scenario in the patent office, namely patent invalidity search and prior-art patent search. Also, the major research directions applied to these tasks were listed including patent query formulation, query expansion, and cross-language patent search. Prior work on patent query formulation and query expansion is described and discussed in detail in the next chapter. The task of multilingual patent search is extensively explored and investigated in Chapter 6.

Finally, patent search was compared to legal search since this represents another important example of a recall-oriented IR application. While there are similarities, the differences and specific challenges of patent search mean that legal search is not considered further in this thesis.

Chapter 3

Prior Work in Patent Search

The previous chapter introduced the topic of recall-oriented IR with a specific focus on patent search. The nature of patents and the challenges of patent search were highlighted. This chapter provides a detailed review of existing work on patent search tasks. The NTCIR and CLEF international evaluation campaigns which adopted patent search tasks are described through highlighting the tasks in their patent tracks, including descriptions of the data used, and the techniques reported by participants in these tracks. Moreover, prior work in patent search in individual research papers is also overviewed and discussed.

This chapter concludes by summarizing the special characteristics of patent search drawn from our review of the existing work.

3.1 NTCIR Patent Retrieval Track

The NII Test Collection for IR Systems Project (NTCIR) is a series of evaluation workshops designed to promote research in information access technologies including IR, question

answering, text summarization, information extraction, etc[10]. NTCIR has been running on an eighteen months cycle since 1999, when it started with NTCIR-1. The current NTCIR for year 2010/2011 is NTCIR-9. One of the activities of NTCIR was the patent retrieval evaluation track, which started at NTCIR-3 in 2002 and remained as one of the tracks until NTCIR-6 in 2007. In later NTCIRs, interest in patents changed to different tasks other than retrieval, such as patent classification and patent translation.

In this section we overview the different patent retrieval tasks in this track during the four NTCIR workshop at which it was included. The first subsection describes the tasks and the data provided, and the following one provides a short summary of the techniques reported for the different tasks over the years.

3.1.1 Task descriptions and patent data sets

Different patent retrieval tasks were investigated through the four times that the patent retrieval task was investigated at NTCIR. These included ad-hoc search, invalidity search, and passage retrieval tasks. Patent retrieval was first included at NTCIR-3 (Iwayama et al., 2003), where the task investigated was an ad hoc patent search task, which was named the "technical survey" task. As discussed in Section 2.4.3, this task does not model a real-life scenario in the patent office, since the task was modelling the scenario of an organization which wants to file a patent on a given idea, and so a technical survey is carried out by searching for relevant patents to investigate the novel aspects of this idea. Nevertheless, it provided a starting point for the exploration of patent retrieval, and for understanding its nature and some of its challenges. The main patent collection contained around 700k Japanese patents. In addition, a collection of 1.7

[10] http://research.nii.ac.jp/ntcir/outline/prop-en.html

million Japanese patent abstracts were provided with manual English translations. The abstract collection was for optional use by participants as a lexical or translation resource. 31 topics with manual relevant assessments were provided in the format described in Section 2.4.3 as a clipping consisting of few words representing an invention idea, and a memorandum which describes the exact information need of the topic. Four translations of the topics were provided in: English, Korean, and traditional/simplified Chinese for an optional cross-language search task. Eight participants submitted runs to the NTCIR-3 patent retrieval task. Precision-recall curves were used for evaluating the participants' runs.

In NTCIR-4 (Fujii et al., 2004), invalidity search was adopted as the patent retrieval task. A collection of 1.7 million Japanese patents was provided to the participants. The topics were a set of rejected Japanese patent applications with the citations list provided by the patent examiners that invalidate the novelty of the patent. The task was to use each claim in a patent topic as a query to find the relevant patents that invalidate the demand of this claim. 110 Japanese topics were provided with manual English translations provided for a cross-language search subtask. Since the citations of the rejected patents represent only a part of the relevant documents set, additional manual assessment was performed on the merged results of all participants to identify any additional relevant documents. Again eight participants participated in this task. For NTCIR-4, mean average precision (MAP) was used as the evaluation metric.

Invalidity search was also investigated at NTCIR-5 (Fujii et al., 2005). However, on this occasion, an additional passage retrieval task was included which required the identification of relevant fragments in each relevant patent that invalidates the claim. The document collection provided for NTCIR-5 was much larger than that used previously, since the organisers provided a collection of 3.5 million Japanese patents. A parallel corpus of English-Japanese abstracts was also provided to enable participants to train their cross-language IR systems. The topic set

contained 1200 rejected patent applications, with English translations of each one for the cross-language search task. Thirteen groups participated in this track at NTCIR-5 for the invalidity search task. Four of these groups participated in the optional additional passage retrieval task. MAP continued to be used for the evaluation of the invalidity search task and was also used for the passage retrieval evaluation.

In NTCIR-6 (Fujii et al., 2007), the data collection from NTCIR-5 was utilized, but with a new set of 1685 topics. The task again included invalidity and passage patent search as in NTCIR-5. NTCIR-6 featured a new English patent retrieval task, where a collection of nearly one million English patents from the USPTO was provided with a set of 3221 English patent topics. The task for the English collection was invalidity search task. Only five groups participated in both tasks at NTCIR-6, with MAP again being used for the evaluation of retrieval effectiveness.

3.1.2 Summary of approaches used by participants

Although the topics in the invalidity search task in patent retrieval are only the claims, which are much shorter than the full patent topics of the prior-art patent search task, most of the approaches introduced at NTCIR for the patent invalidity search extracted a limited number of terms from the topics to form the patent queries. These approaches used language processing and machine learning tools in order to perform effective term extractions for the queries; these tools included: morphological analysers, lexical resources, semantic indexing, and support vector machines (Fujii et al., 2005; Fujii et al., 2007). The effectiveness of these approaches differed significantly, where the MAP achieved ranged from 0.03 to 0.2 for the Japanese patent retrieval tasks. However, the MAP for the English patent retrieval task in NTCIR-6 was less than 0.1. This shows that the performance can differ from one language to another in patent search. However, it

can be noted that the average values of MAP are considerably lower than those usually achieved in standard ad-hoc search tasks, which indicates the challenge of the patent search task.

The approach reported at the NTCIR patent retrieval track which achieved the highest results among all participants is described in (Fujii, 2007a). Fujii (2007a) used a morphological analyser to extract only the nouns from the patent claim in the patent invalidity search task to be used in the query. He used BM25 (Robertson and Walker, 1994) as the retrieval model, while applying query expansion to the formulated query from the top 10 retrieved documents.

For the cross-language search subtasks in NTCIR, most of the reported work utilized the provided parallel Japanese-English abstracts to create translation dictionaries. The techniques used for creating the dictionaries were based on word co-occurrences, random indexing vector-space techniques for creating cross-language thesaurus, or classified dictionaries based on IPC of Japanese-English parallel abstract corpus (Iwayama et al., 2003; Fujii et al, 2007). MT techniques were not examined in NTCIR, since at that time MT systems were less commonly used in IR at that time. The effectiveness of the cross-language experiments reported at NTCIR was comparable to the monolingual retrieval tasks. This high level of cross-language effectiveness may stem from the nature of the tasks which were artificial, since the topics were manually translated to a different language and the objective was only to automatically re-translate them back to the original language. This is not the case for a real-life search scenario in a patent office, where a patent application is originally in one language and the objective is to find relevant documents in other languages. The CLEF-IP track addresses this problem directly by using real-life multilingual patent search tasks (Roda et al., 2009; Piroi, 2010).

3.2 CLEF-IP Track

The Cross Language Evaluation Forum (CLEF) was founded in 2000 as an evaluation campaign to promote research in IR for European languages. Over the years, many different multilingual IR tasks have been explored by participants in CLEF. In 2009, a new evaluation track was introduced to CLEF, referred to as the Intellectual Property track or CLEF-IP (Roda et al., 2009). CLEF-IP was included as a track in CLEF with the purpose of encouraging researchers to develop methods to improve search quality in the intellectual property domain in general and specifically in patent retrieval.

The work in this thesis is developed using the data and tasks provided by the CLEF-IP track. In this section, the nature and properties of the data will be described with more emphasis on the data provided in 2010, since it was larger and better organized than that used in 2009 as described later.

3.2.1 Task description

The task investigated in the CLEF-IP track is prior-art patent search, where the objective is to find relevant prior-art patents to a set of patent applications that acted as the search topics.

For CLEF-IP 2009, the topics were a set of published granted patents in the period after the collection period (Roda, et al., 2009). Considering published patents as the topics was not the best choice by the organizers, as they realized later (Piroi, 2010), since this kind of patent document is the final version of the patent application after the patent office has completed its job of finding additional related work to the invention, and the patent applicant has updated the patent application for the final version of the granted patent. This was corrected in the following year, where the CLEF-IP 2010 topics were actual patent applications. Thus in 2010, the initial version of a patent application submitted to the patent office was considered as the topic, which

is the same scenario as in real-life (Piroi, 2010). For this reason we focus on the 2010 collection and topics for our studies described later in this thesis.

The typical scenario in the patent office is to receive a patent application claiming to describe a novel invention with some initial citations to related work. The patent applicant hopes that this will show that the invention described in their patent application is novel. The patent examiner takes this application and searches for possible prior-art work that can invalidate the invention or at least that has important related work and should be cited by the patent. The final relevant results of the search are used as the citations in the final published version of the patent document. For this reason, the final list of citations in the revised patent was taken as the set of relevant documents for each topic in the CLEF-IP track (Graf and Azzopardi, 2008; Roda et al., 2009; Piroi, 2010) and was hidden from the participants. The task in CLEF-IP was to find these citations automatically from the patent application through IR approaches.

3.2.2 Data collection

The collections provided for the CLEF-IP tasks are a set of patents filed with the EPO with application dates prior to 2000 for the CLEF-IP 2009 collection and prior to 2002 for CLEF-IP 2010 collection. The patent collections consisted of one million patents in 2009 and 1.35 million patents in 2010. The 2009 documents collection is a subset of the 2010 collection.

Each patent in the collection consisted of multiple versions of documents in XML format labelled as A1, A2 … B1, and B2. The "A" letter refers to different versions of patent applications. The number with the letter refers to the version; thus A1 refers to the initial version of a patent application submitted to the patent office. The "B" versions refer to granted patents. Each of these versions contains some updates to the text, citations, and claims of the previous one. For all the experiments in this thesis, different versions of a single patent were merged into

one single document as suggested by the track organizers. The content of each section in the merged document was taken from the latest available version, since some of the final versions consisted of only the sections which had their content updated rather than all the patent sections. Some of the patents had incomplete versions of documents, which led to the presence of some patents in the collection with many missing content fields. The problem of missing data is in some cases so significant that some of these patents consist only of the title.

The patent collections contain material in three different languages: English (EN), German (DE), and French (FR). However, most of the patents have some parts translated into all three languages. The translation into these three languages is one of the special properties of the patent published by the EPO, where the granted published version of a patent ("B" versions) should contain the claims section manually translated into all three languages. In addition, all the patents have the title in the three languages. These translations are aligned on the paragraph level in the XML file. The description section of all patents is always provided in the original submission language only. This section forms more than 50% of the text in each EPO patent document.

The track organizers provided training topics for participants to test their approaches during the development of their retrieval systems. They also provided test topics to be used for submission to the track. The topics provided were English, German, and French patent applications. As mentioned earlier, the patent topic is a full patent document. The length of a patent topic reaches thousands of words, where the average number of the terms in a topic after stop word removal was 3,554 terms/topic for CLEF-IP 2010 topics. This highlights the large difference between topics of recall-oriented tasks and other standard IR tasks.

Table 3-1 provides some information about the data provided in CLEF-IP in 2009 and 2010. Since the IR experiments presented in this thesis will be based on the CLEF-IP 2010 collection, additional details of the 2010 patent collection are provided in Figure 3.1 and Figure 3.2.

Table 3-1: Main differences between patent collections provided by CLEF-IP in 2009 and 2010

Collection	CLEF-IP 2009	CLEF-IP 2010
No. of documents	1,958,956	2,680,604
No. of patents	1,022,388	1,331,158
Application date	Prior to 2000	Prior to 2002
No. of topics	500 training 500, 1000, 10000 test	300 training 500, 2000 test
Topic nature	Granted patent	Patent application

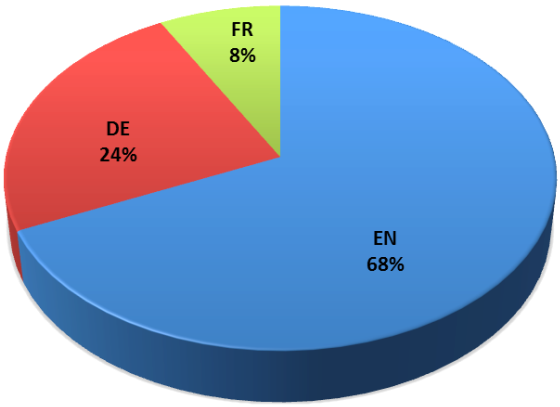


Figure 3.1: Percentage of English, German, and French patents in CLEF-IP 2010 collection

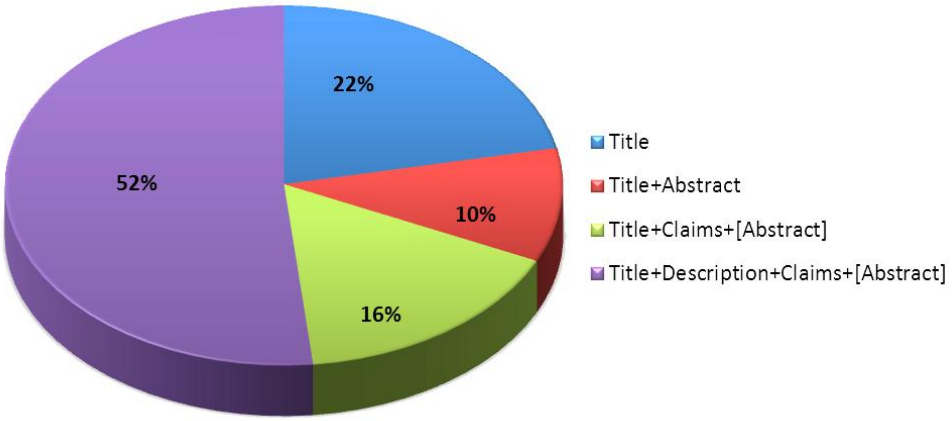


Figure 3.2: Completeness of the presence of English text in the CLEF-IP 2010 patent collection

Figure 3.1 shows the percentage of the English, German, and French patents in the CLEF-IP 2010 collection. Since some patents in the collection do not contain all sections, and since some of the non-English patents do not contain translations into English, Figure 3.2 presents the distribution of the missing English sections in the patents. Figure 3.2 shows the amount of the English content present in the patents in the 2010 collection, where only 52% of the patents in the collection were complete English documents. 16% of the collection included the titles and claims sections only, while some of them contained the abstract section as well. These patents are not complete patent documents, but at the same time, they are not short because of the presence of the claims section which contains most of the important information about the invention disclosed. 32% of the patents do not include the description or the claims sections in English, while most of them included the titles only, which means that the retrievability of these patents is expected to be very low, since they contain only a very small number of words (Bashir and Rauber, 2010). The overall aim of Figure 3.2 is to show that the documents in the patent collection are not homogenous since many of them are in some respect incomplete.

3.3 Approaches Used by Participants in CLEF-IP 2009 and 2010

The approaches used by participants in the CLEF-IP tasks did not differ greatly between 2009 and 2010 (Roda et al., 2009; Piroi, 2010). Most of the work done by the participants focused on the conversion of the long patent application documents into a usable queries, or in other words, the main area of investigation of most of the participants was query formulation. Here we summarize the approaches used by the participants in CLEF-IP 2009 and 2010, while giving more focus to the ones which achieved the highest results. In the next chapter, full detail of our own participation in these two tracks is provided.

3.3.1 CLEF-IP 2009

In CLEF-IP 2009, 15 participants submitted more than 40 runs to the patent prior-art search task. Several metrics were used for the evaluation of the submitted runs, including: MAP, recall at different cut-offs, precision at different cut-offs, and nDCG (Carterette et al., 2008). However, the MAP was used for ranking the best runs in the track. Excluding the best run, which achieved a much better result than the others, the participants' MAP values ranged from 0.0031 to 0.1145, which shows the low retrieval effectiveness of the CLEF-IP task, which is similar to the results observed for the NTCIR patent retrieval tasks. Most of the work tried to utilize the text in various sections of the patents to formulate the query. The results from participants showed that extracting query terms from the description section of the patent topic achieves better results than extracting the terms from other sections. Most of the participants found that using the patent classification (IPC) for filtering the results by constraining them to have common classifications to the patent topic improves the results with respect to both precision and recall (Roda et al., 2009). Also, from the reported results, it appears that the retrieval model used does not have a strong impact on the results. Most of the standard IR models were used by the participants including: TF/IDF (Baeza-Yates and Ribeiro-Neto, 2010), BM25 (Robertson and Walker, 1994), BM25F (Robertson et al., 2004), and language modelling (LM) (Song and Croft, 1999). The results showed that the method of term selection from the patent topic for the query had the highest impact on results (Roda et al., 2009).

As stated earlier, the patent collection contained documents in multiple languages (English, French, and German). Two approaches were used by most of the participants to index this multilingual collection (Roda et al., 2009):

- The first approach indexed only the English content of the patents, where all the content of the English patents was indexed and where available the English translations of the

French and German patents were also indexed. The shortcoming of this approach is that the content of the non-English patents which is not translated into English is excluded.

- The other approach indexed the full patent collection while including all the English, French, and German content in one multilingual index. This approach is not the best solution, since a translation stage should have been used to put all the content of the collection in one language rather than putting the content in its original language. However, translating the non-English content was unrealistic due to the large size of the text that needs translation which requires a long time to train an MT system and translate the patents.

Regarding the topics in English, French, and German, there were two main approaches taken:

- Participants who only indexed the English content, used the English content of the topics for search, since the topics were granted patents and English translations existed for the title and claims sections
- Participants who used the multilingual index, also used the multilingual content of the topic in constructing the queries. In this approach the query content in each language is expected to match the corresponding document content in the same language. However, this is not entirely a correct approach, since a patent topic in one language has relevant patents in all the three languages.

The results reported in the CLEF-IP 2009 track showed that the groups who followed the first approach achieved better performance than those who used the other one (Roda et al., 2009). This suggests that using the multilingual content in this way for search and indexing is not the best solution in patent search. A more detailed discussion about this approach and cross-language patent search in general is provided in Chapter 6.

The best run in CLEF-2009 achieved 0.27 MAP, which is considerably higher than any other submitted run (Lopez and Romary, 2009). The participant group of this run built a sophisticated and very advanced system that utilized a very important feature of patents, which was not utilized by any other participants. This feature is the presence of some citations within the text of the description section of the patent topic, see Figure 2.3. Some of these citations are part of the relevant documents set of the patent topic. In effect, it is like part of the answer is found within the question.

Lopez and Romary (2009) extracted these citations from the description section and used them to generate what they called a "working set" of patents from the patent collection to search a smaller portion of the patent collection rather than searching the full collection. The "working set" to be searched is formed from all the patents in the collection which have similar classification, inventors, or citations to the patent topic and the extracted citations. Figure 3.3 presents the system architecture of the patent search system developed by Lopez and Romary (2009). Their method included multiple indexing of the patents, where they created a separate index for each language in the patent collection (English, French, and German). They also used indexes for English phrases and English concepts generated from various external resources including Wikipedia. These indexes were searched several times using several retrieval models to produce different lists of results. The retrieval models used included LM with KL divergence and Okapi BM25. Later they merged all the retrieved ranked lists from these multiple indexes and retrieval models using machine learning techniques, namely, support vector machines (SVM) (Joachims, 2002). Finally, they used SVM for re-ranking the merged list of results using the metadata of the patent topics as the features.

All the components found in Lopez and Romary's (2009) system seek to model the real-life activities performed by the patent examiner in the patent office. Although the system showed

high effectiveness for patent search, it has a high complexity in its architecture and requires a large number of resources. In the next chapter, we compare this system to a much simpler one developed by ourselves and show that both systems are comparable in their retrieval effectiveness for patent search.

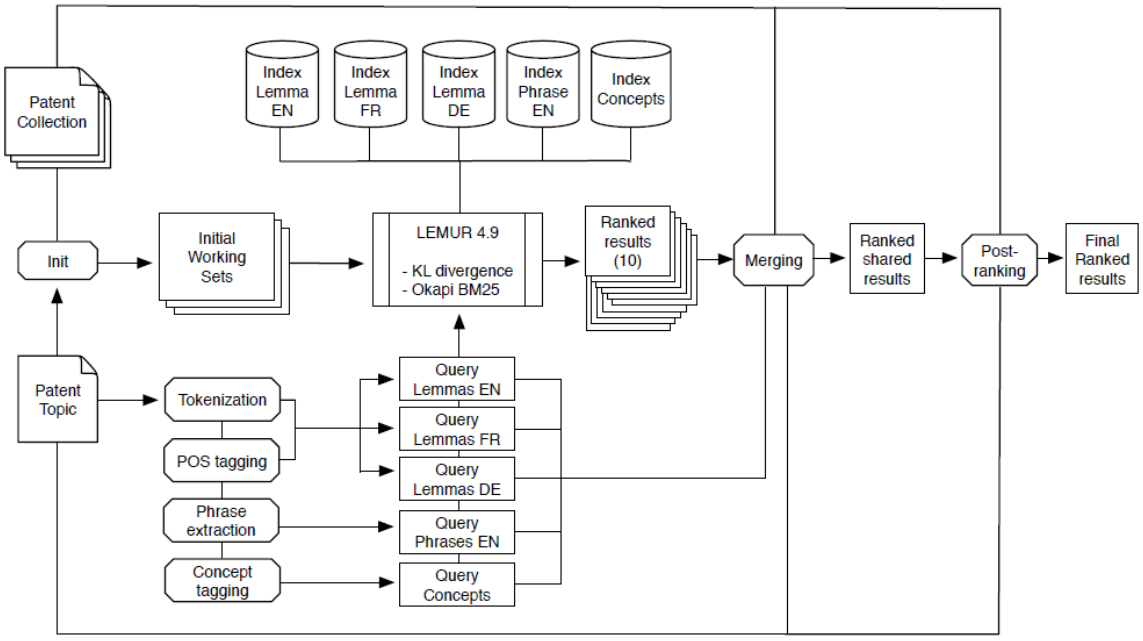


Figure 3.3: Patent retrieval system architecture in (Lopez and Romary, 2009)

3.3.2 CLEF-IP 2010

In CLEF-IP 2010, nine participants submitted 25 runs. The results in this year were very similar to these observed in 2009, where the MAP values ranged from 0.007 to 0.2645. Again the best result was achieved by the same group as in 2009 (Lopez and Romary, 2010). However in this year, our group utilized the citations in the patent topic in a much simpler way and achieved the second best result of 0.203 MAP (Magdy and Jones, 2010c). Full details of our participations in

CLEF-IP are described in detail in the next chapter. The third best participant achieved a MAP of 0.1404 (Piroi, 2010).

As mentioned earlier, the topics for CLEF-IP 2010 were patent applications rather than granted patents as in 2009. Therefore, non-English patent applications did not contain any English translations in any section except the title and the abstract. Hence a translation step was required for these topics. In addition, the citations present within the description text are the initial list of citations included by the patent applicant. The majority of these citations are not included by the patent examiners in the final list of citations in the granted patent which are taken as the relevance set. This means that citation extraction methods in this case are not expected to be as useful as in 2009.

Most of the participants used very similar techniques to those used in 2009, where most of the sections in the patent topic were utilized to formulate the query (Piroi, 2010).

To enable search for the non-English topics, most of the participants used Google translate[11] to translate the topics into English, and then used the translation for formulating the search query (Piroi, 2010). Some participants, who continued using the mixed languages index, formulated the query from the topic in its source language, but this approach achieved lower results than those who used Google translate. One participant utilized the English translations of the title and abstract sections only of the non-English topics to formulate the query. However, this latter approach led to very poor results compared to the others (Piroi, 2010).

Lopez and Romary (2010) who achieved the best result in the CLEF-IP 2010, used a very similar system to the one they used in 2009. However, in their participation in 2010, they added a keyword extraction block to the system architecture in Figure 3.3 to search their indexes with a

[11] http://translate.google.com/

limited number of keywords. The results of this search were merged with the other result lists using an SVM. Also, they applied *patent family* look up to the extracted citations from the description section of the patent topic in order to find the corresponding EPO patents to the citations from outside the EPO. Full details of their method for CLEF-IP 2010 are provided in Section 4.3, where we compare out approach to theirs.

Additional research has been reported on the CLEF-IP 2010 data subsequent to the track submissions (Mahdabi et al., 2011). In this work only the English topics of CLEF-IP 2010 were studied. The objective of this research was to find the best retrieval models and the most useful sections of patent topics for query extraction. The research did not take into account the citations within the patent topic description so that the focus was on the retrieval approach. This is a reasonable approach, since not all patent applications contain citations, so effective retrieval methods are needed where citations are not present. Their analysis showed that the patent description section is the best one for extracting the query terms, which supports the results of existing reported work (Roda et al., 2009; Piroi, 2010; Magdy and Jones, 2010c). In addition, they used a modified version of weighted log-likelihood (Meij et al., 2009) as the query generation model. Their results were comparable to those achieved by other participants in CLEF-IP 2010 when no citation extraction is used (Piroi, 2010). In fact, their results were worse than our run when no citations extraction was used (Magdy and Jones, 2010c), where they achieved 0.1240 MAP, while we achieved 0.1399 MAP without using a special model for query generation.

3.4 Additional Patent Search Tracks

Two further tracks examining patent search outside the scope of our investigation are described in overview in this section.

3.4.1 CLEF-IP 2011 track

CLEF-IP continued running in 2011 and included the same patent data collection form the EPO as was used in 2010, but with the addition of a new set of patents from WIPO. The inclusion of this new material meant that the collection contained more than 1.5 million patents (Piroi et al., 2011). The WIPO patents also contain patents originally in English, French, and German, and similar to the EPO patents, the claims and titles contain translations in all three languages. However, the names of some fields in the XML files of the WIPO patents differ slightly from the EPO patent sample in Figure 2.3, thus additional updates to the parser were required to process the patent XML documents for this data. In addition, a new image retrieval task was introduced to the CLEF-IP track in 2011, where a small set of patent documents were provided with their images for the task. This task is similar to the prior art search task, but the prior art patents should be found based on images and text of a query patent.

The CLEF-IP 2011 task is neither included in our study nor reviewed in this thesis since it was introduced after the completion of the experimental work reported in this thesis.

3.4.2 TREC-CHEM track

Another track that investigated patent search was the TREC-CHEM track (Lupu et al, 2009a; Lupu et al., 2009b). This is part of the Text REtrieval Conference[12] (TREC), which is the largest and oldest IR evaluation campaign. TREC-CHEM, or the chemical IR evaluation track in TREC, was introduced at TREC in 2009. Its objective is to address the challenges in chemical patent IR. The provided test collections consist of full-text chemical patents and research papers from the

[12] http://trec.nist.gov/

Royal Society of Chemistry, UK. The aim is to identify how IR methods should be adapted to facilitate more effective retrieval on texts containing chemical names and formulae, where these chemical formulae represent the key contribution of a patent application. The investigated subtasks in the TREC-CHEM track are prior-art search task and technology survey (ad-hoc) search task. The patent dataset used for this task consists of chemical patents from three sources: EPO, WIPO, and USPTO. In addition, a collection of academic papers from the chemistry domain was also provided for the technology survey task. The approaches introduced for the patent search in the track were very similar to those used in CLEF and NTCIR; however there was some additional work for chemical named entity recognition and parsing by some of the participants to handle chemical formulae effectively (Lupu et al., 2009a).

Although the track is concerned with patent retrieval, we did not include it in our study since its focus is on the development of methods for matching chemical formulae as well as building effective systems for recall-oriented search tasks. Studying this task would require additional work to develop an effective IR system for chemical patents including chemical formulae, which is not the main focus of our study in this thesis.

3.5 Additional Prior Work on Patent Search

The NTCIR patent retrieval track and the CLEF-IP track have provided sets of data collections for different patent retrieval tasks. This has initiated additional research on patent retrieval even outside the tracks. Here we summarize part of this research according to the research direction.

3.5.1 Exploiting patent structure and content for search

Fujii (2007b) used a text-based approach in conjunction with citation-based retrieval methods in invalidity patent search applied to the USPTO patent collection consisting of nearly 1M patents,

which was provided by the NTCIR-6 patent retrieval track (Fujii et al, 2007). The system utilized a similar approach to PageRank used in web search (Brin and Page, 1998), where webpages which get more links from other webpages get higher score of ranking. In the same way, if a patent is cited by a large number of other patents, this cited patent is possibly a foundation of those citing patents and is, therefore, important. This sophisticated approach improved the retrieval effectiveness to 0.0811 MAP from a baseline of 0.0712 MAP, an increase of 14%.

Later, Xue and Croft (2009a) used the same collection to test several methods for patent prior-art search. The research showed that some fields in the topics (which are full patents) can have better retrieval effectiveness when used as the query. This study showed that the "background summary" section in patents is the most useful source of terms when used in the search query. The best result achieved in their research when using the "background summary" section as the query was 0.094 MAP, which was 119%, 42%, and 28% better than when using the patent "title", "claims", and "abstract" sections respectively. This research highlights one of the challenges in patent retrieval mentioned earlier, namely, the diversity of the structure of patents depending on the patent office from which it originates. For example, the "background summary" field only exists in the US patents, and is absent from patents issued by the EPO (although, it can be seen as corresponding to the "description" section in EPO). This diversity means that it is difficult to have a universal retrieval approach for patent search. In further work reported in (Xue and Croft, 2009b), additional features were used in the search process. A learning to rank approach using machine learning was applied to combine these features. These additional features included term search and phrase search using language modelling as the retrieval model (Strohman et al., 2004), patent classification features, and some low level features such as *tf* and *idf* of terms. They managed to further improve the MAP to 0.108, a 15% increase.

In (Takaki et al., 2004) the long patent query was automatically analysed and divided into query subtopics for a Japanese patent collection in a patent invalidity search task from NTCIR-4. Each subtopic in the query was used to retrieve relevant documents with a relevance score. The relevance scores were weighted by the importance of each subtopic element and integrated to determine the final ranked document list. The importance of subtopics was calculated as the summation of the specificity of the query term in the subtopic. The specificity of a query term was calculated using entropy, which is the deviation degree of the appearance of the term in each subtopic. Using this technique, they managed to improve the retrieval effectiveness in an invalidity patent search task from 0.1375 MAP when not performing the subtopic extraction to 0.1484, an 8% increase.

In other research by Mase et al. (2005), different techniques were applied on Japanese patents from NTCIR-4, where a two-stage strategy for patent retrieval was applied. In the first stage, the claim part of the query patent was used to retrieve the top 1,000 patents. In the second stage, several techniques were used to re-rank the top 1,000 patents including text analysis and retrieval methods using the claim structure. Their evaluation results showed that the method used is effective, but the degree of effectiveness is not stable on all topics. However, for the topics where MAP is not improved, they noticed that the retrieval ranks were improved for 70% to 80% of the retrieved relevant documents. This observation by Mase et al. (2005) highlights that MAP can give a misleading indication of the change in retrieval effectiveness for patent search, since there can be a large improvement in the ranking of the relevant documents retrieved at low ranks in the results list without this being reflected in the MAP value. We study the adequacy of using MAP for evaluating recall-oriented search tasks in general and particularly patent search in Chapter 5.

3.5.2 Toward enriching patent queries

A number of query expansion (QE) techniques have been introduced in the field of IR with the objective of improving retrieval effectiveness. In general these operate by adding additional descriptive detail of the user's information need into the original search query. QE techniques are based on providing additional terms to the original user's query, which typically are short in most IR applications. Many approaches have been proposed and explored for the selection of these additional terms and how they are weighted in combination with the original query terms. The expansion terms can be selected from a feedback process (Rocchio, 1971; Salton and Buckley, 1990), or from external sources such as Wikipedia (Xu et al., 2009), dictionaries (Collins-Thompson and Callan., 2005), or query logs (Wen et al., 2002). Expansion can be per term such as using WordNet (Liu et al., 2004) or per query as in the case of relevance feedback (Cao et al., 2008; Billerbeck and Zobel, 2004; Rocchio, 1971).

The main assumption when expanding the original query with additional terms is that the added terms increase the probability of matching more relevant documents, which can improve the retrieval effectiveness. However, this assumption is not always valid, since the added terms can lead also to retrieving more non-relevant documents. For this reason research into QE techniques typically focuses on expanding the queries with "good" terms that lead to improvement in the overall performance of the retrieval system (Billerbeck and Zobel, 2004; Cao et al., 2008). QE in general has diverse impact on retrieval effectiveness depending on the application and the expansion method itself, where it sometimes improves the retrieval effectiveness and sometimes degrades it (Cao et al., 2008).

The main objective of QE is to overcome the problem of mismatch between queries and relevant documents by increasing the number of retrieved relevant documents or improving the rank of relevant documents already retrieved with the original query. QE is effective in many

applications, where queries are short and may be not describing the user's information need well. Although the situation for patent queries is exactly the opposite, where the query is a full document, there is often an insufficient match between the content of patent queries and relevant patents (Roda et al., 2009). This has led researchers to explore the use of various QE techniques to overcome this problem in patent search.

Existing research in QE techniques for patent search has so far not demonstrated any stable improvement in IR effectiveness. The following overview highlights the main studies in this area and their findings.

One of the initial studies that examined pseudo relevance feedback (PRF) for patent search is described in (Kishida, 2003). PRF performs QE by assuming that the top ranked retrieved documents from an initial search run are relevant. In this research study a novel mechanism for PRF was introduced and compared to the standard Rocchio method (Rocchio, 1971). Kishida (2003) developed a term selection formula for terms from the top retrieved based on the Taylor formula of the linear search functions. The main feature that distinguish the Taylor formula from other term selection formulae is using the document retrieval scores to give higher weights for terms extracted from documents at higher ranks. Experiments were carried out on the NTCIR-3 patent retrieval task, but none of the feedback techniques introduced in this work led to any significant improvement in the retrieval results. Kishida (2003) mentions that the reason may be that all terms from the documents assumed relevant were used for expanding the original query without any selection process to identify significant terms.

In NTCIR-4, another approach to utilizing QE through PRF to improve the retrieval effectiveness was explored in (Itoh, 2004). Itoh (2004) utilized the BM25 retrieval model while enabling and disabling PRF. The results showed that retrieval effectiveness when enabling PRF

was improved for a few topics while being degraded for many others. He could not give clear analysis to why this happened for this task.

In the patent invalidity search task in NTCIR-5, another group tried to utilize the PRF approach in a different way by only reweighting the terms of the query based on comparing the hierarchical structure of the patent classification (IPC) of the initial retrieved documents to that of the query (Takeuchi et al., 2005). They extracted keywords and calculated their frequencies from the search result using similarity between IPCs in the search topic and those of each patent document in the collection. They used *tf*, *idf*, and *tf-idf* for calculating the similarity between IPCs of topics and documents. Then, they modified the weights of the keywords considering those appeared relatively frequently in the retrieved documents. Again, this technique did not lead to any significant improvements to the final search results.

Another QE method was explored in NTCIR-5 for invalidity search by adding terms to the query from the patent topic itself instead of the initially top retrieved documents (Konishi, 2005). The technique attempted to expand the query, which is a patent claim, using additional explanatory text of the claim from the description section of the patent topic. The challenge was to locate the part in the description section that describes the given claim. They used morphological analysis and pattern matching techniques to identify these relevant parts and appended them to the query claim. Their technique achieved significant improvement over the baseline. However, a disadvantage of this QE technique is that it is specifically designed for the patent invalidity search task, and cannot be generalised to other tasks such as the prior-art search task, where the query is the full patent application including the claims and the description sections.

Another investigation explored the use of PRF to improve the retrievability of patents in patent search rather than improving the retrieval effectiveness directly (Bashir and Rauber,

2010). The problem addressed in this research was that some patents have a low chance of being retrieved by any query due to their very vague vocabulary or the small amount of text in some of these patents, which are mainly described by images. The objective for this research was to enrich the patent queries with additional terms using the PRF method to improve the retrievability of patents in the collection. They succeeded in significantly improving the patent retrievability for a collection of 54,353 patents from the USPTO, which was evaluated using the Gini coefficient that measures the degree of bias in the retrievability of documents of a collection. However, they did not test their approach on any patent search task to examine how this would actually affect the retrieval effectiveness for patent search.

3.6 Summary

This chapter has reviewed some of the prior work in patent search tasks. The NTCIR patent retrieval task and the CLEF-IP task were described including the data collections provided, the subtasks, and the main approaches used by participants in each track. We have emphasised the CLEF-IP collections in this review, since it forms the basis of the core experimentation in this thesis. We also presented additional prior work done on different patent retrieval tasks using the data provided by the NTCIR and CLEF-IP tracks, but outside of the contribution to the tracks themselves. This included work on building effective queries from the long patent topics, and different query expansion techniques for patent search. The reviewed work has demonstrated some of the features and the challenges of the patent search, which can be stated as:

1. Retrieval effectiveness in patent search is considerably lower (around 0.1 MAP) compared to other more common IR applications and tasks, where a MAP of 0.3-0.4 is often quite realistic. This illustrates the challenge of the task.

2. Although patent search is a recall-oriented task, MAP has been the commonly used score for evaluating patent search in all the examples reviewed here. This is perhaps surprising since recall is considered very important for this task and MAP is a precision focused metric that has weak reflection of the system recall.
3. It has been shown that extracting citations from a patent topic's description section improves the retrieval effectiveness significantly. This is something that should be taken into consideration while building any effective patent retrieval system.
4. There have been some attempts to improve the retrieval effectiveness of patent search through applying different query expansion techniques for the patent queries. However, none of these has led to a significant improvement in the retrieval effectiveness, which further demonstrates the special nature of patent search that makes many standard IR techniques ineffective for improving retrieval.
5. The currently common approaches used for the translation stage in the research in cross-language patent retrieval are typically based on using free online MT systems, such as Google translate, that are not trained on patent data[13], and have limited access to be able to translate the large patent content. Other solutions included using multilingual queries and multilingual indexes for search, or ignoring patent content in other languages altogether. This means that efficient and optimized patent translation methods should be developed to improve cross-language patent search.

The next chapter presents our contribution to the CLEF-IP prior-art patent search task as part of our investigation of the challenges of patent search. In the later chapters, we present the main contributions of this thesis to the patent search task and recall-oriented IR in general.

[13] In November 2010, Google signed an agreement with the EPO to provide translation services to the EPO

Chapter 4

Understanding Prior-Art Patent Search Using CLEF-IP

The previous chapter overviewed prior work in patent search tasks and summarized the state-of-the-art of the main research directions in patent search. In this chapter, we present details of our participation in the CLEF-IP prior-art patent search task as part of our exploration of patent retrieval. We participated in CLEF-IP for two successive years in 2009 and 2010 to explore recall-oriented IR as represented by patent prior-art search. We use our best result achieved for this task, which is comparable to the state-of-the-art (Magdy et al., 2011), as the baseline for the remainder of the experiments reported in the thesis. Our participation in 2009 (Magdy et al., 2009) mainly concerned exploring the patent search task through many experiments focused on query formulation. We achieved the seventh position among 15 participants when ranking with MAP, and the fourth position when ranking with recall. In 2010 (Magdy and Jones, 2010c), we used what we learned from our participation in 2009 and implemented a simple and straightforward method for search to achieve the second best ranking among 25 submitted runs.

In this chapter we also compare our best result with the overall best submission to in the CLEF-IP 2010 track (Patric and Lopez, 2010), and demonstrate that both results are comparable to each other under the application of the same system variables (Magdy et al., 2011).

Finally, we explore query expansion methods (QE) for the patent prior-art search task and propose a novel method for QE that is significantly more efficient and effective than reported work for QE methods for patent search.

4.1 Our Participation in CLEF-IP 2009

In the first year of our participation in CLEF-IP (Magdy et al., 2009), our experiments tested various techniques in the patent retrieval process, including query formulation, structured indexing, weighted fields, and document filtering. Some methods did not contribute to improved retrieval effectiveness, such as structured indexing, while others showed some improvement in retrieval effectiveness, such as extracting query terms from the description section and applying filtering to results based on patent classifications. Query formulation was shown to be the key to achieving better retrieval effectiveness. We tested different query formulations for the patent topic by combining different short sections of the patent topic with different weights vs. using the patent description section as the query. Our experiments showed that longer queries achieve better retrieval results.

4.1.1 Developing the patent search system

The patent retrieval system that we built for our participation in CLEF-IP 2009 forms the basis of the system used for the experiments reported in the remainder of this thesis. This system was extended for our participation in 2010 based on what we learned in 2009. Our patent retrieval system is described in detail in the following subsections.

- **Patent text extraction**

We described in Section 3.2.2 of the previous chapter that a collection of 2 million patent XML documents representing different versions of 1 million patents were provided by the CLEF-IP track in 2009. In our system, different versions for each patent were merged into one document as recommended by the track organisers. In order to index the text of the patents in the collection, an XML parser was developed to extract the text from different sections of patent, which are represented as fields in the structured patent XML file.

Only the English text of the patents was indexed to create the index for our participation in CLEF-IP 2009. This strategy was adopted by many of the participants in the CLEF-IP track (Roda et al., 2009). All the sections of the English patents were indexed, including the "title", "abstract", "description", "claims", and "classifications". For the French and German patents, only the sections with English translation were indexed, including the "title", "abstract", "claims", and the language independent "classification" field. This means that non-English patents did not have their "description" section indexed since it does not contain a translation into English. This can lead to reduced retrievability for these patents. We did not include non-English content since we did not wish to build a multilingual index at this stage of our work. Results from participants in CLEF-IP 2009 showed that the approach of multilingual index is less effective than using only the English content for indexing (Roda et al., 2009). The alternative to avoid building a multilingual index and to index the full patent contents would have been to translate the description sections of the non-English patent into English. This was unrealistic, due to the computational requirement of machine translation systems leading to translation being too slow to be completed in a practical timeframe (Parton et al., 2008).

- **Text pre-processing**

Patent text contains many formulae, numeric references, and patent-specific words (such as *method*, *system*, or *device*) that can cause a negative effect on the retrieval process. Thus, pre-processing to the text was performed to remove predefined stop words[14], digits, and patent-specific stop words.

Patent-specific stop words were identified individually for each patent section (field) separately as suggested in (Sarasua and Corremans, 2000). To obtain the field stop words, the field frequency of each term was calculated separately for each field. The field frequency for a term "T" in field "X" is the number of fields of type "X" across all documents containing the term "T". For each field, all terms with field frequency higher than 5% of the highest term field frequency for this field were selected as stop words. The value 5% was selected subjectively based on our observation of the data. For example, for the "title" section, the following words were identified as field specific stop words: *method*, *device*, *apparatus*, *process*; for another section such as "claims", the following words were identified as field stop words: *claim*, *according*, *wherein*, *said.*

Following the elimination of stop words, Porter stemming (Porter, 1980) was applied to all remaining words.

- **Structured Indexing**

The Indri search toolkit (Strohman et al., 2004) was used for indexing and searching the patent collection. The Indri retrieval model[15] is a combination of the language model (Ponte and Croft, 1998) and the inference network retrieval frameworks (Turtle and Croft, 1991). The Indri

[14] http://members.unine.ch/jacques.savoy/clef/index.html

[15] http://ciir.cs.umass.edu/~metzler/indriretmodel.html

toolkit has the capability of creating a structured index; therefore we maintained the patent structure in the Indri index to keep the patent fields structure. This structured index allows search of specific fields in the document, or search of the full document as a whole. It also allows different weights to be assigned to each field while searching.

For example, given a query Q that is constructed from three different query parts: Q_1, Q_2, and Q_3; each of these parts can be used to search the different fields: $field_X$, $field_Y$, and $field_Z$ in the structured index while giving triple weight to $field_Z$. The Indri query would take the following form:

$$\text{Query: } 1\times(Q_1).field_X \; 1\times(Q_2).field_Y \; 3\times(Q_3).field_Z$$

This means to search $field_X$ for query "Q_1", $field_Y$ for query "Q_2", and $field_Z$ for query "Q_3" while giving searching $field_Z$ triple the weight of searching any of the other fields. Any of these fields can be a subsection of the others. The retrieval score is calculated as the geometrical mean of the retrieval scores of searching each of the field individually. Therefore in the example shown, the final retrieval score would be calculated as:

$$Score_{retrieval} = \sqrt[5]{Score_X \cdot Score_Y \cdot (Score_Z)^3}$$

where $Score_x$ is the retrieval score for searching "Q_1" in $field_X$, and similarly for $Score_Y$ and $Score_Z$.

The main advantage of Indri that motivated us to use it as the retrieval tool in our experiments is its rich query language which allows different features in queries, such as: structured queries, weighted query terms, filtering of results, phrase matching, synonyms, and weighted synonyms. A description of the Indri syntax is provided at the end of the thesis in Appendix B.

Figure 4.1 presents a sample of a structured patent document in TREC format that is indexed by Indri. As shown in Figure 4.1, "Desc1" and "Claim1" are sub-fields of the description "Desc"

and claims “Claims” fields respectively. “Desc1” is the first paragraph in the description field; typically it contains useful information about the field of the invention and what the invention is about. “Claim1” is the first claim in the claims sections, and typically describes the main idea of the invention in the patent. The field “Class” contains the IPC classification information of the patent.

```
<DOC>
     <DOCNO>patent number</DOCNO>
     <TEXT>
          <TITLE>title</TITLE>
          <CLASS>3rd level classification</CLASS>
          <ABSTRACT>abstract</ABSTRACT>
          <DESC>
               <DESC1>1st sentence in description</DESC1>
               Rest of patent description
          </DESC>
          <CLAIMS>
               <CLAIM1>1st claim</CLAIM1>
               Rest of patent claims
          <CLAIMS>
     </TEXT>
</DOC>
```

Figure 4.1: Structured text for a patent in TREC format

- **Query formulation**

As stated previously, query formulation was our main focus for this task in CLEF-IP 2009. Our main objective was to identify the best sections in the patent topic to extract the text from, and to examine the use of structured queries to search the structured index with different weights assigned to each field. The nature of queries to be formulated in patent search differs significantly than that of queries in other IR tasks, such as web search or ad-hoc search, which are typically consist of only a few words. Patent search queries are generally expected to be very long; they do not have to take the form of a search phrase or keywords, but a set of terms with the frequencies acting as weights; they can contain filtering constraints of based on metadata;

they can contain synonyms; and so on. A sample formulated query patent search is presented in Appendix B.

Sets of terms in a query taken from the individual fields of the patent topic were used to search the patent index with each set searching the corresponding field in the structured index to discover the best sections for formulating a patent search query. Various combinations of fields were employed with different weight combinations. Filtering search results was examined by constraining the retrieved documents to share at least one common IPC class with the patent topic. The top three classification levels of the IPC classifications were only used for the filtering process, which include the IPC subject, class, and subclass (example: "A01B" in Figure 2.2, page 18). Only the top three levels of IPC were used instead of the full IPC code, since our preliminary experiments on the training data provided in CLEF-IP 2009 showed that this improves both the precision and recall of the system. Adding deeper levels of the IPC ("main group" and "subgroup") leads to lower recall, which does not align with the objective of the patent retrieval task that is recall-oriented. We also examined using bigram phrases (formed of two terms) in the patent topic to be added to the formulated query. Bigrams appeared more than once were used in the query. The inference network implemented in Indri allows search for phrases without the need to build a phrase index (see Appendix B).

The patent topic text was pre-processed using the same procedures as used for indexing. General and field-specific stop words and digits were removed and stemming was applied to the remaining words. In addition, all short terms consists of less than three letters, were also removed, which we noticed to be mostly symbols or letters referring to figures or equations in the patent document that do not relate to the topic of the patent and can act as noise in the query.

Since the topics were granted patents that all contained translations of the title, the abstract, and the claims, we used only the English parts to formulate the query for consistency with the

indexing phase. This meant that all non-English patent topics did not include the description section in the submitted query. Two sets of query formulation experiments were conducted. The first set focused on using the short text fields to create the query from the patent topic. These short text fields were: “Title”, “Abstract”, “Desc1” (first line in description), “Claim_main” (first sentence in first claim), “Claim1” (first claim), and “Claims” (full list of claims). The second set of experiments tested using the full patent description section as the query. The objective of both sets of experiments was to assess the most valuable parts for use in the retrieval process, and to examine the possibility of minimizing the amount of query text by using only the short fields with the aim of minimizing the retrieval time while maintaining retrieval quality. The second set of experiments was performed only for the English patent topics, since non-English patents did not include translation of the description section.

The 10,000 patent topics provided by the track were formulated in the Indri query language for search, and the Indri retrieval model was used for retrieval experiments.

- **Citation extraction**

We also performed an additional experiment after the track submissions based on the finding of the benefits of using the extracted citations from the patent topic description section reported in (Lopez and Romary, 2009). A simple regular expression Perl script was used to extract citations from the description sections of the topics.

For the 10,000 test patent topics provided in CLEF-IP 2009, 36,742 patent citations were extracted from the patent topics, but only 11,834 patents citations were to patents existing in the patent collection. The 11,834 patent citations were extracted from 5,873 patent topics, leaving 4,127 topics with no citations extracted from them. Only 6,301 citations were found to be relevant. Therefore, when considering only these extracted citations as a results list to the topic set, the MAP value was 0.182, and the recall was 0.2. This means that 20% of the relevant

documents could be extracted from the description text of the patent topics without performing any retrieval. We tested appending the retrieval results to this list of extracted citations after removing the duplicates, since this list of extracted citations is very short compared to the retrieval results list, and it had a higher precision than the retrieval run.

4.1.2 Submitted runs and results for CLEF-IP 2009

We submitted three runs to the CLEF-IP 2009 track. We submitted what we believed to be the best runs based on the results of many pilot runs carried out on the training topics set provided in CLEF-IP 2009, which contained 500 topics. In the pilot runs, we found that some query formulation methods are ineffective compared to others. For example, we found that ignoring the structure of patent documents and topics is better than searching each field of patents with the corresponding section in the query. The reason for this finding is analysed in Section 4.1.3. We also tested several combinations of short sections of patents to construct the query. We submitted the combination which achieved the best result with the training topics for the CLEF-IP 2009 task.

Based on our experiments with the training topics, we applied the following setup to the three submitted runs in CLEF-IP 2009:

1. The patent documents were treated as a full document, neglecting their structure.
2. Stop word filtering and Porter stemming were applied to documents and topics.
3. Queries were formed from short fields by giving weights to the terms of each field as follows: 5×Title + 1×Abstract (English topics only) + 3×Desc1 (English topics only) + 2×Claim_main + 1×Claims.
4. Term bi-grams with a frequency in the text higher than one were used in query. The text of the fields: “Title”, “Abstract”, “Desc1”, and “Claim_main” was used for extracting the

term bi-grams. The inference network implemented in Indri allows search for phrases without the need of building a phrase index, see Appendix B.

The three submitted runs were:

- Run 1: No filtering was performed to the results.
- Run 2: Filtering was performed for retrieval results based on the top three levels of the IPC classification. At least one common class needs to exist between topic and retrieved documents.
- Run 3: The same as 2nd run, but with removing all query terms consisting of less than three letters.

Table 4-1: Results of runs submitted to CLEF-IP 2009 measured by recall and MAP

Run #	Recall	MAP
Run 1	0.544	0.097
Run 2	0.624	0.107
Run 3	0.627	0.107

Table 4-1 reports the results of our three submitted runs to CLEF-IP 2009. As shown in table, filtering the results by IPC classes improves the results significantly for both MAP and recall ("Run 2" compared to "Run 1"). Significance test was performed using Wilcoxon statistical significance with a confidence level of 95% (Hull, 1993). "Run 3" is statistically indistinguishable from "Run 2". However, filtering the short terms speeds up the retrieval process, which leads to a more efficient retrieval system.

Our best submission, "Run 3", achieved the seventh and the fourth ranks when compared by MAP and recall, respectively.

In addition to the submitted runs in CLEF-IP 2009, we performed two experiments subsequent to our official submission as mentioned earlier by examining the use of full patent

description for constructing the query, and appending the extracted citations to the top of the results list. These results were reported in our final version of the paper on CLEF-IP 2009 (Magdy et al., 2009), and are shown in Table 4-2.

Table 4-2: Recall and MAP for the two query formulation techniques tested with and without adding extracted citations

	No Citations		With Extracted Citations	
	Recall	**MAP**	**Recall**	**MAP**
Short Fields (Run 3)	0.627	0.107	0.660	0.200
Description	0.627	0.119	0.668	0.209

Table 4-2 reports the results for the best run of using short fields for query formulation ("Run 3") vs. using the patent description section. The results are reported with and without applying citation extraction from the topic description field. From Table 4-2, it can be seen that using the description text for searching was on average 11% better than using the best combination of the short fields from the perspective of precision, which was statistically significantly better than using short fields. Furthermore, combining the results with the extracted citations from the text led to a huge improvement in the MAP, where these citations were taken as the top ranked results in the final list and the results from the searching then appended below them after filtering out duplicates. Our best result in Table 4-2 would have achieved the best second run among all submitted runs in the track when compared by MAP and recall.

The citation extraction in patent search is further studied in this thesis in Section 4.3, by comparing our simple method of appending the retrieval results to the extracted citations to the more advanced approach described by Lopez and Romary (2009; 2010).

After obtaining the results shown in Table 4-2, we carried out a quick experiment to examine the combination of the description section and the short fields of the patent topic for formulating the query. The result achieved was found to be statistically indistinguishable from using only the

description section for search when compared by both MAP and recall. However, the retrieval process was slower when using all the sections of the patent topic in search, which gives an advantage to using the description section alone for formulating the query.

4.1.3 Analysing the challenge of the task

We conducted an analysis to understand the reasons behind the low score values achieved in patent search. Since IR mainly relies on word matching between topics and documents for search, we measured the degree of word matching between topics and relevant documents compared to the word matching between topics and non-relevant documents. The standard assumption is that documents relevant to a topic tend to have relatively higher word match with the topic than non-relevant ones (Robertson and Jones, 1994). We examine the validity of this assumption in this section.

For the analysis, our patent search results for the training topic set for CLEF-IP 2009 were analysed to try to better understand the reasons behind the low retrieval effectiveness for the patent retrieval task. In order to analyse this problem, the overlap between the short fields of each topic in the training data and its relevant patent set was computed. We did not include the description section for analysis, since it is very long compared to the other sections, and the possibility of finding many overlapping terms is high. We further checked the overlap between these fields of the topics and the corresponding ones of the top five ranked non-relevant documents for each query. The reason behind selecting top five is that the average number of relevant documents for the topics in 2009 was between 5 and 6. The overlap was measured using two measures: 1) the cosine between term vector of each corresponding fields of the two compared patents; and 2) the percentage of zero overlap (no shared terms) between

corresponding fields of the two compared patents. The matching was applied on the pre-processed version of the text, where stop word removal and stemming had been applied.

Figure 4.2 presents the cosine similarity between topic and relevant and non-relevant documents, and Figure 4.3 presents the percentage of corresponding sections in topic and relevant documents which have zero overlap between their terms.

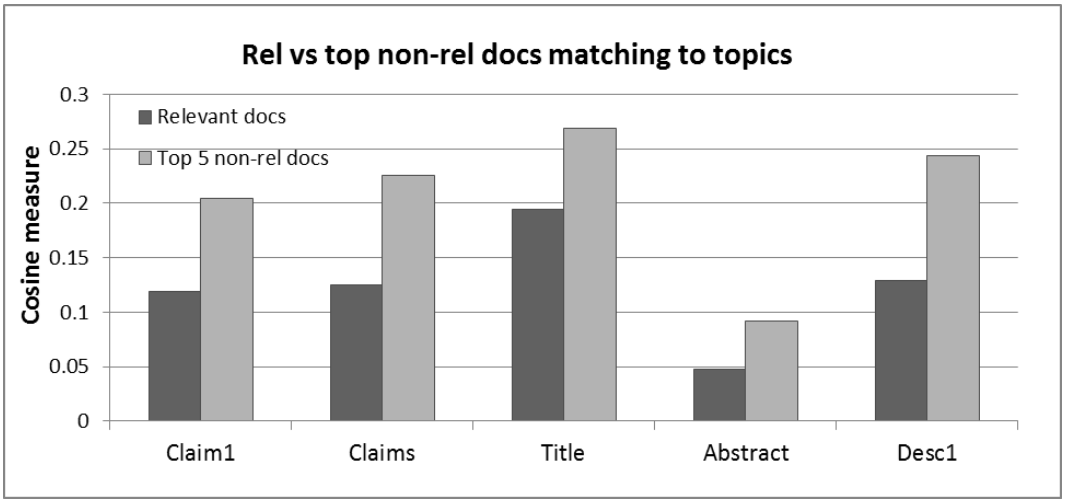


Figure 4.2: Cosine measure between fields of topics and the corresponding ones in the relevant and top 5 retrieved documents

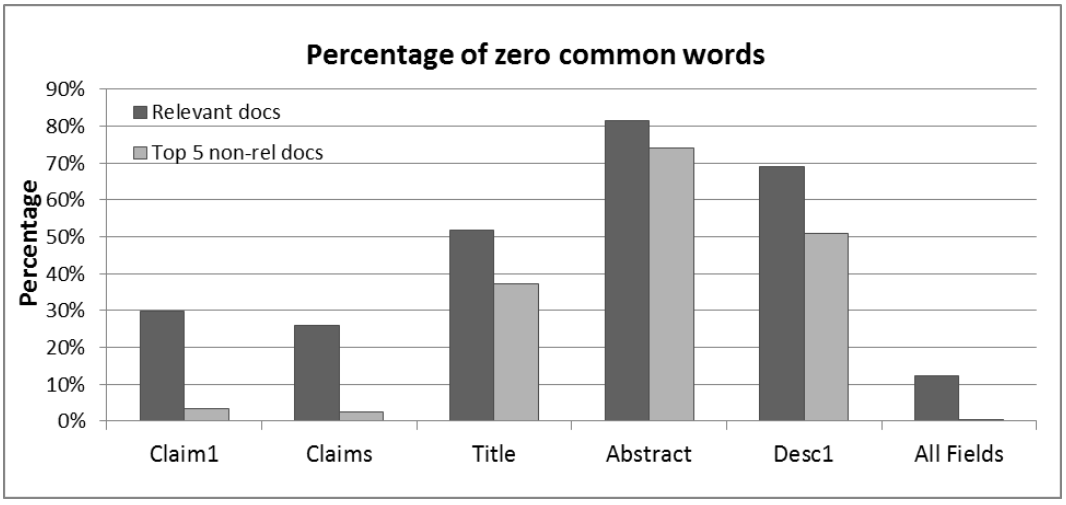


Figure 4.3: Percentage of fields with zero common (shared) words between those of the topics and the corresponding ones in the relevant and top 5 retrieved documents

Figure 4.2 shows that the cosine between the top ranked non-relevant documents and the topic is nearly twice as high as that for the relevant documents for all fields. The same pattern is shown in Figure 4.3, where surprisingly, 12% of the relevant documents for topics have no shared words in any of the short fields of the topics. Both figures show that the word matching between topics and their relevant documents is very low compared to other non-relevant documents. This observation highlights one feature in patent search, where the relevance between topics and documents depends on idea matching rather than word matching. This feature was explicitly stated by an expert patent practitioner in (Adams, 2011), where he noted that very different words can be used to describe very similar inventions.

Our analysis in this section also explains the reason behind our previous finding that using each topic section to search the corresponding section in the patent documents achieves lower retrieval effectiveness compared to applying unstructured search. This can be interpreted that allowing terms from different sections of a topic to search the full patent documents has a higher chance of finding matching terms in sections which do not have to correspond to the section where the terms came from.

4.2 Our Participation in CLEF-IP 2010

4.2.1 Retrieval system configuration

Based on our participation in 2009 and the findings of our subsequent analysis, in CLEF-IP 2010 we adopted a simple and straightforward retrieval approach for the patent retrieval task (Magdy and Jones, 2010c). The same pre-processing as used in 2009 was applied to the documents and the topics. The filtering step based on IPC classification was also applied, and citation extraction was performed. We were able to extract 2,307 citations from 771 patent topics, out of 2,000

topics provided for the track. The patent retrieval system used in CLEF-IP 2010 is presented in Figure 4.4.

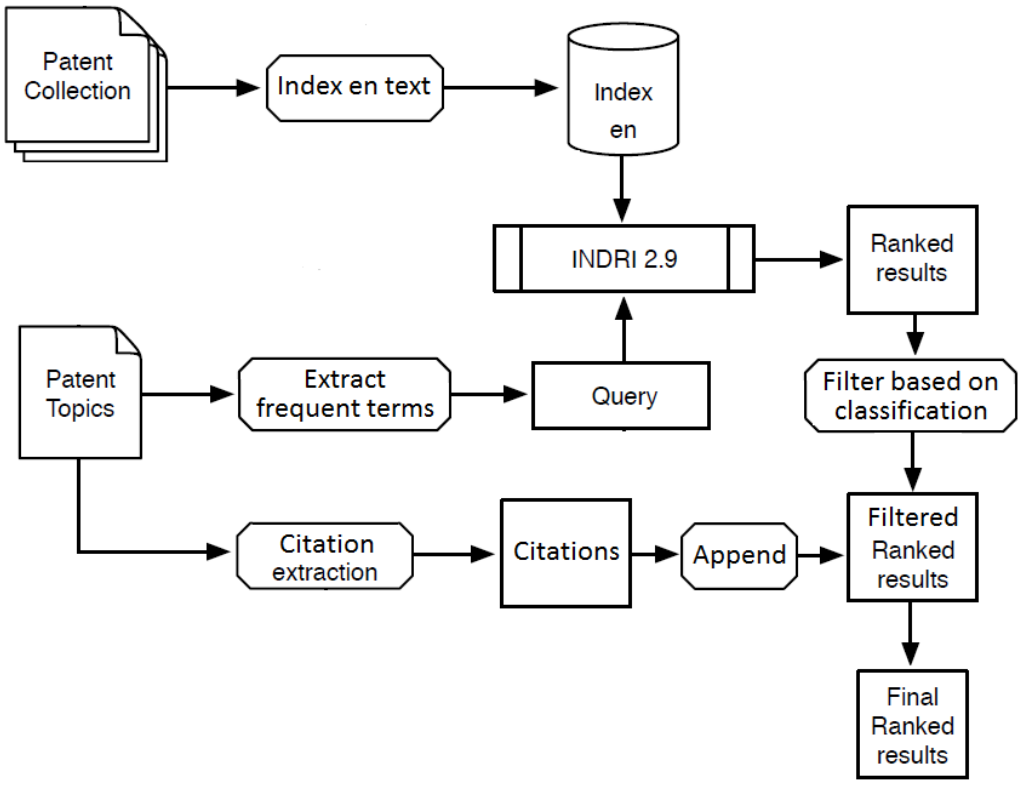


Figure 4.4: DCU Retrieval system used for the patent search task in CLEF-IP 2010

Since CLEF-IP 2010 used patent applications as the topics, there were no English translations for the non-English patents. Thus we used Google translate to translate the full patent applications into English, and then used the translation to formulate the queries. Therefore, the patent topic used for search was always in English. Queries were formulated from the description section of the patent application, since we found that this achieved the best result in our experiments for the CLEF-IP 2009 task. In addition, we filtered out all the terms in the description section which appeared only once. The effect of these infrequent terms was tested and found to be insignificant on the retrieval effectiveness, but their inclusion doubles the speed

of the retrieval. In addition to the terms from the description section, bigrams that appeared more than three times in the remaining sections were added to the query, which was found to improve the retrieval effectiveness.

The configuration of our patent retrieval system in CLEF-IP 2010 shown in Figure 4.4 can be described as follows:

1. English content of the patent collection was indexed.
2. Pre-processing of topics and documents text was performed by applying Porter stemming and filtering out: standard and field-specific stop words, digits, and short terms (one or two letters).
3. Non-English topic patents were translated into English using Google translate.
4. Queries were formulated from the description section, and terms that only appeared once in the description section were removed to speed up the retrieval process.
5. Terms bigrams appearing more than three times in the remaining sections (title, abstract, and claims) combined were added to the query.
6. Filtering was applied based on the top three levels of the IPC classification.
7. Citations were extracted from the description section of the patent topic and used in the top of the retrieval results. Retrieval results were appended to this after removing any duplicates.

4.2.2 Submitted runs and results

We submitted three formal runs to the CLEF-IP 2010 track for the 2,000 topics provided by the track; of which 520 were German, 134 were French, and the rest were English (Piroi, 2010). Our results are shown in Table 4-3, where “IR” represents the results for the pure retrieval run without using the extracted citations; “Cit” represents the list of extracted citations without any

retrieval results; and "IR+Cit" represents the final results after appending the retrieval results to the extracted citations after removing any duplicates.

Table 4-3: MAP, recall for the three submitted runs in CLEF-IP 2010

Run #	MAP	Recall
IR	0.1216	0.5700
Cit	0.1120	0.1187
IR+Cit	0.2029	0.6180

Table 4-3 shows results for our three submitted runs in CLEF-IP 2010. The table shows two extreme runs, namely the "IR" run and the "Cit" run. The "IR" run achieved high recall and moderate precision. On the other hand, the "Cit" run achieved low MAP and very low recall. Although the MAP of the "Cit" run is the lowest in Table 4-3, the "Cit" run has a very high precision, since this score is the average for the full topics set, while only 771 topics out of the 2000 had citations extracted. This means that the MAP for these topics alone is 0.3. The last run "IR+Cit" achieved the highest recall and precision since it comes from a simple merging of the two previous runs.

Table 4-4 presents the same results in Table 4-3, but for each language separately. The results are reported for our three runs for the English, French, and German topics.

Table 4-4: MAP, recall for the three submitted runs in CLEF-IP 2010 reported for English, French, and German topics separately

	English topics		French topics		German topics	
Run #	**MAP**	**Recall**	**MAP**	**Recall**	**MAP**	**Recall**
IR	0.1399	0.5886	0.0870	0.5168	0.0841	0.5354
Cit	0.0594	0.0658	0.2827	0.2941	0.2048	0.2110
IR+Cit	0.1825	0.6160	0.2981	0.6426	0.2360	0.6170

Table 4-4 shows very different performance for the English vs. the non-English topics. It is clear that the retrieval system for the English topics was better than the French and German topics. However, the citation extraction was much more effective in extracting citations referring to relevant documents from the description section of the French and German patent topics compared to the English ones. The explanation for the first observation that the retrieval for English topics was better than French and German may be from the translation process applied to the non-English topics, which is expected not to be perfect. However, this large difference in the MAP values between both sets of topics is not found in the recall values, which means that the top ranked documents of the English topics have higher precision than the non-English ones, but both sets of topics have comparable numbers of retrieved relevant documents. For the citation extraction for both sets of topics, one possible explanation is that the non-English topics tend to cite more relevant patents in their description section than the English topics. This is a possible explanation for the results, since the citation extraction technique we used is language independent and cannot be affected by the language of the topic. For the run where retrieval and extracted citations were merged, the results for all languages are very similar when compared by recall. However, the high precision of the "Cit" run still affected the MAP values of the non-English topics which are higher than the English topics.

Our "IR+Cit" run was the second best run among 25 submitted ones. The best run was from the same participant who introduced the method presented in Figure 3.3 (Lopez and Romary, 2009; Lopez and Romary, 2010). Our run and the best run were the only two ones to utilize the extracted citations, but this was done in very two different ways for the two runs. The next section compares the approaches of these two runs in more detail to show the relative benefits of each of them for the prior-art patent search task.

4.3 Simple vs. Sophisticated Approaches for Patent Search

When available, extracted citations have been demonstrated to have a very beneficial effect on retrieval effectiveness for patent search. For the CLEF-IP 2010 task two different approaches achieved good results: our approach which represents a simple and straightforward retrieval method ("IR+Cit" run in CLEF-IP 2010), and a more sophisticated approach (Lopez and Romary, 2010). Lopez and Romary (2010) used external resources to enrich the extracted citations. This had the effect of nearly tripling the number of the citations. In order to compare these approaches in a fair way, we performed an analysis of both methods after using the same set of citations with both techniques (Magdy et al., 2011). The purpose of this study was to see if the simple approach can have comparable retrieval results to the much more sophisticated one for patent search when similar resources are available to both systems.

4.3.1 Advanced citation extraction methodology

For a fair comparison of the two retrieval systems, the citation extraction algorithm used in (Lopez and Romary, 2010) was applied to extract the set of citations.

The citation extraction algorithm used in (Lopez and Romary, 2010) comprises four steps:

1. **Identification of reference strings**: The text of the patent description section is first extracted and reference blocks (citation blocks) are identified in the text body by a specific Linear-Chain conditional random field model. An example reference block is "JP5-16281" shown in Figure 2.3 on page 20.
2. **Parsing and normalization of the extracted reference strings**: The reference text is then parsed and normalized in order to obtain a set of bibliographical attributes in a fixed format. References to patents are parsed and normalized in one step by a Finite State Transducer (FST) which identifies (i) if the patent referred to is a patent application or a patent

publication, e.g. “A” or “B” version, (ii) a country code, e.g. “EP” or “JP”, (iii) a number, and (iv) a kind code, e.g. “A1”, “A2”, or “B1”.

3. **Consolidation with online bibliographical services**: Different online bibliographical services are then accessed to validate and to enrich the identified reference. For patent references, OPS[16] (Open Patent Service1) is utilized, which is a web service provided by the EPO for accessing the Espacenet patent databases. This step allows the retrieval of the patent numbers from a reference to a patent application number.
4. **Family lookup**: For the citations extracted from the patent topics, in cases where the citation is a non-European patent, the OPS is accessed for *patent family* information in an attempt to identify a corresponding European patent.

The first two steps of this technique correspond to our simple Perl script for citation extraction. However, this technique is much sophisticated in the parsing part since it covers more patent reference patterns including non-EPO patents. Nevertheless, the retrieval effectiveness achieved using this technique for extracting the EPO patent references was not much better than ours. Their extraction technique identified 2,946 citations from the 2,000 topics, while ours extracted 2,307 citations. The most effective addition of their technique was the identification of the patterns of patent references from outside the EPO and the steps 3 and 4, which used the online services for *patent family* lookup to find the corresponding patents in the EPO. These two steps increased the number of the citations that exist in the patent collection from 2,946 to 7,706.

Patent family lookup operates as follows: we noted in section 2.2.4 that the typical scenario when filing a patent is to submit it into multiple patent offices in different languages. The patent applications to different patent offices corresponding to one invention are called the *patent*

[16] http://www.epo.org/searching/free/ops.html

family. When referencing an invention in a new patent application, usually only one of the members in the *patent family* is included. Steps 3 and 4 in this algorithm seek to find a corresponding EPO patent for extracted citations that refer to patents from external patent offices. If a corresponding patent in the EPO is found in the *patent family*, then it is added to the extracted citations list.

The extracted set of citations is simply appended in the beginning of the retrieval results list in our approach after removing any duplicates, see Figure 4.4. However, in Lopez and Romary's approach, this list is used as part of what they refer to as the "working set" which also includes the patents of common citations, inventors, and classifications. Instead of searching the full patent collection, they search the "working set" for relevant documents several times using multiple retrieval models and indexes, see Figure 3.3.

4.3.2 Experimental setup

The extracted citations were used for both the algorithms shown in Figure 3.3 and Figure 4.4 for the CLEF-IP 2010 patent search task. The English topic set of 1,348 topics from CLEF-IP 2010 was used for the experiments. As noted in the last section, applying the citation extraction algorithm identified 7,706 citations. These were actually found in the description sections of only 728 patent topics, while the remaining 620 topics did not yield any extracted citations.

For the analysis, we divided the topics into these two subsets: topics which have extracted citations and topics which do not have citations. The objective was to compare our simple approach to Lopez and Romary's (2010) more advanced one in the situations when citations could be or could not be extracted from the patent topics. The MAP score was used to compare the retrieval effectiveness of the results. A Wilcoxon significance test with *p-value* of 0.05 was used to test the significance of the results (Hull, 1993).

4.3.3 Results

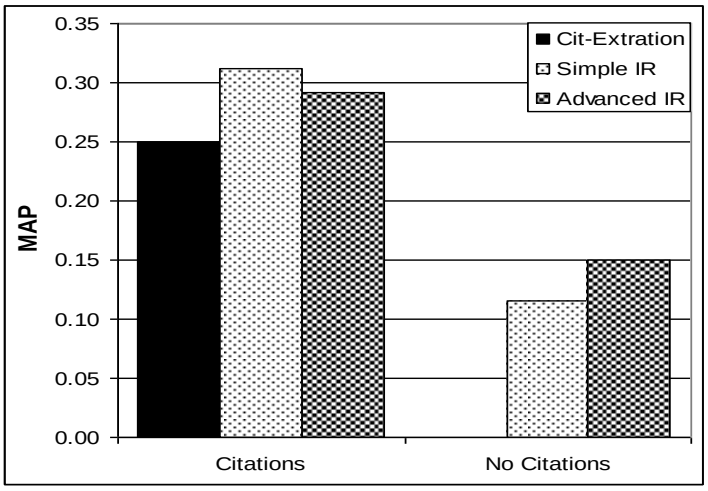


Figure 4.5: Retrieval results for simple and complex IR approaches with CLEF-IP prior-art search task, when citations could be and could not be extracted

Figure 4.5 shows the retrieval effectiveness of the simple and advanced IR techniques for the patent prior-art search task measured by MAP. Results are reported for the two topic sets when citations could and could not be extracted. The retrieval effectiveness when only extracted citations are used without any kind of IR is reported as a baseline. From Figure 4.5, it can be seen that the extracted citations achieve higher retrieval effectiveness than either IR approach when no citations could be extracted. Also it is clear that both systems achieved nearly double the retrieval effectiveness for topics that contain patent citations within their text. Comparing both systems on their performance, it was found that when citations exist, the simple IR approach was statistically significantly better than the sophisticated approach. However, when no citations could be extracted as an initial step, the complex approach was significantly better. This observation leads to the conclusion that when a patent application includes directly cited prior-art patents, simple search approaches are sufficient to achieve good retrieval results.

To further support this finding, the result list of the complex approach was appended to the extracted citations list after removing duplicates in the same way as the final step of the simple approach. This approach gave a MAP of 0.300 which is still lower than the results achieved by the simple approach.

4.3.4 Conclusion

The aim of this study was to compare two approaches to the patent prior-art search task. The first approach is characterized by its simplicity and low resource requirements, while the second one is more sophisticated, using an advanced level of content analysis. The experimental results show that the simple search approach is statistically better than the sophisticated one when initial citations are provided. This is the situation for 54% of the test collection used in our experiments. The observation that simple IR approaches can in many cases achieve similar results to sophisticated ones suggests that the additional linguistic processing does not add value when initial citations are present in the patent application. This finding can indicate that research in patent search still requires further investigation to better understand the retrieval process in patent prior-art search in order to develop more effective and efficient methods for this task.

4.4 Exploring Query Expansion Techniques for Patent Search

While patent search queries are very long, it was demonstrated that there is insufficient overlap between most of the corresponding sections of patent topics and their relevant documents in Figure 4.2 and Figure 4.3. We reported in Section 3.5.2 that there have been several attempts to improve the overlap between patent queries and relevant documents through applying query expansion (QE) techniques. We showed that none of these techniques succeeded in achieving a significant overall improvement in patent search over baseline results.

In this section we investigate QE techniques experimentally and demonstrate the effect of a standard QE technique on the prior-art patent search task for the CLEF-IP 2010. Furthermore, we introduce a novel QE technique inspired by the structure and content of the EPO patents that contain parallel translations. These are used to create a set of synonyms of terms (SynSet) which we use later as an expansion resource for the terms in the query. This technique shows significant improvement to the patent retrieval results when compared using MAP.

The same 1,348 English topics from the CLEF-IP 2010 used in the previous section were used for the experiments. However, we did not apply citation extraction since the focus is on the retrieval technique itself, i.e. the "IR" run in CLEF-IP 2010 represents our baseline, see Table 4-4. The MAP for the baseline run for the 1,348 English topics is 0.1399 (Magdy and Jones, 2010c; Piroi, 2010). The formulated queries of the baseline were expanded using different QE techniques in an attempt to improve the retrieval effectiveness.

4.4.1 Examining standard PRF in prior-art patent search

Although the methods explored in previous work, reviewed in Section 3.5.2, on pseudo relevance feedback (PRF) for patent search were not effective for improving retrieval effectiveness (Kishida, 2003; Itoh, 2004; Takeuchi et al., 2005), we apply PRF to the CLEF-IP 2010 task to see if this finding can be replicated for this task and data using our system. This is important since the reported results are for a patent invalidity search task, not the prior-art search task investigated here.

The PRF implemented in the Indri search toolkit (Strohman et al., 2004) was used in our experiments. In this investigation, different numbers of documents and terms for the feedback process were examined. The feedback mechanism used in Indri is an adaption of Lavrenko's relevance model (Lavrenko and Croft, 2001). The default weighting between the original query

and expansion terms in Indri is 1:1, which was used in our experiments, where the final retrieval score is calculated as the geometrical mean of the original query retrieval score and the expanded terms retrieval score[17]. The numbers of documents that we tested for the feedback process were {1, 5, 10, 20}, and the number of terms tested were {10, 20, 30, 50}, which are the standard numbers of terms and documents used in the feedback using Indri (Strohman et al., 2004).

Table 4-5 reports the results of applying PRF to the CLEF-IP 2010 patent search task. The results in Table 4-5 show that the best PRF run led to a significant degradation in retrieval effectiveness compared to the baseline. A possible explanation for this result is that the initial performance of the baseline is relatively low, which means that the top ranked documents used for PRF are mostly non relevant, meaning that QE is likely to add noise terms leading to degradation in the retrieval effectiveness.

Table 4-5: The effect of using PRF with different number of terms and documents on retrieval effectiveness measured by MAP

	Terms / Docs	10	20	30	50
MAP BL = 0.1399	**1**	0.046	0.065	0.076	0.082
	5	0.037	0.053	0.062	0.072
	10	0.031	0.046	0.053	0.061
	20	0.026	0.036	0.042	0.049

Table 4-5 shows that the results of PRF are better when fewer documents are used for retrieval, and when more feedback terms are added to the original query. Therefore, we examined a more extreme case of the number documents and terms used in feedback. We examined using only one document and 300 terms for the PRF. We used 300 terms since this is

[17] http://ciir.cs.umass.edu/~metzler/indriretmodel.html#prf

the average number of unique terms for a patent topic, meaning that the query will be expanded with a similar number of terms to the original query. The result for this run was 0.117 MAP, which is still significantly lower than the baseline.

Our results in this section align with previously reported results in (Kishida, 2003; Itoh, 2004; Takeuchi et al., 2005). Our attempt and the previous attempts by other researchers failed to improve the retrieval effectiveness for patent search through PRF. However, we cannot generalize it to conclude that PRF or QE in general cannot be effective for patent search, since this can be only for the approaches examined, but there can be different setup for PRF to improve the retrieval effectiveness. In the next section, we introduce a new method for applying QE to patent queries, which leads to a significant improvement in the retrieval effectiveness for patent search.

4.4.2 Query expansion using an automatically generated SynSet

Here a novel method of QE is proposed based on an automatically generated set of synonyms and related words. The idea for automatically generating the synonym set (SynSet) originates from the characteristics of the CLEF-IP patent collection, where some sections in the patents are translated into three languages: English, French, and German. The idea is to use these parallel manual translations to create possible synonyms sets. Although the idea is based on the presence of this data, this approach is potentially applicable to other IR applications where parallel multilingual corpora of a domain close to the data collection are available.

- **Approach**

Related work on automatically building a SynSet from a word-to-word translation model is described in (Wang and Oard, 2006), where automatically generated synonyms are used in conjunction with WordNet and translation models to enhance CLIR. In our approach, a word-to-

word translation model is used to create a SynSet for QE in monolingual search. For a word in one language f which has possible translations to a set of words in another language $\{e_1, e_2 \dots e_n\}$, this set of words can be considered as synonyms or at least words related to each other, since they come as translations of the same word. The probability of e_1 being a synonym of word e_2 can be computed as shown in Equation (4-1).

$$P(e_1|e_2) = \sum_{i=1}^{n} P(f_i|e_2) \cdot P(e_1|f_i) \quad (4\text{-}1)$$

where:

$p(e_1|e_2)$ is the probability that e_1 is a synonym of e_2, $\{f_1, f_2 \dots f_n\}$ are possible translations for word e_2

$p(f_i|e_2)$ is the probability that f_i is a translation of e_2

$p(e_1|f_i)$ is the probability that e_1 is a translation of f_i

- **Extracting parallel corpus**

Here we describe how the parallel corpus was extracted from the EPO patents in order to use it for generating the SynSets. Although our current experimentation required the extraction of a parallel corpus in only two languages, we aligned and extracted parallel sentences in all three languages, since they are available. These extracted parallel corpora are also used for the experiments on multilingual patent search described later in Chapter 6.

Almost all the patents in the collection contain the title in the three languages, since these translations should be provided to the EPO from the first patent application (the A1 version of the patent). However, only some of the patents in the collection contain the claims in all three languages, since the claims translations is only provided in the granted version (the B versions of the patent). It was easy to align the title section parallel translation, since it is only a short field in

the patent XML with usually only a small number of words. Identifying the parallel sentences in the claims section was a more difficult task, since the claims section is formed from several separate claims, each one of which is formed from smaller parts (called "claim-text" in the patent XML file), see Figure 4.6 below and Figure 2.3 on page 20.

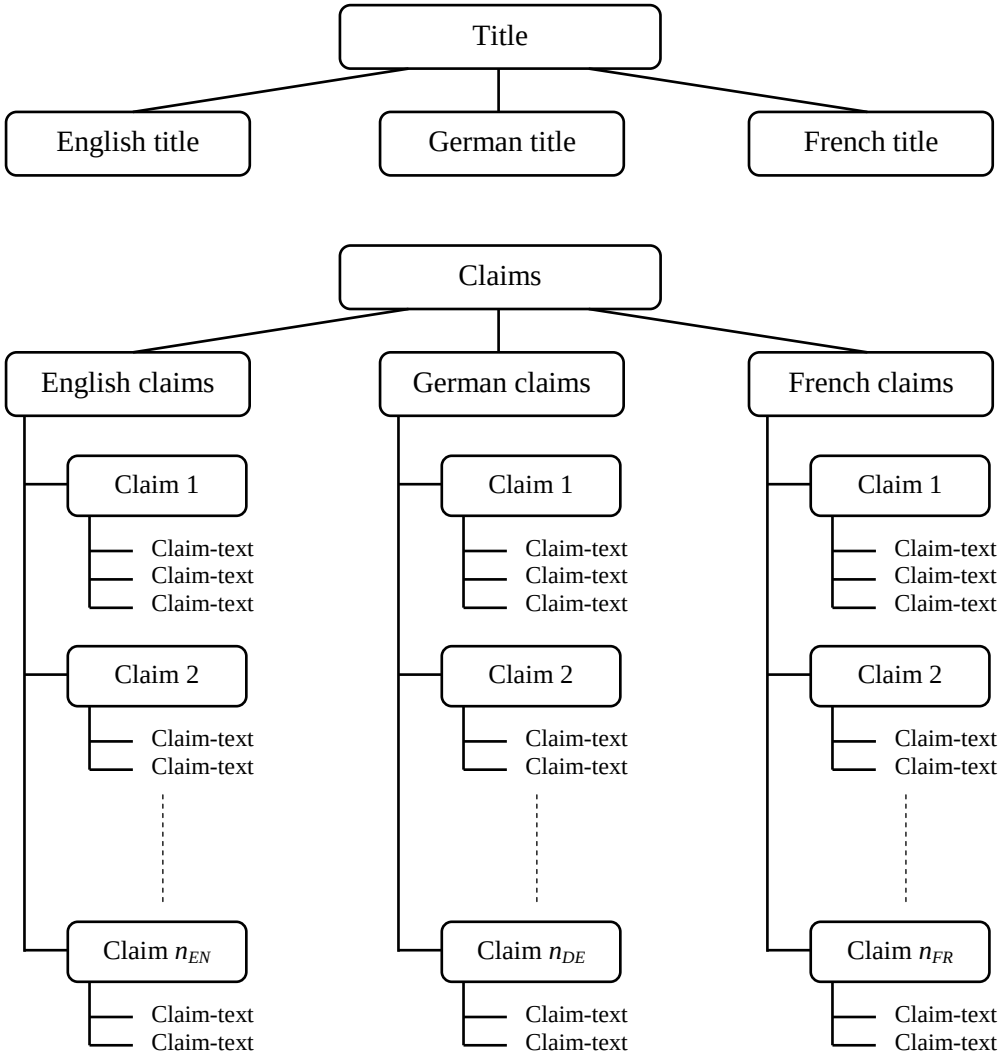


Figure 4.6: Structure of patent sections which contain parallel translations

Sometimes the number of claims does not match for the three languages, and sometimes, even if the number of claims matches, the number of sentences ("claim-text") for a given claim

does not match for all three languages due to rephrasing in translations. To overcome these problems, only the parts of the claims section which have the same number of claims and claims parts were used to create the parallel corpus, since this is an indication that they are parallel. Any other parts where the number of claims or claim parts does not match were ignored.

We applied several additional processing steps to the sentences, including:

1. Deleting parallel sentences in the three languages if any two sentences have significant difference in their length, where the maximum allowed ratio between the lengths of two parallel sentences was 1:9 as recommended in (Stroppa and Way, 2006).
2. Applying text pre-processing to the text of the three languages including stemming and filtering out of punctuation, digits, and stop words, since stop words and the surface form of words have insignificant value during the retrieval process.
3. Deleting very long sentences that contain more than 60 terms, since long sentences require a long time to align at the word level. These long sentences represent less than 5% of the extracted corpus.

For the stemming and stop word removal in step 2 above, the Snowball[18] toolkit was used for applying the stemming to the English, French, and German text. Snowball contains state-of-the-art implementations of stemming algorithms for many European languages (e.g. Porter stemming for English). The stop word lists[19] used for each of the three languages contained: 571, 463, and 603 stop words for English, French, and German languages respectively.

A set of more than 8.15M parallel sentences in the three languages was extracted from the CLEF-IP 2010 patent collection (referred to later as 8M corpus). 1.33M sentences were extracted

[18] http://snowball.tartarus.org/

[19] http://members.unine.ch/jacques.savoy/clef/index.html

from the title section of the patents and the remainder from the claims section. The average length of the English sentences after stop word removal was 14 words.

- **Generating SynSets**

The automatically generated 8M sentence parallel corpus was aligned using Giza++, which uses IBM model-4 for word alignment (Och and Ney, 2003). The output of Giza++ is a translation dictionary where a word in one language is associated with possible translations in the other language with translation probabilities that sum to one. For this work, we used only the English and French translations, since we needed only one language aligned with the English to create English to English synonyms. We also preferred using French over German since German text contained many word compounds that can lead to inaccurate 1:1 word alignment. Giza++ was used to create two translation dictionaries: English to French and French to English. Equation (4-1) was applied to both generated dictionaries in order to create an English to English mapping with weights (probabilities). This is referred to from now on as the SynSet.

The SynSet contains a set of synonyms and related terms, including the original term itself. By checking the generated SynSet, we found it to be subjectively reasonable, although it contained some noisy terms but generally these have low probabilities. In order to reduce the presence of such noisy synonyms, pruning was applied by removing all terms with low probabilities. We selected a pruning threshold probability of 0.1 based on our observation of the generated SynSet. The probabilities of removed terms were added to the probability of the original term based on Equation (4-2).

$$P(e_x|e_x)|_{pruned} = P(e_x|e_x)|_{original} + \sum_{\forall P(e_i|e_x)<0.1} P(e_i|e_x) \tag{4-2}$$

Applying Equation (4-2) led to many terms not having any synonyms other than themselves (i.e. $p(e_x|e_x) = 1$). A further pruning step was applied which removed SynSet entries for all terms that appeared less than 20 times in the 8M sentence training set. This pruning step was applied since we noticed that these terms did not have sufficient training instances to produce reliable SynSet entries. Some samples of the produced SynSets after applying the pruning stages are shown in Table 4-6 (note that terms are in their stemmed form).

Table 4-6: Example of SynSet entries. The probability of each synonym is in brackets

Term	SynSet
motor	motor (0.63), engin (0.37)
weight	weight (0.86), wt (0.14)
travel	travel (0.67), move (0.19), displac (0.14)
color	color (0.56), colour (0.25), dye (0.19)
link	link (0.4), connect (0.18), bond (0.17), crosslink (0.13), bind (0.12)
cloth	fabric (0.36), cloth (0.3), garment (0.2), tissu (0.14)
tube	tube (0.88), pipe (0.12)
area	area (0.4), zone (0.23), region (0.2), surfac (0.17)

The resulting SynSets, as shown in Table 4-6, contain abbreviations (e.g. *wt*), spelling variation (e.g. *colour*), and related terms that can even have higher weight than the original term (e.g. *cloth*/*fabric*). Using these terms for QE aims to help overcome the term mismatch problem between queries and documents.

4.4.3 SynSet QE impact on retrieval effectiveness

The generated SynSet was used to expand the 1,348 English queries of patents from the CLEF-IP 2010 task.

In order to test the effect of using the automatically generated SynSet on the retrieval effectiveness when used in patent search QE, two experiments were conducted. The first one used the probability associated with the SynSet entries as a weight for each expanded term in the query (Wsynset). Therefore, each term was replaced with its SynSet entries, which contain the original term plus additional related terms, with the probability of each term in the SynSet acting as a weight to the term within the query. The second experiment ignored this associated probability and used uniform weighting for all synonyms of a given term (Usynset). The Indri "wsyn" and "syn" operators were used to allow the presence of weighted synonyms and uniformly weighted synonyms in the patent query; see Appendix B. Table 4-7 reports the retrieval results.

Table 4-7: Effect of using the automatically generated SynSet for QE on the retrieval effectiveness for the English topics in CLEF-IP 2010 measured by MAP

	value	% change
Baseline	0.1399	NA
Wsynset	0.1440	+2.9%
Usynset	0.1402	+0.2%

The results in Table 4-7 show some improvement over the baseline. The results achieved when using the weighted SynSet method (Wsynset) were found to be statistically better than the baseline when compared using MAP. However, when checking the result per topic, it was found that only 39% of the topics were improved, 35% degraded, and the rest almost unchanged (change of MAP<5%). This shows that the algorithm is useful for some topics, but harmful for

others. Nevertheless, this result is considered positive when compared to the state-of-the-art in QE for patent search, since, to the best of our knowledge, no significant improvement has been reported for the prior-art patent search task using QE in previous work.

Another advantage of using SynSet QE over PRF is the retrieval speed. It was found that the query length after expansion is increased by only 60% and the retrieval speed is increased by the same ratio. For PRF, the retrieval is performed twice: once to retrieve the top relevant documents to be used for extracting the expansion terms, and again after adding these expansion terms to the query. In addition, the term selection process itself requires a large amount of time for patents since the documents are large. This makes the retrieval time in PRF for patent search significantly greater than the baseline and makes SynSet method a more efficient QE approach.

4.4.4 Conclusion

In this section, we introduced a novel algorithm for QE that showed a significant improvement to patent search effectiveness when measured by MAP. The presented QE technique utilized one of the special features of the EPO patents which is the presence of parallel sentences in the patent documents. These parallel sentences were processed in an automatic way to generate a list of related words, which we named a SynSet. Despite the significance of the improved retrieval effectiveness over the baseline, we found that not all the topics were improved by the SynSet QE technique. We did not further investigate the instances of success or failure for two reasons: 1) the overall result of the technique was significantly better than any reported QE technique; 2) this is not the main focus of our study in this thesis, which is rather an attempt to further understand the challenges of the patent search in order to be able to build more effective tools that can help in improving the retrieval effectiveness for patent search in general not for the CLEF-IP task in specific.

Another use of Synsets that can help in facilitating the search job in the patent office is exploiting them as a lexical resource for direct use by patent examiners to suggest possible related terms when formulating the search queries manually.

4.5 Summary

This chapter has presented our contributions to the CLEF-IP prior-art patent search task, which is the core task investigated in this thesis as a representative of recall-oriented tasks. The main objective of this chapter was to show what we have learned about the patent search task and how this was done by investigating its nature and challenges experimentally.

By the end of this chapter, the first set of research questions about the patent search nature has been addressed. These questions are:

- What are the differences between recall-oriented search tasks and other standard IR tasks, and how can these be demonstrated?
- What are the properties and configuration of a retrieval system to achieve high retrieval effectiveness for patent search?
- Is applying standard retrieval techniques for patent search as effective as for general IR tasks?

From our participation in CLEF-IP for two successive years in 2009 and 2010, we have learned the following:

1. Patent structure cannot be used in a simple way to improve retrieval effectiveness over an unstructured search baseline.
2. Indexing the English parts of multilingual EPO patents is sufficient to achieve good retrieval results.

3. The description section in the patent is the best individual section to be used for formulating the query in our experiments, where it achieved the best retrieval results.
4. Filtering the retrieval results based on patent classification, specifically the top 3 levels of classification in our case, improves the retrieval results for both MAP and recall.
5. Citation extraction from the description section of the patent topic, significantly improves the results. However, this is not applicable to all topics, since some patent applications do not contain initial patent citations within their description section.

Furthermore, this chapter presented a comparison between two very different approaches for the prior-art patent search task; the first is a simple search approach representing our contribution to the task, and the second is a much more complex approach. The comparison showed that the simple approach can achieve better results than the sophisticated one when initial citations are present in the patent topic. This finding highlights how research for the patent search task is in its early stages, where advanced approaches which try to model the working procedures in the patent office achieve similar quality to simple and straightforward ones requiring no sophisticated processing of the content.

The final part of this chapter investigated applying QE to the prior-art patent search task. We confirmed previous reported work that a standard PRF method is not effective for QE for patent search. We introduced a novel QE technique for patent search based on automatically generated synonyms sets (SynSet) that proved to be significantly more effective and efficient than the standard PRF method, and achieved significantly better results compared to the baseline when evaluated using MAP. In addition, the SynSets can be used by patent examiners in the patent office as a lexical resource to assist them in formulating effective patent queries.

Although our contributions to the CLEF-IP 2010 task presented in this chapter form part of the state-of-the-art for this task, we do not consider them to be the main contributions of this thesis for the following reasons:

1. We know that the CLEF-IP tasks are the most recent patent search task to be investigated in the research community, and that it studies a real-life task in the patent office, which is prior-art patent search. However, we believe that this task does not perfectly model the real-life scenario in the patent office, where prior-art patent search is an interactive search task that is performed over multiple search sessions. This means that our contribution to the task is a contribution to the CLEF-IP task itself not the interactive patent search task that is performed in the patent office.
2. We wanted to make our contributions in this thesis general enough to be applicable to any patent retrieval task and even any recall-oriented search task. This motivated us to work on developing new methods to serve the patent search task rather than developing the state-of-the-art patent search system for a task that does not perfectly model the real-life problem.

Therefore, our best results achieved in this chapter are taken as baselines for experimentation in the next chapters. The set of runs submitted to CLEF-IP 2009 by all participants forms the experimental data in the next chapter for developing a more effective evaluation methodology for patent search. The best result achieved from our participation in CLEF-IP 2010 forms the baseline for the experimentation reported in Chapter 6.

In the next chapters, we examine the topic of evaluation in patent search and recall-oriented search in general. The evaluation of the reported techniques has so far used the standard MAP metric. However, since this is a precision-focused metric, it does not seem to be the best metric for use in the evaluation of this kind of task. In the next chapter, we demonstrate this

theoretically and experimentally, and propose a new evaluation metric for these tasks. In the later chapter, we work on developing a novel method for creating an effective and efficient translation system for cross-language patent retrieval in order to overcome the high computational cost and large resource requirements needed for this task. We believe that these two contributions significantly serve patent retrieval and recall-oriented search without being especially developed for a patent retrieval system. The evaluation and efficient translation are seen to be general enough to be applied to any patent retrieval system and possibly any recall-oriented task, can contribute to improving the retrieval effectiveness of such systems.

Chapter 5

Evaluation in Recall-Oriented IR

Effective evaluation is a key element of research in IR. This requires the use of appropriate and meaningful evaluation metrics for specific IR tasks. As shown in previous chapters, the most commonly adopted metric for evaluation of the patent retrieval task since its introduction to the research community has been the standard MAP score. Despite the fact that recall is the objective of this task, which is mentioned explicitly in most of the research papers on the topic (Fujii et al., 2004; Mase et al., 2005; Xue and Croft, 2009b; Roda et al., 2009), when it comes to evaluation, a precision focused metric has been adopted, albeit sometimes standard recall is used as well (Xue and Croft, 2009b; Roda et al., 2009). Viewing recall-oriented tasks purely in terms of measuring recall is actually rather simplistic. In practice the user's effort expended in the search is often also a key consideration. Thus it can be important for an evaluation metric to take account not only of the recall, but also the user's effort in identify relevant items as reflected in the ranks at which they are retrieved (Bonino et al., 2010), which actually means that at least an element of precision and quality of ranking of the results should form part of the evaluation.

This chapter describes a study analysing the behaviour of current evaluation metrics when applied to recall-oriented IR tasks (Magdy and Jones, 2010a). The results of this analysis are

used to motivate the proposal of a novel evaluation metric which combines recall with the quality of ranking of the retrieved relevant results. This allows us to distinguish between systems of similar recall by giving higher scores to systems with better ranking of relevant documents. The new metric is demonstrated to have a 0.87 correlation to recall and 0.66 correlation to precision, which demonstrates how it reflects both recall and precision with more emphasis on recall. Additional theoretical analysis shows that this new metric would also work well for other recall-oriented IR applications such as legal search where the number of relevant documents is typically very large. Also, the robustness of the score is examined when relevant judgements are incomplete and is shown to have high robustness when compared to MAP or recall. This new metric has been adopted in the CLEF-IP track to evaluate the submitted runs of participants (Piroi, 2010).

This chapter initially surveys background on IR evaluation scores, and explores the effectiveness of the current IR evaluation scores for measuring system performance for recall-oriented IR applications. Later, the chapter introduces the new evaluation metric (PRES), and explores its behaviour by use of illustrative examples and by testing it on 48 runs submitted to CLEF-IP 2009, in addition, the behaviour of PRES for other recall-oriented tasks is investigated. It also reports the PRES values for some of the work presented in the earlier chapters. Finally, the robustness of PRES is examined when relevance assessments are incomplete.

5.1 Evaluation Metrics in IR

5.1.1 Mean average precision (MAP)

While many evaluation metrics have been proposed for ad hoc type IR tasks, by far the most popular in general use is mean average precision (MAP) (Baeza-Yates and Ribeiro-Neto, 2010).

The standard scenario for use of MAP in IR evaluation is to assume the presence of a collection of documents representative of a search task and a set of test topics (user queries) for the task along with associated manual relevance data for each topic. The relevance data for each topic is assumed to be a sufficient proportion of the documents from the collection that are actually relevant to that topic. "Sufficient" here relates to the fact that the actual number of relevant documents for each topic is unknown without manual assessment of the complete document collection for each topic. Several techniques are available for determining sufficient relevant documents for each topic (Tague et al., 1981; Buckley et al., 2006).

As its name implies, MAP is a precision metric, where precision focuses on how many documents of the retrieved ones are relevant, see Equation (5-1), rather than checking how many of the relevant documents are retrieved, as is measured by recall, see Equation (5-2). Equation (5-3) shows the definition of average precision (AP) for a given topic, and MAP is the mean of AP taken over all topics in the test collection. From this definition, it can be seen that the impact on MAP of locating relevant documents later in the search of a ranked list is very weak, even if very many such documents have been retrieved, since MAP is a function of the multiplicative inverse of the rank of the relevant documents, see Equation (5-1). This is why MAP emphasizes returning a greater number of relevant documents earlier in the ranked list of results. Thus while MAP gives a good and intuitive means of comparing systems for IR tasks emphasising precision, it will often not give a meaningful interpretation for recall focused tasks. A detailed analysis of the behaviour of MAP is described in (Moffat and Zobel, 2008).

$$Precision(r) = \frac{rel\_ret}{r} \tag{5-1}$$

$$Recall = \frac{rel\_ret}{n} \tag{5-2}$$

$$AP = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{n} \tag{5-3}$$

where:

rel_ret: number of relevant document retrieved in the ranked list

r: the rank

rel(r): a binary function of the document relevance at a given rank, where *rel(r)*=1 when document at rank *r* is relevant and *rel(r)*=0 otherwise.

P(r): precision at a given cut-off rank, i.e. *Precision*(*r*)

n: the total number of relevant documents

5.1.2 Alternative evaluation metrics in IR

Many other IR evaluation metrics have been developed which are designed to reflect other features of IR behaviour or for other types of IR tasks. For example, mean reciprocal rank (MRR) for known-item retrieval and question answering (Voorhees and Tice, 1999), normalized discounted cumulative gain (nDCG) for web search (Carterette et al., 2008), mean average generalized precision (MAgP) for structured document retrieval (Kamps et al., 2007), geometrical mean average precision (GMAP) for robust search (Voorhees, 2005), and others.

Similar to MAP, these IR evaluation metrics focus on measuring effectiveness at retrieving relevant documents earlier rather than on the system recall. While this is sufficient and reasonable for precision focused tasks, it is not suitable for tasks where the objective is to find "all" relevant documents and in particular if the objective is to find all relevant documents with

minimum effort for the user. In this type of application, the user is willing to exert much effort to go deeper in the list in order to find relevant documents. Additionally, for recall-oriented IR applications the maximum number of documents to be checked by the user (the cut-off of the retrieved results) is also very important, since it has a direct impact on the cost of user effort and on recall. The maximum number of documents to be checked by the user is completely overlooked by most of the metrics considered so far, but features as a variable in measures such as the f-score (Rijsbergen, 1979). The f-score combines recall with precision, and has been used for legal IR (Oard et al., 2008); although this score includes recall, it has the problem that the number of documents to be retrieved is not fixed, since it is calculated at a different cut-off for each query. This means that the number of results checked for each query can vary significantly.

5.2 Current Evaluation Scores for Recall-Oriented IR Tasks

A simple solution to measuring performance in a recall focused IR task is to evaluate the recall. However, as noted previously, the problem of doing this is that it fails to reflect how early a system retrieves the relevant documents and thus the user effort involved. Although recall is the objective for such applications, the retrieval score should be able to distinguish between systems that retrieve relevant documents earlier from those that retrieve them later. To overcome this problem the f-score can be used, but at a fixed number of retrieved documents. However the same problem will arise, since applying it at the same cut-off for two systems that retrieved the same number of relevant documents, the f-score will be the same, since precision and recall of the two systems will be equal. This situation arises since the f-score is designed for classification tasks (Rijsbergen, 1979), but for recall-oriented IR applications, the problem is viewed as a ranking problem with a cut-off of a maximum number of documents to be checked N_{max}.

A simple modification for using the f-score is to calculate it as a combination between the recall and the average precision (AP) instead of using the absolute precision, shown in Equation (5-4). Such a modified f-score will reflect the system recall in addition to its average precision, which will distinguish between systems of similar recall, but different ranking of relevant documents. However, while this captures the recall, it will have the same disadvantages for recall focused tasks with respect to *AP* which were noted earlier.

$$F'_{\beta} = \frac{(1+\beta^2)\cdot(AP \cdot R)}{\beta^2 \cdot AP + R} \tag{5-4}$$

where:

AP: average precision of a topic

R: recall at a given number of retrieved documents

β: weight of recall to precision

Table 5-1: Performance of different scores with different IR systems

	Ranks of rel. docs	***AP***	***Recall***	F_1	F'_1	F'_4
system 1	{1}	0.25	0.25	0.0192	0.25	0.25
system 2	{50, 51, 53, 54}	0.0481	1	0.0769	0.0917	0.462
system 3	{1, 2, 3, 4}	1	1	0.0769	1	1
System 4	{1, 98, 99, 100}	0.2727	1	0.0769	0.429	0.864

Table 5-1 shows an illustrative example of how different metrics perform with four different IR systems when searching a collection for a single query. In this case it is known that there are four relevant documents, and it is assumed that the user is willing to check the top 100 documents retrieved by each system. Table 5-1 reports five different evaluation metrics: *AP*, recall, F_1, F'_1, and F'_4. Since F'_1 does not focus on the recall, because it gives similar weights to

AP and recall, and recall is the objective of recall-oriented applications. F'_4 is reported to emphasize recall by giving recall four times the weight of *AP* ($\beta = 4$ in Equation (5-4)).

In Table 5-1, system 3 is the prefect result with all relevant documents retrieved at the top ranks. System 1 has the lowest recall, while system 2 has moderate performance retrieving all relevant documents in the middle of the ranked list; system 4 has fair performance since it ranks one relevant document at rank 1, but achieves 100% recall only after checking the full list of 100 top results.

From the table it can be seen that *AP* for system 1 is much higher than for system 2, which is unfair, since system 2 retrieved all relevant documents in the middle of the list, but system 1 failed to retrieve more than one relevant document in the full list. The same situation arises when comparing system 4 to system 2, even though both systems retrieved the full list of relevant documents, system 2 has done so at much higher ranks than system 4. The recall and F_1 scores fail to differentiate between systems 2, 3, and 4, even though these systems have very different behaviour.

Initial inspection suggests that F'_4 might be a good representation of the system performance, since it ranks system 1, 2, and 3 in a logical way according to their performance for a recall-oriented task. However on deeper analysis, it can be seen that system 4 is evaluated to be nearly twice as good as system 2, even though, while it retrieves a relevant document at rank 1, no further relevant documents are found until the end of the list, and that while system 2 failed to return any relevant documents among the first half of the list, all relevant documents are retrieved by rank 54. For two systems such as 2 and 4 for a recall-oriented task with users willing to check the first 100 documents, system 2 will give more confidence to the user that there is little chance of finding further relevant documents after rank 100; since the presence of low ranked relevant documents in system 4 may suggest that further ones are more likely to be

present. Hence, F'_4 fails to evaluate system 2 and system 4 fairly from the perspective of a recall-oriented application in practical usage.

This exploration of the usage of these scores for the evaluation of recall-oriented search motivates the need for a more meaningful metric that can measure different aspects of a recall-oriented IR system, which includes the overall system recall and the amount of effort required to locate the relevant documents (Bonino et al., 2010). In addition, such a metric should be tuned according to the amount of effort that the user is willing to exert for locating the relevant documents in a ranked list. For example, in patent search, the patent examiners are willing to check hundreds of documents (Azzopardi, 2010), while in legal search, the users can check thousands of documents (Tomlinson et al., 2007). Therefore, such a metric should be a function of the amount of effort the user is willing to exert in search.

5.3 PRES (Patent Retrieval Evaluation Score)

In this section, we introduce our new metric for evaluating recall-oriented search tasks in general and examine it one patent search. We call the score *Patent Retrieval Evaluation Score* (PRES). We provide the detail of the development of the score and examine its applicability in illustrative examples and real data from the CLEF-IP track.

5.3.1 Evolution of R_{norm} to PRES

- **Normalized Recall (R_{norm})**

One of the proposed IR evaluation metrics that has never found its way into wide usage is normalized recall (R_{norm}) (Rocchio, 1964; Robertson, 1969; Rijsbergen, 1979), shown in Equation (5-5). This score was developed by Rocchio as a metric that can reflect in one number the precision-recall curve, with the requirement to rank all documents in the collection according

to relevance to a query (Rocchio, 1964; Robertson, 1969). This metric measures a system's effectiveness in ranking documents relative to the best and worst ranking cases, where the best ranking case is retrieval of all relevant documents at the top of the list, and the worst is retrieving them only after retrieving the rest of the collection. Figure 5.1 shows an illustrative graph of the calculation of R_{norm}, where it is the area between the actual and worst cases divided by the whole area. Full derivation of the R_{norm} equation is presented in Appendix A.

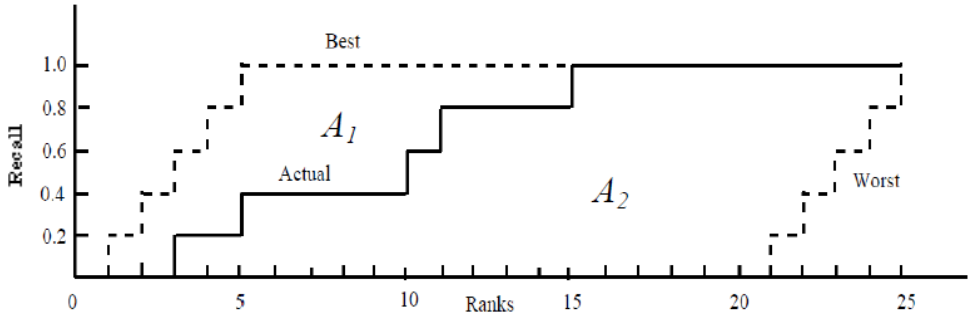


Figure 5.1: Illustration of how R_{norm} curve is bounded by the best and worst cases (Rijsbergen, 1979)

$$R_{norm} = \frac{A_2}{A_1 + A_2} = 1 - \frac{\sum r_i - \sum i}{n(N-n)} \tag{5-5}$$

where:

A_1, A_2: areas shown in Figure 5.1

r_i: the rank at which the ith relevant document is retrieved

N: collection size

n: number of relevant docs

Normalized recall can be seen as a good representative measure for recall-oriented IR applications, since its value is greater when all relevant documents are retrieved earlier. However it requires ranking of the full collection. Applying R_{norm} on collections of very large numbers of

documents is infeasible, since it is impractical to rank a collection of potentially many millions of documents. In addition, some relevant documents may have no match to the query leading to them not being retrieved at all.

One solution to address this problem is to consider any relevant documents not retrieved in the top N_{max} to be ranked at the end of the collection. Applying this assumption enables the calculation of R_{norm}, but leads to its value being nearly equal to the system recall at a cut-off of N_{max}. For example, for a collection of tens of thousands of documents and when retrieving the top 1000 documents; if recall at 1000 equals 50%, R_{norm} with the previous approximation will equal 49.99%, see Figure 5.2. Therefore, R_{norm} can be considered as a score for evaluating recall-oriented applications but only for small collections when it is practical to rank all documents in the collection. However, with large collection sizes, R_{norm} becomes impractical to calculate.

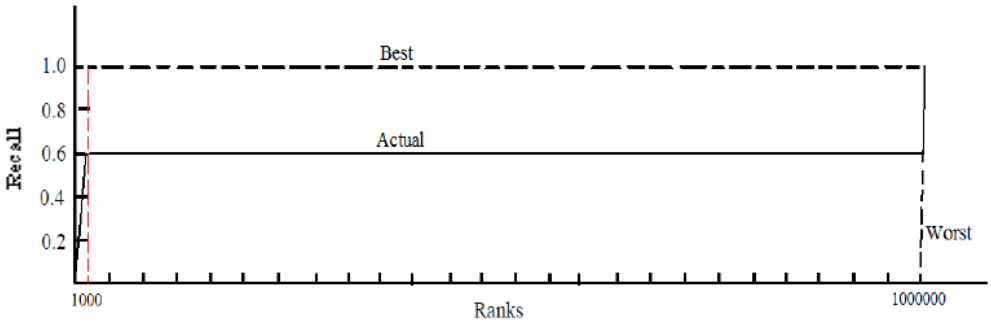


Figure 5.2: Illustration of R_{norm} curve behaves with large document collections

In the next part, we present a modification to the definition of R_{norm} while preserving its basic characteristics to create a novel score that is practically suitable for measuring the quality of retrieval effectiveness of a recall-oriented search task.

- **PRES**

Here we introduce the new score "*Patent Retrieval Evaluation Score*" (PRES), which is based on the same idea as the R_{norm}, but with a different definition for the worst case. The new

assumption for the worst case is to retrieve all the relevant documents just after the maximum number of documents to be checked by the user (N_{max}). The motivation for this assumption is that getting any relevant document after N_{max} leads to it being missed by the user, and getting all relevant documents after N_{max} leads to zero recall, which is the theoretical worst case scenario. Applying this assumption in Equation (5-5), N is replaced with $N_{max}+n$, where n is the number of relevant documents. Any relevant document not retrieved in the top N_{max} is assumed to be the worst case. Figure 5.3 presents the PRES curve and illustrates how it behaves with the new definition of the worst case scenario. For example, for a retrieved ranked list for a topic with 10 relevant documents (n = 10) and for which the user is willing to check the top 100 documents (N_{max} = 100); the best case will be finding the 10 relevant documents at ranks {1, 2, … 10}, and the worst case will be finding them in the ranks {101, 102, … 110}, which means the user misses all the relevant documents. Assuming retrieval of only 7 relevant documents in the top 100, then the missing 3 relevant documents will be assumed to be found at ranks {108, 109, 110}.

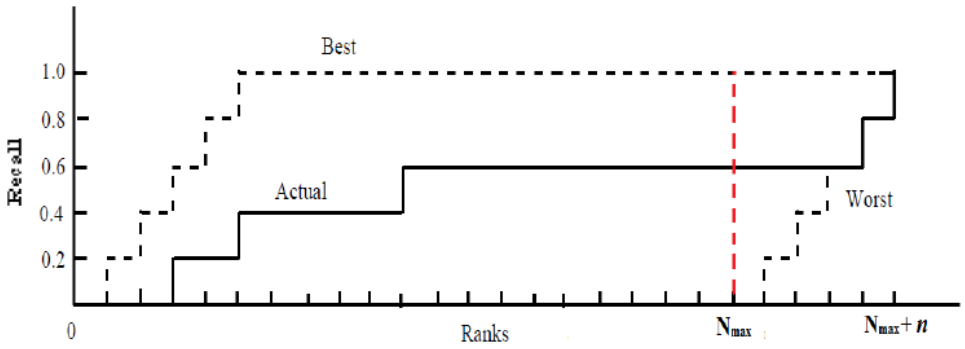


Figure 5.3: PRES curve is bounded between the best case and the new defined worst case

The PRES formula can be derived by applying the new definition of the worst case scenario to Equation (5-5) as follows:

$$PRES = R_{norm}\Big|_{N=N_{\max}+n}$$
$$= 1 - \frac{\sum r_i - \sum i}{n(N-n)}\Bigg|_{N=N_{\max}+n}$$
$$= 1 - \frac{\sum r_i - \sum i}{n \times N_{\max}}$$

but,

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

then,

$$\therefore PRES = 1 - \frac{\sum r_i - \frac{n(n+1)}{2}}{n \times N_{\max}}$$

which can be written as follows:

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{\max}} \tag{5-6}$$

Equation (5-7) shows the direct calculation of the summation of the ranks of relevant documents in the general case when some relevant documents are missing from the top N_{max}.

$$\sum r_i = \sum_{i=1}^{nR} r_i + nR(N_{\max} + n) - \frac{nR(nR-1)}{2} \tag{5-7}$$

where:

R: Recall (number of relevant retrieved documents in the first N_{max} documents)

From Equation (5-6), it can be seen that PRES is a function of the recall of the system, the ranks of the retrieved relevant documents, and the maximum number of results to be checked by the user. For a given N_{max}, *PRES* behaves as shown in Figure 5.4. For recall = *R*, the PRES value ranges from *R*, when retrieving all relevant document at the top of the list, to nR^2/N_{max} when retrieving them at the bottom of the list.

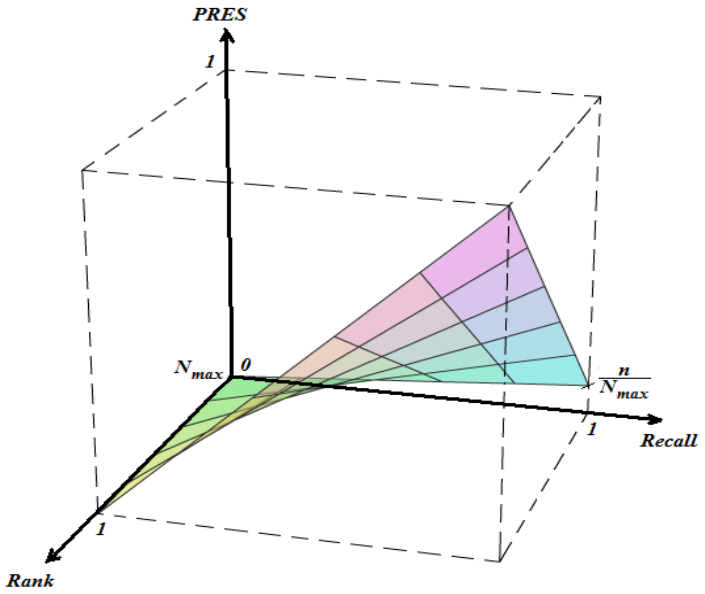


Figure 5.4: PRES behaviour with variation of rank of relevant document and recall

To understand the behaviour of PRES when retrieving relevant document at different ranks, we assume the special case where the number of relevant documents for a topic is one (n=1). In this case PRES will have a linear characteristic. Figure 5.5 shows the difference between PRES and MRR (which is equivalent to MAP when n=1) performance with different ranks for the case where n=1. It can be seen that PRES has a linear decay as the relevant document is retrieved lower in the ranked list, while MRR has an exponential decay, which is a strong punishment to the score value when retrieving the relevant document lower in the ranked list regardless of the number of documents the user is willing to check.

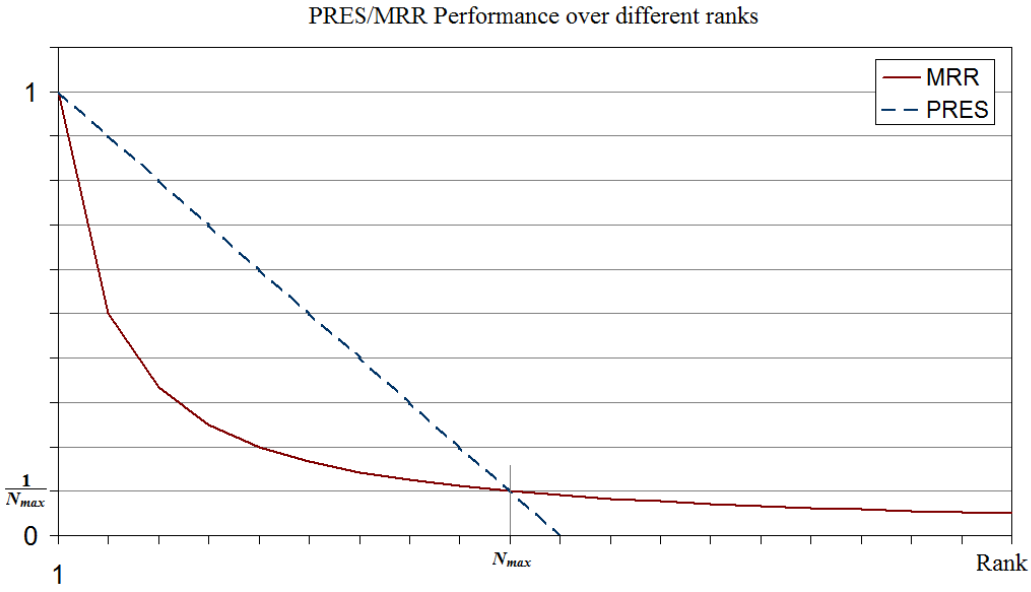


Figure 5.5: PRES vs. MRR at different ranks when n=1 and N_{max} = 10

5.3.2 Theoretical differences between R_{norm} and PRES

Normalized recall was first proposed by Rocchio in 1964 (Rocchio, 1964) as an IR evaluation metric independent of the cut-off value of the retrieved documents, since it requires returning all documents of the collection ranked by relevance. In 1969, Robertson (1969) showed that R_{norm} is the same as the area under the recall-fallout curve (operating characteristic curve), which makes R_{norm} equal to the probability of pairwise error in ranking, and which leads to R_{norm} = 0.5 for random ranking of documents in the collection. This is not the case for PRES, where the PRES value is directly dependent on the cut-off value. Furthermore, random ranking of documents will eventually lead to PRES = 0 for the current common collection sizes, since the probability of finding a relevant document = n/N, where N is typically millions or billions of documents, and a practical value of n never exceeds thousands of documents for the extreme cases, such as legal search.

Normalized recall was a suitable evaluation measure at the time it was introduced, but with current collection sizes and types of applications R_{norm} is an impractical measure for operational use. PRES can be viewed as an IR evaluation measure that has the characteristics of the classic R_{norm}, but with a different meaning. PRES is designed specifically for recall-oriented applications to emphasize the system quality in retrieving the most significant relevant documents as early as possible within a specific number of results in a ranked list.

5.3.3 Example of PRES behaviour

Table 5-2: Performance of PRES with different IR systems

	Ranks of rel. docs	***AP***	***Recall***	**PRES**
system1	{1}	0.25	0.25	0.25
system2	{50, 51, 53, 54}	0.0481	1	0.51
system3	{1, 2, 3, 4}	1	1	1
system4	{1, 98, 99, 100}	0.2727	1	0.28

Table 5-2 shows how PRES performs with the sample examples presented in Table 5-1. From Table 5-2, it can be seen that PRES is a better representative measure for the system performance than the other metrics, since it combines system recall and average ranking of relevant documents. As shown, the four systems are ranked more logically according to PRES from a recall-oriented point of view, where system 3 is the best, system 1 is the worst, and system 2 is better than system 4 since its average quality of ranking for the relevant documents is better.

Some samples of topics from one run of the CLEF-IP 2009 track are presented in Table 5-3 with the maximum number of results to be checked by the user N_{max} = 1000. In Table 5-2 and Table 5-3, PRES is shown to be always less than or equal to recall depending on the quality of ranking of the relevant documents relative to N_{max}. For example, getting a relevant document at

rank 10 will be very good when N_{max}=1000, good when N_{max}=100, but bad when N_{max} = 15, and very bad when N_{max}=10. Systems with higher recall can achieve a lower PRES value when compared to systems with lower recall but better average ranking of relevant documents. This is shown clearly in Table 5-3, where one topic with 67% recall has 63.6% PRES because of good ranking of the relevant documents (41 and 54 among 1000), and one topic with 100% recall has 52.5% for PRES because of the moderate ranking of relevant items where 60% of them are below rank 500 out of 1000.

Table 5-3: AP/R/PRES performance with real samples of topics

Ranks of rel. docs	*n*	*R*	*AP*	PRES
{98,296}	41	0.05	~ 0	0.039
{23,272,345}	6	0.5	0.01	0.394
{2,517,761}	6	0.5	0.085	0.288
{660,741}	3	0.667	0.001	0.201
{41,54}	3	0.667	0.021	0.636
{1,781}	3	0.667	0.334	0.407
{1,33,354,548,733,840,841}	7	1	0.157	0.525
{32,35,46}	3	1	0.051	0.964

Comparing PRES to average precision (*AP*) for the samples in Table 5-3, it can be seen that *AP* is more sensitive to finding relevant documents at very high ranks regardless of the number of documents to be checked by the user. However, PRES is more sensitive to the average ranking of the relevant retrieved documents as a whole relative to the maximum number of documents the user is willing to check. The last sample topic in the table has a PRES of 96.43% even though relevant documents are not ranked in the top 10 or even 20 results. The reason for this is that N_{max}=1000, and the ranks {32, 35, 46} are considered relatively good compared to this number. Nevertheless, when calculating PRES with N_{max}=100, the PRES value for the last sample topic in the table will be 64.33% which represents the average ranking of the relevant documents relative to the maximum number of documents to be checked.

5.3.4 Applying PRES for patent retrieval

In this section, we examine the application of PRES to the evaluation of real runs from the patent prior-art search task at CLEF-IP 2009. The objective is to see how PRES compares to the main evaluation metrics reported at CLEF-IP 2009, namely MAP and recall.

PRES was evaluated for 48 submissions from 15 participants at the CLEF-IP 2009 Patent Track (Roda et al., 2009). Table 5-4 shows the score for each submission in MAP, recall, and PRES. Participant IDs are anonymized and the number of topics for each participant used was a randomly selected subset of 400 topics out of the official 500 in order to further mask participant identities and to avoid violating the privacy of any of the participants. For all topics, N_{max} = 1000 was used, since it is the maximum number of results that were allowed for the track submissions. The average number of relevant documents per topic was nearly 6 (n_{avg} = 6).

Table 5-4: MAP/recall/PRES for 48 submissions in CLEF-IP

Run ID	MAP	Recall	PRES	Run ID	MAP	Recall	PRES	Run ID	MAP	Recall	PRES
R01	0.077	0.530	0.434	R17	0.067	0.584	0.463	R33	0.085	0.457	0.379
R02	0.087	0.617	0.499	R18	0.033	0.656	0.490	R34	0.082	0.427	0.354
R03	0.084	0.609	0.497	R19	0.105	0.600	0.529	R35	0.114	0.572	0.496
R04	0.053	0.219	0.213	R20	0.003	0.051	0.040	R36	0.108	0.553	0.480
R05	0.000	0.020	0.011	R21	0.266	0.760	0.691	R37	0.114	0.572	0.494
R06	0.000	0.016	0.009	R22	0.028	0.256	0.200	R38	0.107	0.553	0.479
R07	0.000	0.012	0.007	R23	0.087	0.728	0.603	R39	0.113	0.575	0.498
R08	0.000	0.016	0.009	R24	0.011	0.069	0.054	R40	0.107	0.560	0.483
R09	0.071	0.454	0.369	R25	0.064	0.492	0.392	R41	0.079	0.547	0.447
R10	0.088	0.533	0.430	R26	0.084	0.511	0.431	R42	0.103	0.555	0.466
R11	0.087	0.489	0.404	R27	0.097	0.514	0.447	R43	0.091	0.575	0.475
R12	0.088	0.534	0.430	R28	0.091	0.514	0.442	R44	0.091	0.574	0.474
R13	0.065	0.508	0.406	R29	0.082	0.436	0.373	R45	0.106	0.616	0.507
R14	0.068	0.467	0.363	R30	0.092	0.559	0.469	R46	0.102	0.611	0.504
R15	0.064	0.434	0.348	R31	0.081	0.568	0.460	R47	0.104	0.589	0.484
R16	0.020	0.197	0.148	R32	0.078	0.476	0.391	R48	0.102	0.587	0.484

From Table 5-4, it can be seen that PRES reflects both the recall and the average quality of the ranking. Run 21 (R21) which achieved the highest MAP and recall also achieved the highest PRES, with the same behaviour being observed for the lowest scoring runs. However, some submissions which achieved high precision but low recall were punished and received a moderate PRES score. This effect can be seen for example by comparing R02 to R11. For systems which achieved high recall but low precision (which reflects bad ranking such as system R18), the PRES score was moderate too.

Figure 5.6 plots the three scores of the same 48 submissions sorted by PRES from low to high values. From Figure 5.6, it can be seen that PRES is a good single score that represents both the precision and recall of each run.

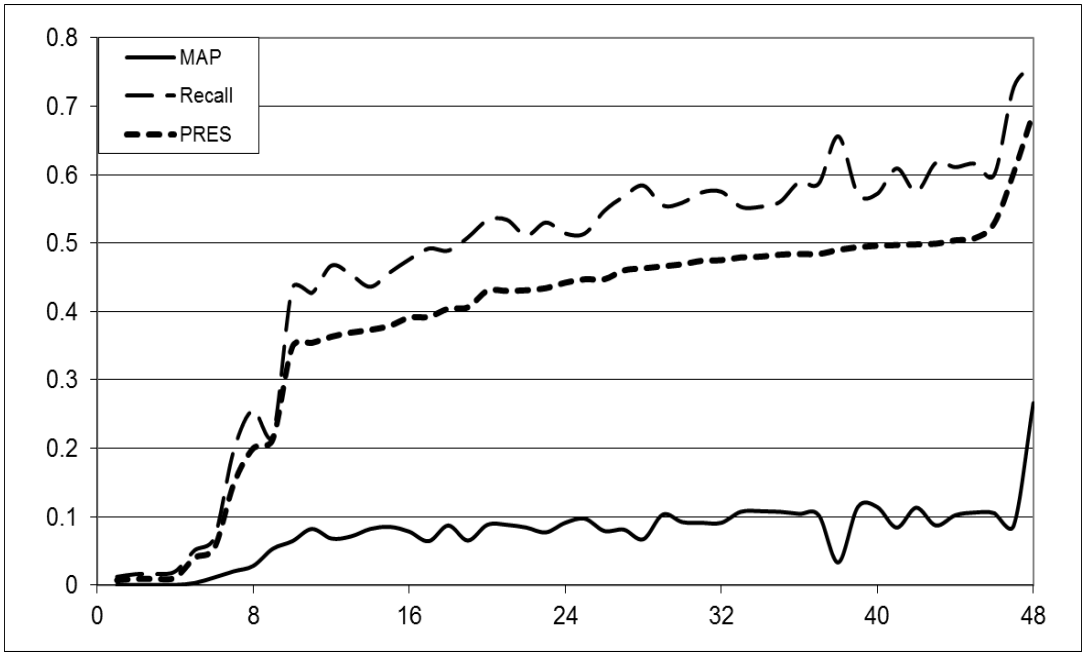


Figure 5.6: MAP/recall/PRES for 48 submissions in CLEF-IP 2009 sorted by PRES

Figure 5.7 shows the change in ranking of the submissions with the three scores. It can be seen that ranking using PRES is more biased towards recall, than MAP. However, this is not always the case, for example R12 has moderate ranking with respect to both recall and MAP, but lower ranking in PRES. This result arises from the fact that MAP is more sensitive to the high ranking of some of the relevant documents, but PRES is dependent on relative average ranking of "All" relevant documents to N_{max}. Figure 5.7 shows that the scores have strong agreement on the ranking of systems with very high or very low performance.

In order to check the agreement of the three scores, pairwise comparison of submissions was carried out between each two runs of the 48 runs for each of the scores. The pairwise comparison for two runs can have one of three results (Buckley and Voorhees, 2000):

1. The first run is statistically significantly better than the second run.
2. The second run is statistically significantly better than the first run.
3. The runs are statistically indistinguishable.

For example, for two runs: R01 and R02, MAP can show that R01 is statistically better than R02, while recall and PRES can show that R01 and R02 are statistically indistinguishable.

The Wilcoxon significance test with confidence level of 0.95 was used for comparing the runs (Hull, 1993). Comparing the 48 runs in a pairwise manner led to 1,128 comparisons. The agreement of scores for each comparison is plotted in Figure 5.8.

From Figure 5.8, it is clear that PRES is an intermediate score between recall and MAP. In addition, in a small number of cases (1%) PRES disagrees when recall and MAP agree. These situations arise in situations such as when recall and MAP agree that system 1 (first run) is better than system 2 (second run), but PRES shows that both systems have the same performance, or when recall and MAP agree that two systems are statistically indistinguishable, but PRES prefers one over the other.

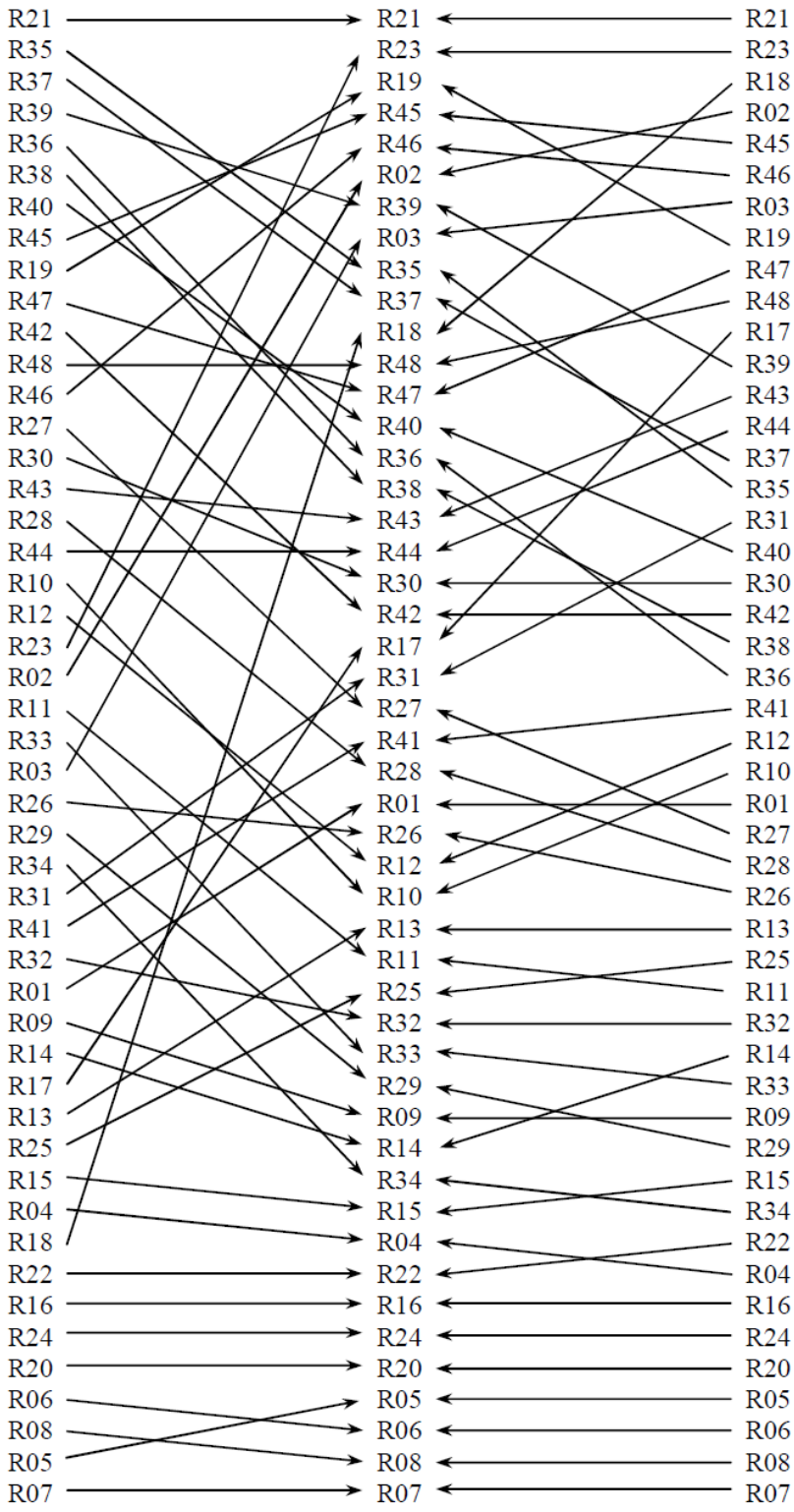


Figure 5.7: Ranking change of 48 submissions according to MAP/PRES/recall

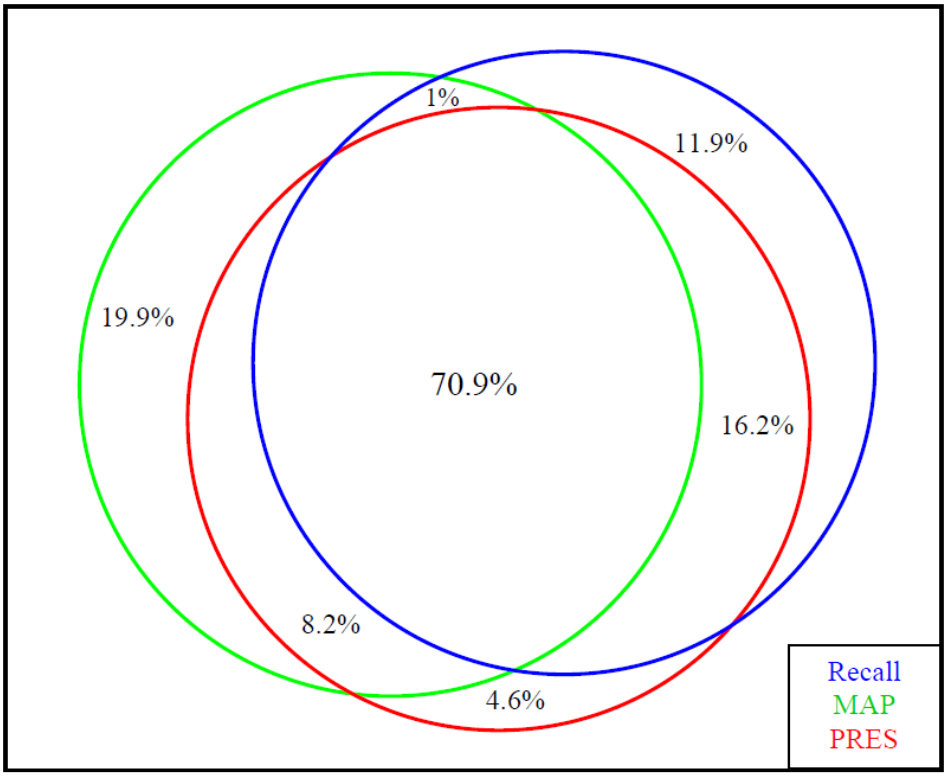


Figure 5.8: Agreement chart of MAP/recall/PRES on pairwise comparison of 48 submissions

Calculating the Kendall's tau correlation between the ranking of runs according to the three scores (Kendall, 1938; Voorhees, 2001), it is found that the correlations are as follows: MAP and recall = 0.56, PRES and recall = 0.87, and PRES and MAP = 0.66. These numbers show that PRES is more correlated to both MAP and recall than MAP and recall are correlated to each other. Moreover, PRES is more correlated to recall than it is correlated to MAP. This emphasizes that the PRES measure lies between MAP and recall with a bias towards recall.

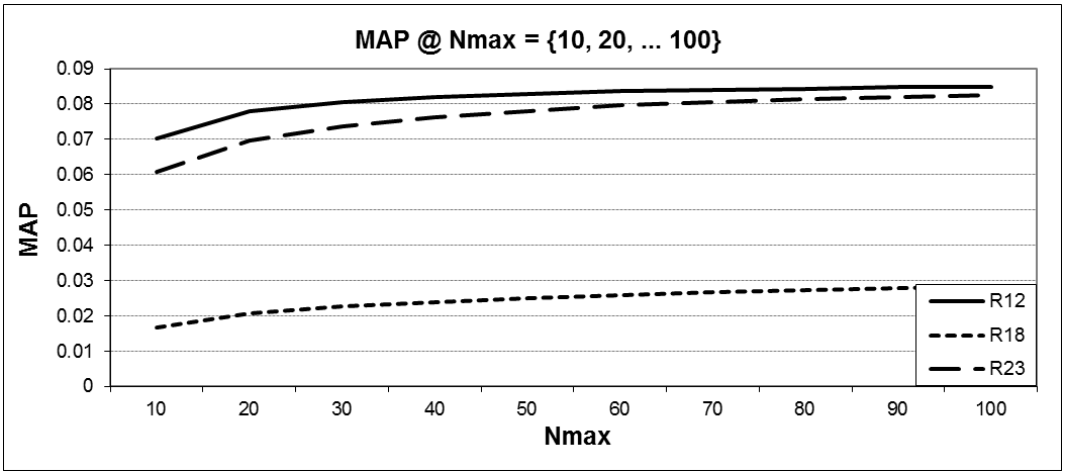
5.3.5 Performance versus different cut-off values (N_{max})

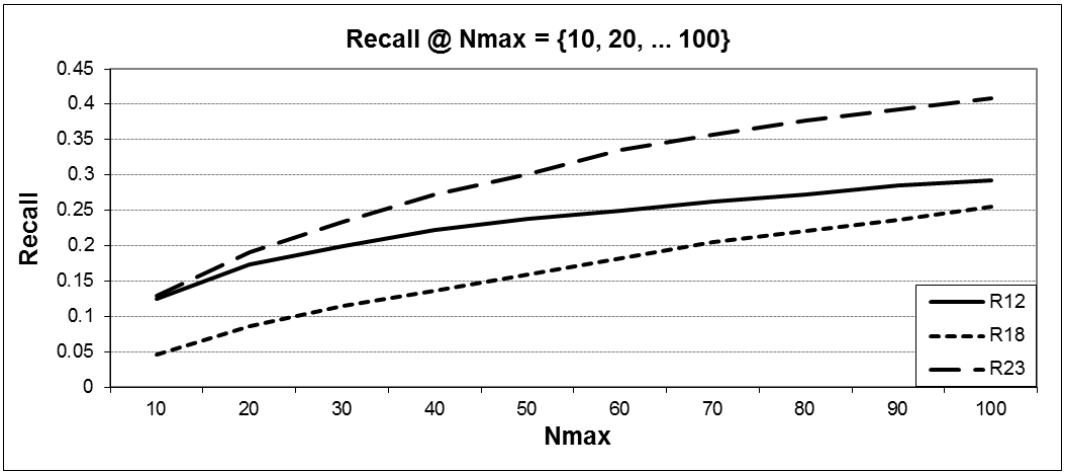
The cut-off value of documents to be checked is one of the key variables that affect the value of PRES. This is the same for recall, since the greater the number of retrieved documents that are to be checked, the higher the possibility of finding more relevant documents. However additionally

for PRES, N_{max} affects its value even if no more relevant documents are found, since for different cut-offs, the relative ranking of relevant documents is different.

For recall-oriented applications, the actual number of documents to be checked by the user is typically higher than other IR applications. This number can exceed a hundred documents in the case of a patent examiner before he/she thinks of reformulating the query (Azzopardi et al., 2010). Different factors can affect the decision to stop checking for relevant documents. For example, the failure to find a relevant document for a while in the results list can make the user decide to stop checking the list any further, or the user can decide to check a fixed number of documents, but when less relevant documents are found while checking the list the user will generally move more quickly through the list, since typically users pass over non-relevant documents more quickly but take more time to check more carefully documents that appear to be relevant to decide whether they are actually relevant. Therefore when less relevant documents are found at the bottom of the list, this can lead to more rapid task completion. For both scenarios the effort the user exerts to find a relevant document will be greater as long as he/she continues to find relevant documents deep in the list. This is the reason why PRES penalizes finding documents deeper in the list of the N_{max} ranked results.

Figure 5.9 shows the effect of changing the value of N_{max} on MAP, recall, and PRES. Three sample runs from CLEF-IP 2009 that represent different performances of patent search systems (R12, R18, and R23) were selected to examine the variation of the three scores at different values of N_{max}. From Figure 5.9, the effect of finding more relevant documents on MAP is very weak regardless of the number of documents to be checked by the user and regardless of the number of relevant documents found deeper in the list. This result was expected for MAP due to the way in which it is defined and calculated, see Equation (5-3). PRES and recall performances look similar in general, however, for the example, when N_{max} = 10, PRES judges R12 to be better than





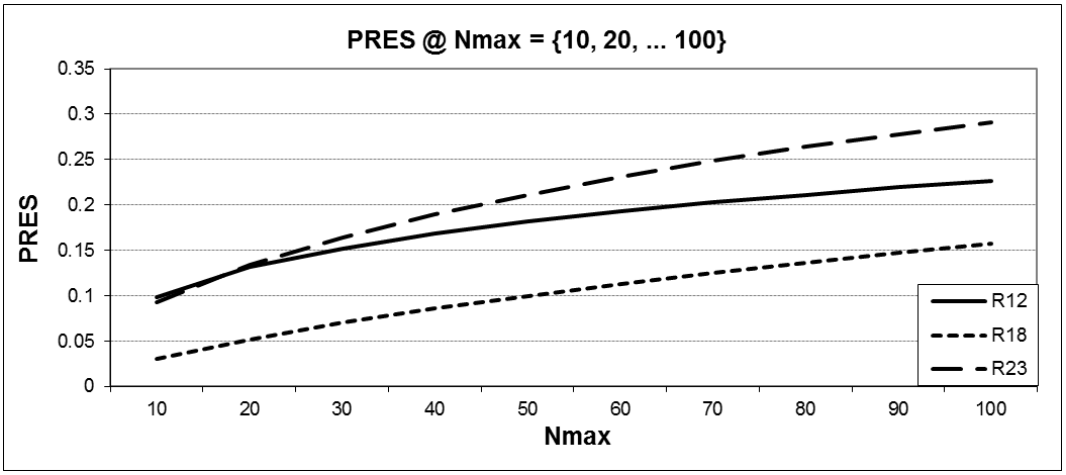


Figure 5.9: MAP/recall/PRES values for different values of N_{max} applied on three sample runs

R23, but recall judges the opposite. Furthermore, for R18 the recall curve with N_{max} has a higher slope than the PRES curve, which means that the effect of finding relevant documents deep in the results list has a stronger impact on recall than PRES. This returns us to the issue of the neglect of the ranking of documents by recall, which is taken into account by PRES, where finding relevant documents deep in the ranked list has a noticeable effect on PRES, but not as strong as finding them on the top of the list.

5.3.6 PRES when $n > N_{max}$

Usually for recall-oriented applications the user will check a number of retrieved results higher than the expected number of relevant documents. However, this scenario can be ignored in some evaluation tracks when the number of relevant documents is very high and the task is to evaluate different IR systems for the ability to find the largest number of relevant documents. This is the exact scenario in the legal search task. The legal track at TREC seeks to evaluate the ability of different systems to retrieve relevant legal documents (Tomlinson et al., 2007). As outlined in section 2.6, the number of relevant documents for a topic can reach tens of thousands. Several metrics and evaluation methods have been proposed to address this issue by estimating the number of relevant documents and the actual system precision and recall (Tomlinson et al., 2007).

In this section, the behaviour of PRES is studied for cases such as this, where the number of relevant documents (n) is higher than the maximum number of documents to be checked by the user (N_{max}), see Figure 5.10.

As shown in Figure 5.10, the best case will never be applicable as retrieving all relevant documents at the top ranks will exceed the cut-off value, and the user will never be able to achieve 100% recall. However, the calculation of PRES in this case can still be applied without

any modification. As described before in section 5.3.1, for a recall R, PRES will range from nR^2/N_{max}, to R. The only difference here is that the maximum applicable R will be N_{max}/n, which is the case when all the retrieved documents are relevant.

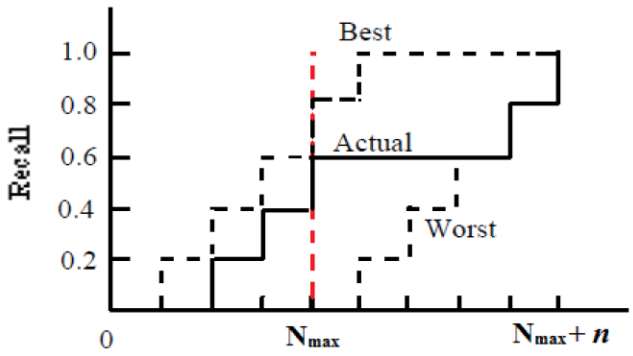


Figure 5.10: PRES curve for situations when $n > N_{max}$

Although the PRES calculation is still applied, the PRES value will have some limitation in expressing the general system performance, since for a recall-oriented task N_{max} should be greater than the expected number of relevant documents, and N_{max} has a direct impact on the value of PRES. Nevertheless, an estimation of PRES ($PRES_{est}$) can be still calculated to approximate the full performance of the system for some special cases as shown Equation (5-8).

$$PRES_{est} = \frac{PRES}{R_{\max}} \tag{5-8}$$

$$R_{\max} = \frac{N_{\max}}{n} \qquad ,(N_{\max} \le n) \tag{5-9}$$

where:

$PRES_{est}$: estimated PRES

R_{max}: maximum possible recall (R_{max} = 1 when $N_{max} \geq n$)

While this provides an estimate of system performance, it is advisable only to use $PRES_{est}$ in evaluation campaigns where there are a large number of runs with a very large number of relevant documents, and it is impractical to evaluate the very long submitted lists of many systems. For an accurate evaluation using PRES, N_{max} should be carefully selected according to the user and application models, and should be higher than *n*.

5.4 PRES in CLEF-IP 2010

PRES was adopted as a metric for the evaluation of submissions to the CLEF-IP 2010 prior-art patent search task (Piroi, 2010). Afterwards, PRES has been reported in many of published works on patent search (Lupu et al., 2011; Ganguly et al., 2011; Magdy et al., 2011; Mahdabi et al., 2011). In this section, some of the reported results for experiments from the previous chapter are revisited to include PRES. Specifically, the results reported in Figure 4.4, Figure 4.5, Table 4-3, Table 4-5, and Table 4-7 are repeated here including PRES values with an analysis giving the interpretation of the results.

5.4.1 PRES for our participation in CLEF-IP 2010

In the evaluation of the CLEF-IP 2010 track, both standard PRES (stands for PRES@1000) and PRES@100 were used to evaluate the submitted runs besides other scores including MAP and recall. Our contribution to CLEF-IP 2010 (Magdy and Jones, 2010c) presented in section 4.2, achieved the second best run using both MAP and PRES scores, which means that the average ranking of relevant results of our system is consistent across the full ranked list. Table 5-5 reports the results from Table 4-3 but adding the PRES values for our runs in CLEF-IP 2010. Racell@100 is also reported in Table 5-5 for comparison with PRES@100. The first placed participant in CLEF-IP 2010 (Patrice and Lopez, 2010) achieved PRES and PRES@100 values

of 0.6187 and 0.3907 respectively. However, and as highlighted in previous chapter, this result was achieved when using a different list of extracted citations that were enriched using patent family lookup. Our best run ("IR+Cit") was still ranked the second best run among the 25 submitted ones using PRES.

Table 5-5: Results in Table 4-3 with PRES values. MAP, recall, recall@100, PRES, and PRES@100 for the DCU runs submitted to CLEF-IP 2010

Run #	MAP	R	R@100	PRES	PRES@100
IR	0.1216	0.5700	0.3036	0.4614	0.2280
Cit	0.1120	0.1187	0.1187	0.1186	0.1176
IR+Cit	0.2029	0.618	0.3846	0.5229	0.3162

5.4.2 PRES for work after CLEF-IP 2010

In section 3.3.2, it was noted that the work reported in (Mahdabi et al., 2011) achieved a lower MAP than ours for the English topics in the CLEF-IP 2010 prior-art patent search task without using citation extraction. They achieved a MAP of 0.1240 compared to our result of 0.1399. However, Mahdabi et al. (2011) reported that they achieve a PRES of 0.485 compared to our result of 0.483, which are very close results. This result means that the systems are highly comparable with regard to retrieving similar numbers of relevant documents which are ranked similarly on average. However, when comparing the systems for retrieval of relevant documents at the very top ranks, then our system is better. For example, moving a relevant document, when N_{max} = 1000 as in this case, from rank 10 to rank 1 will affect the MAP value significantly, but will have an insignificant impact to PRES, since ranks 1 and 10 are considered very good when compared to 1000 documents to be checked. Therefore, as a conclusion, it can be understood that the systems described in (Mahdabi et al, 2011) and ours have very comparable performance in patent search for the English topics when no citation extraction is performed.

5.4.3 Evaluating simple vs. sophisticated patent search approaches using PRES

In Section 4.3, MAP was used to compare the retrieval effectiveness of our simple approach to the more sophisticated one of the first participant in CLEF-IP 2010 when a similar list of extracted citations were used (Magdy et al., 2011). Figure 5.11 shows the PRES values for both approaches when citations could and could not be extracted from English topics of CLEF-IP 2010. Here the PRES metric agrees with MAP, which was presented before in Figure 4.5, where the simple approach is better when initial citations exist in the patent topics, and the sophisticated approach is significantly better when no citations exist.

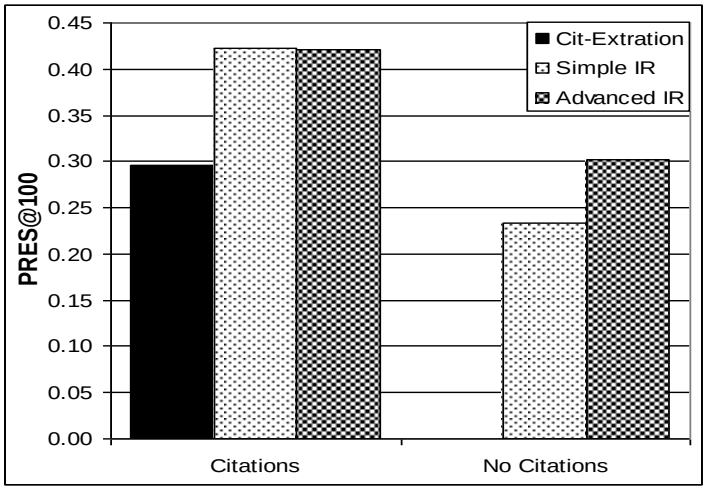


Figure 5.11: PRES values for simple and complex IR approaches with CLEF-IP prior-art search task, when citations could be and could not be extracted

5.4.4 Evaluating QE methods using PRES

Table 5-6 and Table 5-7 show Table 4-5 and Table 4-7 from Section 4.4 extended to include PRES. These tables report the results of applying PRF and SynSet QE to the CLEF-IP 2010 English patent queries. From Table 5-6, it can be seen clearly that PRES aligns with MAP in showing that the PRF method examined in Section 4.4.1 is not an effective approach for QE for

prior-art patent search, the PRES score is reduced by a similar amount compared to the baseline as MAP. Even for the additional run we examined by applying PRF using only one document for extracting 300 feedback terms achieved a degradation in PRES similar to MAP, where the PRES score achieved was 0.370 compared to a baseline of 0.486.

Table 5-6: The effect of using PRF with different number of terms and documents on retrieval effectiveness measured by MAP and PRES

	Terms / Docs	10	20	30	50
MAP BL = 0.1399	**1**	0.046	0.065	0.076	0.082
	5	0.037	0.053	0.062	0.072
	10	0.031	0.046	0.053	0.061
	20	0.026	0.036	0.042	0.049
PRES BL = 0.486	**1**	0.197	0.239	0.258	0.279
	5	0.196	0.234	0.247	0.265
	10	0.190	0.222	0.235	0.251
	20	0.178	0.205	0.216	0.232

Table 5-7 shows that the SynSet QE approach is not effective for patent search when measured using PRES. This is in contrast to measurement using MAP which shows significant improvement in results. Although the reduction in the PRES score compared to the baseline was only -0.7% for the Wsynset run, this slight change was found to be statistically significantly worse than the baseline. This result means that this technique on average improved high rank precision, and degraded the recall and the overall ranking of relevant documents across the full results list. For a recall-oriented task such as patent search, this result is considered a negative outcome. When checking the percentage of topics improved by PRES, it was found that only 33% of the topics were improved while 40% were degraded and the rest unchanged (<1% change in PRES value). The new reading for the results using PRES aligns with most of the published

work on different QE techniques to patent search, as noted in previous chapters, which have never shown a stable improvement in the retrieval effectiveness (Kishida, 2003; Itoh, 2004; Takeuchi et al., 2005).

Table 5-7: Effect of using the automatically generated SynSet for QE on the retrieval effectiveness for the English topics in CLEF-IP 2010

	MAP		PRES	
	value	% change	value	% change
Baseline	0.1399	NA	0.486	NA
Wsynset	0.1440	+2.9%	0.485	-0.7%
Usynset	0.1402	+0.2%	0.480	-1.7%

The results in this section demonstrated the value of having PRES as a metric for the evaluation of recall-oriented search tasks such as patent search. It was shown how MAP can be a misleading score for some of the reported results, where it does not always agree with PRES on the effectiveness of some techniques for improving the retrieval effectiveness of patent search.

The PRES score is potentially an effective metric for measuring the quality of an IR system for a recall-oriented search task. Measuring retrieval performance adequately is an essential matter during the development of an effective IR system, and this is what PRES is mainly designed for: to help in designing more effective recall-oriented retrieval systems, especially patent retrieval systems.

5.5 Metrics Robustness with Incomplete Relevance Judgements

We mentioned in Chapter 3 that the relevance set of patents to a topic in the prior-art patent search task are taken from the final citations list in the patent topics (Graf and Azzopardi, 2008; Roda et al., 2009; Piroi, 2010). The main problem with taking the final citations list in patents as the relevance assessments is that these citations usually are not a complete set of the relevant

existing patents. This situation arises since patent examiners typically stop searching for relevant patents once they have identified a list of relevant patents invalidating the patent invention in hand. Thus, it is very important, when using an evaluation metric for patent search or recall-oriented search in general, to be sure that this score metric will be consistent in evaluating different systems even if only a subset of the actual relevance set have been identified.

In this section, we explore the robustness of the PRES score where relevance judgements are not guaranteed to be complete, which is the case in realistic patent search situations (Roda et al., 2009). In the following analysis, we examine the robustness of PRES and compare it with that of MAP and recall.

5.5.1 Experimental setup

The same experimental data used for evaluating PRES, the set of 48 runs submitted to CLEF-IP 2010 and the relevance judgments, was used for the robustness experiments. In order to investigate the robustness of the evaluation metrics to incomplete relevance judgements, several versions of the relevance judgements *qrels* files were created by selecting different fractions of the judgements *f-qrels*. The original *qrels* provided by the track are assumed to be the full 100% *qrels*[20]; other versions representing fractions of the *qrels* were generated by selecting a certain fraction value (*f*) of the relevant documents for each topic in the topic set. In (Bompad et al., 2007), a similar setup was used, but with two main differences. The first is that here we focus on a recall-oriented patent retrieval task, where missing any relevant document in the assessments should be considered harmful for a fair evaluation of systems. The second difference lies in the

[20] Although of course this is actually known not be the case since exhaustive manual analysis of the collection has not been carried out.

nature of the data collections used in the studies. In (Bompad et al., 2007), TREC collections characterized by a high average number of relevant documents per topic were used. This situation allowed the study to test many *f-qrels*, where *f* ranged from 0.01 to 0.9. In the current experiments, the topics are a set of patent applications which are characterized by having a relatively low average number of relevant documents, which we noted before to be less than 6 relevant documents per topic on average. This low number does not allow such large variation in the values of *f*.

To conduct the robustness experiments, four fraction values of the *qrels* were used (*f* = 0.2, 0.4, 0.6, and 0.8). For each *f* value, three *f-qrels* were generated by selecting the fraction of relevant documents at random, hence the three versions are always different. This produced a set of 12 *f-qrels*. The objective was to compare the ranking of the runs according to each evaluation metric using these *f-qrels* against their ranking when using the full *qrels*. Kendall's tau correlation (Kendall, 1938) was used to measure the change in the ranking when relevance judgements are incomplete. The higher the correlation for smaller values of *f*, the more robust the metric is to the incompleteness of relevance judgements.

Two values of cut-offs were used, the first is the one reported in the CLEF-IP 2009 track itself, i.e. 1000 results for each topic. The second cut-off value is 100, which can be seen as a more realistic value for a patent retrieval task since this is the order of the number of documents typically checked for relevance by a patent examiner for each topic (Azzopardi et al., 2010).

5.5.2 Experimental results

Table 5-8 shows the Kendall tau correlation values for the three scores at different cut-offs and for the different subsets of the relevance set for each value of *f* (% *qrels*). Figure 5.12 and Figure 5.13 plot the worst-case values of the correlation for the three scores for cut-off values of

100 and 1000 respectively. Analysis of the results in Table 5-8 , Figure 5.12 and Figure 5.13 reveals several findings:

- MAP has a much lower Kendall tau correlation when compared to recall and PRES, especially for lower values of *f*. This result surprisingly shows that MAP, the most commonly used metric for patent retrieval evaluation, is the least reliable one when there is no guarantee of the completeness of the relevance judgements.
- Recall and PRES have nearly identical performance with incomplete judgements with slightly better performance to PRES for lower values of cut-off.

Table 5-8: Correlation between the ranking of 400 topics from 48 runs with different percentages of incomplete judgements and different cut-offs for MAP, recall, and PRES

		Cut-off = 100			**Cut-off = 1000**		
% *qrels*	**Sample**	**MAP**	**Recall**	**PRES**	**MAP**	**Recall**	**PRES**
20%	1	0.50	0.94	0.94	0.58	0.93	0.93
	2	0.71	0.88	0.92	0.70	0.92	0.89
	3	0.59	0.90	0.90	0.66	0.90	0.90
	avg.	0.60	0.90	0.92	0.65	0.92	0.91
40%	1	0.93	0.92	0.93	0.76	0.96	0.96
	2	0.71	0.93	0.93	0.87	0.95	0.93
	3	0.75	0.91	0.92	0.84	0.94	0.92
	avg.	0.79	0.92	0.93	0.82	0.95	0.94
60%	1	0.66	0.95	0.96	0.82	0.98	0.98
	2	0.66	0.95	0.97	0.82	0.98	0.97
	3	0.65	0.96	0.96	0.90	0.97	0.98
	avg.	0.66	0.95	0.97	0.85	0.98	0.97
80%	1	0.79	0.96	0.97	0.96	0.98	0.98
	2	0.94	0.97	0.99	0.94	0.97	0.98
	3	0.75	0.98	0.98	0.84	0.98	0.99
	avg.	0.83	0.97	0.98	0.91	0.98	0.98
100%	NA	1	1	1	1	1	1

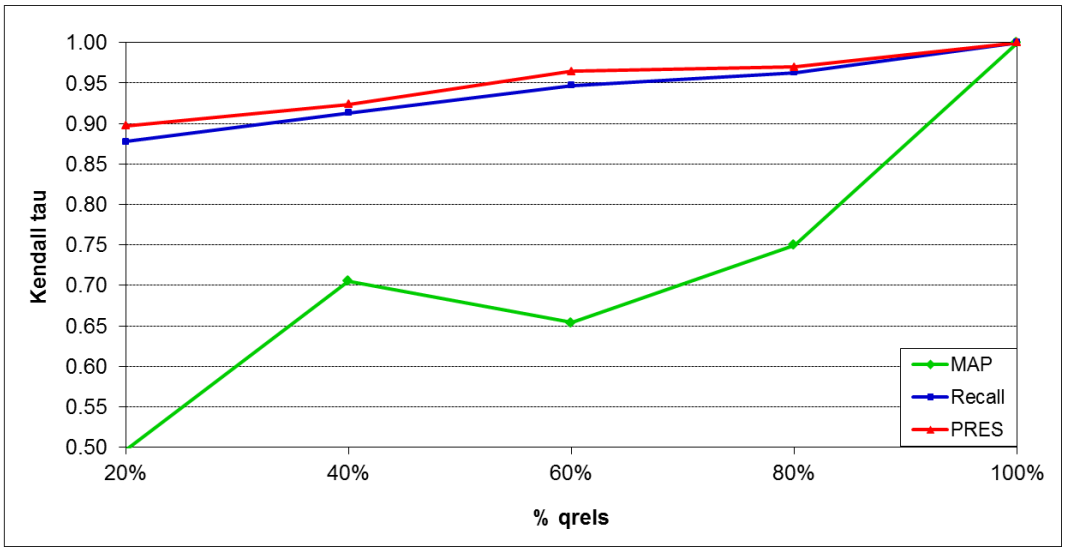


Figure 5.12: Lowest Kendall tau correlation values for MAP/recall/PRES for cut-off = 100 when relevance judgements are incomplete. %qrels = percentage of present qrels

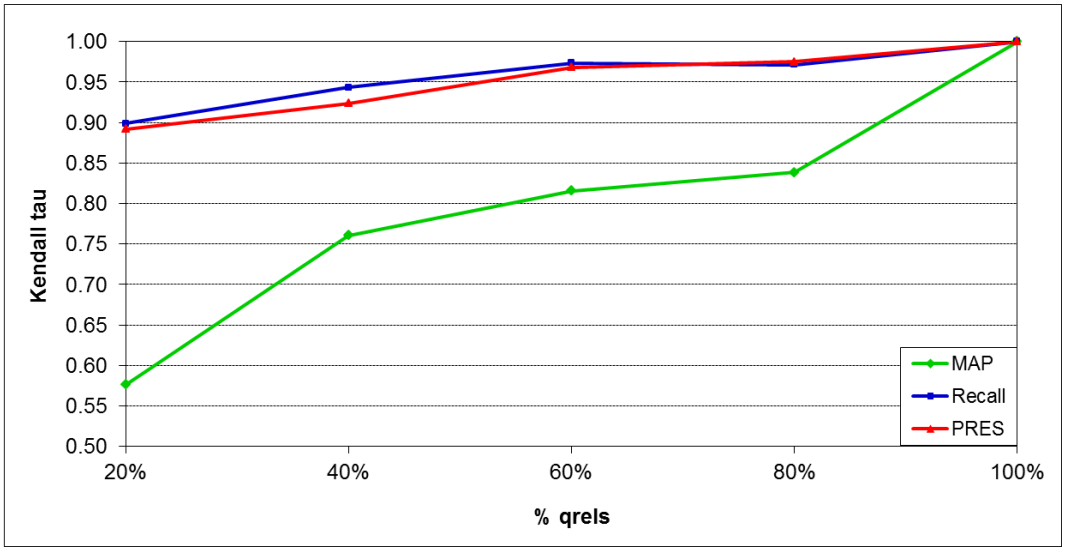


Figure 5.13: Lowest Kendall tau correlation values for MAP/recall/PRES for cut-off = 1000 when relevance judgements are incomplete. %qrels = percentage of present qrels

The study of Voorhees (Voorhees, 2001) determined rankings to be nearly equivalent if they have a Kendall tau correlation value of 0.9 or more, and to have a noticeable difference for Kendall tau correlation less than 0.8. According to the results found in our investigation, recall and PRES will have an equivalent ranking for systems even with only 20% of the relevance

judgements. However, MAP may have a noticeable change of system ranking even if only 20% of the judgements are missing.

A drop in the curve of correlation of MAP for cut-off of 100 can be seen in Figure 5.12 when % *qrels* = 60%. This may arise due to the randomness used in selecting the fraction of relevant document from the *qrels*.

Although PRES and recall have similar performance with the incomplete judgements, the metrics cannot be claimed to be the same. The experiments here test only the stability of the metrics with respect to completeness of relevance data. However, additional factors should be taken into consideration when considering the suitability of an evaluation metric, in this case the metric's ability to distinguish between the performance of different systems in a fair way. Bearing all these factors in mind, PRES can be considered as the more suitable evaluation metric for patent retrieval since it has been shown to have a greater ability to rank systems in a recall-oriented IR environment.

5.6 Summary

This chapter has addressed the second set of research questions in this thesis about the evaluation of the recall-oriented IR tasks represented in patent search. These questions are:

- What features of performance should be measured for a recall-oriented IR application such as patent search?
- Are the current evaluation methods and metrics, which are currently being used for patent search and recall-oriented IR applications, appropriate to evaluate this type of application?
- Can a more meaningful metric be developed for these tasks?

- How can the suitability and reliability of such a novel metric be established to ensure that it reflects real system performance and ensure that it is robust to different system variables?

These questions have been addressed through a study on the evaluation of recall-oriented applications and the proposal, development, testing and analysis of a novel evaluation metric called PRES designed specifically for these applications. The score is a refinement of normalized recall. PRES reflects system recall combined with the quality of ranking of retrieved relevant documents within the maximum numbers of documents to be checked by a user. It has been tested and compared to the most widely used IR metrics for a patent retrieval task. Illustrative samples and real data examples demonstrated the effectiveness of the new score. The PRES value varies from R to nR^2/N_{max} according to the average quality of ranking of relevant documents; hence it can be seen as a function of system recall, ranking of relevant documents, and the maximum number of documents to be checked by a user (which directly affects the recall and relative ranking).

A study of the robustness of the PRES, MAP and recall scores has been presented. The aim of the study was to test the consistency of the performance of these three evaluation metrics which are currently used for patent retrieval evaluation when the relevance judgement set is incomplete. Fractional values of the *qrels* with different samples were used to conduct the experiments. Kendall tau correlation was used to measure the consistency of the ranking of systems. Results show that the most commonly used score for evaluating patent retrieval, MAP, is the least stable of these evaluation metrics, since it shows the least consistency in ranking different runs when the relevance judgements are incomplete. PRES and recall both have very robust performance even when only small portions of the relevant judgements are available.

Considering the strong performance of PRES for evaluating recall-oriented IR applications, in addition to its high robustness; PRES can be recommended as a standard score metric for evaluating recall-oriented IR applications, especially patent retrieval.

Chapter 6

Multilingual IR for Patent Retrieval

The need to search multilingual content in order to satisfy the user's information need is a major challenge in various IR environments. The international nature of patents means that search typically requires a multilingual search component to complete the search task, since the objective in patent search is to find all relevant documents. These documents often come from various patent offices working in different languages. One special concern in the multilingual search for patents is the huge amount of text that needs to be translated for the search process, since the query often takes the form of a complete patent application, which usually needs to be fully translated into each document language. In recent years machine translation (MT) has become established as the dominant technique for translation in cross-language information retrieval (CLIR). This has largely come about due to the increased availability of high quality MT systems, which usually achieve better retrieval effectiveness than dictionary-based translation (DBT) methods (Levow and Oard, 2005). Current statistical MT (SMT) methods lend themselves well to patent translation since patent offices often publish patent content with parallel translations which can be used for MT system training. For example, patents from the EPO contain parallel content in English, French, and German. However, translation using MT is

time consuming and resource intensive for cross-language patent retrieval (CLPR), because of the very long patent topics can run to tens of pages. In addition, a very large parallel training corpus is usually needed to achieve acceptable translation quality (Stroppa and Way, 2006; Wang et al., 2006). The translation time and the large resources required for CLPR using standard MT methods have not received significant attention to date. Besides, some language pairs have very limited suitable training resources available, meaning that it is not possible to train an effective SMT system for these language pairs leading to low translation quality, and consequentially usually low retrieval effectiveness.

In this part of the study, a novel technique for adapting MT for CLIR is presented which addresses the high computational cost and resource requirements of MT for CLPR. The technique is demonstrated to be more than 20 times faster than standard MT techniques in both the training and decoding phases when tested on the patent search task from the CLEF-IP 2010. Retrieval effectiveness using the new translation method is shown to be statistically indistinguishable from results obtained using standard MT. Furthermore, the retrieval effectiveness is found to be statistically significantly better than standard MT techniques when only a small amount of training data is used to train the MT system. In addition, the new MT approach enables document translation for CLPR in a practical amount of time with a resulting improvement in retrieval effectiveness. While document translation has been shown to be useful in CLIR, its application using current MT approaches has been impractical due to the very large amount of time required for translation of patent documents (Chen and Gey, 2004a; Parton et al., 2008).

6.1 Background

6.1.1 Cross-language information retrieval

CLIR is concerned with searching a collection of documents that are in a different language from the user's query (Nie, 2010). The key challenge of CLIR is crossing the language barrier between the query and the documents. Two main approaches are available for CLIR: translation of documents to the query language prior to the search stage or query translation to the document language at search time (Parton et al., 2008; Oard, 1998). In practice the latter is the most common, since it is more practical in operational systems; moreover it provides more flexibility to expand the query with multiple possible translations for each term to help prevent problems that can arise from incorrect translations (Darwish and Oard, 2003; Wang and Oard, 2006).

Considering the translation stage itself, two common techniques have been used for query translation in CLIR: bilingual dictionaries and MT systems (Oard and Diekema, 1998). Bilingual dictionaries are sets of entries of words in one language and possible translations in the other language. MT systems are optimised for translating whole sentences from one language to another, while preserving the target translated sentence in a correct morphological, semantic, and syntactic form. MT has become the most commonly used technique for translation in CLIR in recent years due to the increasing availability of high quality MT systems. Much CLIR research now uses freely available online tools such as Google translate, Bing translate[21], and Yahoo Babel Fish[22]. Furthermore, some open source SMT libraries are also available freely for research purposes, e.g., MaTrEx (Stroppa and Way, 2006) and Moses (Koehn et al., 2007). There are

[21] http://www.microsofttranslator.com/

[22] http://babelfish.yahoo.com/

other translation techniques for CLIR such as corpus-based translation which applies statistical analysis of words or phrases in parallel or comparable corpora in different languages to obtain probabilities of translations, but these have not received widespread attention due to the general lack of suitable data sources (Sheridan and Ballerini, 1996; Oard and Diekema, 1998).

6.1.2 Cross language patent retrieval

As discussed in Chapter 3, CLPR has always featured as one of the tasks in existing patent evaluation campaigns (Fujii et al., 2004; Roda et al., 2009; Piroi, 2010). The typical procedure adopted is to translate the query into the target language of the collection using one of the available free MT systems, and then perform search in the document language. Thus, this research has treated the translation stage as a black box without any control over the translation process. In addition, little attention has been directed towards the time taken for the translation process. This is a significant issue for the very large topics typically encountered in patent search, for which the query translation time can actually be significantly longer than the search time. The valuable feature of the presence of parallel translations for a large number of international patents is generally neglected by most CLPR researchers. For example, the parallel translations in the EPO patents and the WIPO patents. This parallel translation data can be used to build translation models for multilingual search of patents.

One recent research study (Jochim et al., 2010) utilized these parallel corpora from EPO patents in order to translate queries for patent search. This research used the data to build domain-specific translation dictionaries rather than using this data for MT training. In this research the word alignment tool Giza++ was used to build a word-to-word translation dictionary for the language pairs English-French and English-German. The highest probability translation for each word was then used in the translation process regardless of the context of the word

being translated. The reported results of this study in (Jochim et al., 2010) for CLPR are considerably lower than those reported when general MT is used (Prior, 2010) when both techniques were tested on the same data collection. This observation shows that MT is a more effective technique for translation in CLIR than simple dictionary-based approaches; this is likely to be especially true for patent search when the queries are much longer and context can help in selecting appropriate translations.

6.1.3 Related work

There is some reported work on adapting the translation stage in CLIR for the purpose of achieving high retrieval effectiveness in search. In the TREC-2002 Arabic/English CLIR task, a bilingual dictionary was built by aligning Arabic and English stems and was provided to the task participants (Oard and Gey, 2002). The idea of aligning stems was based on the fact that only stems are valuable in IR, which is similar to what is proposed in our study. However, in the TREC 2002 work, stop words were not filtered out which led to their presence as possible translations in the bilingual dictionary. Similar work was presented in the same track by Franz and McCarley (Franz and McCarley, 2002), where they aligned stems instead of words for what they called the convolutional model for CLIR, which integrates the retrieval and translation models into one model. The approach used SMT-like technology, but only for predicting the "bag of words" from which an English translation of a given document can be composed rather than for actually generating translations. This approach achieved better retrieval results than when standard MT was used for translation.

Other research has been reported which aims to improve the quality of the translations for CLIR. The majority of this work has focused on finding better candidate translations using dictionary-based translation (DBT) approaches (Darwish and Oard, 2003; Levow et al, 2005;

Wang and Oard, 2006). In (Darwish and Oard, 2003), a probabilistic structured query method was adopted to allow searching with weighted candidate translations of query terms. This approach showed its effectiveness over using only the highest probable translation. A more advanced approach was introduced later (Wang and Oard, 2006) that combined bidirectional translation and synonyms to generate better dictionary-based candidate translations achieving higher retrieval effectiveness over results reported in (Darwish and Oard, 2003). Little further work has been reported on DBT for CLIR in recent years where the MT approach has come to dominate due to the significant improvement of the quality of MT systems. For CLIR tasks, MT is straightforward to use and usually achieves high retrieval effectiveness.

In this chapter, we adapt the current "high quality" MT systems for CLIR to be more efficient in computational and resource requirements, while maintaining its retrieval effectiveness and demonstrate its utility on CLPR.

6.2 A Novel Translation Method for CLIR Based on MT

This section presents the proposed novel method for MT for CLIR. The objective is to achieve high retrieval effectiveness for CLPR using lower resource and computational requirements than are needed by standard MT systems while maintaining or improving search effectiveness.

6.2.1 Basic concept

The basic idea of the new approach is to train an MT system for translation of topics or documents in CLIR using training data pre-processed for IR. The pre-processing of IR data uses the standard stages performed by most of IR systems; specifically: case folding, stop word removal, and stemming (Manning et al., 2009). These three operations aim to improve retrieval efficiency and effectiveness by matching different surface forms of words. While these are

standard processes in IR, for MT, applying these operations within the MT process would be destructive to the quality of the translated output. For example, the translated sentence "*he are an great idea to applied stem by information retrieving*" instead of "*It is a great idea to apply stemming in information retrieval*" would be considered a very bad translation from an MT perspective. However, from an IR perspective this output is fine since it contains all the information needed for the retrieval process. Both sentences will appear the same after IR pre-processing as "*great idea appli stem informat retriev*".

Our hypothesis in this chapter is that training an MT system using corpora pre-processed for IR can lead to similar or improved translated text from the IR perspective, which can consequently lead to better retrieval effectiveness. In addition, the training of the MT system is expected to be much faster and more efficient since a large proportion of the training text represented by the stop words will be removed and the remainder will be normalized creating a smaller vocabulary, and that a smaller processed training corpus can be as effective as a larger unprocessed one for translation in CLIR.

6.2.2 MT training and decoding

Figure 6.1 presents the workflow of the proposed CLIR system. The upper part represents the MT training phase which produces the translation model used for the translation step in the CLIR. The new "Text Processing" step introduced for both languages in the parallel corpus works by applying the standard IR pre-processing for documents and queries. The resulting translation model is in the "Processed" form, where words are in their stemmed form and no stop words are present. For consistency, the terms "Processed" and "Text Processing" in the remainder of this chapter refers to "case folding", "stop word removal" and "stemming".

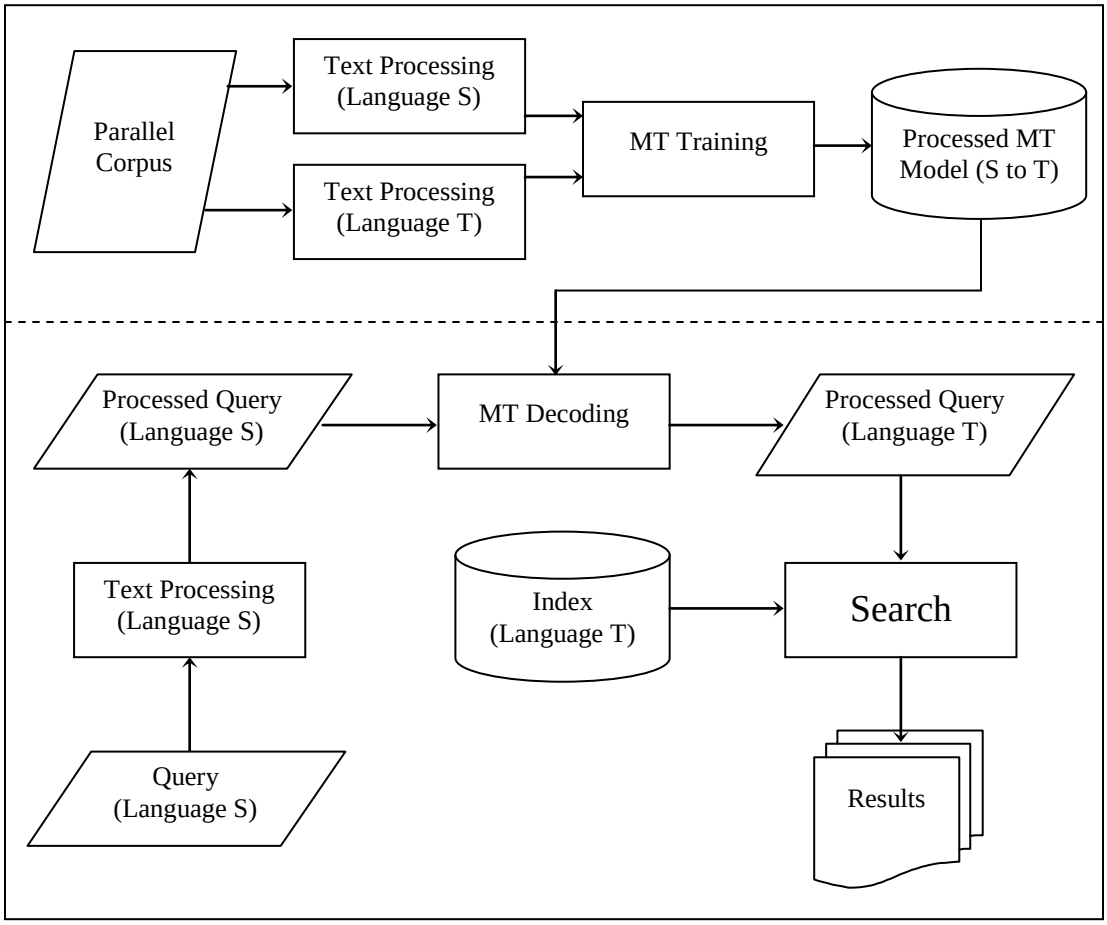


Figure 6.1: Workflow of the proposed CLIR system

For the query translation in the CLIR search when using MT, a query in source "S" language is translated into target "T" language; the translated query is then processed in language "T" to be used for search (Nie, 2010). Actually, when using MT for CLIR, longer queries are preferable since they tend to be more grammatical, therefore better translation can be achieved using an MT system taking context into account, and consequently better retrieval effectiveness (Gao et al., 2001). The novel translation approach introduced here is shown in the lower part of Figure 6.1. It can be seen that the "Text Processing" step has been moved to be a step prior to translation instead of a posterior step in the standard CLIR workflow. Therefore, the processing is applied to the source language query which produces a much shorter input with a reduced vocabulary to be

translated using the processed MT model. The output from the translation process is in its processed form, and therefore no additional processing of the query is required. This query is used directly to search the index of documents and produce a list of retrieval results. Although for the novel CLIR system in Figure 6.1 the processing to query is applied prior to the translation, the context of the query text is still maintained, since all terms exist in order, but in stemmed form and with no stop words among them.

6.3 Experimental Investigation

The following experimental investigation is designed to test three dimensions of the proposed approach. The first is to examine the effect of processing the words before the MT step on the quality of the translated text, which will be reflected in the retrieval effectiveness. The second is to investigate the efficiency of the proposed translation process in terms of the computational requirements for the training and decoding (translation) phases when compared to translation using standard MT. However, more emphasis is given to the decoding time for query translation since it is the online processing time for translating the query which is generally more significant to the user. The third dimension examines the effect of using a limited amount of training data on the retrieval effectiveness.

The non-English topics in the CLEF-IP 2010 task are used for the experimentation. Retrieval effectiveness is measured using MAP and PRES. In our analysis, we focus on PRES since it is designed for measuring retrieval effectiveness in patent search. We include MAP results here in order to allow direct comparison of our work with previously reported results for this task that did not report PRES (Roda et al., 2009; Jochim et al., 2010). Significance is tested using a 2-tailed t-test and Wilcoxon tests with *p*-value 0.05 (Hull, 1993).

6.3.1 Test data

The cross language search task in CLEF-IP 2010 was adapted for use in our experiments (Piroi, 2010). The collection consists of 1.35M patents from the EPO with 69% of them totally in English and 31% in German or French. However, as noted earlier, the German and French patents have some sections manually translated into English, including the patent title, abstract, and claims. For initial experiments, only the English text of the collection was indexed to create an index of only the English content of documents, which was the same approach we used in our contributions in CLEF-IP tasks (Magdy et al., 2009; Magdy and Jones, 2010c). The CLEF-IP 2010 track provided two sets of topics; 300 training topics of which 89 were German, 15 were French, and the remainder were English; and 2000 test topics of which 520 were German, 134 were French, and the rest were English (Piroi, 2010). For our experiments, the 89 German training topics and the 134 French test topics were selected to have a close number of topics for each language (520 German topics would be too many and 15 French topics would be too few for the experiments).

Regarding the parallel corpus used for training the MT systems, the same parallel corpus we extracted from the EPO patents in the CLEF-IP 2010 collection was utilized, see Section 4.4.2. This corpus contained more than 8M parallel sentences in three languages: English, French, and German. The sentences used were processed by removing digits, punctuation, and applying conversion to lower case.

For the stemming and stop word removal in our experiments in this chapter, the same setup used in Section 4.4.2 for processing the parallel sentences before generating the SynSets was used. The Snowball toolkit was used for stemming, and the same stop word lists for English, French, and German were used here.

6.3.2 The MaTrEx MT system

The MT experiments were performed using the MaTrEx (Machine Translation using Examples) MT system developed by the machine translation group at Dublin City University (Stroppa and Way, 2006).

The default configuration built within the MaTrEx system was used for the experiments. The configuration of the workflow of the MT system was as follows:

1. Building Language Model: a standard trigram LM was built from the target language corpus.
2. Building translation model was done through:
    a. Word alignment using GIZA++, which learns the translation tables of IBM model-4 for word alignment (Och and Ney, 2003).
    b. Build lexical translation tables for bidirectional translation for both languages.
    c. Build aligned statistical phrases, using the *grow-diag-final* algorithm to produce the final phrase aligned table (Stroppa and Way, 2006).
    d. Build lexicalized reordering model.
    e. Build the generation models, where forward and backward probabilities are computed.

The generated language model and translation model were used in the decoding step to produce the highest probable translation based on the context of the sentence.

6.3.3 Baselines construction

The same query formulation method presented from our participation in CLEF-IP 2010 was applied here, see Section 4.2. Queries were formulated from the description section of the patent application after filtering out all terms in the description section which appeared only once, in addition to the bigrams that appeared more than three times in the remaining sections combined, namely: title, abstract, and claims sections.

Two baseline runs were carried out for each language: the first one used the 8M extracted sentences to train the MaTrEx MT system (Stroppa and Way, 2007) to create two translation models (French→English) and (German→English) using the standard MT training technique. The translation models were then used to translate the 89 German topics and 134 French topics into English, prior to searching the collection. The second baseline used Google translate to translate the German and French topics into English then search the collection, which is the same as our submitted runs in CLEF-IP 2010.

Table 6-1 shows the MAP and PRES values for each of the baselines for the French and German topics. From these results it can be seen that for the French topics, the Google and MaTrEx MT systems achieved similar retrieval effectiveness. However for German topics Google translate achieved lower performance for both MAP and PRES. This can be attributed to the many unusual word compounds found in the German text that require a training corpus from a similar domain in order to be translated effectively.

Regarding the Google baseline, it should be noted that the result of the German topics in Table 6-1 is different than that reported for our participation in CLEF-IP 2010 in Table 4-4, since the result reported here is for the 89 German training topics, while the result reported in Table 4-4 is for the 512 German test topic submitted to the track. However, the results for the French topic are the same in both tables, since both are for the 134 French test topics.

Table 6-1: Baseline runs for the 89 German topics and 134 French topics

	French		**German**	
	MAP	PRES	MAP	PRES
Google	0.087	0.413	0.067	0.466
MaTrEx	0.085	0.413	0.075	0.487

Although Google translate is fast since it is powered over a cluster of very powerful machines, the translation for our experiments took several days since it only allows the translation of a limited number of sentences at a time. For MaTrEx, it was found that the average translation time for the French patent topics (contains 7,058 words on average) was 31 mins and for the German patent topics (contains 3,571 words on average) was 12 mins on a server machine (Intel Xeon quad-core processor, 2.83GHz, 12MB cache, and 32GB RAM). However, the average search time using all the translated text as a query was 42 secs for French topics and 14 secs for German topics on a desktop machine (Intel Core2Due, 3GHz, 6MB cache, 3GB RAM). This highlights the importance of developing faster translation techniques for patent search topics.

6.3.4 Dictionary-based translation baselines

Although MT is currently the most commonly used technique for translation in CLIR and is the main focus of this chapter, we also report results for an investigation of dictionary-based translation (DBT) of search queries for two reasons. The computational and development costs of DBT are significantly lower than those for MT, while good CLIR results have been reported for DBT methods (Darwish and Oard, 2003; Levow et al, 2005; Wang and Oard, 2006). Thus, it is interesting to know if comparable retrieval effectiveness can be achieved using DBT for CLIR patent search. Furthermore, very limited investigation has been reported for this technique for the CLEF-IP prior-art patent search task to date. Only straightforward usage of the technique was reported in (Jochim et al., 2010) where poor query formulation for patent topics led to low retrieval effectiveness compared to a monolingual baseline.

In our experiments, DBT was applied to the French and German topics using Giza++. The experiments here tested use of the highest probable translation as in (Jochim et al., 2010) and

using multiple weighted translations as in the approaches described in (Darwish and Oard, 2003; Wang and Oard, 2006). The experiments using the approaches described in (Darwish and Oard, 2003; Wang and Oard, 2006) showed that working with large numbers of low probability translations yields low effectiveness and higher computational cost. Based on this, we imposed a cumulative probability threshold for translations of 0.99 as suggested in (Wang and Oard, 2006) and 0.6 which achieved the best results in (Darwish and Oard, 2003). However, both cumulative probabilities led to some words with large numbers of possible translations in some cases consisting of thousands of terms, which can correspond to the vague and ambiguous meaning of some terms in the patent text. Therefore, to reduce the computational cost for our experiments, we used a cumulative probability of 0.99 and also applied a hard threshold to allow not more than *N* candidate translations for a given term. We tested some values for *N* between 2 to 10. The Indri "wsyn" operator was used to allow the presence of multiple weighted alternatives for each term in the query, see Appendix B.

Table 6-2: Retrieval effectiveness when using DBT for the 89 German topics and 134 French topics

	French		**German**	
Translation/word	MAP	PRES	MAP	PRES
1	0.078	0.384	0.061	0.449
2	0.066	0.379	0.055	0.453
3	0.060	0.362	0.049	0.428
5	0.056	0.338	0.045	0.407
10	0.038	0.288	0.045	0.357

Table 6-2 reports the results of using DBT for CLPR using the same data used for training the MaTrEx MT system (8M parallel corpora). All the results were found to be statistically worse than using the MaTrEx MT, which confirms that MT is a better method for translation in CLPR. Surprisingly, it was found that adding more alternative translations in the query leads always to worse results, which contradicts many reported results in CLIR, including (Darwish

and Oard, 2003; Wang and Oard, 2006). In addition, we found that the search time significantly increases when using multiple translations since the query length increases significantly. The average time of search when N=10 for the French topics was 19 mins (28 times slower than when using one translation/term), and for the German topics was 10 mins (45 times slower). These results show the advantages of using MT in CLPR, where it is better at selecting the correct translation based on the rich context in patent topics rather than using weighted candidate translations, which creates more ambiguity to the long query. DBT methods were reported to be effective for standard ad hoc CLIR tasks, where the queries are typically a small number of words that lack context to assist in selecting the proper translation of a term in an MT system. This may be why adding multiple weighted translations for query terms of these tasks showed effectiveness in the retrieval results.

6.4 Experiments with the New CLIR MT Approach

The same training dataset of parallel sentences was used to train the MaTrEx MT system again, but after pre-processing the data by removing stop words and applying stemming ("processed MT"). This was then compared to the standard MT system without pre-processing the data ("ordinary MT"). Two additional MT training setups were also investigated to understand the effect of each pre-processing stage in the processed MT. The system "stemmed MT" trained the MT system using the same data but after applying only stemming without removing the stop words; and "stopped MT" applied the training using the parallel corpus after removing stop words but without applying stemming to the words. The idea behind these two additional MT training sets was to investigate the effect of each of stemming and stop word removal individually on the efficiency and effectiveness of the MT in the CLPR task. Table 6-3 summarizes the MT training setups used in our experimentation.

Table 6-3: Pre-processing applied to different MT systems in our experimentation

MT	Pre-processing applied to the training corpus
Ordinary MT	Punctuation and digits filtered out, and text in lower case
Stemmed MT	Ordinary MT + applying stemming to text
Stopped MT	Ordinary MT + filtering out stop words
Processed MT	Ordinary MT + applying stemming to text + filtering out stop words

A number of subsets of the 8M training data set of different sizes were selected at random and used to train the MT system to explore the behaviour of the MT systems and CLPR performance when less training examples are available, which will be the case in practice for some language pairs or translation environments. In addition to the full 8M training set, subsets of the following sizes: 800k, 80k, 8k and 2k sentences were extracted at random from the full corpus and used to train the MT systems for additional experiments.

6.4.1 Experimental results

Figure 6.2(a) and Figure 6.2(b) present the retrieval effectiveness when translating the French and German topics using the ordinary MT system compared to using the alternative MT systems for different sizes of training data, evaluated using MAP and PRES. It can be seen that the difference in retrieval effectiveness using these translation methods is not significant compared to each other for almost all training sizes. However, with small size training sets (2k), it is found that the processed MT and the stemmed MT achieved significantly better retrieval effectiveness than the ordinary MT and the stopped MT when compared using PRES for both query languages. In addition, for the French topics when using processed MT, results remain statistically indistinguishable from Google translate for training sizes 8M, 800k, and 80k. However, for the ordinary and other MT systems, the 80k training set translation led to a retrieval result that is

statistically worse than Google translate when compared by PRES. These results highlight the effectiveness of stemming on MT in CLPR when smaller sizes of training data are available.

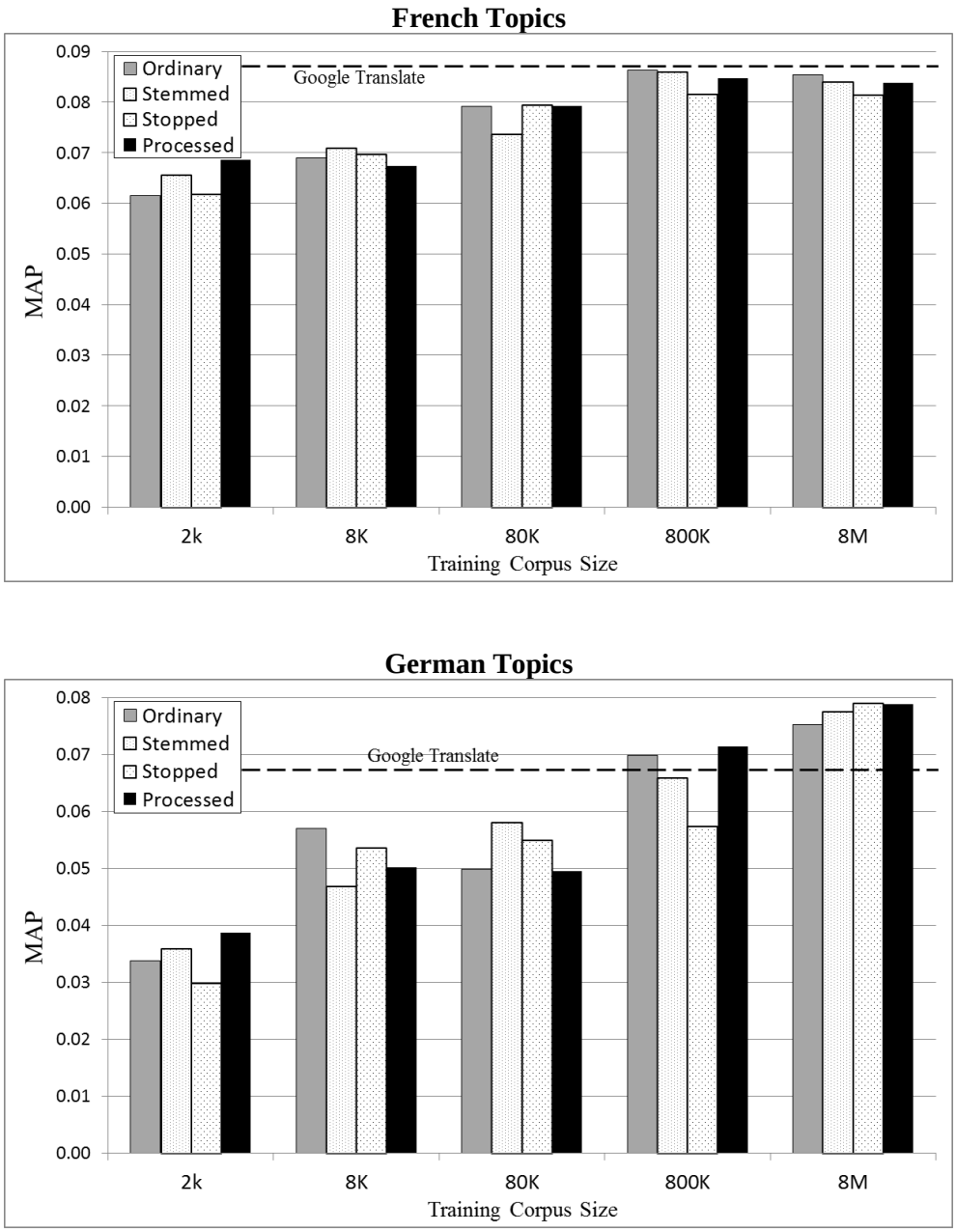
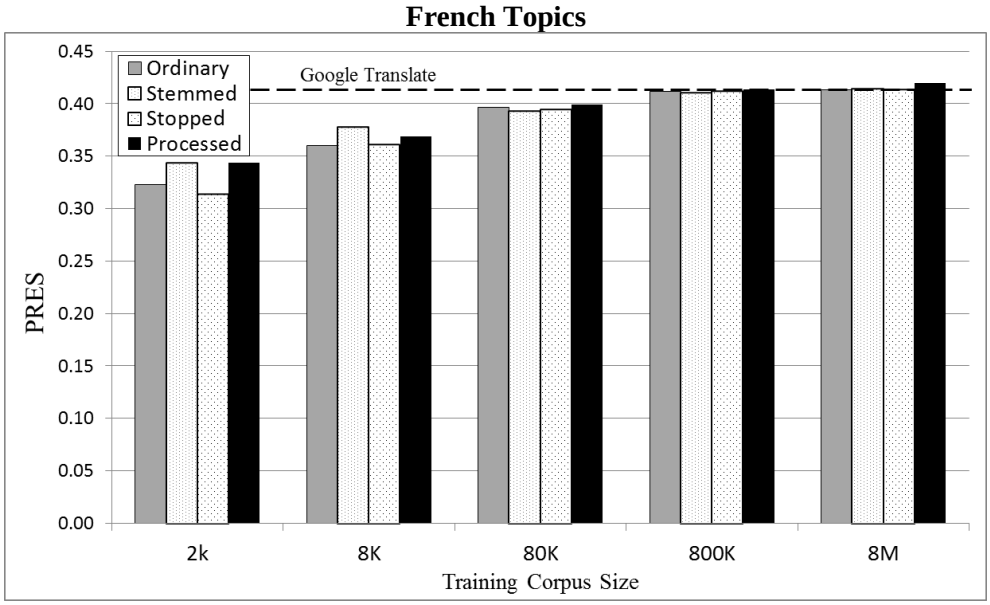


Figure 6.2(a): MAP for French and German topics when using the four MT systems



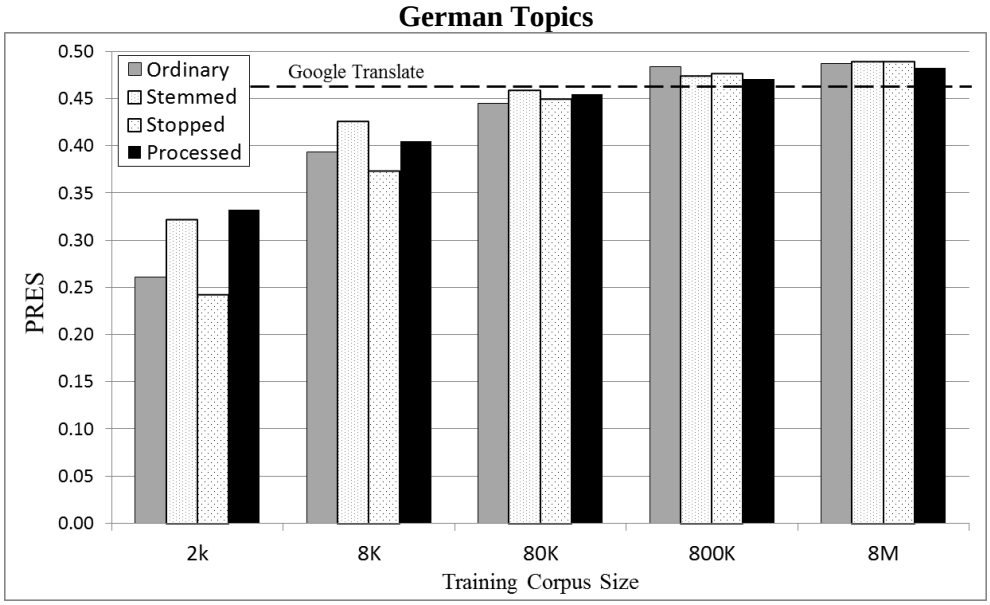


Figure 6.2(b): PRES for French and German topics when using the four MT systems

Figure 6.2: Retrieval effectiveness for French and German topics compared when using ordinary MT, stemmed MT, stopped MT, and processed MT for the cross language patent search task

One of the main reasons for these results is the presence of out-of-vocabulary (OOV) terms when attempting to translate words which do not appear in the MT training data. To explore the issue of OOV terms while translating the patent topics, the OOV percentage for each method is reported in Figure 6.3. The OOV rate of the stemmed MT is the same as that of the processed MT, and the OOV rate for the stopped MT is the same as that of the ordinary MT. Hence, Figure 6.3 reports only the OOV rates for the ordinary and processed MT systems. It can be seen that the stemming helps to overcome some of the OOV terms, which leads to the presence of a translation. Also, it can be seen that for small size training sets, the ordinary translation approach suffers from a large percentage of OOVs, while the processed MT and stemmed MT systems overcome part of this problem. The German topics suffer from higher OOV than the French ones due to the presence of word compounds in German.

The second main benefit of the new approach to translation is shown clearly in Figure 6.4, which compares the average decoding time required to translate a patent topic into English using these different MT systems. It can be seen that the processed MT and stopped MT systems are at least 5 times faster than the ordinary MT and stemmed MT systems when using the same training parallel corpus. In addition, with smaller sizes training data sets, the speed of decoding using the MT systems with no stop words reaches up to 23 times faster than the MT systems which include stop words. In fact, the decoding time needed for the processed and stopped MT systems when it is trained with 8M parallel sentence is less than the decoding time required for the ordinary system when it is trained with only 2k examples. This result demonstrates the strong impact of filtering out stop words before the translation process on the translation speed. Figure 6.4 also shows that the fastest MT system is the stopped MT, which is slightly faster than the processed MT; and the slowest MT system is the stemmed MT, which is slightly slower than the ordinary MT. This result shows that stemming leads to slightly slowing down the translation speed.

However, this is not comparable to the effect of stop word removal which leads to superior speeding up of the translation time.

Similar results to those for decoding time shown in Figure 6.4 were obtained for the training time of the MT systems. The training time for the processed MT and stopped MT systems was 5 to 15 times faster than the training time for the ordinary MT system.

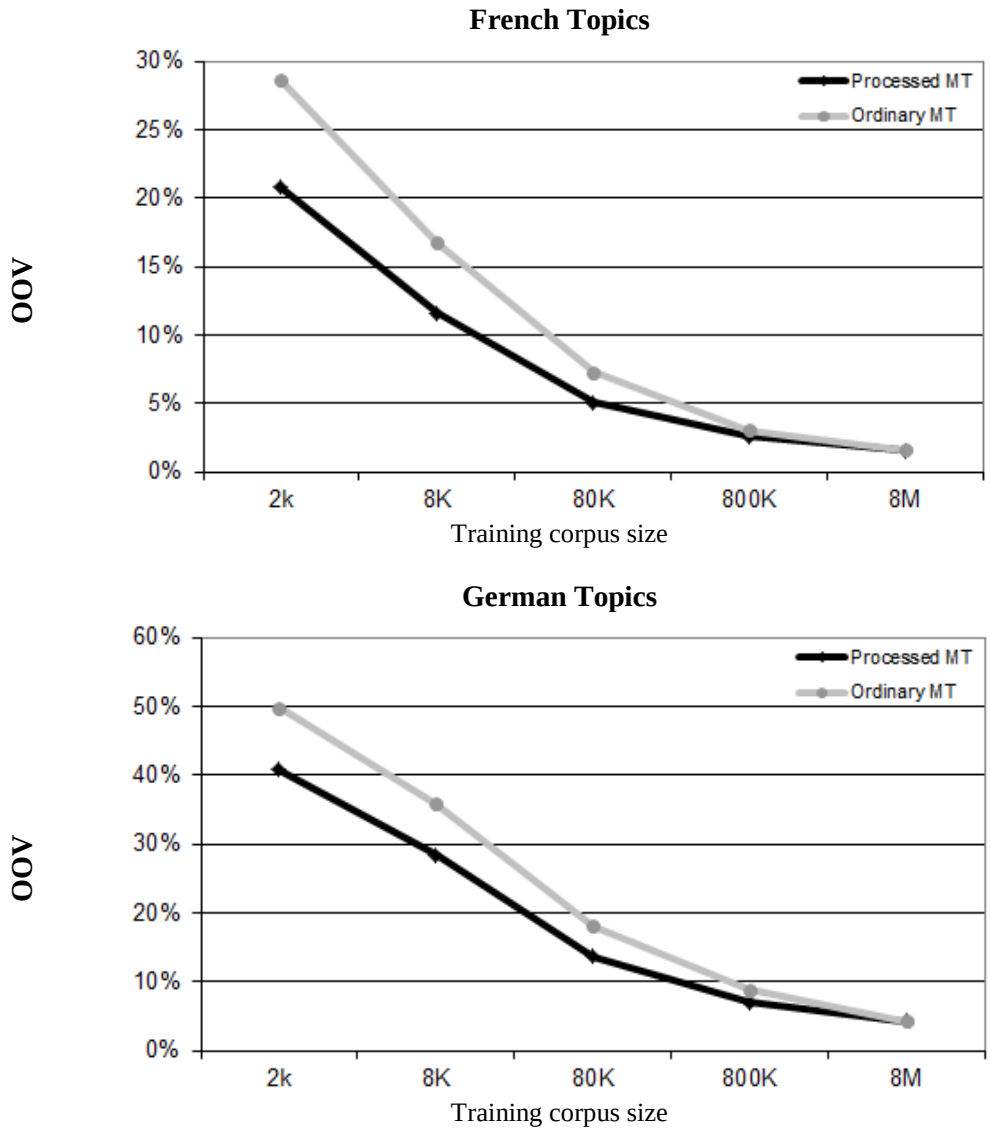


Figure 6.3: Out of vocabulary (OOV) rates of French and German topics with different sizes of training data sets compared for processed and ordinary MT

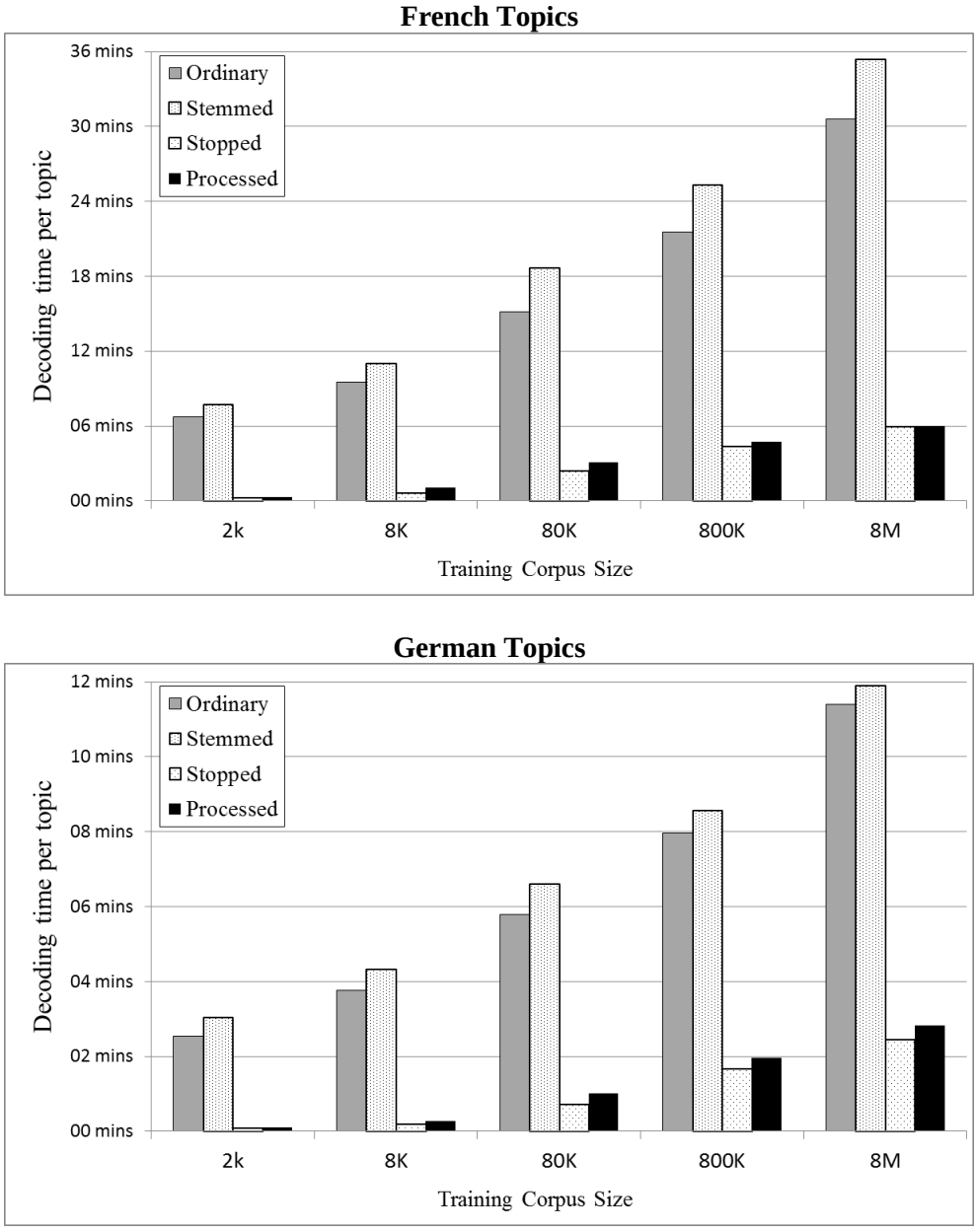


Figure 6.4: Average decoding time for translating French and German topics with different sizes of training data sets compared for ordinary MT, stemmed MT, stopped MT, and processed MT

All the values of the results reported in Figure 6.2, Figure 6.3, and Figure 6.4 for the ordinary and processed MT systems are presented in Table 6-4 for a precise comparison between the MT systems. The significant changes in the retrieval effectiveness between the systems and Google translate are marked in the table.

Table 6-4: Retrieval effectiveness, OOV, and decoding time for French and German topics compared when using ordinary MT compared processed MT for the cross language patent search task. Underlined values indicate that the result is indistinguishable from Google translate, and '*' indicates that processed MT is statistically better than ordinary MT

		Google	2k	8K	80K	800K	8M	
MAP	Processed MT	0.087	0.069	0.067	0.079	0.085	0.084	French Topics
	Ordinary MT		0.062	0.069	0.079	0.086	0.085	
PRES	Processed MT	0.413	0.343*	0.369	0.399	0.414	0.419	
	Ordinary MT		0.323	0.360	0.396	0.412	0.413	
OOV (%)	Processed MT	NA	20.7%	11.6%	5.0%	2.6%	1.6%	
	Ordinary MT		28.6%	16.8%	7.3%	3.0%	1.6%	
Decoding time (*mm:ss*)	Processed MT	NA	00:19	01:05	03:06	04:44	06:03	
	Ordinary MT		06:43	09:30	15:09	21:31	30:35	
MAP	Processed MT	0.067	0.039	0.050	0.050	0.071	0.079	German Topics
	Ordinary MT		0.034	0.057	0.050	0.070	0.075	
PRES	Processed MT	0.466	0.332*	0.405	0.455	0.471	0.483	
	Ordinary MT		0.260	0.394	0.445	0.484	0.487	
OOV (%)	Processed MT	NA	40.7%	28.3%	13.6%	7.0%	4.2%	
	Ordinary MT		49.8%	35.8%	18.0%	8.9%	4.2%	
Decoding time (*mm:ss*)	Processed MT	NA	00:07	00:17	01:01	01:58	02:49	
	Ordinary MT		02:33	03:46	05:47	07:58	11:24	

6.4.2 Discussion

The results in this section lead to two main findings as follows:

1. Stemming the text before translation leads to improved retrieval effectiveness when only a limited parallel corpus is available to train the MT system. However, the effect of stemming on retrieval for large training corpora is not significant. In addition, stemming leads to a slight slowing down of the translation speed, which was found to be between 10%-20% slower than when no stemming is applied.
2. Stop word removal before translation leads to a large speeding up of the translation system for any size of training data without having a significant effect on the retrieval effectiveness.

These two findings show the importance of applying both stemming and stop word removal together, which is the "processed MT" approach, which achieves both effectiveness and efficiency in the translation and retrieval processes in CLPR.

The effect of stemming was analysed by checking the OOV rates to understand the reason behind the improved retrieval effectiveness for limited training data, which showed that stemming overcomes a significant proportion of the OOV terms that will not be translated if no stemming is applied. For example, if only the word *played* appeared in the training data as the surface form for the term *play*, this means that any other form of the word will not be translated if it appears in the sentences to be decoded by the MT system, such as: *play*, *plays*, *playing*. When applying stemming, all these terms will be normalized to the term *play*, and will be translated regardless of the surface form that appears in the text to be translated.

Regarding the processing time when applying stemming which appears to be slightly slower than when no stemming is applied, this is explained by the translation tables created for the stemmed terms which are expected to be larger than for words, since the entries of the words:

play, *played*, *plays*, and *playing* will be combined to only one entry which is *play*. This creates some additional confusion for the MT system to select the proper translation based on the context, which requires additional time. The positive thing is that this additional time was found to not be significant.

For the effect of stop word removal, removing the stop words from the text reduces the amount of text to be translated by nearly half. However, the gain in speed for the translation process is much more than the double (5 to 23 times). The reason for this comes from the special nature of stop words, where the MT takes a longer time to translate them in order to select the proper translation in the proper position, since they are the most confusing terms to be translated by an MT system. This arises due to the wide variation in the use and behaviour of pronouns between languages.

One of the observations from the results reported in Figure 6.4 is the difference in the average translation time for a French patent compared to that of a German patent. This arises from the length of the patents, where the French patents are nearly double the length of the German patents on average because of the compounds in German, which also leads to a higher percentage of OOV terms in the German-English translation process that speeds up the translation since no translation is examined for OOV words.

This section can be concluded by noting the overall positive impact of using processed MT on both the efficiency and effectiveness of the translation and retrieval. In the next section, we use this high quality and fast MT to translate the patent documents, which explores an approach that has always been seen as impractical with ordinary MT systems due to the translation time required to translate large amounts of text.

6.4.3 Document Translation for CLPR

As described in the retrieval task overview in section 6.3.1 and as shown in Figure 3.1 on page 43, 31% of the evaluation patents in the collection are German and French patents. In the experiments reported so far, only the sections of these patents that have manual English translations have been indexed. Thus the sections of the patents which are only available in the original languages of French and German were not indexed for search. This was the standard approach adopted by many participants in the CLEF-IP 2009 and 2010 (Roda et al., 2009; Piroi, 2010). However, this approach has a significant drawback since for a large proportion of these patents, only the title field has an English translation. This means that these documents are very short (only small number of words in the title are available for search) and hence the documents have a very low chance of being retrieved, see Figure 3.2 on page 43, which presents the amount of English content present in the collection. Some of the submitted runs used a multilingual index formed by indexing the English, French, and German text into a single index without translation and then searched the collection with patent topics in their original languages in an attempt to exploit the non-English content to improve search effectiveness, but the results of these runs were lower than those submitted using only the English language text (Piroi, 2010). Other later trials attempted to improve the results by using multilingual queries through translating patent queries into the three languages (Jochim et al., 2010), while indexing the patent documents in their original languages. However, this approach also showed lower results than those reported in this chapter and those reported at CLEF-IP 2010, where the best achieved MAP scores for the German and French topics were 0.04 and 0.056 respectively (Jochim et al., 2010) (the PRES score was not reported in this research). This low result may stem from the multilingual query approach itself and also from using translation dictionaries which fail to utilize context in the translation process.

Previous studies in different IR applications using document translation in CLIR have shown that it can improve retrieval effectiveness, particularly when combined with query translation (Chen and Gey, 2004a; Parton et al., 2008). The main hindrance to continued research into document translation for CLIR has been the impractical translation time required to translate the documents. In (Chen and Gey, 2004a), an "approximate fast translation" for documents was applied. This method was based on using an MT system to translate only the unique terms in the document collection without taking account of their context. The top translation for each term was used to replace the original term in the documents. However, ignoring the context was found to lead to low quality translation. In (Parton et al., 2008), a sophisticated SMT system (DARPA Gale MT) was tested to translate Chinese and Arabic document collections into English in a translingual IR task. However, in this work it was estimated that the time needed to translate the full retrieval collection would exceed 30 years. This excessive translation time led them to drop many of the steps in the SMT process in order to speed up translation. However, they comment that dropping these steps led to poor translation.

In this section, we use the processed MT method to translate the claims and abstract sections of the French and German patents into English where such translations do not exist in the original documents in order to enrich the documents with this information. This resulted in all patents having a comparable document length and amount of information regarding the patent content. Our main objective in this experiment is to explore whether the increased speed of our new approach to MT enables practical document translation for CLIR. As shown in the results in Figure 6.2 and Figure 6.4, an acceptable quality of translation can be achieved for the purpose of CLPR at a superior high speed. Hence, we apply the processed MT method using the 2k and 8k translation models, which were found to be very fast, to translate the missing French and German parts to enable them to be added to the index.

We checked the number of non-English patents in the collection which miss any of the abstract and claims sections in English. Our analysis showed that nearly 44% of the French and German patents miss at least one of the abstract or the claims sections. This resulted in translation of non-English sections from 137k German patents and 47k French patents. In total 1.15M German sentences and 390k French sentences were translated. The sizes of the plain text to be translated for the German and French were 343MB and 128MB respectively, and 211MB and 70MB after stemming and stop word removal respectively. After translation, the size of the English text for the German patents in the collection increased from 353MB to 545MB when using the 2k model and to 541MB when using the 8k model. For the French patents, the size increased from 127MB to 192MB when using the 2k model and to 190MB when using the 8k model. The difference in the sizes of the translated texts generated using the two models arises due to the differences in the percentage of untranslated OOV terms for each translation model. Table 6-5 presents the values of the amount of French and German text to be translated.

Table 6-5: Amount of French and German content to be translated and added to the patent collection index

	French	**German**
Number of patent enriched with translations	47k	137k
Number of sentences translated	390k	1.15M
Size of text to be translated	128MB	343MB
Size of text to be translated after processing	70MB	211MB
Increase in patents content after adding translated text	51%	54%

Table 6-6 reports the time taken to translate the German and French documents into English using the new processed MT approach with the 2k and 8k translation models on the server machine described earlier. In addition, the estimated translation time if the ordinary MT were

used is reported too based on experiments reported in Figure 6.4. As shown in Table 6-6, it would be unrealistic to use the ordinary MT system to translate this amount of text, since the time required for translation even when using our very small 2k translation model is more than one month. This becomes even more unrealistic when considering the slightly larger 8k translation models where the estimated translation time will exceed two months. The degree of reduction in the processing time is highlighted clearly in this context when the amount of data to be translated is very large. Using the 2k model the translation time is reduced to only 2 days, and for the 8k model this is reduced to less than one week. In addition, it is expected that the retrieval effectiveness would be better or at least the same when compared to using ordinary MT, based on results in Figure 6.2. This significant reduction in the translation time makes document translation a practical component of CLPR.

Table 6-6: French and German documents translation time with 2k and 8k translation models using processed MT vs ordinary MT (estimated)

Training model		**Processed MT**	**Ordinary MT** (*estimated*)
2k	German	1day 4hrs 46mins	27days 20hrs 1min
	French	22hrs 47mins	9days 17hrs 49mins
	Total	**2days 3hrs 33mins**	**37days 13hrs 50mins**
8k	German	3days 7hrs 26mins	44days 19hrs 6mins
	French	3days 4hrs 22mins	24days 22hrs 15mins
	Total	**6days 11hrs 48mins**	**69days 17hrs 21mins**

Figure 6.5 shows the retrieval effectiveness, PRES and MAP, for two new sets of results when searching the patent collection after adding the translated parts to the German and French patents using the 2k and 8k translation models. All sets of translated queries were used to search the two new collections, including Google translate and processed MT system with different sizes of translation models. Translation of the query set using ordinary MT is not included here since its performance has already been shown to be similar or lower to that of the processed MT.

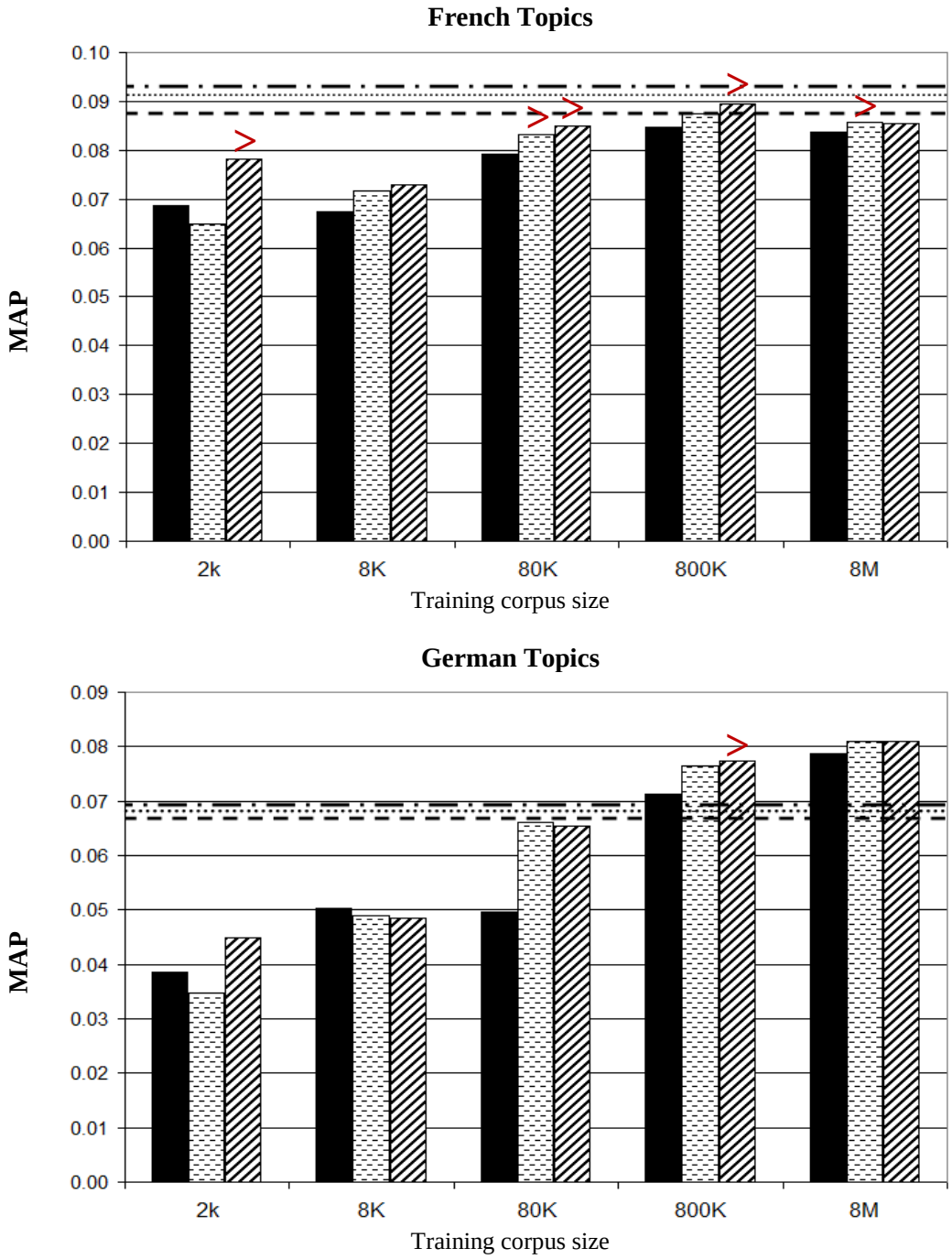


Figure 6.5(a) MAP for French and German topics after adding missing parts to collection

'>' refers to statistical significant improvement in results. Dotted horizontal lines represents queries translated by Google translate.

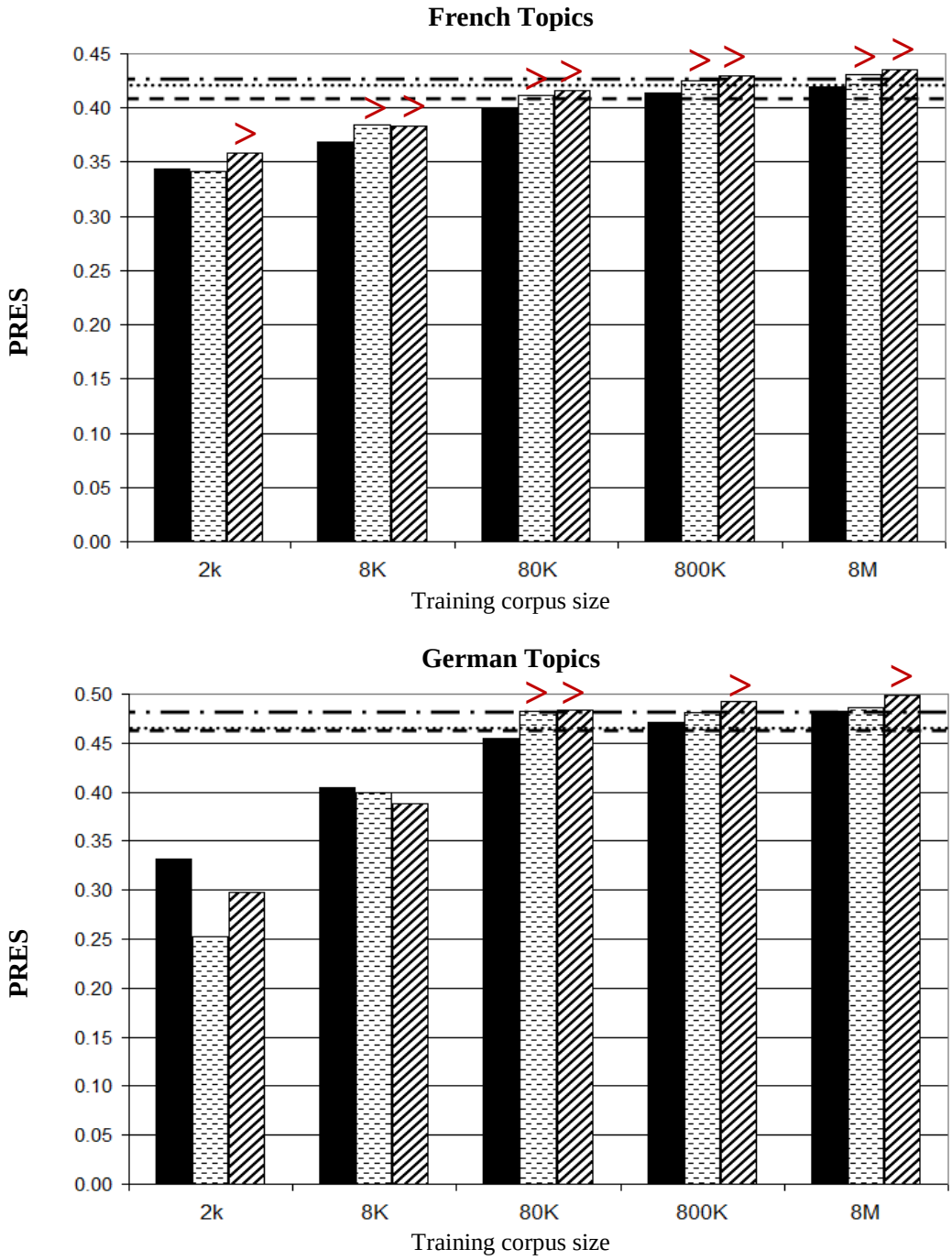


Figure 6.5(b) PRES for French and German topics after adding missing parts to collection

Figure 6.5: Retrieval effectiveness for French and German topics after adding translated missing parts to the French and German documents compared when only English parts are indexed.
'>' refers to statistical significant improvement in results. Dotted horizontal lines represents queries translated by Google translate.

Figure 6.5 compares the new results to the earlier ones in Figure 6.2 when only the English text of the patents was indexed. Comparing the results using MAP shows that there is some statistically significant improvement in results for the French query sets, but very limited improvement or no improvement at all for the German ones. However and as mentioned earlier, we are focusing more on the PRES score since it better reflects the objective in a patent search task. When comparing PRES values, a statistical significant improvement in the retrieval effectiveness can be seen for most of the query sets for both the French and German topics. Considering the French topics, adding the translated parts to the non-English documents showed a large improvement in all the query sets (including those translated by Google translate) except when both the queries and documents are translated with the 2k model. For the German topics, significant improvements only occurred when documents are translated with the 8k model and queries are translated with 80k or higher models (including Google translate). The logical explanation of this observation is the impact of the high percentage of OOV terms in both the translated queries and documents for German resulting from the smaller training set. It can be seen that a high OOV rate in the translation model in both documents and queries leads to some matching of untranslated terms, which is like using mixed language queries as reported in (Jochim et al., 2010, Piroi, 2010), which showed that using mixed language queries and documents leads to an unstable effect on the retrieval effectiveness, where it can sometimes improve the results, but often degrades it, which is consistent with the results reported here.

One of the conclusions of this study is that poor and fast translation models, such as 2k and 8k, can be used with the processed MT approach, for translating multilingual document collections given that a better translation model will be used to translate the queries. The terms “poor” and “better” depend on the language pair. For example, only the 2k model need be

considered poor for French, whereas both the 2k and 8k models are considered poor for German since they have a higher OOV rate because of word compounding.

In overall conclusion, the results of the experiments in this section show that the processed MT approach makes document translation techniques a realistic component in CLPR systems. Depending on the document collection size, and the estimated time for translation, a convenient translation model size can be selected.

Regarding applying document translation for multilingual patent search, it has been shown that significantly better results can be achieved even with using small translation models for translating the documents in a reasonable amount of time, but in this case, queries are recommended to be translated with a better translation model to achieve significantly better results. After adding the translated French and German text to the index, our best result for the French and German topics in this chapter are PRES 0.436 and 0.499, which are, to the best of our knowledge, the highest achieved scores for these topics for the CLEF-IP 2010 dataset when using IR techniques without inclusion of citation information. Nevertheless, including the translated content in the index file would be expected to lead to further improvements in results obtained in the previously reported work by different participants, including those which use citation extraction.

6.5 German Decompounding

All the results reported for CLPR in this chapter for the German topics showed lower effectiveness in retrieval when compared to French topics. We determined this to be due to word compounding in German, which leads to higher OOV rates, where the trained MT do not have these terms in their dictionaries, leading to having them not translated. This impacts negatively on the retrieval effectiveness. In this section, an investigation of German decompounding is

presented in a trial to improve the retrieval effectiveness for the German topics especially for the case where only a limited amount of data is used for training the MT system[23]. We apply a decompounding algorithm that is trained using part of the patent corpus to test the effect of German topic decompounding prior to translation. The effect of decompounding on document translation is not presented here since this would require rerunning all the experiments even for the French topics, since French topics have relevant German documents, and decompounding is not the main focus in our study.

6.5.1 Decompounding algorithm

There has been little research on decompounding for patent search and for training MT systems. Koehn and Knight (Koehn and Knight, 2003) train decompounding for MT using knowledge from parallel corpora, preventing incorrect decompounding when there is a one-to-one correspondence between two words in different languages.

In our experiments, decompounding German words is realized using an approach previously employed in domain-specific CLIR (Chen and Gey, 2004b). Words are decomposed with respect to a corpus (in contrast to (Chen and Gey, 2004b), our corpus does not only contain base forms), choosing the decomposition with the smallest number of words and the highest decomposition probability. The decomposition probability is defined as:

$$p(c) = \prod_{i+1}^{n} p(w_i) \tag{6-1}$$

where the probability of a constituent word is computed as

[23] This study was joint work with Dr. Johannes Leveling, a native German speaker in DCU

$$p(w_i) = \frac{ctf_{w_i}}{\sum_{j=1}^{N_w} ctf_{w_i}} \quad (6\text{-}2)$$

where:

ctf_{w_i} is the collection frequency of word w_i

N_w is the number of unique words.

The decompounding approach identified compounds formed by concatenation of constituents, allowing for an additional *s* between words (so called Fugen-*s*) and an omitted *e* (*e*-elision). Note that we specifically did not consider additional Fugen-elements such as *en* between two constituents, as removing these might result in a change of meaning (Chen and Gey, 2004b).

A training collection was created by combining the 3M English sentence corpus from the Leipzig corpora list[24] with a random sample of 800k sentences from the German patents in the CLEF-IP collection.

We evaluated the decompounding algorithm based on a gold standard corpus (GSC) of 2000 sentences randomly extracted from German patents. The GSC was manually annotated with the correct decomposition of words by Dr. Johannes Leveling. It contains 27,932 unique words and 318k words in total. We found that there seem to be frequent spelling errors in the patent texts, possibly resulting from the OCR source of some of the documents. Spelling errors were also manually decompounded in the GSC. In addition to spelling errors, 12.7% of the word forms in the annotated corpus are chemical formulae or substance names. This indicates the domain-specific nature of patents. In the GSC, chemical formulae were decompounded only when the

[24] http://corpora.uni-leipzig.de/

head noun is a German word. For example, *Methylrest* (methyl radical) has the head noun *Rest* (radical).

The decompounding method achieved 95.0% accuracy (the percentage of correctly decompounded words) measured over all words in the annotated GSC and 81.4% accuracy for unique words. Decompounding the GSC increases the total number of words by 16.3%, while the number of hapax legomena (words occurring only once) is reduced by 48.8%, compared to the original GSC. This illustrates that compounding is a productive process in German.

6.5.2 Results with decompounded German text

Text decompounding was applied on the German MT training corpora and on the German topics before translation. Figure 6.6 shows the effect of decompounding on German topics prior to translation on the retrieval effectiveness of CLPR. As shown, it is clear that decompounding improves the retrieval effectiveness when a limited amount of training data is used, namely, 2k, 8k, and 80k. For large training corpora, no significant change was found, which could be anticipated, since the OOV rate for large training corpora is low, and many of the compounds in the German topics are covered in the translation model, as shown in Table 6-7. For limited training corpora, splitting German compounds into their constituent words gives them a higher chance of being found in the translation model (which is decompounded too), leading to a translation of the term and consequently to better retrieval results.

Table 6-7: OOV rates while translating German topics with and without decompounding

Training Corpus	2k	8k	80k	800k	8M
Baseline (compounded)	40.7%	28.3%	13.6%	7%	4.2%
Decompounded	19.7%	10.3%	3.1%	1.3%	0.7%

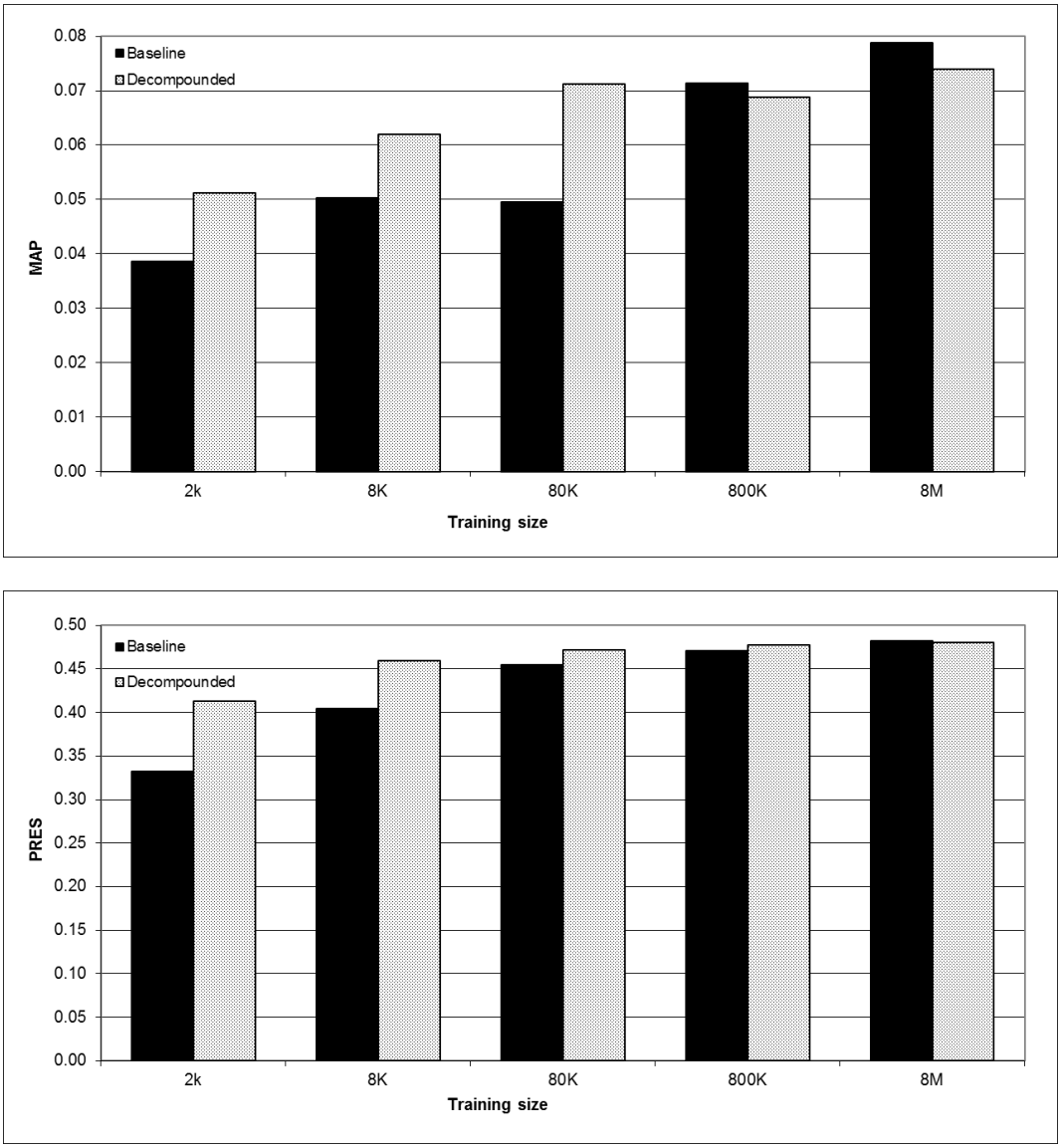


Figure 6.6: Effect of German decompounding on CLPR

6.6 Summary

In this chapter, the final set of research questions about CLIR in patent search was addressed and investigated. These questions are:

- Patent search is often multilingual by nature; what are the differences between cross-language patent search and cross-language information retrieval for standard ad-hoc search tasks?
- Can the translation process in cross-language patent search be optimized to be more efficient with the long queries typically used in patent search?
- How can the retrieval effectiveness be maintained or even improved while increasing the efficiency of the cross-language patent retrieval system?

These questions were addressed through the introduction and evaluation of a novel technique for training MT systems for the purpose of CLIR. Although the technique mainly comprises a re-ordering of the workflow of the steps in CLIR, the impact was shown to be significantly more efficient in the resources and computational requirements of the MT process.

Initially we demonstrated that MT is a more effective technique for translation in CLPR than DBT methods. We showed that using DBT with different setups, by using one or multiple translations, does not achieve comparable results to MT.

The proposed translation technique for CLIR was tested on the recall-oriented patent search task that requires a large amount of training data for conventional statistical MT and for which the query translation time that can reach more than 50 times the search time. Experimental results showed that processing the text by case folding, stop word removal, and stemming before MT training and decoding leads to speeding up of the translation process up to 23 times. In addition, this technique proved to be much more effective when only a limited amount of training resources were used for a given language pair. Our analysis shows that stemming is responsible for the improved retrieval results when limited training data is available, and that stop word removal is responsible for speeding up the translation process. This shows the importance of having both techniques applied before translation. The proposed MT system

allows the preparation of only a couple of thousands of parallel sentences for MT training to produce significantly higher retrieval effectiveness than ordinary MT. Furthermore, we have shown that the very large reduction in translation time using this approach makes document translation in CLIR a more practical proposition with improved retrieval effectiveness.

Finally, a test of the impact of German decompounding on the retrieval effectiveness of CLPR shows strong impact on retrieval quality when only a limited amount of training data is used for the translation system, since decompounding has the effect of reducing the OOV rate. Our results also show that decompounding is not useful when larger amounts of training data are available for training the MT system, where the OOV rate is low even with the presence of compounds.

Chapter 7

Conclusions and Future Directions

7.1 Topic of the Thesis

This thesis has presented a study on recall-oriented IR tasks as represented by patent search, where the objective is to retrieve all possible relevant items. The nature and challenges of patent search tasks were comprehensively discussed and explored. The objective of studying patent search as an IR task was to propose and examine new technologies which have the potential to improve the effectiveness and efficiency of the work of patent examiners.

This chapter summarizes the contributions of this study and outlines potential directions for future work.

7.2 Contribution of the Thesis

The majority of the work reported in patent search to date has focused on developing effective patent retrieval methods. Much of this work has focused on the conversion of the long patent topic into an effective query to search the patent collection (Roda et al., 2009; Piroi, 2010; Xue and Croft, 2009a; Xue and Croft; 2009b). The work performed in this area has included

formulating the patent query from specific patent sections, combination of sections, weighting the terms with different schemes, segmenting the patent query topic into subtopics and searching with each individually then merging the results, and so on. Less focus in existing work has been directed to other aspects in patent search, such as meaningful evaluation for patent search and cross-language search.

In the first chapters of the thesis, we overviewed the patent search task and described our contribution in the development of a simple effective patent retrieval system. This work included our participation in the CLEF-IP task for two successive years, where we achieved top ranks among the participants in the second year of our participation. Furthermore, additional work was presented comparing our approach to the best run in CLEF-IP 2010, where we proved that our approach to retrieval is comparable to the best run. Also, we performed some attempts to improve the retrieval effectiveness of patent search through applying query expansion (QE) techniques, including a novel method based on QE using an automatically generated synonym set (SynSet), which showed a significant improvement to the retrieval effectiveness when compared by MAP. However, we later demonstrated that this improvement is not effective for patent search task when measured by our own PRES evaluation metric. Although our contribution to building effective patent search systems is considered one of the best in patent search, we do not consider this to be part of the main contribution of the research in the thesis, since our system is designed for the prior-art search task in CLEF-IP, which does not perfectly model the interactive search scenario for the prior-art patent search in the patent office. Rather, we consider it as a baseline to the main contributions presented in the later chapters.

The research presented in this thesis is different from much of this existing research in patent search. In our work, we sought to address some of the problems in recall-oriented search and patent search in particular which have not received much attention in most of the existing

reported research. The most frustrating issue we noticed in the state-of-the-art work was the absence of a proper evaluation metric for this kind of task. Most of the published research on patent search confirms that the task is recall-oriented, but the evaluation of this research uses MAP to measure the improvement or degradation of retrieved results. This is surprising since MAP is primarily a precision-oriented retrieval metric. The inadequacy of MAP as an evaluation metric provided our initial motivation to develop an evaluation metric that measures the performance of a recall-oriented IR system in a meaningful way for real-world patent search tasks. Thus we developed PRES, which measures both the system recall and the quality of ranking of the relevant documents depending on the user's potential effort for checking a ranked results list. PRES was used in CLEF-IP 2010 evaluation campaign as one of the standard metrics used to evaluate patent retrieval effectiveness (Piroi, 2010). In addition, it has been reported in many research publications including the first book on challenges of patent search (Lupu et al., 2011).

The second issue, which we found to be neglected by most of the research community for the patent search tasks, was the translation stage for CLIR tasks in patent search. We believe that this stage should be treated differently for patent search tasks, since the amount of text to be translated, even when translating the query, is very considerable when compared to other cross-language retrieval tasks. Moreover, the patent text is special and domain-specific training resources are required for the translation stage, which is not available for many language pairs. This led us to focus on finding a solution for the resource and computational requirements for this task, with the awareness that any successful practical retrieval system should be efficient as well as being effective. Thus we developed what we called the "processed MT" approach, which represents a simple but novel idea to transfer the IR processing components (stemming and stop word removal) to the standard MT system, leading to a large increase in speed (more than 20

times faster) and significant improvement in retrieval effectiveness when limited MT training resources are available.

7.3 Answers of the Thesis Research Questions

In this section, we review the research questions introduced at the beginning of the thesis in Section 1.1.1, and consider how these have been addressed in our study.

7.3.1 The special nature of patent search and recall-oriented IR

- *What are the differences between recall-oriented search tasks and other standard IR tasks, and how can these be demonstrated?*

We have highlighted the main differences between recall-oriented and precision-oriented IR including the nature of the search task, the objectives, and the users. We demonstrated this with the clear example of patent search. We discussed the special nature of patent text and the challenges of the search task. We reviewed some of the prior work of different patent search tasks and studied the prior-art patent search task ourselves. We demonstrated the challenges of this task which include: the long length of the query compared to other tasks, the low retrieval effectiveness of the search results, the term mismatch problem between topics and relevant documents, and the issues of multilingual search for the task.

- *What are the properties and configuration of a retrieval system to achieve high retrieval effectiveness for patent search?*

We performed different experiments to explore the best setup for a patent retrieval system. This included testing: different query formulations, text pre-processing, structured indexing, pseudo relevance feedback (PRF), filtering results based on classification, and information extraction. Our experiments led us to find the best setup of these techniques, and showed the

failure of others. We noticed that an effective query for patent topic is to use the full description section of the patent application as the query. We found that stemming, stop word removal, and filtering out digits and short terms increases the retrieval speed without affecting the retrieval effectiveness negatively. In addition, we found some techniques to be ineffective in the experiments we performed, such as using structured indexing and applying PRF. We noticed the high impact of extracting the reference citations from the description section of the patent topic on improving the retrieval effectiveness significantly. However, we did not consider this as a general rule since it is not available for many patent topics. Moreover, we compared our system, which is simple and straightforward, to the state-of-the-art system in CLEF-IP 2009 and 2010, which is much more sophisticated. We showed that both systems are comparable in achieving good retrieval results when extracted citations are used.

- *Is applying standard retrieval techniques for patent search as effective as for general IR tasks?*

We demonstrated this to be untrue for some standard techniques in general IR tasks, such as PRF. Using PRF with similar setup as other IR applications led to a significant degradation to the retrieval effectiveness in our experiments. This shows that it is not straightforward to use similar techniques from other IR applications directly in patent search. The different nature of patents requires further investigation to identify the best setup for these techniques in order to achieve positive results. In addition, we showed that the standard evaluation metrics for other IR tasks are not the best choice for patent search. We also showed that the use of multiple translations when applying DBT methods for CLPR leads to degradation in the retrieval effectiveness which contradicts with what is reported for standard ad hoc search tasks.

We also demonstrated that some standard IR techniques are effective for patent search which include: stemming, stop word removal, and using a standard retrieval model (LM in our case).

7.3.2 Recall-oriented IR evaluation:

- *What features of performance should be measured for a recall-oriented IR application such as patent search?*

The straightforward answer to this question is recall. However, our investigation for this part and our study of the published research on this topic showed that while recall is the main objective for this task, precision is still important too. Thus the overall objective is to find as many as possible of the relevant document at the highest possible retrieval ranks.

- *Are the current evaluation methods and metrics, which are currently being used for patent search and recall-oriented IR applications, appropriate to evaluate this type of application?*

The comprehensive investigation presented in Chapter 5 showed that the metrics currently primarily used for patent search, namely MAP and recall, and other metrics used for different IR applications do not suit this kind of application. These metrics lack the ability to measure both the recall and ranking quality while capturing the effort exerted by user to identify the relevant documents.

- *Can a more meaningful metric be developed for these tasks?*

We developed *patent retrieval evaluation sore* (PRES), which measures the system recall and quality of ranking in one metric. PRES is a function of the maximum numbers of documents to be checked by a user (N_{max}). N_{max} is assigned depending on the user and the retrieval task, where in some applications the number of checked documents can be only 10 whereas in other IR tasks the user is willing to check hundreds or even thousands of documents as in the case of patent search and legal search.

- *How can the suitability and reliability of such a novel metric be established to ensure that it reflects the real system performance and assure that it is robust to different system variables?*

We successfully investigated the adequacy and the robustness of PRES through many experiments on 48 different runs submitted to the CLEF-IP track 2009. These experiments showed that PRES is a meaningful and suitable metric for this task when compared to the currently used metrics for patent search. Moreover, it was proven that it is highly robust when relevant judgements are incomplete, which is the case for the patent search task.

7.3.3 Multilingual patent search:

- *Patent search is often multilingual by nature, are there differences between cross-language patent search and cross-language information retrieval for standard ad-hoc search tasks?*
- *Can the translation process in cross-language patent search be optimized to be more efficient with the long queries typically used in patent search?*

Our study in the first chapters of the thesis and especially in Chapter 4 showed that the best query formulation for the patent prior-art search task comes from using a large portion of the patent application itself rather than extracting a small number of words. Even for invalidity search, the topic is a long claim formed from several sentences. We noticed that CLIR using these long queries requires high computational resources for the translation process. We showed that the translation time for these very long queries when using conventional high quality MT systems is more than 50 times the search time. Moreover, domain-specific training data is required, which was demonstrated by the low performance of translating the German topics that contain word compounds by Google translate compared to the MaTrEx MT system that is trained on patent data. This motivated us to find more effective and efficient solutions to the translation stage in patent cross-language search.

- *How can the retrieval effectiveness be maintained or even improved while increasing the efficiency of the cross-language patent retrieval system?*

We developed a novel approach for translation in CLIR using MT systems based on a simple modification to the workflow of a CLIR system, shown in Figure 6.1. The modification applies standard IR pre-processing of stemming and stop word removal to the text before the translation process rather than after it. This modification in the workflow led to speeding up the translation time up to 23 times without affecting the retrieval effectiveness. Moreover, the "processed MT" system, which represents the novel translation stage in the modified CLIR workflow, requires less training resources to achieve significantly better retrieval results than standard MT systems, which is very useful in cases where only limited training data is available for a language pair.

7.4 Revisiting the Hypotheses of the Thesis

The study in this thesis has successfully proved that the two hypotheses introduced in the first chapter in the thesis to be true, which are:

H1. A deeper understanding of the nature and the requirements of patent search and recall-oriented IR tasks in general can lead to an improved understanding of the utility of existing evaluation metrics for these tasks, and the proposal and implementation of a novel metric specifically developed for patent search that addresses these requirements.

We have proved this hypothesis to be true, where we successfully developed PRES and proved its adequacy with this class of application.

H2. The special nature of patent topics, which are considerably long and domain-specific, leads to high cost requirements for effective cross-language search. It is hypothesized that machine translation methods used in cross-language patent search can be adjusted and optimized to significantly improve the efficiency and effectiveness for the high cost cross-language search for such task.

We successfully introduced the "processed MT" which was proved to be significantly more efficient in its computational and resources requirements than standard MT techniques for CLIR in the patent search task.

7.5 Possible Future Directions

The main focus of the work in this thesis is its contribution to the recall-oriented IR domain represented by patent search. Despite our experimentations being focused on patent search in particular, the algorithms and approaches presented in this thesis are potentially general enough to be applied to different IR tasks. Our motivation did not focus on just improving the state-of-the-art in patent search rather than opening new research tracks for recall-oriented search tasks. The research in our study has the potential for different future directions of research, including possible improvements to the techniques presented, and the application of these techniques in different IR tasks. Potential future work includes:

- The query formulation in patent search still needs much research to develop an effective algorithm for modelling the patent topic into a search query. As demonstrated in our thesis, simple approaches for formulating the patent query from a topic achieves comparable results to more sophisticated and advanced approaches (Magdy et al., 2011). This result means that additional work and investigations are required to develop more effective models for query formulation and document indexing in patent search, since this search task is based on "idea" matching rather than "word" matching. Future directions on this topic should take into consideration the interactive nature of the patent search task, where prior-art patent search is performed over search sessions in the patent office. Thus, it is important to model this interactive search in future investigations in patent search.

- The theoretical and experimental adequacy of PRES was well demonstrated on patent search based on tens of runs by participants in an evaluation campaign for patent search. However, it would be very interesting to bring PRES to actual practice by direct consultations with professional patent experts. Such a study should have a practical and theoretical analysis of the user model represented by PRES. This can be similar to the study presented for MAP in (Robertson, 2008). A practical study of the patent examiners' requirements for an ideal patent retrieval system and a detailed understanding to the patent examiners' stopping model for checking the retrieved results of such retrieval system should assist in applying some potential modifications to the implementation and design of PRES to further improve the modelling of the actual patent retrieval system performance. This may include integrating additional factors to the metric, such as: time of task completion, maximum number of non-relevant retrieved documents to be checked by the patent examiner before rewriting a new query, and the level of experience of the patent examiner.
- We showed theoretically that PRES can be applied to legal search by calculating $PRES_{est}$. Despite our theoretical analysis, real sets of runs are needed in order to explore its practical behaviour on this type of data. This is very important since N_{max} for some topics in legal search will be much less than the actual number of relevant documents. Therefore, it should be practically justified that the $PRES_{est}$ approximation is acceptable for this kind of situation. In addition, the user model of the legal search task should be studied as well to verify the practical adequacy and meaningfulness of the PRES score for this task.
- Another potential feature that can be added to PRES is enabling PRES to evaluate recall-oriented tasks with graded relevance judgements, where some documents can be considered more relevant than others. This is valid for patent search tasks, which are described in Section 2.4.3 on page 29, where a prior-art patent search task has different types of relevant

documents that can invalidate the patent application or just describe prior-work. This feature could be similar to the use of graded relevance in the nDCG score (Carterette et al., 2008).

- Regarding the translation work for CLIR, the motivation for this work stemmed from the nature of patent cross-language search, which has a high computational cost and requires a large amount of domain-specific MT training data. The long length of the patent topics causes the high computational cost of the translation process. The same problem exists for other tasks, such as cross-language duplicate document detection. In duplicate document detection, the full document is used to search for other similar documents. Applying this across languages requires high computational cost for translation. "Processed MT" may be an ideal solution for such a task. It would be very interesting to examine this practically. Similarly, the "Processed MT" approach can be very useful for automatically linking news articles across languages, where an article in one language can be fully translated and used to search for the corresponding article in another language at very high speed. Translating a full article using standard MT system would be very slow and inefficient.
- There is a strong potential for using the "processed MT" in different IR applications, especially for language pairs where limited MT training resources only exist. "Processed MT" was shown to produce significantly better results when limited training resources are available. It would be interesting to test the translation approach on languages such as Indian languages, where there are very limited or even no MT training resources for these languages. The forum for information retrieval evaluation (FIRE) evaluation campaign[25] tasks could be used for examining this approach, where there are many CLIR tasks for English and different Indian languages including Hindi, Bengali, and Marathi.

[25] http://www.isical.ac.in/~clia/

- It is interesting to explore the performance of the new MT approach even for standard ad hoc tasks and web search, where queries are typically short. Of course the objective will not be reduced translation time, since it is already insignificant. Rather, the main objective will be testing the approach when queries lack context information, which can lead to low MT performance. After the pre-processing of the training data, the MT model would be trained using this kind of sentences which are improperly constructed.
- For the usage of "processed MT" for document translation, a more well founded approach to selecting the sentences used to build the translation model rather than selecting them at random as was done in our experiments should be explored. The reason behind selecting parallel sentences at random in our study was to simulate the situation when limited resources are available for a pair of languages and there is no opportunity to select the set of training sentences. However, an algorithm designed for selection of appropriate sentence pairs could be used to build a more effective translation model for the purpose of fast and accurate document translations.
- Regarding the work done for QE using the SynSet, additional analysis of the success and failure of the technique should be applied to seek to better understand the instances where this approach fails to improve the retrieval results. Also, further pruning methods in SynSet creation could be explored, since there may be some terms that degrade retrieval effectiveness when used for expansion which could be eliminated using alternative pruning methods.
- Another different methodology for applying PRF to patent search is to use feedback for query reduction instead of expansion. We participated in a study for query reduction based on PRF for patent long queries in a very recent research study (Ganguly et al., 2011). The reduction of the query was based on keeping only the original terms that match the top

retrieved documents well and filtering out the other terms. The hypothesis for this work was that query terms which do not contribute to the retrieval of the top documents can be noisy and harmful. We applied PRF query reduction to the same CLEF-IP prior-art patent search task investigated in this thesis. We observed a small improvement to the baseline when we reduced the query to 75% of its original length. However, this improvement was still not significant (Ganguly et al., 2011). This result motivates further investigation to better filter out the harmful terms to attempt to achieve a significant improvement in retrieval effectiveness.

7.6 Closing Remarks

We believe that the work presented in this thesis has opened potential new research directions for recall-oriented tasks represented in patent search, and can potentially be extended into many other IR tasks. The central focus of our work focuses on contributions to improving retrieval effectiveness in patent search, and in addition, we believe that is could be applied successfully in contributions to problems in recall-oriented IR tasks in general, which have not received enough research attention to date. We hope that this work will act as a starting point for other researchers to continue investigations on the problems we addressed, trying to further improve the techniques presented in this thesis and find additional applications for them.

Bibliography

Adams S. A Practitioner's View on PaIR. In *Proceedings of the fourth International Workshop on Patent Information Retrieval (PaIR'11) at the Twentieth Annual International ACM Conference on Information and Knowledge Management (CIKM 2011)*, Glasgow, Scotland, (2011)

Azzopardi L., H. Joho and W. Vanderbauwhede. A Survey on Patent Users Search Behaviour, Search Functionality and System Requirements. *IRF Report* 2010-00001, (2010)

Baeza-Yates, J., and Ribeiro-Neto, B. Modern Information Retrieval: The Concepts and Technology Behind Search. *ACM Press Books*, (2010)

Bashir S. and A. Rauber. Improving Retrievability of Patents in Prior-Art Search. In *Proceedings of 32nd European Conference on Information Retrieval (ECIR 2010)*, Milton Keynes, U.K., (2010)

Billerbeck B. and J. Zobel. Techniques for Efficient Query Expansion. In *Proceedings of the 11th Symposium on String Processing and Information Retrieval (SPIRE 2004)*, Padua, Italy, (2004)

Bompad T., C.-C. Chang, J. Chen,R. Kumar, and R. Shenoy. On the Robustness of Relevance Measures with Incomplete Judgements. In *Proceedings of the 30th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, The Netherlands, (2007)

Bonino D., A. Ciaramella, and F. Corno. Review of the State-of-the-Art in Patent Information and Forthcoming Evolutions in Intelligent Patent Informatics. In *World Patent Information, Vol.32, Issue 1, pages: 30-38*, (2010)

Brin S. and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer Networks, Vol. 30*, (1998).

Buckley C., A. Singhal, M. Mitra, and G. Salton. New Retrieval Approaches Using Smart. In *Proceedings of 4th Text Retrieval Conference (D.K. Harman, ed.), NIST Special Publication 500-236*, (1996)

Buckley, C., and Voorhees, E. M. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, Athens, Greece, (2000)

Buckley, C., Dimmick, D., Soboroff, I., and E. Voorhees. Bias and the Limits of Pooling. In *Proceedings of the 29th ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, Seattle, Washington, USA, (2006)

Cao G., J.-Y. Nie, J. Gao, and S. Robertson. Selecting Good Expansion Terms for Pseudo-Relevance Feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, (2008)

Carterette, B., Bennett, P. N. Chickering, D. M., and Dumais, S. T. Here or There: Preference Judgments for Relevance. In *Proceedings of 30th European Conference on Information Retrieval (ECIR 2008)*, Glasgow, Scotland, (2008)

Chen, A. and F. Gey: Combining Query Translation and Document Translation in Cross-Language Retrieval. In *Proceedings of the CLEF 2003: Workshop on Cross-Language Information Retrieval and Evaluation*, (2004a)

Chen A. and F. C. Gey. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding. In *Information Retrieval, Vol. 7, Issue 1-2, pages: 149-182*, (2004b)

Collins-Thompson K. and J. Callan. Query Expansion Using Random Walk Models. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2005)*, Bremen, Germany, (2005)

Connett-Porceddu, M., D. Ashton, N. Bacon, N. dos Remedios, C. Nottenburg, S. Okada, G. Quinn, Y. Wei, and R. Jefferson. Analysis of Trends in Search and Retrieval of Intellectual Property-Related Information. *IP Australia*. (2005)

Croft, W. Bruce, and John Lafferty (eds.). Language Modeling for Information Retrieval. *Springer, The Information Retrieval Series, Vol. 13*, (2003)

Cronen-Townsend S., Y. Zhou, and W. B. Croft. Predicting Query Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, Tampere, Finland. (2002)

D'hondt E. Lexical Issues of Syntactic Approach to Interactive Patent Retrieval. In *Proceedings of the 3rd BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2009)*. Padua, Italy, (2009)

Darwish K., D. W. Oard. Probabilistic Structured Query Methods. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, Toronto, Canada. (2003)

Doi H., Y. Seki, and M. Aono. A Patent Retrieval Method Using a Hierarchy of Clusters at TUT. In *Proceedings of the 5th NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan. (2005)

European Commission, Pharmaceutical Sector Inquiry, Preliminary Report (DG Competition Staff Working Paper), 28 November (2008)

Franz, M. and McCarley, S. 2002. Arabic Information Retrieval at IBM. In *Proceedings of TREC-2002*. (2002)

Fuhr N. Probabilistic Models in Information Retrieval. In *Computer Journal Vol. 35 Issue 3, pages: 1-21,* (1992)

Fujii, A., Iwayama, M., and Kando, N. Overview of Patent Retrieval Task at NTCIR-4. In *Proceedings of the 4th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, Tokyo, Japan. (2004)

Fujii, A., Iwayama, M., and Kando, N. Overview of Patent Retrieval Task at NTCIR-5. In *Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, Tokyo, Japan. (2005)

Fujii, A., Iwayama, M., and Kando, N. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, Tokyo, Japan. (2007)

Fujii, A. Integrating Content and Citation Information for the NTCIR-6 Patent Retrieval Task. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, Tokyo, Japan. (2007a)

Fujii, A. Enhancing Patent Retrieval by Citation Analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, the Netherlands. (2007b)

Fukui M., S. Higuchi, Y. Nakatani, M. Tanaka, A. Fujii, T. Ishikawa. Applying Hybrid Query Translation Method to Japanese/English Cross-Language Patent Retrieval. In *Proceeding of ACM SIGIR 2000 Workshop on Patent Retrieval*, Athens, Greece, (2000)

Gao J., J-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. Improving Query Translation for Cross-Language Information Retrieval Using Statistical Models. In *Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR 2001)*, New Orleans, Louisiana, USA, (2001)

Ganguly D., J. Leveling, W. Magdy, and G. J. F. Jones. Query Reduction based on Pseudo-Relevant Documents. In *Proceedings of the 20th Annual International ACM Conference on Information and Knowledge Management (CIKM 2011)*, Glasgow, Scotland, (2011)

Graf, E., and Azzopardi, L.: A Methodology for Building a Patent Test Collection for Prior Art Search. In *Proceedings of the Second International Workshop on Evaluating Information Access (EVIA 2008)*, Tokyo, Japan, (2008)

Hersh W. and Over P. TREC-8 Interactive Track Report. In *TREC-8*, (1999)

Hersh W. and Over P. TREC-9 Interactive Track Report. In *TREC-9*, (2000)

Hull, D. 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 (SIGIR 1993)*, Pittsburgh, Pennsylvania, USA. (1993)

Itoh H. NTCIR-4 Patent Retrieval Experiments at RICOH. In *Proceedings of the 4th NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan, (2004)

Iwayama M., Fujii, A., Kando, N., and Takano, A.: Overview of Patent Retrieval Task at NTCIR-3. In *Proceedings of the 3rd NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering*. Tokyo, Japan (2003)

Jansen B. J., A. Spink. An Analysis of Web Searching by European AlltheWeb.com Users. In *Information Processing and Management Journal, Vol.41 Issue 2, pages: 361-381*, (2005)

Jochim C., C. Lioma, H. Schütze, S. Koch, and T. Ertl. Preliminary Study into Query Translation for Patent Retrieval. In *Proceedings of the 3rd International Workshop on Patent Information Retrieval (PaIR '10)*, Toronto, Canada, (2010)

Joachims T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. *Kluwer Academic Publishers Norwell, MA, USA*, (2002)

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., and Robertson, S. INEX 2007 Evaluation Measures. In *Proceedings of Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*. Dagstuhl Castle, Germany. (2007)

Kando N. What We Shall Evaluate? -- Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and Patent Attorneys. In *Proceeding of ACM SIGIR 2000 Workshop on Patent Retrieval*, Athens, Greece, (2000)

Kando N. and M. Leong. Workshop Patent Retrieval: SIGIR 2000 Workshop Report. In *ACM SIGIR Forum, Vol. 34, Issue 1, pages: 28-30*, (2000)

Kendall M. A New Measure of Rank Correlation. In *Biometrika, Vol. 30, Issue 1-2*, (1938)

Kishida K. Experiments on Psuedo Relevance Feedback Method Using Taylor Formula at NTCIR-3 Patent Retrieval Task. In *Proceedings of the 3rd NTCIR Workshop on Evaluation*

of Information Retrieval, Automatic Text Summarization and Question Answering. Tokyo, Japan (2003)

Koehn P. and K. Knight. Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics (EACL 2003)*, Stroudsburg, PA, USA (2003)

Koehn P, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of* the *Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, Prague, Czech Republic, (2007)

Konishi K. Query Terms Extraction from Patent Document for Invalidity Search. In *Proceedings of the 5th NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan. (2005)

Krier M. and F. Zacc. Automatic Categorization Applications at the European Patent Office. In *World patent Information, Vol. 24, Issue:3, pages: 187-196*, (2002)

Lavrenko, V. and Croft, W.B. Relevance-Based Language Models. In *Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (*SIGIR 2001)*, New Orleans, Louisiana, USA, (2001)

Larkey L. A Patent Search and Classification System. In *Proceedings of the 4th ACM Conference on Digital Libraries (DL-99)*, Berkeley, CA, USA, (1999)

Leong, M. K.: Patent Data for IR Research and Evaluation. In *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, Tokyo, Japan, (2001)

Levow G-A, D. W. Oard, and P. Resnik. Dictionary-Based Techniques for Cross-Language Information Retrieval. In *Information Processing and Management: an International Journal*, *Vol.41, Issue 3, pages: 523–547*, (2005)

Liu S., F. Liu, C. Yu, and W. Meng. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, Sheffield, UK (2004)

Lopez P. and L. Romary. Retrieval Models and Regression Models for Prior Art Search. In *Proceedings of the CLEF 2009: Workshop on Cross-Language Information Retrieval and Evaluation*, Corfu, Greece, (2009)

Lopez P. and L. Romary. Experiments with citation mining and key-term extraction for Prior Art Search. In *Proceedings of the CLEF 2010: Workshop on Cross-Language Information Retrieval and Evaluation*, Padua, Italy, (2010)

Lupu M, F. Piroi, X. Huang, J. Zhu, J. Tait. Overview of the TREC 2009 Chemical IR Track. In *TREC 2009*, (2009a)

Lupu M., J. Huang, J. Zhu, and J. Tait, TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC. In *SIGIR Forum, Vol. 43, Issue 2*, (2009b)

Lupu M., K. Mayer, J. Tait and A. J. Trippe. Current Challenges in Patent Information Retrieval. *Springer, Vol. 29*, (2011)

Lv Y. and C. Zhai. Adaptive Relevance feedback in Information Retrieval. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China, (2009)

Magdy W., J. Leveling, and G. J. F. Jones. Exploring Structured Documents and Query Formulation Techniques for Patent Retrieval. In *Proceedings of the CLEF 2009: Workshop on Cross-Language Information Retrieval and Evaluation,* Corfu, Greece, (2009)

Magdy W. and G. J. F. Jones. PRES: a Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*. Geneva, Switzerland, (2010a)

Magdy W. and G. J. F. Jones. Examining the Robustness of Evaluation Metrics for Patent Retrieval with Incomplete Relevance Judgements. In *Proceedings of CLEF 2010: Conference on Multilingual and Multimodal Information Access Evaluation*, Padua, Italy, (2010b)

Magdy W. and G. J. F. Jones. Applying the KISS Principle for the CLEF- IP 2010 Prior Art Candidate Patent Search Task. In *Proceedings of CLEF 2010: Workshop on Multilingual and Multimodal Information Access Evaluation*, Padua, Italy, (2010c)

Magdy W., P. Lopez, and G. J. F. Jones. Simple vs. Sophisticated Approaches for Patent Prior-Art Search. In *Proceedings of the 33rd European Conference on Information Retrieval (ECIR 2011),* Dublin, Ireland, (2011)

Mahdabi, P., M. Keikha, and S. Gerani. Building Queries for Prior-art Search. In *Proceedings of the Second Information Retrieval Facility Conference (IRFC),* Vienna, Austria. (2011)

Mase H., T.M., Ogawa, Y. Proposal of Two-Stage Patent Retrieval Method Considering the Claim Structure. In *ACM Transactions on Asian Language Information Processing* (*TALIP) Vol. 4, Issue 2, pages: 186-202*, (2005)

Manning C. D, P. Raghavan, and H. Schutze. Introduction to Information Retrieval. *Cambridge University Press.* (2009)

Meij E., W. Weerkamp, and M. de Rijke. A Query Model Based on Normalized Log-Likelihood. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China, (2009)

Moffat, A., and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. In *ACM Transactions on Information Systems (TOIS), Vol. 27, Issue 1, pages:1:27*, (2008)

Nie, Jian-Yun. Cross-Language Information Retrieval. *Morgan & Claypool Publishers*, (2010)

Oard, D. W. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup AMTA*, Pennsylvania, USA, (1998)

Oard, D. W. and Diekema, A. R. Cross-Language Information Retrieval. *Annual Review of Information Science ARIST, Vol.33, Issue 4*, (1998)

Oard, D. W. and F. Gey. The TREC-2002 Arabic/English CLIR Track. In *Proceedings of TREC-2002*. (2002)

Oard, D. W., Hedin, B., Tomlinson, S., and Baron, J. R. Overview of the TREC 2008 Legal Track. In *Proceedings of TREC 2008*, (2008)

Och, F. J. and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics, Vol.19, Issue 1, pages: 19-51*, (2003)

Osborn M. and T. Strzalkowski. Evaluating document retrieval in patent database: a preliminary report. In *Proceedings of the 6th ACM International Conference on Information and Knowledge Management (CIKM 1997)*, Las Vegas, Nevada, USA, (1997)

Papineni K., S. Roukos, T. Ward, and W. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania, (2002)

Parton K., K. R. McKeown, J. Allan, and E. Henestroza. Simultaneous Multilingual Search for Translingual Information Retrieval. In *Proceedings of 17th ACM International Conference on Information and Knowledge Management (CIKM 2008),* California, USA, (2008)

Piroi F. CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In *Proceedings of the CLEF 2010: Workshop on Cross-Language Information Retrieval and Evaluation,* Padua, Italy, (2010)

Piroi F., M. Lupu, A. Hanbury, and V. Zenz. CLEF-IP 2011: Retrieval in the Intellectual Property Domain. In *Proceedings of the CLEF 2011: Workshop on Cross-Language Information Retrieval and Evaluation,* Amsterdam, Netherlands, (2011)

Ponte, J. M. and Croft, W. B. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR 1998),* Melbourne, Australia, (1998)

Porter M.F. An Algorithm for Suffix Stripping. *Program, Vol. 14, Issue 3, pages:130-137,* (1980)

Robertson S. E. The Parametric Description of the Retrieval Tests. Part 2: Overall Measures. In *Journal of Documentation, Vol. 25, Issue 2, pages: 93-107,* (1969)

Robertson S. E. and K. Spark Jones. Simple, Proven Approaches to Text Retrieval. *Technical Report 335, Cambridge University Computer Laboratory*, (1994)

Robertson S. and S. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994),* Dublin, Ireland, (1994)

Robertson S., H. Zaragoza, and M. Taylor. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2004),* New York, USA, (2004)

Robertson, S. A New Interpretation of Average Precision. In *Proceedings of the 31st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, (2008)

Rocchio J. Performance Indices for Document Retrieval Systems. In *Information storage and retrieval, Computation Laboratory of Harvard University, Cambridge*, MA, (1964)

Rocchio J. The SMART Retrieval System: Experiments in Automatic Document Processing. *Prentice-Hall, Inc. Upper Saddle River, NJ, USA*, (1971)

Roda G., Tait J., Piroi F., and Zenz V. CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In *Proceedings of the CLEF 2009: Workshop on Cross-Language Information Retrieval and Evaluation*, Corfu, Greece, (2009)

Saito M. The Search System which uses Japanese-English concordance table PATOLIS-e. In *Proceeding of ACM SIGIR 2000 Workshop on Patent Retrieval*, Athens, Greece, (2000)

Salton G., C. Buckley. Improving Retrieval Performance by Relevance Feedback. In *Journal of the American Society for Information Science (JASIS), Vol.41, Issue 4, pages:288-297*, (1990)

Salton G., J. Allan, and C. Buckley. Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of the 16th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 1993)*, New York, USA, (1993)

Sarasua L. and G. Corremans. Cross Lingual Issues in Patent Retrieval. In *Proceeding of ACM SIGIR 2000 Workshop on Patent Retrieval*, Athens, Greece, (2000)

Schmidt H., K. Butter, and C. Rider. Building Digital Tobacco Document Libraries at the University of California San Francisco Library/Center for Knowledge Management. *D-lib Magazine, Vol. 8, Issue 2*, (2002)

Sheridan P. and Ballerini M. Experiments in Multilingual Information Retrieval using the SPIDER System. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996)*, Zurich, Switzerland, (1996)

Song F. and W. B. Croft. A General Language Model for Information Retrieval. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, Berkeley, USA, (1999)

Spink A., B. J. Jansen , D. Wolfram , T. Saracevic. From E-Sex to E-Commerce: Web Search Changes. In *Computer, Vol.35, Issue 3, pages: 107-109*, (2002)

Strohman T., D. Metzler, H. Turtle, and W. B. Croft. Indri: A Language Model-Based Search Engine for Complex Queries. In *Proceedings of the International Conference on Intelligence Analysis*, VA, USA, (2004)

Stroppa, N. and A. Way. 2006. MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, (2006)

Tague J., Nelson, M., and Wu, H. Problems in the Simulation of Bibliographic Retrieval Systems. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1981)*, Cambridge, UK, (1981)

Takeuchi H., N. Uramoto, and K. Takeda. Experiments on Patent Retrieval at NTCIR-5 Workshop. In *Proceedings of the 5th NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan. (2005)

Takaki T., A.F., Ishikawa, T. Associative Document Retrieval by Query Subtopic Analysis and its Application to Invalidity Patent Search. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM 2004)*, Washington, DC, USA, (2004)

The Sedano Conference. The Sedano Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. *The Sedano Conference Journal*, (2007)

Tomlinson S., Oard, D. W., Baron, J. R., and Thompson, P. Overview of the TREC 2007 Legal Track. In *Proceedings of TREC 2007*, (2007)

Turtle, H. and Croft, W.B. Evaluation of an Inference Network-Based Retrieval Model. In *ACM Transactions on Information System, Vol. 9, Issue 3, pages: 187-222*, (1991)

Van Rijsbergen, C. J. Information Retrieval, 2nd edition. *Butterworths*, (1979)

Verberne S., E. D'hondt , and Oostdijk N. Quantifying the Challenges in Parsing Patent Claims. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval AsPIRe'10*, Milton Keynes, UK, (2010)

Voorhees E. M. Using WordNet for text retrieval. In WordNet, an Electronic Lexical Database, *MIT Press*, (1998)

Voorhees E. M.: Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, New Orleans, USA, (2001)

Voorhees, E. M. The TREC robust retrieval track. In *ACM SIGIR Forum, Vol. 39, Issue 1*, (2005)

Voorhees, E. M., and Tice, D. M. The TREC-8 Question Answering Track Evaluation. In *Proceedings of TREC 1999*, (1999)

Wang J., D. W. Oard. Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, Seattle, Washington, USA, (2006)

Wang, W., Knight, K., and Marcu, D. Capitalizing machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, New York, USA, (2006)

Wen J., J. Nie, and H. Zhang. Query Clustering Using User Logs. In *ACM Transactions on Information Systems (TOIS), Vol. 20, Issue 1, pages: 59-81*, (2002)

World International Property Organization (WIPO), frequently asked questions (FAQs). http://www.wipo.int/patentscope/en/patents_faq.html, last visited 28, June 2010.

World International Property Organization (WIPO), International Patent Classification (IPC). http://www.wipo.int/classifications/ipc/en/, last visited 24, August 2011.

Xue X., and Croft, W.B. Transforming Patents into Prior-Art Queries. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, Boston, MA, USA, (2009a)

Xue X., and Croft W. B. Automatic Query Generation for Patent Search. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China, (2009b)

Xu Y., G. J. F. Jones, B. Wang. Query Dependent Pseudo-Relevance Feedback Based on Wikipedia. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, Boston, MA, USA, (2009)

Appendix A: Derivation of R_{norm}

Equation (2) in Chapter 5 is derived from the Figure 5.1, where it was noted that:

$$R_{norm} = \frac{A_2}{A_1 + A_2} = 1 - \frac{\sum r_i - \sum i}{n(N-n)}$$

where:

A_1, A_2: areas shown in Figure 5.1.

r_i: the rank at which the i[th] relevant document is retrieved

N: collection size

n: number of relevant docs

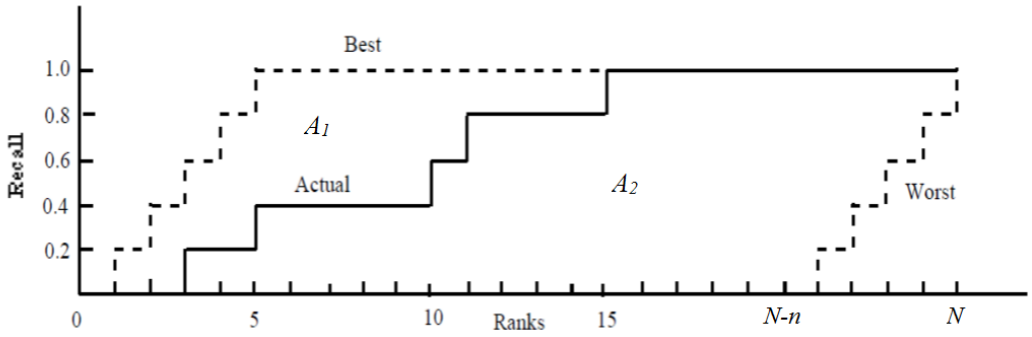


Figure 5.1 Illustration of how R_{norm} curve is bounded by the best and worst cases (Rijsbergen, 1979)

In this part we show the derivation of how this formula is generated from the graph. Figure 5.1' shows how the value of A_1 and A_2 can be calculated from the R_{norm} graph presented in Figure 5.1.

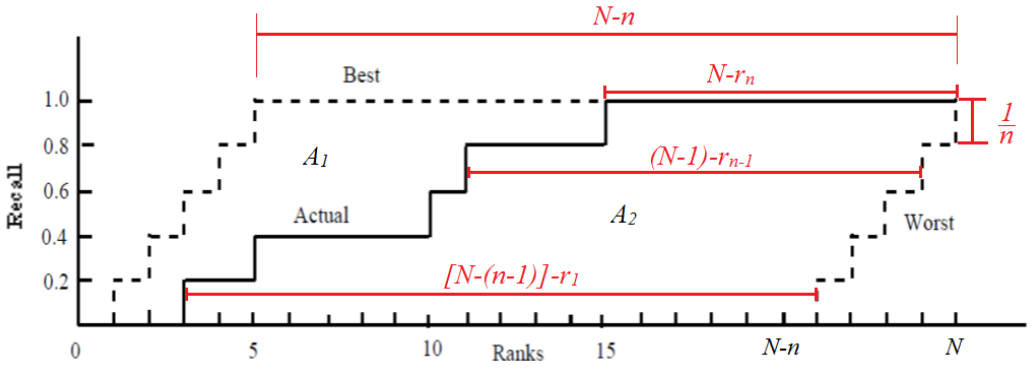


Figure 5.1' How A_1 and A_2 can be calculated from R_{norm} graph

From Figure 5.1', the areas A_1 and A_2 values are

$$A_1 + A_2 = n \times (N-n) \times \frac{1}{n}$$
$$= N - n$$

$$A_2 = \sum_i [N-(n-i)-r_i] \times \frac{1}{n}$$
$$= \frac{1}{n} \times \sum_i [(N-n)-(r_i - i)]$$
$$= \frac{1}{n} \times [\sum_i (N-n) - \sum_i (r_i - i)] = \frac{1}{n} \times n(N-n) - \frac{1}{n} \times \sum_i (r_i - i)$$
$$= (N-n) - \frac{1}{n} \times (\sum_i r_i - \sum_i i)$$

$$\therefore R_{norm} = \frac{A_2}{A_1 + A_2} = \frac{(N-n) - \frac{1}{n} \times (\sum r_i - \sum i)}{N-n}$$
$$= 1 - \frac{\sum r_i - \sum i}{n(N-n)}$$

Appendix B: Indri Query Language

```
<query>
  <number>PAC-132</number>
  <text>
  #filreq(#syn(F24J B23K F28D F28F F24D).class #weight(35 radiat 32
  aluminum 29 tube 28 copper 27 plate 25 weld 21 heat 14 alloy 14
  water 13 pipe 13 illustr 10 embodi 10 space 8 channel 7 panel 7
  shape 7 steel 6 laser 6 fin 6 beam 6 fig 5 object 4 puls 4 suitabl
  4 horizont 4 thick 4 convect 4 transfer 4 distanc 4 increas 4 view
  3 tabl 3 refer 3 secur 3 thermal 3 serpentin 3 achiev 3 vertic 3
  conduit 3 detail 3 diamet 3 paint 3 assembl 3 power 3 configur 3
  transport 3 distribut 3 side 3 limit 3 absorb 3 medium 3 acut 2
  shown 2 press 2 angl 2 conduct 2 direct 2 energi 2 mention 2
  infrar 2 emiss 2 meander 2 transvers 2 corrug 2 surround 2 face 2
  recess 2 solar 2 focus 2 long 2 lead 2 order 2 temporarili 2
  mechan 2 accord 2 overcom 2 section 2 size 2 curv 2 path 2 hot 2
  capac 2 develop 2 manufactur 2 appli 2 bond 2 descript 2 drawback
  2 perman 2 #1(laser beam) 2 #1(copper tube) ))
  </text>
</query>
```

In this appendix a sample Indri query for the CLEF-IP 2010 patent search task is presented in order to ease replication of this work. As noted in Section 4.2, all terms in the description section of a patent topic which occurred more than once were included in the query; in addition bigram phrases which appeared more than once in the title and abstract sections combined were included. All terms are in their stemmed form since the collection text is stemmed as well. An example query is shown above. This is one of the shortest queries in the CLEF-IP 2010 task. The average length of patent queries in this task is much larger than the presented one, where the average

number of unique terms of English topics in CLEF-IP 2010 is 317 unique terms, and the longest topic reaches 1633 unique terms.

In the Indri query language:

- #filreq(x y) operator is used to apply filtering while retrieving; where it does not retrieve any documents which do not contain "x", and it ranks documents based to the query "y".
- "#syn" operator treats all terms between its brackets as synonyms, where finding any of them is considered a hit. In the example shown, all the classifications for the patent topic are listed in the "#syn" operator in order to allow the retrieval of documents which share at least one of the listed classifications, and filter out anything else using the "#filreq".
- "#weight" operator allows giving weights to different terms without the need of repeating the term in the query. This way is more efficient than using the patent sentence with terms repeated in different locations.
- "#1" operator matches words inside brackets as an exact phrase.

Additional operators are used for constructing Indri queries in the described work in this thesis. The "#wsyn" operator is the same as the "#syn" operator, but can give different weights to each term used as a synonym. For example: *#wsyn(0.63 motor 0.37 engin)* searches for either the term "motor" or the term "engine" in the documents, but gives higher weight to the term "motor".

Appendix C: Publications List

The following publications are based on the work appearing in this thesis:

- Magdy W., J. Leveling, and G. J. F. Jones. Exploring Structured Documents and Query Formulation Techniques for Patent Retrieval. In *Proceedings of the CLEF 2009: Workshop on Cross-Language Information Retrieval and Evaluation,* Corfu, Greece, Springer, pages: 410-417, (2009)
- Magdy W. and G. J. F. Jones. A New Metric for Patent Retrieval Evaluation. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe'10) at 32nd European Conference on Information Retrieval (ECIR 2010)*, Milton Keynes, U.K., (2010)
- Magdy W. and G. J. F. Jones. PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. In *Proceedings of the Thirty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, Geneva, Switzerland, ACM, pages: 611-618 (2010)
- Magdy W. and G. J. F. Jones. Examining the Robustness of Evaluation Metrics for Patent Retrieval with Incomplete Relevance Judgements. In *Proceedings of CLEF 2010: Conference on Multilingual and Multimodal Information Access Evaluation*, Padua, Italy, Springer, pages: 82-93 (2010)

- Magdy W. and G. J. F. Jones. Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task. In *Proceedings of the CLEF 2010: Workshop on Cross-Language Information Retrieval and Evaluation,* Padua, Italy, (2010)
- Magdy W., P. Lopez, and G. J. F. Jones. Simple vs. Sophisticated Approaches for Patent Prior-Art Search. In *Proceedings of the Thirty-Third European Conference on Information Retrieval (ECIR 2011),* Dublin, Ireland, Springer, pages: 725-728 (2011)
- Magdy W. and G. J. F. Jones. Should MT Systems be Used as Black Boxes in CLIR?. In *Proceedings of the Thirty-Third European Conference on Information Retrieval (ECIR 2011),* Dublin, Ireland, Springer, pages: 683-686 (2011)
- Leveling J., W. Magdy, and G. J. F. Jones. An Investigation of Decompounding for Cross-Language Patent Search. In *Proceedings of the Thirty-Fourth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011),* Beijing, China, ACM, pages: 1169-1170 (2011)
- Ganguly D., J. Leveling, W. Magdy, and G. J. F. Jones. Query Reduction based on Pseudo-Relevant Documents. In *Proceedings of the Twentieth Annual International ACM Conference on Information and Knowledge Management (CIKM 2011),* Glasgow, Scotland, ACM, pages: 1953-1956 (2011)
- Magdy W. and G. J. F. Jones. An Efficient Method for Using Machine Translation Technologies in Cross-Language Patent Search. In *Proceedings of the Twentieth Annual International ACM Conference on Information and Knowledge Management (CIKM 2011),* Glasgow, Scotland, ACM, pages: 1925-1928 (2011)
- Magdy W. and G. J. F. Jones. A Study on Query Expansion Methods for Patent Retrieval. In *Proceedings of the fourth International Workshop on Patent Information Retrieval*

(PaIR'11) at the Twentieth Annual International ACM Conference on Information and Knowledge Management (CIKM 2011), Glasgow, Scotland, ACM, pages: 19-24 (2011)

In addition to the publications, a patent was filed based on part of the work presented in this thesis, namely the cross-language search system described in Chapter 6.

- Magdy W. and G. J. F. Jones. A Method of and a System for Retrieving Information. *Filed by Dublin City University in the US Patent Office, Application No. 61/433,784, January 2011*