Towards Effective Cross-Lingual Search of User-Generated Internet Speech

Ahmad Khwileh

B.Tech., M.Tech.

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: Professor Gareth J. F. Jones

September 2018

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 the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

(Candidate) ID No.: 12211299

Date:

Contents

List of Figures

List of Tables

Abstract

Ahmad Khwileh

Towards Effective Cross-Lingual Search

of User-Generated Internet Speech

The very rapid growth in user-generated social spoken content on online platforms is creating new challenges for Spoken Content Retrieval (SCR) technologies. There are many potential choices for how to design a robust SCR framework for UGS content, but the current lack of detailed investigation means that there is a lack of understanding of the specific challenges, and little or no guidance available to inform these choices. This thesis investigates the challenges of effective SCR for UGS content, and proposes novel SCR methods that are designed to cope with the challenges of UGS content. The work presented in this thesis can be divided into three areas of contribution as follows.

The *first* contribution of this work is critiquing the issues and challenges that influence the effectiveness of searching UGS content in both mono-lingual and cross-lingual settings. The *second* contribution is to develop an effective Query Expansion (QE) method for UGS. This research reports that, encountered in UGS content, the variation in the length, quality and structure of the relevant documents can harm the effectiveness of QE techniques across different queries. Seeking to address this issue, this work examines the utilisation of Query Performance Prediction (QPP) techniques for improving QE in UGS, and presents a novel framework specifically designed for predicting of the effectiveness of QE.

Thirdly, this work extends the utilisation of QPP in UGS search to improve cross-lingual search for UGS by predicting the translation effectiveness. The thesis proposes novel methods to estimate the quality of translation for cross-Lingual UGS search. An empirical evaluation that demonstrates the quality of the proposed

method on alternative translation outputs extracted from several Machine Translation (MT) systems developed for this task. The research then shows how this framework can be integrated in cross-lingual UGS search to find relevant translations for improved retrieval performance.

Dedication

To my Lord (Allah): My Lord! Inspire and bestow upon me the power and ability that I may be grateful for Your Favours which You have bestowed on me and on my parents, and that I may do righteous good deeds that will please You, and admit me by Your Mercy among Your righteous slaves.

- (Quran (Surah An Naml) : Chapter 27, Verse 19)

Acknowledgments

First of all, I would like to express my sincere gratitude, respect and appreciation to my supervisor Dr.Gareth Jones for his valuable guidance and tremendous support throughout the course of my PhD studies. Being a part-time PhD student, Dr Gareth offered me with flexibility to meet him after working hours, during weekends and bank holidays which helped me to progress on my research just like other regular students in the lab.

My sincere thanks must go to the member of my thesis advisory panel, Dr.Andy Way, for his great support and guidance. I am also most grateful to the examiners of this thesis, Dr.Gregory Grefenstette and Dr.Jennifer Foster for their insightful suggestions and constructive feedback on improving this work.

I would like to express a big thanks to my friend and mentor, Dr.Rami Ghorab, for introducing me to this research course, and consistently providing his valuable guidance, tremendous feedback and unfailing encouragement and support over the years of my study.

I would to like to thank the past and present members of our research group at DCU who provided me with their enormous help and support during this research : Dr.Haithem Afli, Dr.Debasis Ganguly, Dr.Walid Magdy, Dr.Piyush Arora, Dr.Keith Curtis and Dr. David Nicolas Racca.

Special thanks must go to all the colleagues and friends at Google from the Search Quality, Publisher Quality and gTech teams who sponsored me to pursue this work and provided their limitless support to successfully finish it.

There are no proper words to convey my deep gratitude to my parents, Yousef Khwileh and Huda Khwileh, my brothers and sisters for all the unfailing emotional support, limitless encouragements and love to be able accomplish this work.

Last but not least, my wife, Malak Alazeez, and my daughter Joana Khwileh, who were there for me when times were tough, stayed by my side, cheered me on, and celebrated each accomplishment. LOVE YOU both.

Chapter 1

Introduction

Increasing amounts of user generated multimedia content are being uploaded to social video-sharing websites such as as YouTube (Youtube, 2017), Facebook (Facebook video, 2017), BlipTv (BlipTV, 2017) and many others. In 2016, YouTube, the predominant social video-sharing site, reported that 300 hours of video content were uploaded every minute in over 75 different languages (YouTube Press, 2016). The ease and the flexibility of multimedia content production, coupled with the low cost of publishing and wide potential reach, are driving an exponential growth in the amount of multimedia content available on the Web.

While much of the uploaded content has both audio and visual elements, for a significant amount of it, the informational content is primarily in the audio stream in spoken form. We refer to this type of content as **User-Generated Speech (UGS)**. Unlike other speech types, UGS content is uploaded to social-media platforms by different producers/uploaders with varying background, interest, style, and recording settings. Some of this UGS content, such as news broadcasts and TV shows, is carefully authored, edited and quality controlled, while others such as videoblogs and personal recordings are not.

Along with the unprecedented increase in the amount of UGS content available online, there is an increasing user demand to access this content. For example, based on recent statistics from YouTube, it has been reported that there are over

a billion Internet users, every day, watch hundreds of millions of hours of videos and generate billions of views (YouTube Press, 2016). The very large amount of content available, together with the very complex and inconsistent structure of this content are creating the need for the development of sophisticated Spoken Content Retrieval (SCR) systems designed to address these challenges in order to enable an effective search over this type of content.

Furthermore, UGS content is often very uneven in quality and topical coverage in different languages. The lack of material in individual languages means that cross-language information retrieval (CLIR) within these collections is required to satisfy the user's information need. Search over this content is dependent on available metadata, which includes user-generated annotations and often noisy transcripts of spoken audio. The effectiveness of CLIR depends on translation quality between query and content languages. Building an effective retrieval framework for such a large scale, highly varied, and multilingual archive of UGS content presents new challenges and exciting opportunities for Information Retrieval (IR) research (Naaman, 2012; Bendersky et al., 2014).

SCR systems require the combination of technologies from Speech processing and Information Retrieval (IR). SCR utilises Automatic Speech Recognition (ASR) systems to generate textual transcripts from spoken audio. In fact, SCR can be considered as the application of IR techniques to the extracted ASR transcripts. Challenges for SCR research vary across application and domains, but the most common one is with regard to dealing with the recognition errors associated with ASR transcripts (Larson and Jones, 2012). ASR is a fully automatic process that comes at a cost, that the final output is far from perfect and contain errors where the system *incorrectly* recognises, or even misses some of words that were presented in the original speech file. Research in ASR has made a considerable progress in recent years towards developing technologies to improve the quality of its output transcripts (Li et al., 2014).

SCR research initially focused on planned, formally edited and structured speech

(Garofolo et al., 1997, 1999a,b, 2000), but has then shifted focus to more informal spoken content produced spontaneously outside of the studio, and in conversational settings (Larson and Jones, 2012; Eskevich, 2014). However, there has been limited SCR research focused on UGS content. The inconsistency in content quality, style, structure, and content available across languages together with the ASR noise in UGS presents challenges for IR systems where previous SCR work did not explore.

This thesis contributes towards addressing and analysing the interdisciplinary SCR challenges for UGS content. More importantly, we put these analytical and research efforts into a practical applications to help IR techniques to effectively process and retrieve this type of content.

The next sections introduce the research challenges and questions of this thesis.

1.1 Research Challenges for UGS Search

Foundational research in SCR focused on professionally-generated speech collections (e.g. Jones et al., 2007; Pecina et al., 2007). The term *professionally-generated* indicates that the spoken documents were generated and recorded in a professional settings, where speech transcripts exhibit consistent theme of style, length, quality and metadata. The scale, dynamics, and decentralisation of UGS content means that the conclusions of earlier studies cannot be applied directly without taking into consideration and investigating these issues.

UGS content often consists of multiple sources of information with varying reliability, such as ASR transcripts and user-created metadata, making it very different from the professionally generated collections that were explored in previous SCR work such as the work reported in (Oard et al., 2006; Pecina et al., 2007; Larson et al., 2009). The distinct characteristics of the noise associated with each data source, such as the informal words using in the metadata, spelling errors and the inconsistency of the metadata, as well as the high variability of length and quality of ASR output, pose significant retrieval challenges that requires special consideration.

The following points indicates the major monolingual retrieval challenges for UGS content.

- *Distribution of the document lengths*: UGS is a non-scripted, informal spoken content generated by different social-media users with varying style and interest. These users can upload a spoken document in any length. Therefore, in UGS, there is no restriction or control on document lengths, and they are found to be highly variable.

- *High variability in ASR quality*: Even if the same ASR system is used to process a UGS collection, the variation in the audio quality, speaking styles and speakers generally leads to significant variability in the reliability of these transcripts.

- *High variability in topical structure*: Along with the high variability in length, documents in UGS can have uniquely varied topical structure. Some documents may cover a single topic or covering multiple closely related or distinct topics that cannot be easily segmented and indexed for SCR usage.

- *Inconsistency and sparseness of the associated user-contributed metadata*: UGS content is often presented with metadata authored by the uploader such as the title or description of the video. These titles may be very short made up of only one or two terms, while descriptions can be long, generic, informal and sometimes incomplete, making their utility for retrieval very varied and unreliable.

This length, structure and quality inconsistencies pose a significant reliability of SCR when tuning, training a search system for a single setting is not possible.

Another key challenge for the effective exploitation of UGS content in a multilingual setting is effective search between the languages of the user queries and the content metadata. From multilingual perspective, the quality of UGS depends solely on the characteristics of the individuals who actually produce or upload videos in

each language. The lack of formal editorial control means that the uploaded videos are typically very varied across languages in terms of audio and metadata quality. Furthermore, *the quantity and topical coverage* of UGS content across different languages is very uneven. This indicates that satisfying an information need for a user of one language can only be achieved by providing relevant content in another language. For example, bilingual Arabic speakers frequently enter Arabic queries for which the only relevant content is in English.

To illustrate this content variation, Figure 1.1 shows an example of the *Google Video*[1] search engine with a simple Arabic query محاضرة في نظم استرجاع المعلومات and the equivalent English query *Information retrieval system lecture*, to satisfy an information need for lectures in the area of IR. For the English query, the video search engine[2] was able to find more than 10,000 matching results in English with all top ranked results being relevant with high quality metadata [3]. While for the Arabic query, the search engine only located 461 matching results with only one of the top 10 results identified as relevant[4].

This, in fact, is a use case for CLIR which seeks to enable users to enter search queries in one language to retrieve relevant content in another one. In CLIR, translation technologies are key to successfully bridging the language gap between a user's query and the relevant content (Oard and Diekema, 1998; Herbert et al., 2011).

Arabic language is currently used by an estimated 420M speakers in different countries, making it one of the most spoken languages on earth [5], and is currently the language with the largest growth in Internet users in the last decade with an estimated 2500% growth. In 2016, there were an estimated 168M Internet users with 45% Internet penetration (Internetworldstats.com, 2017). However, the Arabic content available online is still minimal, estimated as being less than 0.1% of the

[1]www.google.ie/video
[2]Retrieved from www.google.com/video on 2017-01-15
[3]This is based on relevant assessment done by a fluent speaker of English
[4]This is based on relevant assessment done by a native speaker of Arabic
[5]https://en.wikipedia.org/wiki/Arabic

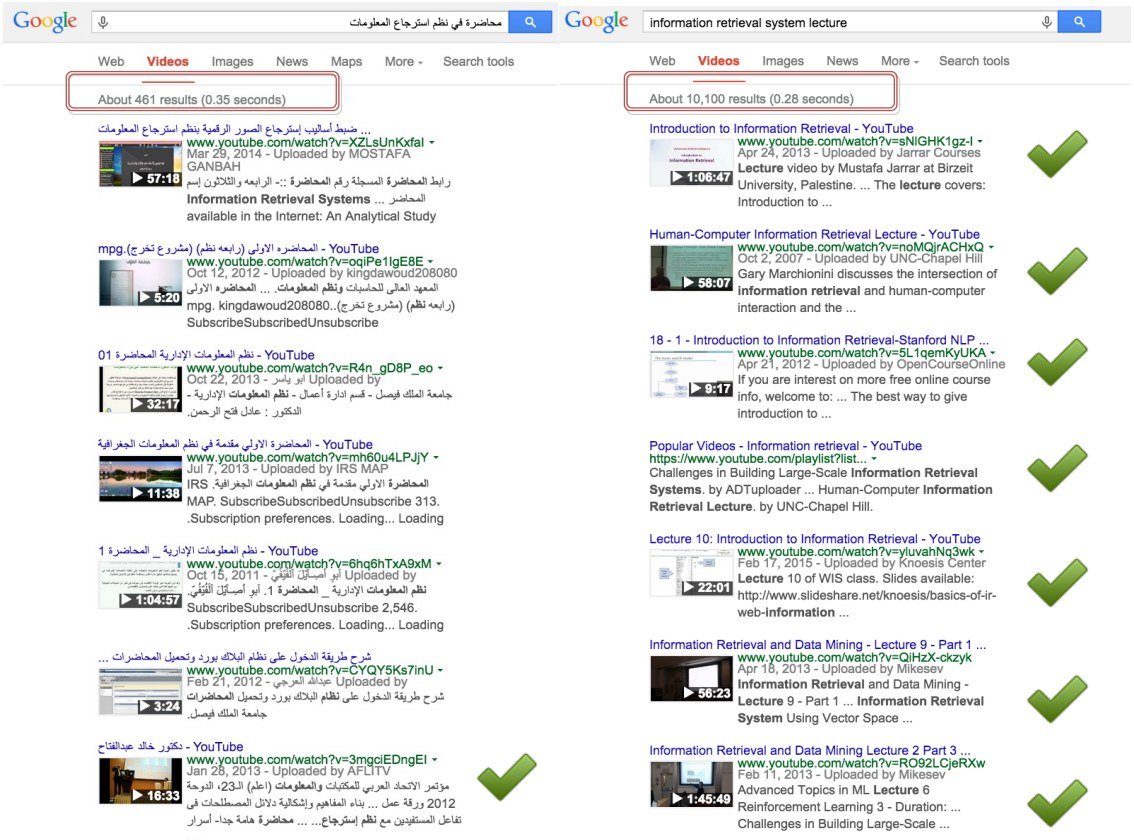

Figure 1.1: Example of the content variation issue across languages: Video search results for Arabic and English queries.

Internet content [6].

Furthermore, some Arabic speakers are actually comfortable reading and listening to languages other than Arabic, notably French and English. They are though frequently unable to express their information need in languages other than Arabic. To support the need of online users such as the Arabic speakers, Search engines are required to provide an effective cross-lingual tools to enable multilingual access over UGS content.

Although employing CLIR to this task may sound like a trivial problem where a translation module could be integrated into the search system, maintaining search effectiveness within a real-world, topically diverse and noisy speech collection can be very challenging (Khwileh et al., 2015, 2016). In this setting, CLIR effectiveness may not only suffer from issues arising from imperfect translation, but also other

[6]https://en.wikipedia.org/wiki/Languages_used_on_the_Internet

challenges that arise from the uncontrolled amount of noise in the UGS content. The main goal of this work is to propose an effective retrieval framework to cope with the problems of imperfect translation and uncontrolled amount of noise of this task.

1.2 Scope of the Work

The work reported in this thesis deals with an informal real-world UGS collection harvested from social video sharing site with specific focus in search. Apart from the scale and the content variation issues, this work studies the effectiveness of UGS modalites that have not been explored before such as automatically-generated speech segments and user-generated metadata.

This research considers tasks such those which were explored in the state-of-the-art adhoc SCR tasks (e.g. Pecina et al., 2007; Jones et al., 2007) using the topical text-based queries that focus on the video topic rather than what visual content it may contain (i.e. car, sky, professor).

A typical video or speech-based multilingual search engine may contain several components that can be studied for UGS as such as:

- Visual content analysis for UGS-based video content.
- ASR and speech processing for UGS.
- Machine Translation (MT) system for enabling CLIR access over this type of content.
- IR System for indexing and retrieving UGS data.

This work does *not* intend to provide any further contribution on the speech/visual content analysis, neither on the MT side for UGS content. Instead we deal with these components as black boxes, and where possible tune and adjust their output for the interest of SCR. Therefore, the research questions of this thesis are targeted towards

studying the interaction between each of the main components (social content, ASR transcripts, MT systems) with the IR component as outlined in the next section. In particular, this research introduces multiple novel IR techniques to address the challenges of SCR over UGS as follows.

- Utilisation of metadata fields and Query Expansion in cross-lingual and monolingual search of UGS. (Khwileh et al., 2016)
- Utilisation of automatically-generated passage-evidence to address the length variation issues of UGS transcripts. (Khwileh and Jones, 2016)
- Query Performance Prediction (QPP) to automatically select the right retrieval settings for UGS. (Khwileh et al., 2017a)

1.3 Research Questions

In this thesis, the special nature of UGS content is studied to understand the challenges of UGS content. Understanding the nature and challenges of UGS leads us to investigating possible effective solutions for improving search over this content. Therefore, our research questions are targeted toward each component of the retrieval framework, starting with understanding the overall retrieval behaviour using IR techniques continuing to what were done in previous work (e.g. Jones et al., 2007; Pecina et al., 2007), then moving towards improving query and document representations to implement IR techniques that are designed for UGS content. This thesis seeks to investigate these issues by answering the following Research Questions (RQ).

- **Research Question 1 (RQ1)** *(Understanding UGS search as an IR task)*:
    1. What are the main challenges that face UGS monolingual and cross-lingual retrieval, and how different are they from other SCR tasks?

2. How do Internet-collected UGS data sources behave in monolingual and cross-lingual retrieval? How do they behave when combined and weighted together using state-of-the-art retrieval frameworks?

- **Research Question 2** (RQ2) *(Improving query representation retrieval)*:

  1. How do traditional query expansion approaches work under such a setting of noisy data collected from Internet videos?
  2. Can we have an effective QE approach that adaptively utilises UGS data sources to expand individual queries in order to improve overall retrieval effectiveness?

- **Research Question (RQ3)** *(Improving document Representation in UGS retrieval)*:

  1. Can automatic speech segmentation be beneficial in improving QE effectiveness for UGS content?
  2. What are the characteristics of the most effective speech evidence (i.e. speaker based or window based speech segments, full ASR document) for UGS retrieval?
  3. Can we develop a technique to predict the most effective speech segmentation for each query?

- **Research Question 4 (RQ4)** *(Towards an Effective CL-UGS retrieval)*:

  1. Can a prediction technique be developed to estimate the translation quality of CLIR for UGS content?
  2. Can we implement an adaptive CLIR technique that is able select the most effective translations in UGS retrieval?

1.4 Contributions of Proposed Research

To the best of our knowledge, current research does *not* provide the proper methods to understand and manage UGS Internet based data for SCR. The aim of this thesis is to bridge this gap and contribute to discover the characteristics of UGS content. The main focus of this work is to develop adaptive IR techniques to be able to deal with uncertain and noisy settings of UGS.

The contributions of this thesis are summarised as follows.

1. An evaluation framework for analysing and understanding the retrieval challenges of UGS content. Published as a full conference paper in (Khwileh et al., 2015).

2. A novel QE technique for effective utilisation of metadata in UGS retrieval, published as full journal article in (Khwileh et al., 2016)

3. A novel prediction technique to estimate the effectiveness of Query expansion in UGS retrieval, published as a full conference paper in (Khwileh et al., 2017b)

4. A novel adaptive QE technique that utilises both automatic segmentation and query performance prediction to deal with the retrieval robustness of UGS. Published as a long conference papers in (Khwileh and Jones, 2016; Khwileh et al., 2017b)

5. A novel prediction technique to estimate the translation effectiveness of cross-lingual UGS retrieval, published as a full conference paper in (Khwileh et al., 2017a).

6. A novel adaptive CLIR technique for UGS retrieval, published as a long conference paper in (Khwileh et al., 2017a).

1.5 Thesis Structure

The rest of this thesis is organised as follows.

- Chapter 2 provides a brief background on fundamental IR and CLIR techniques that is used in this thesis. The chapter starts by providing an overview of IR systems and their evaluation. Then introduces the standard retrieval models in IR such as the VSM, BM25 and DFR. Chapter 2 also provides an overview of some well established IR tools such as QE and QPP, which are heavily used in this thesis. Finally, the chapter presents an overview of CLIR, as well as the most well established technique to CLIR such as query and document translation approaches CLIR.

- Chapter 3 presents a survey from the related work of SCR, as well as an overview of the current state-of-the-art advances in CLIR research for related tasks.

- Chapter 4 describes basic components of IR evaluation framework. In addition, the chapter explains the data collections, the query test sets, as well as the experimental settings used to perform the research presented in this thesis.

- Chapter 5 describes the research and experimental investigation we conducted to answer RQ1 of this thesis. In particular, this chapter presents the experimental investigation of the challenges of UGS retrieval (RQ1), as well as the initial experiments on utilisation of metadata fields.

- Chapter 6 describes the research investigation of standard QE methods for UGS retrieval to address RQ2 of this thesis. Chapter 6 also presents the field based QE that is proposed to utilise UGS fields in QE for UGS retrieval.

- Chapter 7 describes the experiments in utilising automatic speech segmentation for UGS retrieval. In particular, Chapter 7 addresses the RQ3.1 and RQ3.2 for analysing the utility of segment and document-based QE for UGS retrieval.

- Chapter 8 presents the proposed technique to predict the most effective evidence for QE in UGS retrieval (RQ3.3).

- Chapter 9 presents the proposed prediction method to estimate translation quality for CLIR. Furthermore, Chapter 9 describes the adaptive CLIR technique that is proposed for this task (RQ4).

- Chapter 10 provides a summary of the presented work, revisits each of the research questions as well as how they were addressed in this thesis. Finally, Chapter 10, outlines the directions for future work.

Chapter 2

Overview of Information Retrieval Methods

This chapter provides the necessary background to understand the subsequent chapters of this thesis. The chapter begins with an overview of the basic information retrieval methods that relevant to this work, and then introduces more advanced methods as Relevance Feedback (RF), Query Performance Prediction (QPP) and Cross Language Information Retrieval (CLIR).

2.1 Information Retrieval

Information Retrieval (IR) is the science of retrieving relevant information to satisfy a user's information need. IR techniques are used in search applications, such as Web search engines, to identify relevant information in order to satisfy user's information needs. The information need is generally expressed as a *query* statement that is a set of words written by the user to obtain some valuable information on a particular topic. IR systems are required to process the query in order to identify and retrieve relevant documents. Correspondingly, IR systems are also required to process and index documents of a given collection to make them accessible via user queries. This is achieved by organising the search collection in such a way that relevant

documents can be retrieved at the search time, in an efficient, and effective manner. The process of converting documents collection into a searchable corpus, or index, is often referred to as *indexing*.

A general overview of the IR process is shown in Figure 2.1. Four major process can be generally found in an IR system, explained as follows.

- *Indexing*, to process the raw documents of a given collection into a search index. A suitable data structures should be utilised to allow efficient, and effective retrieval of documents from a collection. The most popular data structure used for this purpose is known as an *inverted index* (Zobel and Moffat, 2006). For each term in the collection, the inverted index contains a term-posting list which lists the documents containing the term together with their associated weights and metadata.

- *Querying*, to process the query in order to understand the underlying information need and being able to locate relevant information in the search collection.

- *Matching*, to match the query against each document in the search index, and retrieve the set of documents that are similar and closely related to the query.

- *Ranking*, to score each of retrieved document based on its estimated relevancy to the query, and rank the retrieved results accordingly.

Among these processes, ranking is arguably the most important and challenging process in IR (Salton et al., 1975). The main challenge in ranking with regard to IR, is how to model the relevance between each document in the collection to a given query in order to be able to rank similar documents accordingly. Hence, the goal of a retrieval model is to rank the matched documents based in on their estimated relevancy to the information need.

IR models use mathematical ranking functions to score the matched documents in descending order based on their relevance. IR researchers suggested several ranking approaches to model the retrieval relevancy as accurately as possible.

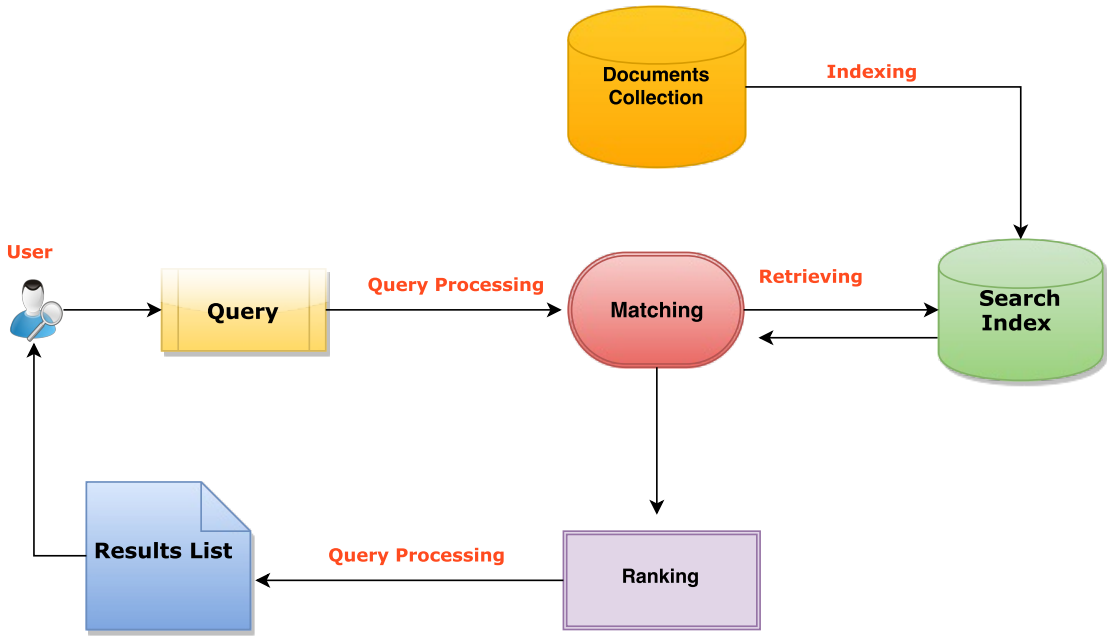

Figure 2.1: An overview of standard information retrieval process.

In the next sections, we review some of the well-known retrieval models from the literature, and the ones used in the experimental work of this thesis.

2.1.1 Vector Space Model

One of the oldest, most well-established of retrieval models is the Vector Space Model (VSM) (Salton et al., 1975). In the VSM, the query q and each document d are represented as vectors over the term space of the entire vocabulary V of the documents collection. The VSM model assumes that the relevance of a document to a query can be estimated by measuring the similarity between their vector representation. A vector representation of the query and documents allows to measure the distance between them within the vector space and hence, being able to induce the similarity between them. The underlying assumption is that the more similar the vectors are, the more relevant they are assumed to be.

The distance between the query and the documents in collection is often calculated by measuring the cosine of the angle (i.e. the ϕ) between two vectors, that is

the dot product of the two normalised vectors, as shown in Equation 2.1.

$$sim_{VSM}(d, q) = \sum_{i=1}^{V} d_i q_i = |d||q| \cos \phi \tag{2.1}$$

One vital process in VSM, is how to construct and generate the query and documents vectors. The VSM model typically utilises a bag-of-word approach to represent each vector, where values in the vector represent the weights for the terms that appears in the document or the query. The process of constructing the term weights for each of the created vectors is referred to as *term weighting*. A better term weighting function leads to a better estimation of relevancy, and hence an improved retrieval model.

The term weight functions of VSM, and other IR models, are typically composed of multiple components or features that are used to calculate the weight of each term. Generally speaking, retrieval models differ by the way these features are extracted and modelled to estimate relevancy. In the next sections, we introduce the main features of the term weighting functions.

1) Term Frequency (tf)

The frequency of a term tf in a document can be used to understand the *topic* of that document. For example, if the word *Internet* is highly frequent in a document, it can be induced that, for any query about the *Internet or the World Wide Web* this document might be relevant. Therefore, if a query term is highly frequent in a document, the weight of that term should be increased to boost that document up to a higher rank in the final list.

However, using the absolute value of the term occurrence does not always provide accurate estimation of its relevancy (Singhal, 1997). For example, if a document A contains the *same* query term being repeated thousands times, this does not necessarily mean that it is more relevant than document B which has the same query term occurred five times.

The weight of term tf is typically calculated using a number of standard functions. The following are the most popular techniques to tf implementation.

- $\frac{1}{2} + \frac{tf}{2\max(tf)}$, which normalises the term frequency values within a range of $[\frac{1}{2}, 1]$ (Salton and Buckley, 1988), where $\max(tf)$ is the maximally occurring term in the collection.

- $1 + \log(tf)$ (Buckley et al., 1993; Singhal et al., 1996), which aims to reduce the weight of highly frequent terms.

2) Inverse Document Frequency (*idf*)

While tf captures how common the term is in the collection, the *idf* represents the *uniqueness* of the term within the collection. Unlike the tf, *idf* captures the global frequency of the term in the collection to measure how important it is for retrieval. For example, the presence of common query terms in a document A such as "like" or "for" does not indicate whether it is relevant or not. *idf* avoids failing in such an issue by measuring Document Frequency (*df*), which is the number of documents that contain these terms.

By definition, *idf*, is the inverse of the *df*, in which it assigns a higher weight to those query terms which have a lower *df* (Salton and Buckley, 1988).

The *idf* measure is typically calculated as $idf(t) = \log(\frac{N}{df(t)})$, where N is the total number of documents in the collection and $df(t)$ is the number of documents in which t occurs (Jones, 1973).

Another common approach to IDF is using the the INQUERY formula which is calculated as follows (Allan et al., 1995; He and Ounis, 2006). $\frac{log2(N+0.5)}{Nt}{log2(N+1)}$ where Nt is the number of documents that contain the query term t, and N is the number of documents in the whole collection.

3) Document Length

The third common component included within term weighting function is the *length normalisation*. Longer documents with wordy text are likely to obtain higher tf values and hence rank higher. By contrast, relevancy is not restricted to long and wordy documents, therefore, IR system should estimate a document's relevancy independent of its length. *Length normalisation* approaches are designed to reduce the effect of the length bias within the term weighting function. Cosine normalisation is a common length normalisation technique proposed within the VSM model that involves reducing the length of each document vector by dividing its components with the magnitude of the vector. In cosine normalisation, the dot product of a document and the query (as shown in Equation 2.1) yields the value of the cosine of the angle between them. However, cosine normalisation was shown to perform poorly for large document collections (Harman, 1994; Singhal, 1997), and was replaced by more robust techniques such as *pivoted length normalisation* (Singhal et al., 1996). Length normalisation has become a dominant part of the IR models in any retrieval model, in the following sections, we introduce other theoretically motivated IR models and their use of term weighting and length normalisation in IR ranking.

2.1.2 Probabilistic Model

Probabilistic IR models are based on the posterior probability of a document d being relevant, given a query q, that is the $P(d = R|q)$ for each document d in the collection.

Retrieved documents are ranked in descending order of their estimated probabilities. This representation is known as the *probability ranking principle* (PRP) (Robertson, 1977) which indicates the following :

If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible

on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

The binary independence model (BIM) is an early version of PRP models that assumes that terms in the collection are all pairwise independent (Robertson, 1977). Probability in BIM is calculated based on the *Boolean assertion* of whether a term is presented in document or not. The BM25 is a later and more developed PRP model that extends the BIM by including information term weighting components (tf,*idf* and length normalisation) that were explained in the previous section (Robertson et al., 1994; Sparck-Jones et al., 2000). The BM25 model scores each document d by summing the *idf* values of the query terms multiplied by a tf function incorporating a document length normalisation factor as shown in Equation 2.2; where $tf(t,d)$ is the term frequency of a term t in document d, L_d is the length of document d, and L_{ave} is the average length of documents computed over the collection. k_1 and b are scalar parameters.

$$sim_{BM25}(d,q) = \sum_{t \in q} \log \frac{N}{df(t)} \times \frac{(k_1+1)tf(t,d)}{k_1(1-b+b\frac{L_d}{L_{avg}}) + tf(t,d)} \tag{2.2}$$

k_1 is tuning parameter for the term frequency contribution, while b is is responsible for determining the degree of the needed length normalisation smoothing.

Divergence from Randomness Models

the Divergence From Randomness (DFR) model is another PRP model, and is based on the *informativeness* of a document. The concept of a document's *informativeness* is measured based on the deviation of its terms frequencies distribution from a random distribution where *the more the divergence of the within-document term frequency from its frequency within the collection, the more the information carried by the word in the document* (Amati, 2003). DFR models estimate the informativeness of each term t in a document d by measuring the divergence of its tf in

the documents from that in the whole collection. DFR framework presents several retrieval models that are explained in details in (Amati and Van Rijsbergen, 2002; Amati, 2003).

The PL2 model is a probabilistic retrieval model using DFR framework. The reason this model is mainly selected in this thesis over other retrieval models, is our data collection and experiments specifications. As described before UGS content has a very large variation in the lengths of the metadata and documents . Previous studies, such as (Amati and Van Rijsbergen, 2002; Amati, 2003), showed that the PL2 model has less sensitivity to length distribution compared to other retrieval models. The PL2 document scoring model is defined in Equation 2.3.

$$Score(d,Q) = \sum_{t \in Q} qt_w . \frac{1}{1+tf_n} (tf_n \log_2 \frac{tf_n}{\lambda} + (\lambda - tf_n). \log_2 e + 0.5 \log_2(2\pi . tf_n)) \quad (2.3)$$

where $Score(d,Q)$ is the score for a document d for all query terms $t \epsilon Q$. λ is the Poisson distribution of F/N; F is the query term frequency every query terms $t \epsilon Q$ over the whole collection, and N is the total number of documents at the collection. qt_w is the query term weight given by $qt_f/qt_fmax; qt_f$ is the query term frequency and qt_fmax is the maximum query term frequency among the query terms. tf_n is the normalised term frequency defined in Equation 2.4, where l is the length of the document d. avg_l is the average length of documents, and c is a free parameter for the normalisation.

$$tf_n = \sum_d (tf. \log_2(1 + c. \frac{avg_l}{l})), (c > 0) \quad (2.4)$$

2.1.3 IR Evaluation

In search applications, it is important to determine the retrieval effectiveness of an IR system to evaluate which system is the most effective for a particular task.

IR systems are generally automatically evaluated by comparing the documents as returned by an IR system for a set of user search queries with *a sample* of relevant

documents identified by the user for each query. The retrieval effectiveness of an IR system is usually measured by ***recall*** that is the number of relevant documents (as assessed by the user) retrieved out of the total number of relevant documents in the collection, and ***precision*** which is the number of documents which are relevant out of the total number of documents retrieved. The recall measures how many of the total known relevant items the system has been able to retrieve for the user, while the precision measures the proportion of the retrieved items that are relevant. One of the most popular precision metrics is Average Precision (AP) and Mean Average Precision (MAP).

Equations 2.5 and 2.6 show the how AP and MAP are calculated for a set of search queries, where the relevant set of documents for a query term $t \ \epsilon \ Q$, is $\{d_1, \ldots, d_m\}$, and N being the total number of relevant documents for the query q available in the collection D, and R_k is the ranked list of documents from d_k to d_1 retrieved by the system in response to the query Q.

$$AP(q) = \frac{\sum_{k=1}^{N} \mathrm{P@R_k}}{N} \tag{2.5}$$

$$MAP(Q) = \frac{\sum_{q \in Q} AP(q)}{|Q|} \tag{2.6}$$

2.2 Relevance Feedback

The previous sections explained the concept of ranking and how IR models are being constructed to produce relevant results to the user. One fundamental problem in IR is that queries are often a poor expression of the user information need. Either because the users do not know enough about the subject to write an effective query, or they do not make the efforts to do so.

Within the IR process, a common approach to improve the initial query is to use Relevance Feedback (RF) (Rocchio, 1971; Carpineto and Romano, 2012). RF involves modifying the query or the search settings at the search time, and gener-

ates further retrieval runs with a *hopefully* improved ranking of the results. This modification process relies on a feedback evidence which is received on the initial results by explicitly asking the user to evaluate the retrieved results.

In the absence of user feedback, the top ranked documents are assumed relevant and used as feedback evidence to modify the query. This type of RF is often referred to as *Pseudo Relevance Feedback (PRF) or Query Expansion (QE)*. The assumed relevant documents are used to refine the search with an adjusted settings and add query terms aiming to return a better ranking of the original result. In particular, RF encompasses two processes as follows.

- Terms re-weighting, where the weight of each query term is adjusted based on their estimated relevancy. Relevancy is estimated based on the explicit feedback evidence, or for QE, it can be using their occurrences within the top ranked documents which are assumed relevant.

- Terms expansion, where a new useful terms are added to the original query to improve the matching process. These terms are extracted as the most-common terms from the top ranked documents. Typically, query term re-weighting is performed for both the newly added terms and the original query terms.

RF techniques address the vocabulary mismatch problem by improving the initial query to be able to locate more relevant documents. RF allows the retrieval of relevant documents even if they do not contain the original query terms (Carpineto and Romano, 2012). For example, if the query *"coffee machines"* is expanded to terms such as *"americano"* or *"espresso"*, the search system would be able to locate new set of documents that are also relevant to the initial query. However, if the newly added terms are not directly correlated with the topic of the initial query, new irrelevant documents can be retrieved and harm the retrieval effectiveness (Mitra et al., 1998a). For example, if the term *"java"* is added to the initial query, it will allow the retrieval of many non-relevant documents about "java programming" and harm the retrieval effectiveness. This poor term expansion is made worse when the

relevant documents have complex topical structure, such as our UGS collection, so that the expansion needs to performed only using the relevant topics to avoid issue. In Chapter 6 and Chapter 7, we show how the QE effectiveness in UGS is hindered by this topic drift issue and propose a new technique to improve QE effectivness in UGS retrieval. The next sections present the most well-established RF approaches from the literature.

2.2.1 RF using Rocchio Method

A classic example for QE is the Rocchio method (Rocchio, 1971), which was developed within the VSM framework. As explained in Section 2.1.1, the VSM assumes that the query and the documents in the collection are represented as vectors. The aim of Rocchio's method is to shift the query vector towards the vector of the feedback documents (which are assumed relevant)[1], and move it away from the the non-feedback ones (which are assumed non-relevant to the query).

The RF method as proposed by Rocchio is shown in Equation 2.7, where α, β, γ and are the weights of the original query vector q, R is the set of feedback documents used for PRF, and NR is the complementary set of non-relevant documents. The values of these parameters are tuned and set empirically for each retrieval task.

$$Q' = \alpha q + \frac{\beta}{R} \sum_{d \epsilon R} d - \frac{\gamma}{|NR|} \sum_{d \epsilon NR} d \tag{2.7}$$

2.2.2 RF in the Probabilistic Model

Within the probabilistic model, a common method to QE is based on the Robertson/Sparck Jones RF approach (Robertson, 1990). The main idea of this approach is to improve the weight of the extracted terms that has high *idf* values and occur frequently in the feedback documents. This approach is based on the relevance weight shown in Equation 2.8, where r is the number of relevant documents in which

[1]The terms **PRF documents**, and **feedback documents**, **Expansion documents**, are used interchangeably throughout this thesis, which is referred to the list of top-ranked documents used for QE

the term t appears, N is the number of documents in the collection, and n is the number of documents in which term t appears, R is the number of known relevant documents.

To perform the QE process, the terms in the top ranked documents are ranked using the Offer Weight (OW) which uses the obtained RW score weight of each term as shown in Equation 2.9. The top ranking terms are then extracted and added to the initial query.

$$RW(t) = log\frac{(r+0.5)(N-R-n+r+0.5)}{(n-r+0.5)(R-r+0.5)} \tag{2.8}$$

$$OW(t) = RW(t) * r \tag{2.9}$$

Another probabilistic RF model that is the QE DFR model (Amati, 2003). RF in DFR has two stages as follows.

- *Firstly*, it applies a DFR term weighting model to measure the informativeness of the top terms in the top ranking document. The main concept of the DFR term-weighting model is to infer the informativeness of a term by the divergence of its distribution in the top-ranked documents from a random distribution.

  The DFR weighting model used in this thesis is called Bo1, which is a parameter-free DFR model uses BoseEinstein statistics to weight each term based on its estimated informativeness. This parameter free model has been widely used and proven to be effective for multiple tasks such as (He and Ounis, 2007; Plachouras et al., 2004; Amati, 2003).

  The weight w of a term t in the top ranked documents using the DFR Bo1 model is shown in Equation 2.10, where tf_x is the frequency of the term in the pseudo-relevant set (top n ranked documents). P_n is given by $F(t)/N$; F is the term frequency of the query term t in the whole collection and N is the

number of documents in the whole collection.

$$w(t) = tf_x . \log_2(\frac{1 + P_n}{P_n}) + \log_2(1 + P_n) \tag{2.10}$$

- *Secondly*, the query term weight qt_w, is further adjusted according to the newly obtained weighting values of $w(t)$ for both the newly extracted terms and the original ones using Equation 2.11, where $w_{max}(t)$ is indicated by the maximum $w(t)$ values among the expanded query terms.

$$qt_w = qt_w + \frac{w(t)}{w_{max}(t)} \tag{2.11}$$

The following example illustrates the QE for the query : *EEE PC 900 Troubleshooting in laptop*

The terms *pc, laptop, mac, us, classrooms* are generated from running the DFR QE to take the top 5 terms from the top 5 documents. Note that the two expansion terms *pc* and *laptop* also appear in the original query, therefore, the weight for each of these terms is boosted to be greater than 1[2]. The new expansion terms (*mac*, *us* and *classrooms*) are added to the original query and their weights are adjusted based on their informativeness and uniqueness in the top n documents versus the whole collection. The term *mac* is predicted as informative and unique so it gets weight greater than 0 since it appears only in the top n documents while the other terms (*classrooms*, *us*) are assigned very low weights close to 0 since they also appear in other documents (non top-n). Using this method, the final expanded and reweighted query is explained as follows.

eee×1.0000, *pc*×1.9211, *900*×1.0000 ,*troubleshoot*×1.0000, *laptop*×1.29882, *mac*×0.2195, *us*×0.000037, *classroom*×0.000049.

Some of the original query terms may *not* appear in the top-terms such as (900, EEE and troubleshoot), in this case the formula in Equation 2.11 only gives them

[2] 1 is the normal weight for any term that appears on the original query

the same weight they would have in a single-pass retrieval settings.

2.3 Query Performance Prediction

Our work proposes several adaptive techniques that are able to automatically adjust the retrieval settings for each query in order to maximise effectiveness in UGS retrieval. We utilise Query Performance Prediction (QPP) methods to implement these adaptive IR techniques.

In Section 2.1.3, we explained how the performance of IR systems can be measured using performance evaluation metrics such as recall and precision. However, these metrics are not always practical to use due to several issues, explained as follows.

Firstly, performance metrics which require a human-based evaluation rely on the assumption that the ground truth data of the relevant-assessment is always available. However, in real-world settings, this information cannot be always for each new query. Furthermore, this human based evaluation is based on estimating the performance for a *sample* of held-out queries. By contrast, this sample is not necessarily representative of how other queries behave. This case can be very common to web search where queries are completely user-generated, and can significantly vary in structure and format.

Secondly, the performance metrics are often reported by taking the average across the obtained performance of the all tested queries. Though, performance for each query can vary where some queries perform significantly better than others. This indicates that, using these query-averaged metrics, the scores of poorly performing queries are typically masked by scores of other query performance (Voorhees, 2004).

QPP methods are designed to provide a query-level estimation of retrieval effectiveness without the need for human-based evaluation (Carmel and Yom-Tov, 2010; Hauff, 2010). The motivation behind QPP research in IR is to develop algorithms to infer the performance of query, and to utilise such an inference to adjust the

retrieval settings to maximise overall effectiveness of the IR system.

QPP methods are generally divided into two families, pre-retrieval and post-retrieval. In pre-retrieval QPP, prediction is based on analysing the query and its difficulty (He and Ounis, 2004; Cronen-Townsend et al., 2006; He and Ounis, 2006; Hauff et al., 2008; Hauff, 2010). Query difficulty is an estimate that defines whether relevant content is hard (hence a low retrieval performance is presumed) or easy (high retrieval performance is presumed) to find given a certain query. In post-retrieval QPP, retrieval results of the query are analysed to estimate its performance (Kurland et al., 2011; Shtok et al., 2012).

2.3.1 QPP Quality Evaluation

The effectiveness of QPP methods is evaluated by studying the correlation between the values obtained by the QPP and the actual performance evaluated using average precision (AP at a cutoff of 1000) values for queries in a given query set (as measured by using relevance-judgements) (Voorhees, 2003a; Carmel and Yom-Tov, 2010; Hauff, 2010).

Correlation is measured using three well-known metrics which are used in the query performance prediction framework for evaluating prediction performance. For these evaluation measures, higher correlation values indicate increased prediction performance.

The most common metric is the Pearson's r correlation coefficient that demonstrates the linear relationship between two variables X and Y by looking the variance VAR and covariance Cov of their distributions.

Pearson's r is defined as follows

$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}.$$

Other metrics such as the Spearman's ρ (sometimes labeled r_s) and Kendalls τ that they are both designed to capture the non-linear relationship between the

studied variables X and Y (Voorhees, 2003a). Nevertheless, although they may differ in definition, previous efforts reported that there is no significant difference in terms of which correlation metric to use for QPP (Carmel and Yom-Tov, 2010; Hauff, 2010; Kurland et al., 2011; Shtok et al., 2012)

2.3.2 Using QPP in UGS retrieval

QPP methods have been utilised to estimate the retrieval effectiveness for text-based IR tasks (Hauff et al., 2008; Hauff, 2010; Shtok et al., 2012). In this thesis, we utilise QPP to develop adaptive IR techniques to deal with the uncertain and noisy setting of UGS retrieval. We analyse the effectiveness of both pre and post retrieval QPP methods to predict the performance of retrieval within this UGS settings. The utilisation of QPP in this thesis is explained as follows.

- In Chapters 6 and 7, we discuss the issues and the main challenges for UGS retrieval and the potential improvement that can be gained by using QPP.
- In Chapter 8, we analyse the effectiveness of several QPP methods for improving QE for UGS retrieval. We propose a novel prediction framework that can be used to develop an adaptive QE approaches for UGS retrieval.
- In Chapter 9, we utilise QPP methods for predicting the performance of translation for CL-UGS retrieval. We also propose an adaptive CLIR approach that utilise QPP to improve the translation quality for this task.

2.4 Overview of Cross Language Information Retrieval

Cross Language Information Retrieval (CLIR) has been an active research area since the mid 1990s. Numerous CLIR methods have been researched, implemented and widely tested in the literature for many domains and across different document

types. CLIR has formed the focus of several workshops which have been held to investigate the problem of CLIR for different tasks.

Starting from text document, where the first workshop on CLIR research was held at the SIGIR 96 [3] conference (Grefenstette, 1998), then moving towards multimedia content (including image (Peters et al., 2004), speech and professionally-generated video content (Federico and Jones, 2004; Federico et al., 2005; Pecina et al., 2007; Larson et al., 2010)), and most recently towards user-generated textual content (Bagdouri et al., 2014; Lee and Croft, 2014).

The research presented in this thesis is the first effort to investigate CLIR over real-world spoken content collected from the Internet. The aim of this work is to investigate the retrieval challenges of CLIR in a UGS settings, and provides directions to deliver robust and effective multilingual access over this content.

The goal of CLIR is to satisfy a user information need expressed as a query in one language using content from another language. CLIR techniques use translation to bridge this language barrier between the query and the indexed content. Translation techniques in CLIR differ mainly in where the translation module is to be placed, either in the query processing or the document indexing stage.

Figure 2.2 shows how CLIR techniques can utilise translation technologies to bridge the barrier between query language (L2) and document language (L1). The Query Translation approach (QT CLIR) is the most common CLIR technique (e.g. Oard and Diekema, 1998; Herbert et al., 2011; Sokolov et al., 2014); where the query is translated to match the index language (L1). This technique is known to be low cost (per translated query) and easy to implement since a translation tool can be used online at retrieval time to translate the query into the document language. However, this approach is very dependent and sensitive to the quality of the query translation for retrieval. Some queries may lack context and semantic content, which makes them harder to interpret and translate reliably.

[3]Special Interest Group on Information Retrieval (1996) Conference : http://sigir.org/sigir1996/

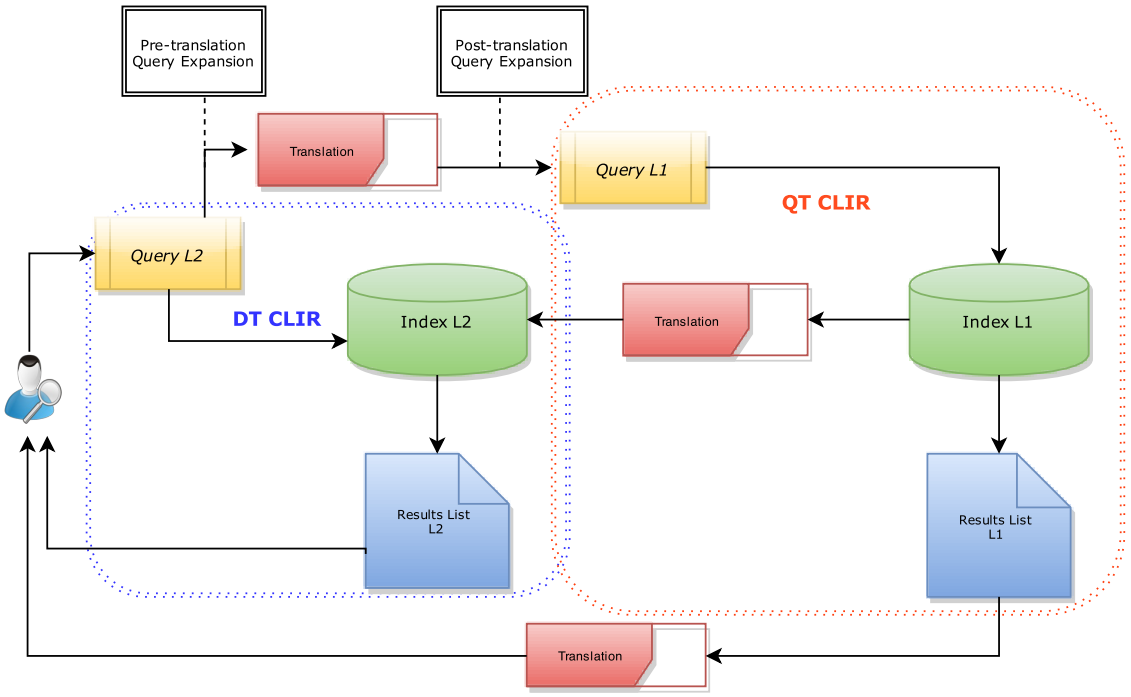

Figure 2.2: An architecture of CLIR system showing the approaches of QT and DT CLIR with the use of both pre- and post-translation QE.

Previous work has explored multiple techniques to overcome these issues either by improving the translation quality using various translation techniques (e.g. Chen et al., 1998; Gao et al., 2001; Varshney and Bajpai, 2014; Lee et al., 2010) or by improving the query itself using query reformulation techniques such as QE in RF process (Carpineto and Romano, 2012). Although noisy, QE is shown to be effective to improve the CLIR performance for multiple tasks (Bellaachia and Amor-Tijani, 2008). In the cross-lingual settings, the query can be expanded before translation to make it easier to process and to translate using pre-translation QE (Ballesteros and Croft, 1997). QE expansion can also be applied after translation (post-translation QE), or using a combination of pre-translation and post-translation QE as shown in Figure 2.2, in order to combat errors induced by the query translation (e.g. Ballesteros and Croft, 1998; Rogati and Yang, 2001).

The alternative to (QT CLIR) is Document Translation (DT CLIR) where all documents in the collection are translated to the query language (e.g. Oard and Hackett, 1997; Lee and Croft, 2014). Several arguments suggest that document

translation should be competitive or superior to QT CLIR for some tasks, due to the fact that it is less sensitive to translation errors. DT CLIR has the advantage that all translation is carried out offline prior to the retrieval, which allows for the possibility of having a more tuned and accurate translation.

Another advantage of DT CLIR is that it does not require any result translation as shown in Figure 2.2, since documents are already translated during index time. However, while DT CLIR has shown to be effective for several tasks, its application in CLIR settings is impractical due to the very large amount of time and resources required for document translation. Particularly, when the document collection is large and search is to be carried out across multiple language pairs. A less common but proven to be an effective CLIR technique, is the Hybrid CLIR approach which utilises both document and query translation approaches, thus allowing the relative advantages of both approaches to complement each other (e.g. McCarley, 1999; Kishida and Kando, 2006; Parton et al., 2008).

In the next section we describe how translation is implemented within CLIR systems.

2.4.1 Translation Technologies in CLIR

Several approaches have been proposed to carry out the translation process within the CLIR framework. The most commonly used ones are bilingual dictionaries and machine translation (MT) (Zhou et al., 2012).

Bilingual dictionaries perform a word-by-word translation using a machine-readable dictionary which has sets of entries of words and their possible translations in the other language (Pirkola et al., 2001). This approach can suffer from issues such as coverage, since some words may not be contained in the machine-readable dictionary, and ambiguity since it relies on a dictionary where many words have multiple possible translations and selecting the correct translation among them is a non-trivial task.

Machine translation (MT) techniques use a trained system to perform an auto-

matic translation of free-text from one natural language to another (Nikoulina et al., 2012; Magdy and Jones, 2014). While MT can also have similar dictionary coverage problems, the creation of single best translation addresses the translation ambiguity issues.

In recent years, MT has become the most commonly used technique in CLIR due to the increasing availability of high quality, and easy to use off-the-shelf MT tools. Most recent CLIR research has used the translation module as a black box without any control over the translation process, making use of the freely available online translation tools such as Google Translate [4], Bing translate[5] and others, which have proven to be effective. For example, in the CLEF evaluation campaigns[6] 2009, the best performing non-Google MT system achieved just 70% of the performance achieved by Google Translate tool (Leveling et al., 2009). Furthermore, there are several open source MT libraries available which have been used for CLIR research. These libraries allow for more tuned and flexible MT training and decoding, e.g. Moses (Koehn et al., 2007a)[7], and MaTrEx [8] (Stroppa and Way, 2006).

2.4.2 Statistical Machine Translation

The task of MT is to take input sentences in source language L and automatically produces output in target language T, where the output is *adequate* and *fluent*. An *adequate* translation of particular text indicates that it carries the same semantic of the original text. While *fluent* translation indicates that it is grammatically understood by native speaker of the target language. While fluency and adequacy are an important for creation of natural language using MT, CLIR is more concerned with accurate lexical and semantic translation to improve the retrieval process.

MT research began in the 1950s (Locke and Booth, 1955), and showed a rapid

[4]http://translate.google.com/
[5]http://www.microsoftranslator.com
[6]http://www.clef-initiative.eu/edition/clef2009
[7]http://www.statmt.org/moses/
[8]http://www.openmatrex.org/

development over the past decades. Early MT research focused on developing rule-based and knowledge-based systems, where linguistic and language experts *manually* create a set of rules to how text in one language could be translated into another language using both structural and lexical transformation of the target language (Koehn et al., 2003).

With the increasing availability of multilingual and bilingual training corpora, as well as the computational storage and power in the 1990s, the focus started to move towards using statistical methods to develop the so-called Statistical Machine Translation (SMT) systems (Shannon, 2001). In SMT, translation models are trained on a large corpus of *parallel data* to learn how to translate, and another larger corpus of *monolingual data* to learn the format and the structure of the target language.

A *parallel corpus*, also known as bitext, is a collection of parallel sentences in two different languages. This data is sentence-aligned where each sentence in one language is matched with its associated translated sentence in the other language. Parallel data is utilised to train a *translation model* that is able to provide an adequate translation of the target language. The co-occurrences of words and text segments, such as phrases in the parallel data, are used to infer translation model between the the target and source languages.

While the *monolingual data collection* is construct a *language model* that is able generate a fluent output of the target language. The sentence matching within the parallel data varies according to how the translation model is constructed on the parallel data. For instance, in phrase-based machine translation, this sentence matching is typically constructed and modelled between continuous sequences of words, whereas in hierarchical phrase-based machine translation or syntax-based translation, more structure and semantics are added to the model (Zhou et al., 2012; Koehn et al., 2007b).

Translation and language models can be trained on different representations of the input and output sentences. Translation models can be trained on a shallow representation of text which considers the sentence as a string of tokens, or a deeper

representation in a tree-like rules that involve more layers of linguistic annotation such as part-of-speech tagging, syntactic and semantic parses. Language Models are often constructed based on shallow representations, such as token of uni or n-grams.

SMT systems generate translations using statistical models whose parameters are trained and tuned using the parallel corpora. The first SMT model was proposed by IBM Brown et al. Brown et al. (1993). This model is based on the concept of the noisy channel coding theory developed by Shannon (2001). Using this model the translation is extracted and ranked using their probabilities as follows.

Given a sentence s in L source language language, the model task is the find best translation s_t in the target language T, so that $p(s_t|s)$ is maximised. Following the Bayes rule, the $p(s|s\_t)$ is estimated using priori probability of $p(s)$. Therefore, the best translation s_t as shown in Equation 2.12.

$$\begin{aligned} s_t &= \underset{s_t}{\operatorname{argmax}}\, p(s_t|s) \\ &= \underset{s_t}{\operatorname{argmax}}\, \frac{p(s|s_t) * p(s_t)}{p(s)} \\ &= \underset{s_t}{\operatorname{argmax}}\, p(s|s_t) * p(s_t) \end{aligned} \tag{2.12}$$

Since s is given, and the propose of this task is to compare possible translation candidate s_t, the denominator P(s) of the Bayes rule is insignificant since it will be always similar. The remaining task is to drive the $p(s|s_t) * p(s_t)$ where $p(s_t)$ is the language model and the $p(s|s_t)$ is the translation model of the task.

A wide range of translation models have been developed in the recent years (see (Lopez, 2008), (Koehn, 2009) and (Bisazza and Federico, 2016) for detailed reviews of SMT approaches). The following list briefly outlines the most popular SMT translation models (Parton, 2012; Ture, 2013).

- *Word-based translation model* : In this model, words (unigrams) are mapped to words. The IBM model 1 (Brown et al., 1993), is an example of word-based models that is considered to be the oldest SMT models. IBM model 1

is a word-to-word translation model that allows insertion and deletion in the translation output.

- *Phrase-based translation model* : In this model, phrases (n-grams) are mapped to phrases during the training, and these phrases are not necessarily of the same length (Koehn et al., 2003). The idea behind the use of n-grams instead of unigram tokens, is to provide more context to reduce translation ambiguity. Figure 2.3 shows the process of Arabic-to-English phrase-based translation. The input is segmented into a number of sequences of consecutive words of phrases, where each phrase is translated into an English phrase, and English phrases in the final output are reordered to match the grammatical rules of English.

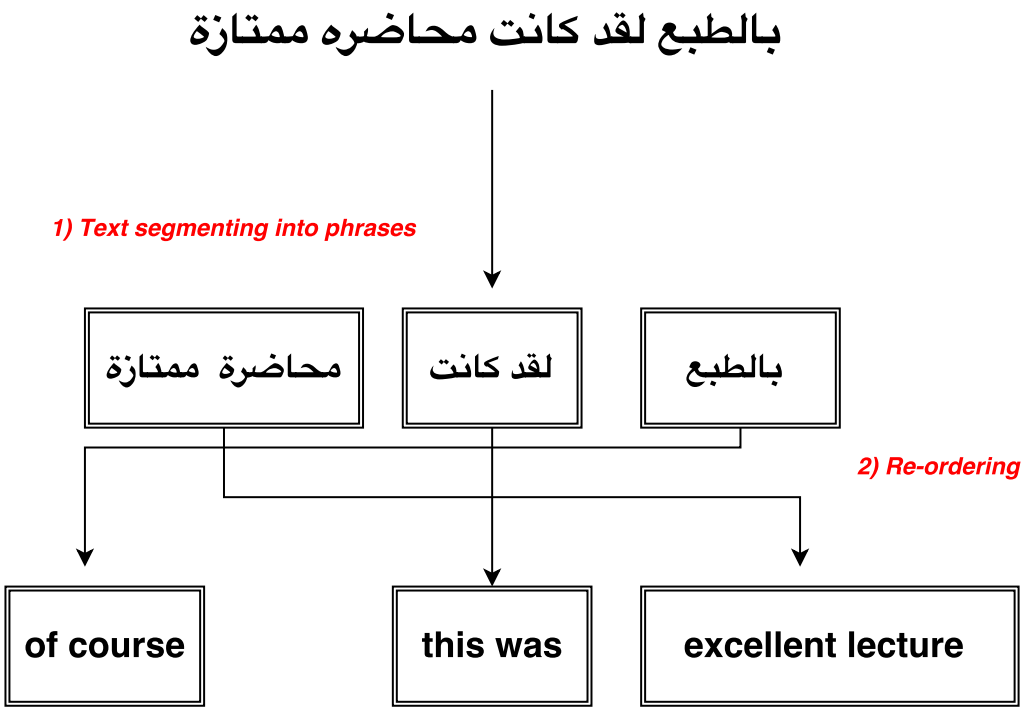

Figure 2.3: An Overview phrase-based translation for Arabic to English

- *Hierarchical phrase-based translation model* : This model generalises the phrase-based model by extracting hierarchical rules that include both phrases and non-terminals (also called syntactic variables). Non-terminals can be defined as any arbitrary n-grams of tokens. Hierarchical phrase-based models learn a

synchronous context-free grammar (SCFG) from the text, whereas the grammar rules are composed of ngrams (phrases) and non-terminals (Chiang, 2005). For instance, an Arabic-English translation rule could map "لي V لا تفعل" to *do not V me*, where V is the non-terminals which can be filled with another phrase translation. The idea behind using non-terminals and hierarchical rules is to allow a wider re-ordering of phrases in the final output. The main advantage of extracting these such rules is to allow a long-distance reordering of phrases, that phrase-based SMT systems cannot produce.

- Syntax-based translation model : This model is an extension of the hierarchical one that extracts the a SCFG rules from the bitext, but the extracted SCFGs are linguistically motivated trees (syntactically well-formed trees) (Liu et al., 2006). In this model, non-terminals are mapped to noun or verb phrases rather than n-grams as in the hierarchical model. The hypothesis behind extracting linguistically correct rules is to allow more fluent output that is human readable. Human-readable output does not only improve the fluency of the translation but also help in post-translation tasks such as translation post-editing (Zhou et al., 2012; Parton, 2012).

2.4.3 Neural Machine Translation

Neural Machine Translation (NMT) is a new exciting and promising MT approach that uses neural networks models to predict translations (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Cho et al., 2014; Tu et al., 2016; Zhang, 2017; Chen et al., 2018; Ott et al., 2018). While many different components need to be trained and tuned separately in SMT(e.g., translation models, language models, reordering models), NMT is just a single large-scale neural network (with millions of artificial neurons) that is designed to model the entire MT process.

Furthermore, unlike SMT models which use pre-defined and engineered features such as linguistically motivated trees and phrases, NMT models automatically ex-

tract and learn features by utilising neural nets that is built using distributed representation of the words such as *Word embedding* where words or phrases from the vocabulary are mapped to vectors of real numbers. Mikolov et al. (2013) proposed two different architectures for distributed word representation training, the *Continuous Bag of Words (CBOW)* architecture which predicts the current word based on the surrounding words, and the *Skip-gram model* predicts surrounding words given the current word. The utilisation of these word embedding approaches enable NMT to incorporate different types of annotations and external knowledge much better than the previously described SMT models.

The translation process adopted in NMT, can be simply summarised with following steps.

- An encoder analyses the given source sentence to build a *semantic* vector (word embedding) which is a sequence of numbers that represents the sentence meaning.

- A decoder, then, processes the sentence vector to produce a translation.

In order to perform these two steps, NMT needs to learn a sequence to sequence model using the semantic vectors of the source and target languages. Recurrent Neural Network (RNN) is a sequential model that is widely used in NMT (Kalchbrenner and Blunsom, 2013). RNN has three layers: an input layer, recurrent and output layers. The key layer here is the recurrent layer which maintains a context (hidden) vector covering previous sequential information about each word. To produce the recurrent layer in RNN, a non-linear function is utilised to compute the weights of input words based on the previous hidden states. The output layer is responsible for generating probabilities for each possible sequence and selecting the one which has the maximum.

The use of RNN or other sequence-to-sequence modelling in NMT enables better generalisation to very long sentences without the need to explicitly store any gigantic phrase tables or language models as in the case of SMT (Luong, 2016). Many other

techniques and aspects of NMT are not covered in this section as they are out of the scope for our this work, however, interested readers are referred to the recent comprehensive tutorial on NMT presented in (Neubig, 2017).

While NMT models have successfully shown that it can outperform SMT for many language pairs, in this thesis, we chose to use the phrase-based SMT model as it was more mature to be integrated in CLIR. Nevertheless, our proposed CL-UGS framework can be easily integrated with the output layer of the NMT or any other MT types. We leave this potential expansion of our work for future research in this area.

2.4.4 Using SMT in CLIR

SMT systems are often utilised using both DT and QT CLIR approaches which were explained in the previous section. The main challenge in MT with regard to CLIR is to find the right corpus to train the MT model so that it is able find effective translations for any query or document (Zhou et al., 2012). Finding the right corpus and domain to train the MT model is particularly important to avoid any translation coverage issues such as Out Of Vocabulary (OOV) issues. At the same time, constructing a parallel corpora that tuned for each CLIR task can be extremely expensive and time-consuming to produce. Research in CLIR, such as the work of Federico and Bertoldi (2002); Darwish and Oard (2003); He and Wu (2008), explored the possibility of using supplemented the MT model by other resources such as bilingual dictionaries to improve the SMT coverage. While this combination has proven effective and outperforms regular SMT systems, coverage issues still presents major issues for CLIR especially for languages where the data resources are limited (Zhou et al., 2012; Darwish et al., 2014).

Other efforts in addressing SMT coverage problem have focused on moving from training using parallel to *comparable* data. A comparable data is a combination of texts in multiple languages that are generated *independently*, but share the same communicative structure, functions and themes (Sheridan and Ballerini,

1996; Abdul-Rauf and Schwenk, 2011). Comparable data utilised to extract contexts, structure by aligning sentences across languages (Shakery and Zhai, 2013). For instance, current off-the-shelf MT tools such as Google and Bing translate are trained on comparable data collected from the internet from each language pair. It has been reported that the reason why these off-the-shelf tools often outperform (for CLIR) any SMT systems that is trained on limited data from available parallel corpus (Leveling et al., 2009; Zhou et al., 2012).

In this thesis, we show that off-the-shelf MT tools are also ineffective to our UGS task for the following reasons.

- Low resource languages such as Arabic, where there is not enough data to train, MT translation output can still suffer from several quality issues (Darwish et al., 2014; Alqudsi et al., 2014; Khwileh et al., 2016).

- MT are often utilised in CLIR using the single-best results (Zhou et al., 2012) due to the convenience of using translation systems as black-boxes in IR. Magdy (2011) showed that CLIR can benefit from looking inside the MT black-box, improvement was gained by aligning the MT pre-processing steps to that which is used CLIR. However, the reported improvement was in terms of improving the overall efficiency only.

This work aims to follow a similar approach by looking inside the MT system and optimise its output for UGS retrieval. This research utilises both open-box and black-box tools to improve the quality of translation in CLIR for UGS content.

In the next chapter, we investigate the effectiveness of both open-box and black-box mt tools for CLIR of UGS to study the impact of translation on the retrieval effectiveness. Later in Chapter 9, we provide a novel approach to predict the translation quality of an MT output for CLIR in UGS retrieval.

2.5 Summary

This chapter provided an overview of IR technologies and their evaluation. It also introduced the main retrieval models in IR such as the Vector Space Model (VSM), Best Matching (BM25) and Divergence From Randomness (DFR) models. The chapter also presented an overview of techniques such as RF and QPP which are utilised heavily in our investigation of IR for UGS content.

Furthermore, this chapter presented a background of cross-lingual search, as well as the most-well established technique to CLIR such as Query Translation (QT) and Document Translation (DT). Finally in this chapter, we provided a brief overview of Statistical Machine Translation (SMT) and how it is being utilised for handling translation for CLIR.

After providing this background overview on the most relevant areas to this research, in the next chapter we provide a detailed review on the previous work in speech retrieval and other techniques that utilised research work related to the investigation of this thesis.

Chapter 3

Information Retrieval for Speech and Multilingual Content

The work of this thesis draws on prior research in Spoken Content Retrieval (SCR) and Cross Language Information Retrieval (CLIR). This research seeks to study the interaction between the translation errors in CLIR, transcription errors in SCR and noise in user-generated content. In the following sections, we provide an overview of relevant existing research in SCR and CLIR.

3.1 Spoken Content Retrieval

In SCR, Automatic Speech Recognition (ASR) technologies are used to derive transcriptions of the speech media in a from of a timely-code linguistic elements. These transcripts serve as the basis and core component of the SCR retrieval process as shown in Figure 3.1.

ASR transcripts generally contain insertion, deletion, and substitution errors where the ASR system has failed to recognise correctly (Jurafsky and Martin, 2009; Li et al., 2014). These errors pose a major reliability issue for SCR, and can have a negative impact on the retrieval effectiveness.

The effectiveness of an ASR system is evaluated using a comparison of the gen-

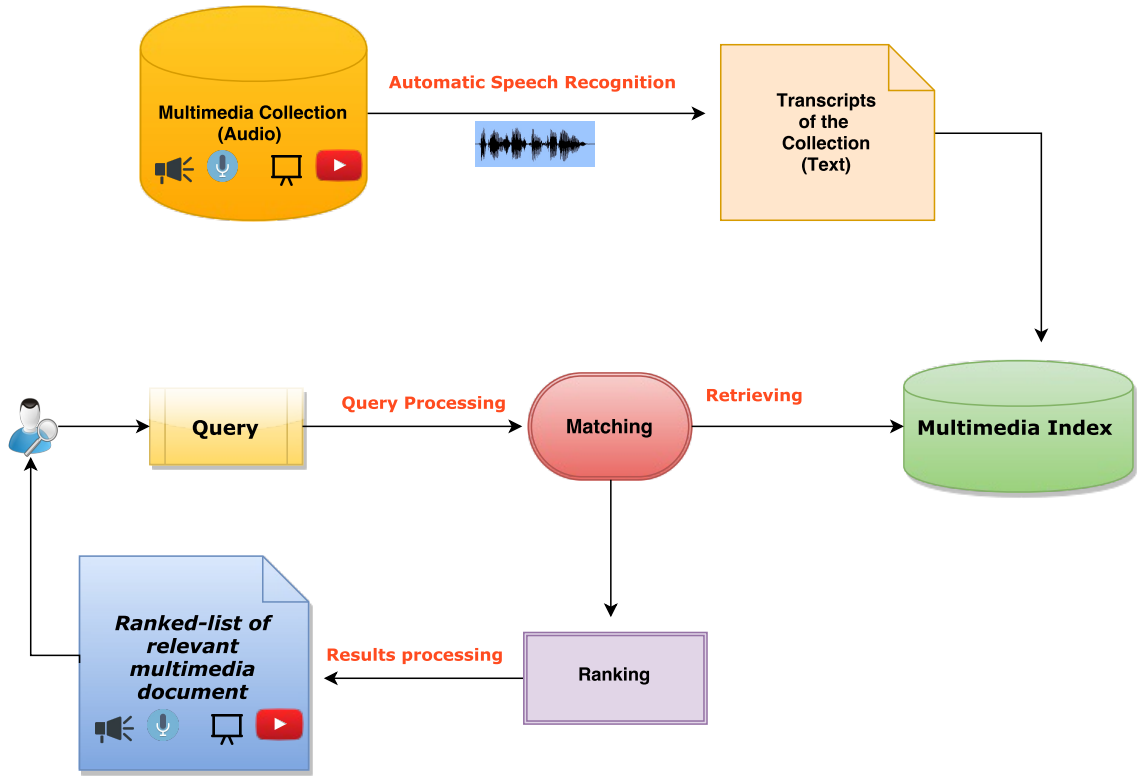

Figure 3.1: An Overview of SCR.

erated transcript against an accurate manual transcript of the spoken content. ASR performance is generally evaluated in terms of *Word Error Rate (WER)*, that is the number of individual words in the transcript that were substituted, inserted and deleted as compared with manual transcript, divided over the overall number of recognised words (Jurafsky and Martin, 2009).

The impact of transcription errors on the retrieval effectiveness is rather complex. It has been reported that this impact is highly dependent on the complexity of the retrieval task itself, and the quality of the extracted ASRs which vary depending on the type of spoken data, whether it is from video lectures, meeting, user generated internet video content or others (Stark et al., 2000; Larson and Jones, 2012; Eskevich, 2014). The degree of errors found in an ASR transcript depends on various characteristics of the speech data being recognised for SCR. The next section provides a review of SCR work on different types of speech.

3.1.1 Prepared speech of high quality: Broadcast news

Broadcast news is a type of speech media that is recorded by trained professional broadcasters, and recorded in high quality settings using professional recording equipment. The speaker follows the pronunciation norms of formal language, and works in quiet and noise-free conditions within a designated studio. Broadcast news documents have consistent structure, with similar length of items, and single focus of topic, with content being *self describing* and background of each news story being fully explained in each broadcast. ASR output for this type of content is generally accurate with an WER of less than 10% leading to a a highly reliable ASR (Eskevich, 2014).

Research in SCR began in the early 1990s with the emergence of early Large Vocabulary Continuous Speech Recognition (LVCSR) technologies. The initial work on SCR involved some basic IR tasks such as keyword spotting, browsing, and accessing small and private broadcast news speech collections (Rose, 1991; James, 1995; Wactlar et al., 1996).

The first formal evaluation of SCR took place within the Text REtrieval Conference (TREC) the Spoken Document Retrieval (SDR) tasks (Garofolo et al., 1997, 1999a,b, 2000). The SDR tasks ran for 4 years (1997 as TREC-6 SDR, 1998 as TREC-7 SDR, 1999 as TREC-8 SDR and 2000, as TREC-9 SDR), and aimed to bring together IR and speech recognition researchers to participate on known-item and adhoc retrieval tasks on benchmark corpus of radio and television broadcast newsware recordings (Garofolo et al., 1997, 1999a,b, 2000).

SDR collections were provided by the Linguistic Data Consortium (LDC) English Broadcast News Speech HUB-4 ASR corpus, a subset of the DARPA Topic Detection and Tracking (TDT) corpus. Each news story was manually segmented into detected story units, and was made available for the SDR's participants (Garofolo et al., 1997, 1999a,b, 2000). SDR tasks allowed different SCR methods and research to be fairly compared to each others. One key finding from the SDR tasks was that Query Expansion (QE) (Woodland et al., 2000) and Document Expan-

sion (DE) techniques (Singhal and Pereira, 1999), are highly effective approaches to compensate the recognition errors of ASR, and improve the performance of SCR.

Experiments at the TREC SDR tracks reported that SDR effectiveness is largely robust to speech recognition errors rates of around 20% WER, and that with help of QE and DE techniques (Singhal and Pereira, 1999; Woodland et al., 2000), retrieval performance using ASR generated transcripts was found to match the one achieved using the accurate manual transcripts of the speech data. Based on this, at the end of TREC SDR tasks, SCR was reported to be a largely *solved problem*, and it was agreed not to dedicate further research efforts to it within the TREC programs. (Garofolo et al., 2000).

However, this conclusion, though in can be considered reasonable for formal and scripted speech content which were studied at the TREC SDR tasks, cannot be applied to more complex informal spoken content such as conversational or UGS speech, which have greater recognition challenges for SCR, arising from the generally high WER rates encountered, and the complex topical structure.

3.1.2 Informal and Conversational Speech

Following the progress of ASR technologies on more challenging types of spoken content, SCR research shifted its focus from the high-quality broadcast speech to more informal, non-scripted conversational speech media such as lectures (Akiba et al., 2008), meetings (Eskevich and Jones, 2014) and interviews (Oard et al., 2006; Pecina et al., 2007).

Informal and conversational speech contains natural non-scripted materials where more than one speaker communicates with through different means. The following sections give an overview of the conversational speech types studied in SCR research.

a) Lectures Speech

Although lectures speech may appear to be identical to broadcast speech since they are both based on typically prepared presentations, lectures are more unscripted

and informal in spoken style, and may include features such as hesitations and mispronunciations. Therefore, ASR for this type of speech content can be more challenging than broadcast ones and hence, less reliable for SCR (Glass et al., 2004).

Much SCR research on lectures speech has mainly focused on solving the out-of-vocabulary words (OOV) problem. OOV is common issue for lecture speech whenever the lecture's presenter or the audience uses terms from specific domain that is not available for in the ASR's training data set(Li et al., 2014).

Previous research in SCR proposed to use additional information source to address this coverage issue. For instance, the work of (Jones and Edens, 2002; Glass et al., 2007; Lee and Lee, 2008) used information that was presented in the meta-data such as the speaker notes and description, texts from the slides, or even the material of the actual lecture, to enrich the ASR training data set. Later in 2011, a SCR benchmark task was introduced at the 9th NTCIR (NII Testbeds and Community for Information Access Research evaluation workshop) as the SpokenDoc track which involved SCR tasks focused in searching a corpus of Japanese lectures (Akiba et al., 2011). SpokenDoc efforts presented several techniques related to the speech processing of the Japanese language for passage retrieval in SCR such as speech segmentation approaches(Eskevich, 2014).

b) Meeting Speech

Speech occurring in meetings is more conversationally free style content, but can include both prepared speech and discussions from multiple speakers. Meetings may also come with prepared agendas and topics, but they include more back and forth discussions from speakers with varying styles, accents, dialogue acts and vocabulary which introduces more recognition challenges for ASR in SCR (Jones et al., 1996; Wrede and Shriberg, 2003).

A main challenge of this type of speech is that the privacy and confidentiality issues related to the content of meetings has prevented the presence of any real-world SCR benchmark datasets for this domain. Instead, SCR research for meeting

speech has mostly concentrated around meetings recorded in laboratories that has been created based on an artificial predefined scenarios (Morgan et al., 2003; Renals et al., 2008).

For example, the AMI and AMIDA projects (Renals et al., 2008) created the AMI meeting corpus consisting of carefully collected and documented meeting recordings, individual slides, minutes, and videos from the meetings. More than 70% of the provided data by these projects was artificial based on meetings carried out according to a pre-written scenario, where a meeting's participants were assigned certain roles and instructed to speak accordingly in the recorded discussions (Renals et al., 2008). While the discussions in these recorded meetings did not relate to a real-life situations, the audio itself was recorded naturally and spontaneously as in any other meeting.

Furthermore, from a SCR perspective, unlike lecture and broadcast speech, it is not clear how a search task of this type of content can be carried out. For example, a user might be looking to find a jump-in point where the discussion about a certain topic started or a decision was made during a meeting, or a certain participant expressed their opinion of a topic being discussed.

Therefore, in order to facilitate effective and efficient access over meeting speech, a careful segmentation of its content is required. Much research in SCR for meeting speech has typically focused on the development of tools for segmenting or summarising the meetings, and how these segments or summaries can be used for searching this type of content. (Renals et al., 2008; Eskevich and Jones, 2014).

c) Interviews Speech

Interview speech recordings are more informal and unstructured that are based on a free and spontaneous conversations. SCR over spontaneous, and conversational informal speech can be characterised as current state-of-art research problem for SCR (Larson and Jones, 2012).

The first interview SCR benchmark dataset was released by the Cross-Language

Evaluation Forum (CLEF) 2005-2007, at the Cross-Language Speech Retrieval (CL-SR) task (White et al., 2005; Oard et al., 2006; Pecina et al., 2007). The CL-SR tasks investigated SCR for a dataset that consisted of interviews with survivors and witnesses of the Holocaust from the Shoah Visual History Foundation collection [1]. In 2008-2009, a video retrieval track was founded within CLEF (VideoClef) (Larson et al., 2009, 2010) that studied new dataset of Dutch TV interviews.

SCR over the CLEF's interview speech datasets revealed that even with ongoing improvements in ASR quality, recognition errors generated from the informal style of this content present significant challenges to the retrieval effectiveness.

Overall these tasks found that SCR over interviews could be significantly improved significantly by including manually-generated metadata, but a careful selection on how these metadata should be combined within the retrieval framework is required to maintain effectiveness (Jones et al., 2007; Pecina et al., 2007).

d) UGS Content

UGS represents the speech media that is available online, and is being *produced and maintained by social media users*. With the development of social media sharing and streaming platforms such as YouTube.com, users are encouraged to be "producers", even though they generally have limited or no background on recording or producing. User can freely express their views and comments by creating a short or long videos using their mobile device and make them publicly available online. Unlike previously described speech, UGS collections have larger scale with huge variations in the quality, style, topics and themes of the speech documents.

In 2010, VideoClef task was developed into an independent multimedia benchmark tasks called MediaEval (MediaEval, 2017) which offered UGS such as the Blip10000 (Schmiedeke et al., 2013) that contains around 15,000 speech media files, and this has been reported to be the largest reported speech collections in SCR research (Eskevich, 2014).

[1] A more detailed review of the CLIR work within CL-SR tasks is presented in Section 3.2.

MediaEval offered different state-of-the-art new multimedia benchmark tasks to be explored within the UGS scale such as event detection, genre tagging. Perhaps the most closely related task to SCR were the MediaEval 2011 Rich Speech Retrieval task (Schmiedeke et al., 2012) and the Search and hyperlinking task at MediaEval 2012 (Eskevich et al., 2012b). Both tasks studied passage-retrieval challenges on UGS where new evaluation frameworks and metrics were introduced (Eskevich et al., 2012c), and automatic segmentation algorithms were proposed (Wartena, 2012; Eskevich et al., 2012a).

However, there has not been any reported element work on adhoc or cross-lingual SCR for UGS content. At the same time, previously introduced methods for SCR are not designed to handle UGS because none of these method made an attempt to address diversity of topics, the varying quality of transcripts and metadata, the length and style variation of this content. These issues require a deep understanding of the challenges and effective techniques from an SCR perspective, which we aim to study in this thesis.

3.2 Cross language Speech Retrieval

CLIR tasks have been explored across different domains and document types (Peters et al., 2012). The most closely related CLIR work to that examined in this thesis was carried out within the Cross-Language Evaluation Forum (CLEF) evaluation campaigns[2], these are outlined in this section.

3.2.1 Cross-Language Spoken Document Retrieval (CL-SDR) tasks

From 2002-2004 the CL-SDR task investigated news story document retrieval using data from the NIST TREC 8-9 SDR with manually translated queries (Federico and Jones, 2004; Federico et al., 2005). The aim of these tasks was to evaluate

[2] `www.clef-initiative.eu/`

CLIR systems on noisy automatic transcripts of spoken documents with known story boundaries which involved the retrieval of American English news broadcasts of both unsegmented and segmented transcripts taken from radio and TV news.

These CLIR tasks were done using topics in several European languages. No metadata was provided in these tasks, but some interesting findings indicate that even with the *manually translated* queries, the best CLIR performance resulted in 15% reduction from the monolingual ones (Federico and Jones, 2004), while using dictionary term-by-term translation, this reduction went up to be between about 40% and 60% which highlights the challenge of CLIR search over these collections (Federico et al., 2005).

3.2.2 Cross-Language Speech Retrieval (CL-SR) task

A more ambitious CL-SR task ran at CLEF 2005-2007 (White et al., 2005; Oard et al., 2006; Pecina et al., 2007). This task examined CLIR for a spontaneous conversational speech oral history collection with content in English and Czech. The task provided ASR transcripts, automatically and manually-generated metadata for the interviews.

The goal for the Czech and English tasks was to develop SCR techniques for monolingual and cross lingual searchers to identify sections of an interview that they would find relevant to their information need. These tasks reported that the use of manual metadata yielded substantial and statistically significant improvement on the retrieval effectiveness. A further investigation was carried on the CL-SR standard collection by Inkpen et al. (2006), who showed that retrieval effectiveness could be improved by careful selection of the term weighting scheme between the ASR and the manual metadata.

Alzghool and Inkpen (2008) also used the test collection of CLEF 2007 CL-SR to present a method for combining results from different retrieval models in order to improve the overall retrieval effectiveness. Alzghool and Inkpen (2008) provided a comparison between both ASR and manual metadata for SCR effectiveness, and

indicated the high superiority of the manual metadata for maintaining the retrieval effectiveness.

Another interesting follow up study, reported by Jones et al. (2007), examined and compared the CLIR effectiveness of each source of evidence included in this collection. The major finding from this work indicated that searching the manually generated metadata gives higher performance in terms of recall and precision over the search of noisy ASR transcripts.

3.2.3 VideoCLEF and MediaEval Tasks

The VideoCLEF task was introduced at the CLEF 2008 and CLEF 2009 tasks. VideoCLEF released a collection of Dutch TV content featuring English-speaking experts and studio guests. Tasks participants were provided with Dutch archival metadata, Dutch speech transcripts, and English speech transcripts (Larson et al., 2009, 2010). VideoCLEF piloted tasks that are not directly related to the research of this thesis, these tasks include performing classification, translation and keyword extraction using either machine-learning or information retrieval techniques.

The Video CLEF tasks were followed by the establishment of the MediaEval benchmarking campaign in 2010 (MediaEval, 2017). The mediaEval tasks made use of the Blip10000 dataset which explored the following features.

- ASR Transcripts.
- Speaker identification for each transcript to enable speaker-based segmentation of the transcript.
- Cues extracted from the visual content.
- The Titles and descriptions which were uploaded by the user for each video.

The Mediaeval tasks which are relevant to the work presented in this thesis are explained as follows.

- *Tagging Tasks (Schmiedeke et al., 2012)* : This task was presented to investigate how to automatically assign genre labels to semi professional user generated (SPUG) videos using different methods, and sets of features. Genre was defined as related to common browsing categories used for Internet video sharing websites, in particular to blip10000 data collection, by blip.tv(BlipTV, 2017). The ground truth data was provided by the genre label which was associated with the video by the uploader. The participants results were evaluated in terms of mean average precision (MAP) (Schmiedeke et al., 2012). This task provided techniques to develop tagging systems based on classification approaches for the blip10000 collection.

- *The Search and Hyperlinking Task (Eskevich et al., 2012b)*: The Search and Hyperlinking Task also used the blip10000 within the MediaEval benchmark. This task was divided into two sub-tasks as follows. *Search* subtask, where participants were provided with a query set for a known-item search (a search for single known item) task which was generated using crowd sourcing platform for the blip10000 collection. This task was a passage search, which aimed to retrieve video segments corresponding to textual or multimedia queries. The participants provided methods to combine ASR transcripts and user-metadata to improve retrieval effectiveness. And the *Linking subtask*, this task also utilised the groundtruth of the search sub-task as anchor videos, where links to other videos were to be generated. Task participants were asked to return a ranked list of video segments which were potentially relevant to the information in this relevant video segment (regardless of whether its relevant to the initial textual query or not) (Eskevich et al., 2013).

Although the activities presented at MediaEval explored various multimedia tasks over the blip10000 collection, there has not been any CLIR elements. Our focus in this work is to study these different SCR approaches for Internet-based user-generated multimedia collections in cross-lingual settings.

Overall, previous research in SCR has focused on running IR/CLIR tasks for professionally-generated speech whether its documentaries, TV shows or interviews with high quality recording, and consistency of length, visual and audio quality across the collections. These collections included manually or automatically created metadata. For example, domain experts following a carefully prescribed format wrote the manually created metadata for CLEF 2005-2007 with consistent speech quality of word error rate of 25% across the collections used (White et al., 2005; Oard et al., 2006; Pecina et al., 2007). The current UGS content on the web has brought new modalities such as user-generated metadata that now play an essential role for effective access of this content (e.g Eickhoff et al., 2013; Filippova and Hall, 2011; Toderici et al., 2010).

Previous efforts in CL-SR were focused on measuring the impact of the ASR accuracy and translation errors on the overall retrieval effectiveness using query translation (QT) CLIR. This could be attributed to several reasons such as the cost and the resources needed for translating the spoken content, as well as the availability of effective MT systems that can be trained and tuned for ASR translations.

The only reported use of Document Translation (DT) CLIR investigation was carried out within TRECVID 2005 and 2006, but were mainly focusing on visual retrieval tasks only that are not related to this research. The focus of TRECVID efforts was to study the visual relevance of video to user queries, and made use of ASR transcripts primarily to support the use of visual features in these tasks. TRECVID tracks included multimedia search tasks of a TV news video collections in Chinese and Arabic (30 hours of Chinese news and 83 hours of Arabic news broadcasts). These videos were accompanied by ASR transcripts which had been machine translated into English. Translation of the imperfect ASR transcripts resulted in having a quite poor noisy evidence for retrieval, forcing participants to focus on visual aspects of content-based retrieval (Smeaton et al., 2006; Over et al., 2005).

3.3 CLIR for Internet-based User-Generated Content

While CLIR for published text has been ongoing with a wide variety of language pairs for many years, recent research has begun to explore CLIR for user-generated content (UGC) text. One example of this work is reported in (Bagdouri et al., 2014) which explored the retrieval of questions posed in formal English across UGC documents of Arabic collected from a forum posts. Bagdouri et al. (2014) employed a DT CLIR approach where they translated the Arabic informal text into English. Their results showed that retrieval performance can be enhanced by applying a text classifier to help the translation of informal content.

Lee and Croft (2014) also experimented with a CLIR task for informal text documents. They developed an CLIR task over a large collection of Chinese forum posts and reported how translation noise is increased by the informal text used in discussion forums. Their proposed approach used a QE method to improve retrieval effectiveness. Their results showed that QE approaches can indeed be useful to reduce the impact of translation errors on the retrieval effectiveness.

UGC has begun to attract considerable research in video retrieval and indexing in the recent years. While none of this work has so far included an element of adhoc search or CLIR, much of it has addressed the main issues of user-generated content in video retrieval. For example, some work has focused on the quality of user-generated metadata for video content analysis and retrieval (Bendersky et al., 2014; Eickhoff et al., 2013; Filippova and Hall, 2011; Toderici et al., 2010). Filippova and Hall (2011) showed how titles, description, user tags and comments can utilised to provide a valuable clues to predict the topic of YouTube videos. Eickhoff et al. (2013) Utilised users comments to extract potential tags and indexing terms for UGS retrieval. Bendersky et al. (2014) utilised various sources including the video metadata, frequent uploader keywords, common search queries, playlist names and Freebase entities to cluster videos based on their predicted topics. Other work

has focused on the quality of audio features within the scale and the dynamics of UGS content (Chelba et al., 2012; Langlois et al., 2010). For example, Chelba et al. (2012) Utilised the sentences and segments such as utterances in the ASR transcripts to build a large scale n-gram language models for speech recognition over user-generated speech.

Moreover, from 2010, the TRECVID [3], the video retrieval benchmark in the multimedia community, provided a collection of Internet videos to be used in several tasks. However, the design of TRECVID tasks have mainly focused on exploiting visual information for applications on the shot-level (such as concept detection), or short video clips (such as event detection).

The known-item search task (KIS) (Over et al., 2011) at TRECVID, the task aimed at exploring the retrieval of visual queries and was included at TRECVID annually from 2010 to 2012. Results from the participants were rather inconsistent from year to year in terms of the retrieval effectiveness of different search approaches, one conclusion being the difficulty of actually setting up such an evaluation task on Internet collections.

3.4 Using QE in SCR

SCR research has proposed multiple techniques to improve the error-prone ASR transcripts for retrieval purposes. For example, Singhal and Pereira (1999) proposed to use a document expansion (DE) approach to alleviate the effect of transcription mistakes on the retrieval effectiveness. Their work tried to recover those words that might have been in the original video but had been mis-recognized by enriching documents with selective terms drawn from highly ranked documents that share the same topic. DE approaches has evolved in recent years and proven very useful for multiple SCR tasks (e.g. Masumura et al., 2011; Ganguly et al., 2013; Lee and Lee, 2014).

[3]http://TRECVID.nist.gov

For CL-SR, QE and DE are often used for recovering from both the translation errors of the query and the transcription errors of the ASR transcript. Several works have explored and proposed document re-ranking using QE/DE techniques for CL-SR, but primarily focusing on professionally generated spoken collections.

For example reported in (Lo et al., 2003; Wang and Oard, 2005; Lam-Adesina and Jones, 2006) investigated the effectiveness of these document re-ranking approaches on video collections which were provided by the CLEF Speech retrieval tracks (White et al., 2005; Oard et al., 2006; Pecina et al., 2007). Lo et al. (2003), proposed a document expansion using external mandarin collection, by adding helpful terms or bi-grams to improve retrieval performance in Mandarin-to-English CL-SDR. Wang and Oard (2005) showed that the CL-SDR effectiveness for French-to-English yielded 79% of monolingual performance when searching manually assigned metadata which was provided by the CLEF SDR collection (Oard et al., 2006).

The main issue for these document re-ranking approaches (whether it is query-based or document-based expansion) is that they are most likely to be challenged by the so-called ***topic drift*** problem; in the case when the newly expansion terms are not relevant to the original query topic and thus negatively impact the effectiveness. This is certainly a common issue when documents in the collection are long and a single document may contain multiple topics as suggested in (Singhal, 1997; Terra and Warren, 2005; Mitra et al., 1998b).

In SCR, topic drift can happen due to the relatively long ASR transcripts where a single spoken document may represent multiple sub-topics. Several techniques have been explored to cope with this issue within the context of video retrieval. For example, the work reported in (Volkmer and Natsev, 2006; Rudinac et al., 2009) explored the use of the visual information for the query/document expansion of the ASR transcripts in order to improve the overall spoken content mono-lingual retrieval within video collections.

In the context of CL-SR, previous efforts such as the work reported in (Lo et al., 2003; Wang and Oard, 2005; Terol et al., 2005) focused mostly on using manually-

created summaries or segments for expansion, which were provided within the CLEF video collections. Unfortunately, having these manually-generated summaries or segments within the current large-scale UGS content is unlikely due to the cost required for creating them.

Instead, as part of our RQ3 efforts (Can speech segmentation be beneficial for UGS retrieval), in Chapter 7, we use automatic segmentation techniques of ASR transcripts such as the one studied by (Wartena, 2012; Eskevich et al., 2012a), and investigate their robustness and effectiveness for QE for UGS content.

3.5 Summary

This chapter reviewed previous research in SCR and CLIR fields by providing an overview of the related work from the literature. This chapter introduced the existing research for SCR for two types of speech media , namely, planned and spontaneous. We provided a review of related monolingual and cross-lingual speech tasks studied previously.

Existing research in SCR has mainly focused on developing IR and CLIR methods for small, highly maintained, professionally created speech collections such as broadcast news and interviews. These collections were presented with manually generated metadata that were written for each document by professional indexers. The presence of such a highly maintained metadata helped previous research to develop effective SCR techniques for these collections (Jones et al., 2007; Pecina et al., 2007). The work presented in this thesis investigates the task of SCR for new emerging type of speech media which is UGS content. Online UGS content is a large scale, noisy, inconsistent and decentralised type of speech content that requires further research. In Chapter 5, as part of our RQ1 (What are the challenges of UGS retrieval?) investigation, we study the challenges of UGS content from SCR and CLIR perspective. Furthermore, in Chapter 9 as part of our RQ4 (Can we build an adaptive CLIR technique for UGS retrieval) investigation we present a novel CLIR approach for

UGS retrieval. The proposed approach utilises QPP methods to adaptively identify the right retrieval settings for UGS content.

This chapter also presented an overview of the previous QE/DE methods that were proposed to address the mismatch problem in SCR. These approaches have been shown to be effective for SCR in professionally generated speech media(Lam-Adesina and Jones, 2006; Wang and Oard, 2005). However, none of these approach has been tested for UGS content. In Chapter 6, as part of our RQ2 (Can Query Expansion techniques be beneficial for UGS retrieval?) investigation , we study the effectiveness of these approaches over the noisy settings of UGS content.

Furthermore, the previous QE techniques in SCR utilised manually segmented or summarised speech, with manually meta-data created by professional indexers. Therefore, applying these approaches within the scale of UGS content is not reasonable due to the cost and time required to create these summaries and segments. Instead, as part of RQ3 (Can speech segmentation be beneficial for improving the effectiveness of UGS retrieval) investigation in Chapter 7, we study the effectiveness of using automatic text segmentation to automatically generate speech segments for QE in UGS retrieval.

Chapter 4

Evaluation Framework For UGS Retrieval

In this chapter, we describe the experimental framework that we use throughout this thesis to conduct our investigations of UGS retrieval. Section 4.1 describes the basic component of an IR evaluation framework. Section 4.2 describes the tools and resources we use in this thesis, Section 4.3 and Section 4.4 describe the evaluation test sets that we study in this work, while Section 4.5 describes the topic sets we use in our task. In Section 4.6, we present initial experiments we conduct in order to develop a retrieval framework UGS content.

4.1 Components of an IR Evaluation Framework

In order to evaluate the effectiveness of IR system for a specific task and perform an experimental investigation, the following components need to be prepared.

- Documents collection : A collection of documents is required to build the search index for the IR task. These documents can be in structured or unstructured representation, and in different types and formats. Each document may contain several fields which can potentially serve as a source-of-evidence to understand what the document is about.

- Query set : A set of topics that represents different information needs to be used for testing the IR system. These topics should be written in the same way a real user would express while trying to find certain information need. IR tasks often assign multiple users to write queries in different style and structure to ensure fair and comprehensive evaluation of the proposed IR systems. For CLIR tasks, queries should be written in a language that is different from the language of the test collection. Prior research in IR suggests that at least a set of 25 queries is required to achieve statistically significance in comparative evaluation of different IR systems for the same task(Buckley and Voorhees, 2000).

- Relevance judgements : In order to evaluate the performance of an IR system, a sample of result set that contains the list of *relevant* documents retrieved from the test collection is required for each query. The relevance of documents to the topics are determined by human assessors. Human assessors are asked to manually evaluate which of these documents are deemed relevant to their query. Ideally, this assessment is required to be performed on each document of the collection. However, this is not reasonable due to scale of the test collection and the cost required to perform that. Instead, a *pooling* technique is usually used to create a pool of documents retrieved by the calibration of different different IR systems(Buckley et al., 2006; Harman, 1993). Pooling is based on the assumption that each retrieval run returns a finite set of documents in response to the query, and that there is a certain amount of overlap between the retrieved documents across these runs.

In the following sections, we explain the preparation of each component for the UGS retrieval task we investigate in this thesis.

4.2 Components for the Experimental Investigation

This section describes the tools and resources used for conducting the experiments reported in this thesis. All experiments are conducted using terrier retrieval engine (Santos et al., 2011). Terrier retrieval engine[1] is a standard open-source IR toolkit providing an implementation for many of the well-established retrieval algorithms and widely used by the IR research community.

The following sections provide an overview of the settings used to perform the experimental investigations reported in this thesis.

4.2.1 Information Retrieval System

The following Terrier components are used to process and index the document collections used in our experimental investigation in this thesis.

- *Tokenisation* : The text is tokenized into individual words. Other special token such as the hyphens, underscores were removed. The default tokenizer in terrier was used, that is the TRECFullTokenizer [2].

- *Stop-word removal* : Common words such as *are*, *the* and others defined as as stop words, were removed from each document. Stop words were removed based on the standard Terrier list.

- *Stemming* : Since each word can have multiple morphological variations that are derived from the same stem, we used the Porter stemmer (Willett, 2006) to extract the stem of each word in our collection.

For our experimental investigation, we built different search indexes based on varying combination of the document fields.

[1] http://www.terrier.org/
[2] http://terrier.org/docs/v4.1/javadoc/org/terrier/indexing/TRECFullTokenizer.html

4.2.2 Retrieval Settings

All retrieval experiments were performed using Divergence From Randomness (DFR) framework explained in Section 2.1.2. DFR models estimate the informativeness of each term t in a document d by measuring the divergence of its tf in the documents from that in the whole collection. DFR framework presents several retrieval models that are explained in details in (Amati and Van Rijsbergen, 2002; Amati, 2003). For our experimental investigation, we use the PL2 model as shown in Equation 4.1.

$$Score(d,Q) = \sum_{t \in Q} qt_w . \frac{1}{1+tf_n} (tf_n \log_2 \frac{tf_n}{\lambda} + (\lambda - tf_n). \log_2 e + 0.5 \log_2(2\pi . tf_n)) \quad (4.1)$$

where $Score(d,Q)$ is the score for a document d for all query terms $t \epsilon Q$. λ is the Poisson distribution of F/N; F is the query term frequency every query terms $t \epsilon Q$ over the whole collection, and N is the total number of documents at the collection. qt_w is the query term weight given by $qt_f/qt_fmax; qt_f$ is the query term frequency and qt_fmax is the maximum query term frequency among the query terms. tf_n is the normalised term frequency defined in Equation 4.2, where l is the length of the document d. avg_l is the average length of documents, and c is a free parameter for the normalisation.

$$tf_n = \sum_d (tf. \log_2(1 + c. \frac{avg_l}{l})), (c > 0) \quad (4.2)$$

As previously explained, the reason behind selecting this model over other available retrieval models [3] is the characteristics of our UGS collection; previous studies, such as (Amati and Van Rijsbergen, 2002; Amati, 2003), have shown that PL2 has less sensitivity to length distribution compared to other retrieval models and works better for experiments that seek early precision, which also aligns with our known-item experiment. PL2 is thus more suitable since our Internet based data collection

[3] Its worth noting that no statistically significant improvement has been found between this model and other models (such as BM25, LM and others) for this task.

has large variation in document and field lengths as shown in Table 4.2.

4.2.3 Experimental Evaluation

All retrieval results are evaluated using Terrier evaluation API [4], which is developed based the standard Trec_Eval tool [5].

The main evaluation metrics used for measuring the adhoc retrieval effectiveness our experiments are MAP (as shown in Section 2.1.3, Equation 2.6) and Recall reported at the k result cut-off of 1000 result items. Throughout this thesis, $Recall@k$ for each query is calculated as shown in Equation 4.3, where ${D_q}^{[rel \bigcap ret]}@k$ is the number of relevant document that are retrieved at k result cut of for query q, and ${D_q}^{[rel]}$ is the total number of relevant documents that are available in the collection for query q.

$$Recall@K = \frac{{D_q}^{[rel \bigcap ret]}@k}{{D_q}^{[rel]}} \tag{4.3}$$

For the known-item experiments, we utilise the evaluation using the MRR metric explained in Section 4.5.1 Equation 4.5. In order to assess the reliability in the comparison of experimental results, statistical significance testing is performed over each of the experiments in this thesis. We employ the Wilcoxon signed-test with a 95% confidence measure for performing the statistical significance of all the reported results (Hull, 1993).

4.3 Blip10000 Collection

In this research we use the Blip10000 collection as the main test set for our UGS retrieval experiments (Schmiedeke et al., 2013). Blip10000 is a collection of Internet videos that were uploaded to the social video sharing site *Blip.tv* [6] (BlipTV, 2017).

[4] http://terrier.org/docs/v4.0/evaluation.html
[5] http://trec.nist.gov/trec_eval/
[6] https://web.archive.org/web/20120331073050/http://blip.tv/

Figure 4.1: Shots extracted from a randomly selected videos in the Blip10000 collection.

The *Blip.tv* platform, similar to other online social-media platforms, allows online users to share their video content freely with minimal or no restrictions on length, topics or format. Users are required to sign-up first to obtain an account by providing basic information such as name, address and email. Users are then required first to accept the *Terms & Conditions* (TC) where they agree to the certain polices of the platform. Such polices exists to prevent users from sharing certain content which intends to abuse, harass, stalk, threaten or otherwise violate the legal rights of others.

Blip's users upload their content hoping to share their content publicly and gain interaction from the audience of that platform. This interaction can be measured in terms of number of views, comments and shares. Users can also opt-in to advertise their content, and potentially generate profit out of their work. The Blip10000 collection used in our experiments is a crawl of Blip.tv pages, where each page represents the content of a user-uploaded video. Each of these pages contains the following element .

- *Title*: A short textual statement contains set of words written by the up-

Table 4.1: Number of videos found in each genre of the Blip10000 collection.

Genre	Number of Videos	Genre	Number of Videos
Art	594	Movies and Television	801
Auto & Vehicals	297	Music and Entertainment	30
Business	148	Personal or Autobiographical	297
Citizen Journalism	22	Politics	1,781
Comedy	519	Religion	742
Conferences and Other Events	148	School and Education	608
General	2,522	Sports	742
Documentary	298	Technology	1,039
Educational	208	The Environment	504
Food and Drink	30	The Mainstream Media	786
Gaming	341	Travel	297
Health	18	Videoblogging	1,692
Literature	89	Web Development	282

loader to indicate the intended title heading of the uploaded video. Most UGS platforms (such as Youtube and Blip.tv) mark this field as mandatory for uploaders. Users are, however, free to pick any title for their file and use any language to express themselves.

- *Video file* : A video file that contains the actual media file the user wish to publish on the Blip.tv platform. This file should be represented as movie that contains both visual and audio materials to express a particular topics. Nevertheless, this movie can be of any quality, and be recorded in any setting whether it is in a professional studio or in a small room using personal camera or mobile device).

- Description: Users may optionally add associated textual metadata including a *description* of what the video is about.

The Blip10000 collection was originally used as the benchmark test set for the MediaEval 2012 Search and Hyperlinking (S&H) task (Eskevich et al., 2012b). This

collection contains the crawled videos together with the associated metadata. Metadata is composed of the titles and descriptions for each video that were provided by the video uploader. Blip10000 consists of 14,838 videos having a total running time of ca. 3,288 hours, and a total size of about 862 GB[7]. These videos were uploaded by a 2,237 different uploaders. Uploaders are registered social users of the blip.tv site who are able to share and consume content on the social site. Note that as part of *Terms & Conditions* of blip.tv site, uploaders have to be a registered user with an age of more than 17 years. Figure 4.1 shows shot examples from different videos of the blip10000 collection. Some of these videos, such as news broadcasts and TV shows are carefully authored, edited and quality controlled, while others such as Vlogs (video-blogs) and personal recordings are not.

The Blip10000 collection covers a 25 different genres (topics) from the following list. *(Art, Autos and Vehicles, Business, Citizen Journalism, Comedy, Conferences and Other Events, Documentary, Educational, Food and Drink, Gaming, Health, Literature, Movies and Television, Music and Entertainment, Personal or Auto-biographical, Politics, Religion, School and Education, Sports, Technology, The Environment, The Mainstream Media, Travel, Videoblogging, Web Development/design).*

The number of videos for each genre in the Blip10000 collection is shown 4.1. Figure 4.2 also shows the percentage of videos for each genre in the collection. These genres were classified by the Blip.tv site (BlipTV, 2017) and provided within the Blip10000 collection (Schmiedeke et al., 2013). Figure 4.2 and Table 4.1 demonstrate the high diversity of topics presented in this collection. This diversity of topics represents the nature of document topics one would find in any social media platform. The most common genres in the blip10000 collection were the *General, Politics, videoblogging and Technology.* It should be noted that videos associated with the *General* category may topically belong to multiple genres. Each of these

[7]The Blip10000 collection can be obtained from:
http://skuld.cs.umass.edu/traces/mmsys/2013/blip/Blip10000.html

genres includes multiple subtopics.

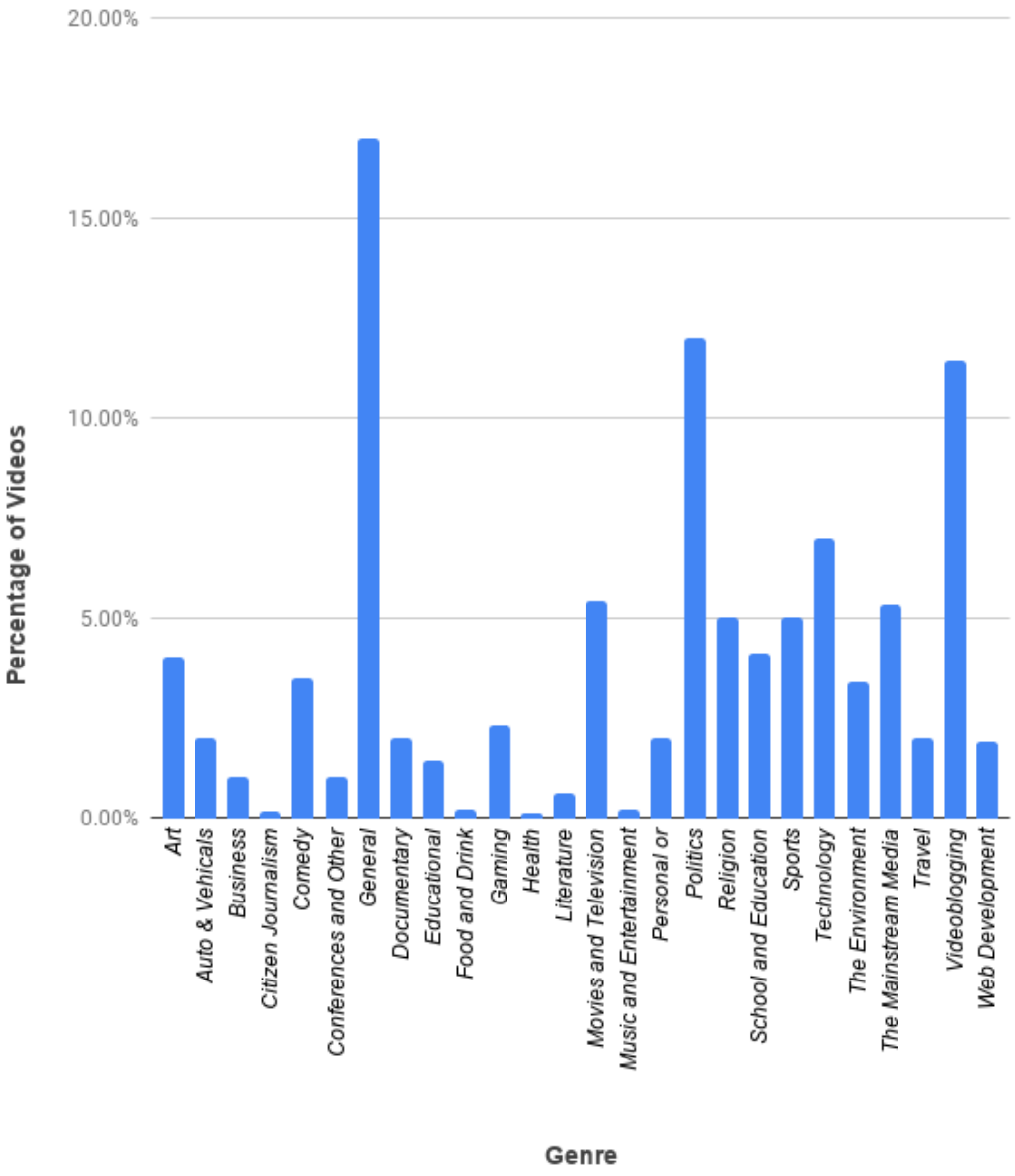

Figure 4.2: Overall genres distribution in the Blip10000 collection.

The following examples are some of the subtopics that are found on the most common genres.

- *Technology*: this genre represents videos about technology related topics, such as devices and software reviews by users or by domain experts, acquiring new technical skills, such as courses and workshop on learning Web design and development. As well as other tech-related topics such as open source adoption and the current initiatives by major tech companies.

- *Politics*: this genre represents videos about government and political related topics such as climate changes, US economy trends, UK trade initiatives, US presidential election, China industrial and trade Policy.

- *General*: this genre represents videos which can apply for multiple genres, such as TV shows interviewing politicians and technology experts on the the same video can apply for both *Politics and Technology* genres. Other example is the event videos which cover government-related summits and activities which can apply for both Event and Politics genres.

- *Videoblogging*: this genre represents videos about users reporting their personal daily activities such as reviewing books they read, cities they visit and the experiences they had.

4.3.1 Blip10000 ASR Transcripts

ASR transcripts were also provided with by the S&H organisers. These transcripts were generated by LIMSI using the LIMSI/Vocapia speech-to-text system[8]. Lamel and Gauvain (2008) used a language identification detector to automatically identify the language spoken in the whole video along with a language confidence score. Each file with a language identification score equal or greater than 0.8 was transcribed in the detected language.

[8]http://www.vocapia.com/news/2011_07_15.html

The LIMSI ASR system was trained using multi-layer perception (MLP) features described in detail in (Le et al., 2010; Lamel and Gauvain, 2008). It is worth noting that our work treats the ASR system as a black-box within the SCR framework. Our proposed techniques are designed to perform able independently of how the ASR system was built and trained[9].

Due to the scale and complexity of the test set, the ASR quality has not been formally reported by the creators of the Blip10000 collection (Schmiedeke et al., 2013). However, Eskevich et al. (2013) estimated the ASR quality of these transcripts using the Word Recognition Rate (WRR) of the relevant segments for 30 queries of the ones provided by the official S&H task. The WRR was found to hugely vary between 40% and 90%, which is realistic state-of-the-art rates for a transcription task of this variability and complexity.

The ASR transcripts together with the associated metdata files were provided as enriched XML files, as shown in Figures 4.3,4.4. For our task, we processed the XML files into searchable structured documents, where each document contains all three fields, ASR transcripts, title and description.

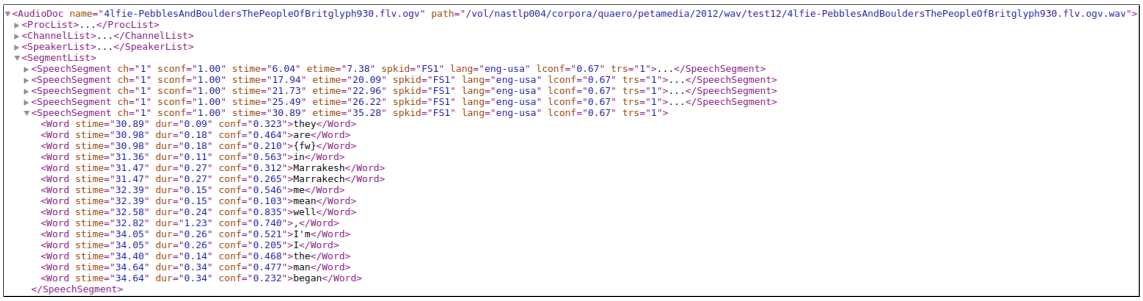

Figure 4.3: Example of the XML representation for ASR transcripts as provided in the bilp10000 collection.

4.3.2 Blip10000 Metadata

The length statistics of the UGS fields are shown in Table 4.2 which shows there is a huge variation in the length distributions across different fields. Table 4.2 also highlights the length variations of individual fields between the videos. For example,

[9]LIMSI/Vocapia did not expose any information about the data used to train their ASR system.

```
<video>
  <title><![CDATA[Showing and hiding layers]]></title>
  <description><![CDATA[Learn how a layer can be easily hidden and shown on the web pages events. Its too easy!
      ]]></description>
  <explicit>false</explicit>
  <duration>206</duration>
  <url>http://blip.tv/file/1003994</url>
  <license>
    <type>Creative Commons Attribution-NonCommercial-ShareAlike 2.0</type>
    <id>5</id>
  </license>
  <tags>
    <string>dreamweaver_8_free_video_tutorials</string>
    <string>free_dreamweaver_tutorials</string>
  </tags>
  <uploader>
    <uid>194769</uid>
    <login>adobedreamer</login>
  </uploader>
  <file>
    <filename>Adobedreamer-ShowingAndHidingLayers182.flv</filename>
    <link>http://blip.tv/file/get/Adobedreamer-ShowingAndHidingLayers182.flv</link>
    <size>3843589</size>
  </file>
  <comments />
</video>
```

Figure 4.4: Example of the XML representation for user-generated metadata as provided in the blip10000 collection

Table 4.2: Length statistics for (measured at the word-level) for Blip10000 fields.

Metric	Title	Desc	ASR
Standard Deviation	3.0	106.9	2399.5
Average Length	5.3	47.7	703.0
Median	5.0	24.0	1674.8
Maximum Length	22.0	3197.0	20451.0
Minimum Length	0.0	0.0	0.0

while one video may have no transcript produced, another may contain over 20K words. The number of sentences found in each field is shown in Table 4.3. Sentences were detected based on the presence of any of these characters ('.', '?', '!') at the end of a token. As it can be seen from Table 4.3, there is also huge quantity variation in terms of sentence distribution across UGS fields. Both Table 4.2 and Table 4.3 show that the speech ASR transcripts have more content (in terms of words and sentences) than both title and description fields which were provided as additional metadata by the uploader. Furthermore, although they were written by the same person (uploader) within the same conditions and settings, numbers from both tables also demonstrate the difference in word length and sentence distribution between the title and description fields. Across all fields, Titles are generally the shortest UGS field with an average length of 5 words.

Table 4.3: Sentence distribution for the Blip10000 fields.

Number of Sentences	Title	Description	ASR
No content	6,731	1,923	5,366
1 sentence	7,730	12,737	206
>1 sentence	376	178	9,208
>5 sentences	0	69	8,219
>100 sentences	0	12	2,393

Table 4.4: FRES scores of the title and description fields in Blip10000 collection.

FRES Score	Titles	Desc	School level	Rating
100-70	286	40	5th, 6th grade	Easy to read.
50-30	172	34	College	Difficult to read.
Less than 30	7,649	12,841	College graduate	Very difficult to read.

In order to assess the quality of the metadata fields, we calculate the Flesch Kincaid readability score (FRES) of these fields to measure the readability of these fields. FRES is a well known and widely used readability test that is utilised to evaluate the complexity of the text in order to determine the number of years of education required for someone to understand it (Kincaid et al., 1975). FRES estimates the complexity based on the words per sentence and syllables per word in a given text as shown in Equation 4.4.

$$FRES = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right) \tag{4.4}$$

Table 4.4 shows the distribution of readability scores for blip10000 metadata (titles and descriptions fields), together with the school level and rating for each score[10]. Results in Table 4.4 show that over 98% of the titles and descriptions are found to be very difficult and highly complex to read. This indicates that these fields were supplemented to the UGS platform with little attention to the readability of the uploaded text. This low level of text readability in these fields can be attributed to the nature of social media platforms which are encouraging users to publish more content with no restriction on style or quality of the uploaded text.

Beside readability, the following are examples of the quality issues found in the Blip10000 metadata.

[10]FRES is calculated based on the implementation found in https://github.com/rossweinstein/readability

- Generic and vague metadata : Figure 4.5 shows an example of video from news broadcast channel called *("GVTV News")* which has all video titles in the following format *("GVTV News NCTV11")*. While this video has over 28 minutes of speech discussing information and issues about multiple political parties in the United States, the supplemented title does not represent what the video is about but rather a vague indication that it is news video. Figure 4.5 also shows that the supplemented description of this video contains only contact information about the channel such as the name of the news channel and a link to their website.

- Missing metadata : as shown in Table 4.3 many videos have missing metadata either the Title or description are left empty. This behaviour is not restricted to certain type of channels or video uploaders. For example, the document shown in Figure 4.6, taken from a semi-professional event channel called "*(Gov2event)*", has the title of *("The Platform for Change: Tim OReilly on Gov 2.0")* while the description is left empty and does not discuss anything about the actual content of the video.

- Out-of-Vocabulary (OOV) words : At their preference, users may choose to write specific abbreviation or certain phrases in the uploaded metadata that are often unique across the collection and not representative of the actual content. For example, several videos titles from the blip10000 collection have words such as *("A0002", "A0219", "D025", "SE#7", "KDE42", "MA123", "RSDC")*. These OOV words are not present in any speech transcripts of the blip10000 since they were not recognised by the ASR system. Other type of OOV words are video titles which rather refer to the system filename of the uploaded video such as *("Movie93", "Movie83", "RR Eps.222", "11-part4", "Episode1")*. Furthermore, some of the metadata included merged words such as *("Adobedreamer", "TimeLapse", "FrontHouse","FootyMorning", "PrivateLessons", "Globalist", "mailbag")*

The quality of metadata found on UGS platforms can be highly variable since it depends on the characteristics and preference of the uploaders, who have a varying

background, interest, writing quality and style. For example, while a professional event broadcast channel (such as the one demonstrated in Figure 4.5 may choose to not pay attention to writing high quality metadata for their uploaded videos, other user channel such as the one showing in Figure 4.7 can have high quality metadata and include specific words that explain the content of the video. The nature of the UGS metadata makes our retrieval task in this work very different from previously studied SCR collections which utilised high quality metadata written by professional indexers (Jones et al., 2007; Pecina et al., 2008).

Filename : *GVTVNEWS-GVTVNews011208NCTV11961.flv.ogv*
Title : GVTV News 01=12-08 NCTV11
Description : Nevada County Local and World news brought to you by Grass Valley Television and anchored by Elisse Garner and Randy Hansen.
This Project is also shown on NCTV Public Access Nevada County Channel 11 through Comcast. Also you can recieve podcast through Apple Itunes just search (gvtv news). If you have any questions or news contact us at news@gvtv.org or call us 530-477-3309.Visit our website at www.gvtv.org .
Video length : 28 minutes

Figure 4.5: Blip10000 document example from GVTV News (a news channel) where generic and vague metadata are provided.

4.4 TREC Standard adhoc Collections

In order to address our research questions, in some sections of this thesis, we carry out our experiments over an additional test data to verify the robustness of the proposed approaches. Therefore, in addition to the Blip10000 collection, we conducted

Figure 4.6: Blip10000 document example from Gov2event (events coverage channel) uploaded with no showing description of the content.

experiments on two challenging and noisy text-based standard collections from the TREC evaluation benchmark, detailed in previous studies (Zhou and Croft, 2007; Shtok et al., 2012; Kurland et al., 2012), as follows.

- Large-scale Web collection *WT10G*[11] (TREC topics 451-550) data collection that contains 1,692,096 web documents.
- *ROBUST*[12] TREC Volumes 4 and 5 minus the Congressional Records (CR) collection (TREC topics 301-450, 601-700), which contains 524,929 news text documents.

We use the title fields of TREC topics as the main topic set for our experiments. A summary of all test collections used in this thesis is shown in Table 4.5.

In the next section we describe the search topic sets we use for our UGS retrieval tasks.

[11] ir.dcs.gla.ac.uk/test_collections/wt10g.html
[12] trec.nist.gov/data/t13_robust.html

Filename: *Aramistech-KeepYourDataSafe832.flv.ogv*
Title : Keep Your Data Safe
Description : In this video I will show you a program named NTI Shadow which has some unique features that make backing up your data with this software easy and worry free. Visit my blog at http://www.aramistech.com/software/keep-your-data-safe for more information about this software and the review on this video.
Video length : 11 minutes

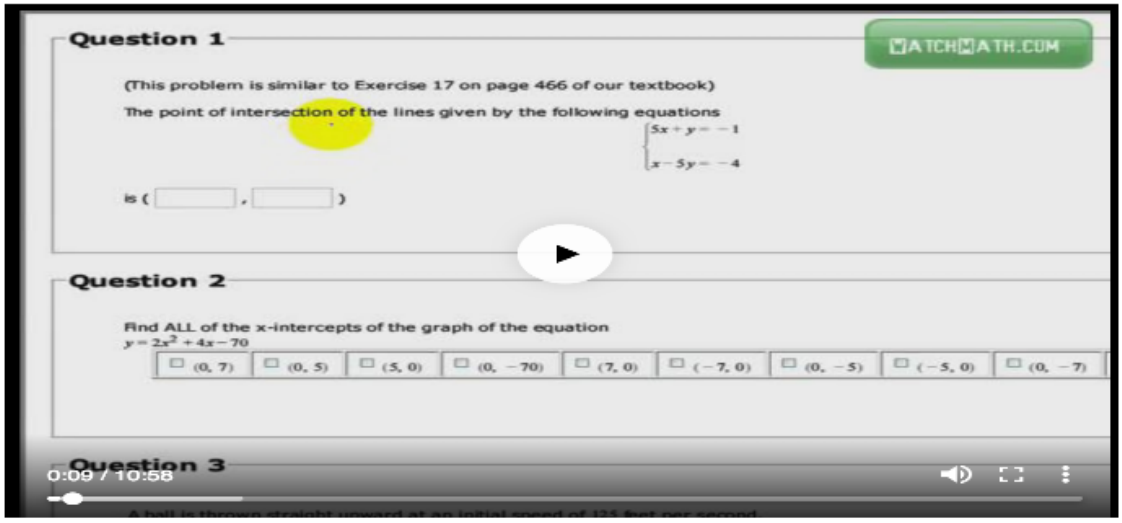

Figure 4.7: Blip10000 document example from Aramistech (user channel focusing on technology) showing high quality metadata uploaded with the video.

4.5 UGS Topic Sets

Our experiments over the blip10000 use the topics provided by MediaEval benchmark task, and an extension to these topics developed within our work to support our experimental investigation. We set up multiple search tasks over the Blip10000 collection to understand retrieval challenges of UGS content from different perspective. We explore two novel UGS search tasks, namely, the known-item and adhoc-search for both long (with an average of 7 words per topic), and short (with an average of 2 words per topic) topics.

Furthermore, cross-lingual topic sets were created in different languages (non-English, namely French and Arabic) to study the UGS challenges from cross-lingual perspective.

The following topic sets were used for our investigation:

- Known-item English Monolingual Search *(Mn-Kn)* query set.

Table 4.5: Summary of the test collections used in this thesis

Collection	*Data*	*Total Documents*	*Topics*
WT10G	WT10g	1,692,096	TREC (451550)
ROBUST	Disk 4&5 - CR	528,155	TREC (301 - 450) & TREC (601 - 700)
Blip10000	Crawled from Blip.tv	14,838	monolingual/cross-lingual queries as explained in Section 4.5

Table 4.6: Length statistics (at word-level) for the topic sets provided by the Mediaeval S&H 2012 task.

Metric	short-query set	Long-query set
Standard Deviation	2.4	7.6
Average Length	5.1	10.8
Median	5	9
Max	12	37
Min	2	5

- Adhoc English Monolingual Search using the *(Mn-Ad)* query set.
- Cross-Lingual Arabic-English adhoc search using the *(Cl-Ar)* topic set which is translated using Google Translate API(Google, 2017).
- Cross-Lingual French-English adhoc search using *(Cl-Fr)* which is translated using Google Translate API (Google, 2017)
- Cross-Lingual Arabic-English adhoc search using the *(Cl-Ar-Moses)* topic set which is translated using Moses translation system (Koehn et al., 2007a).
- Cross-Lingual French-English adhoc search *(Cl-Fr-Moses)* translated using Moses translation system (Koehn et al., 2007a)

Section 4.5.1 describes the topic set we use for the known-item search, while Section 4.5.2 describes how the adhoc and CLIR topic sets are created for this task.

4.5.1 Known-item Search - (Mn-Kn) query set

A topic in Known-item search indicates that it was written to find a single previously seen relevant (the *known-item*). Therefore, a search system is required to retrieve

the single known item, and to rank it as highly as possible.

The S&H task (Eskevich et al., 2012b) was a known-item search task constructed over the Blip10000 collection. The task provided 60 English topics collected using the Amazon Mechanical Turk (MTurk) crowd-sourcing platform MR. Each topic contains a full query statement (long-query) and a terse web type search query (short-query). The length statistics of these topic sets are shown in Table 4.6. For our investigation, we use both topic sets to give a better understanding of retrieval behaviour for both the monolingual and CLIR tasks. We use the long-query set for known-item search Mn-Kn, while we edited the short-query into an adhoc queries as will be later explained in Section 4.5.2.

We evaluate our investigations for these known item topics using the standard metric for this task, that is the Mean Reciprocal Rank (MRR) metric computed as shown in Equation 4.5, where $rank_i$ indicates the rank of the relevant known item for the ith query is found.

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \tag{4.5}$$

Similar to other known-item experiments, we also chose to define the recall as the number of times the relevant-item was found across the set of queries (Büttcher et al., 2010). The recall is reported by default at the standard TREC1000 results cut off (1000 results).

4.5.2 Adhoc Search - Mn-Ad Topic set

In order to explore the adhoc search for our UGS retrieval task, we created an adhoc version of the S&H's short query set by asking two assessors to adjust each topic into a more general form. This was done by removing specific terms that are related to one specific relevant item and re-writing the whole query into more natural adhoc form; for example the query "Troubleshooting the EEE PC 900 Laptop" was changed to "Troubleshooting PC and Laptops". Our new adhoc topic set is referred to as *Mn-Ad* in this thesis.

To create the relevance judgements for these adhoc queries, a *pooling method* was developed by combining different result lists created using different IR methods (Buckley et al., 2006) . We ran each query using 6 different retrieval models TFIDF using implementation of Robertson and Sparck Jones (1976), Okapi BM25 (Robertson et al., 1998), PL2 (Amati, 2003), language modeling (LM) Himestra (Hiemstra, 2001), DLH13(Amati, 2003)) in different indexes. These indexes included combinations between the textual metadata associated with each video (title and descriptions) and the ASR transcripts. Retrieval runs were produced using the Terrier retrieval platform[13] which will be later explained in Section 4.2.1. Stop words were removed based on the standard Terrier list, and stemming performed using the Terrier implementation of Porter stemming. We then used the NTCIR pooling script[14] to generate a pool combining the top 30 results for each query.

We adjusted *Relevation*[15], an open source text-based IR relevance judging system introduced in (Koopman and Zuccon, 2014), to embed videos together with their with their metadata (descriptions) and assigned two assessors fluent in English to evaluate the results of each query as shown in Figure 4.9. Both assessors work as a part-time reviewers for annotation tasks, one female assessor in at the age of 24, and another male assessor at the age of 31. For each query, the list of pooled videos results were retrieved. Assessors were asked to go through each of the results items and play the video, read the description and mark their judgement on whether video is relevant to the query or not. As some queries are too general and may overlap with multiple topics, we chose to add another option which is "somewhat relevant" to flag these results[16].

Figure 4.8 shows a video result item with their judgement for the query "Troubleshooting PC and Laptops". Since some of the queries are too general, we chose The agreement level between the assessors was 93%. We produced a relevance file

[13]http://www.terrier.org/
[14]http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html
[15]https://github.com/ielab/relevation
[16]Based on initial feedback from the assigned assessors

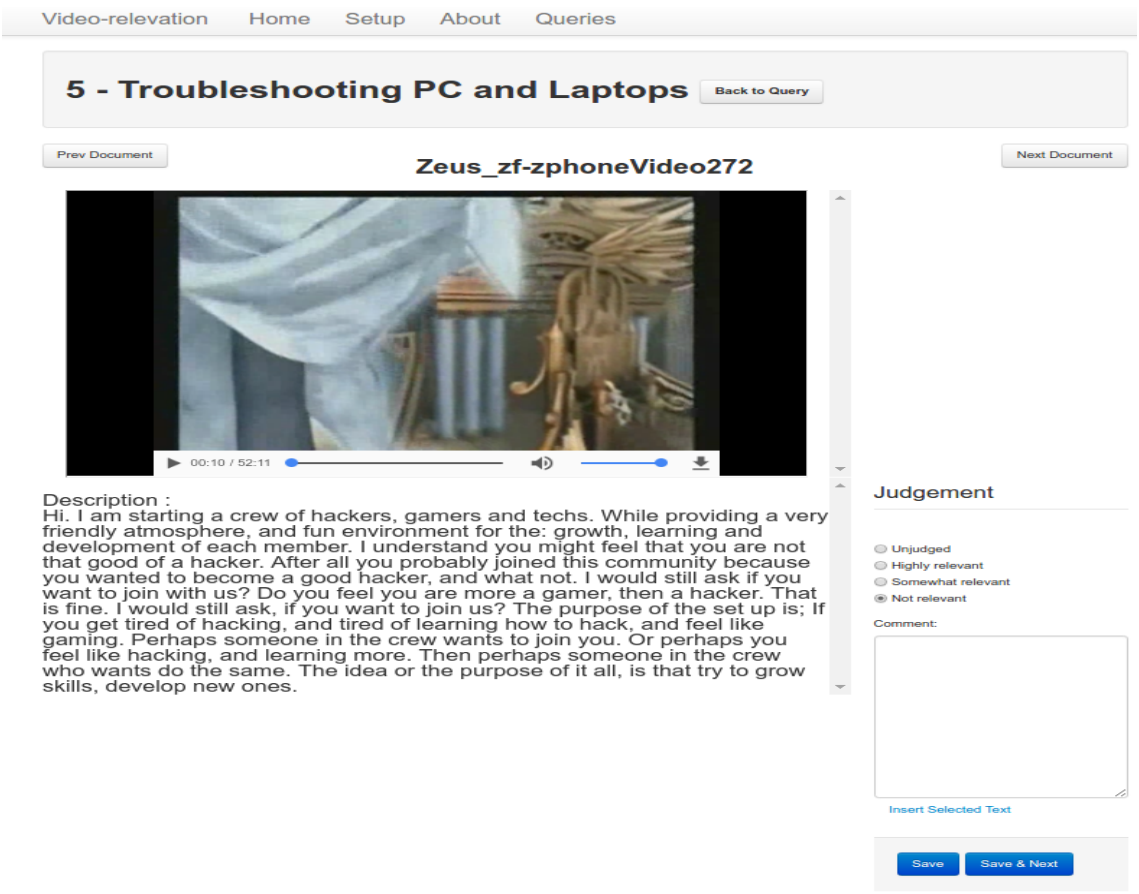

Figure 4.8: A Video Evaluation page Relevation IR relevance Judging Screen-shot.

containing each topic together with the list of videos selected as relevant[17] by both reviewers. Each topic has between 7-13 relevant videos with an average of 9 relevant items per topic, this number depends mostly on the difficulty of each topic and the availability of relevant documents in the collections. In the next section we explain how we developed a cross lingual version of these queries for our UGS task.

4.5.3 CL-UGS Topic sets

To create the CLIR topic sets, we extended the adhoc monolingual (Mn-Ad) topics by giving them to Arabic and French native speakers, who are also fluent in English, and asking to write their own queries using their native languages about each of these topics. The question we asked is, *"How would you create your query based on this*

[17]To be qualify for relevance, the video has to be marked by both reviewers as high or somewhat relevant.

information need?".

Having these topics being expressed in two languages (in Arabic and French) allows us to draw better conclusions about the CLIR performance for this task from different perspective.

To set up the CL-UGS task, query translation was utilised to bridge the language gap between the non-English CLIR topic set and the English documents set. The topic sets were translated using two state-of-the-art MT techniques, as follows.

- Off-the-shelf MT tool : We used the Google translate API[18] to translate these query sets back into English and generate the two query sets *(Cl-Ar) and (Cl-Fr)*.

- Open-box MT system : We used Moses MT (Koehn et al., 2007a) open-source to setup two translation systems namely, Arabic-to-English and French-to-English. We utilise these two systems to generate both *(Cl-Ar-Moses)* and *(Cl-Fr-Moses)* query sets.

As would be expected, MT translation produced errorful versions of the original monolingual ones. In addition to the much anticipated deletion/insertion errors, there were also Named Entity Errors (NEEs) and Out-Of-Vocabulary (OOV) errors that both MT systems did not translate correctly. In Section 4.6, we show how these translation edits impact the retrieval effectiveness of the MT translated queries compared to the monolingual ones.

[18]https://developers.google.com/translate

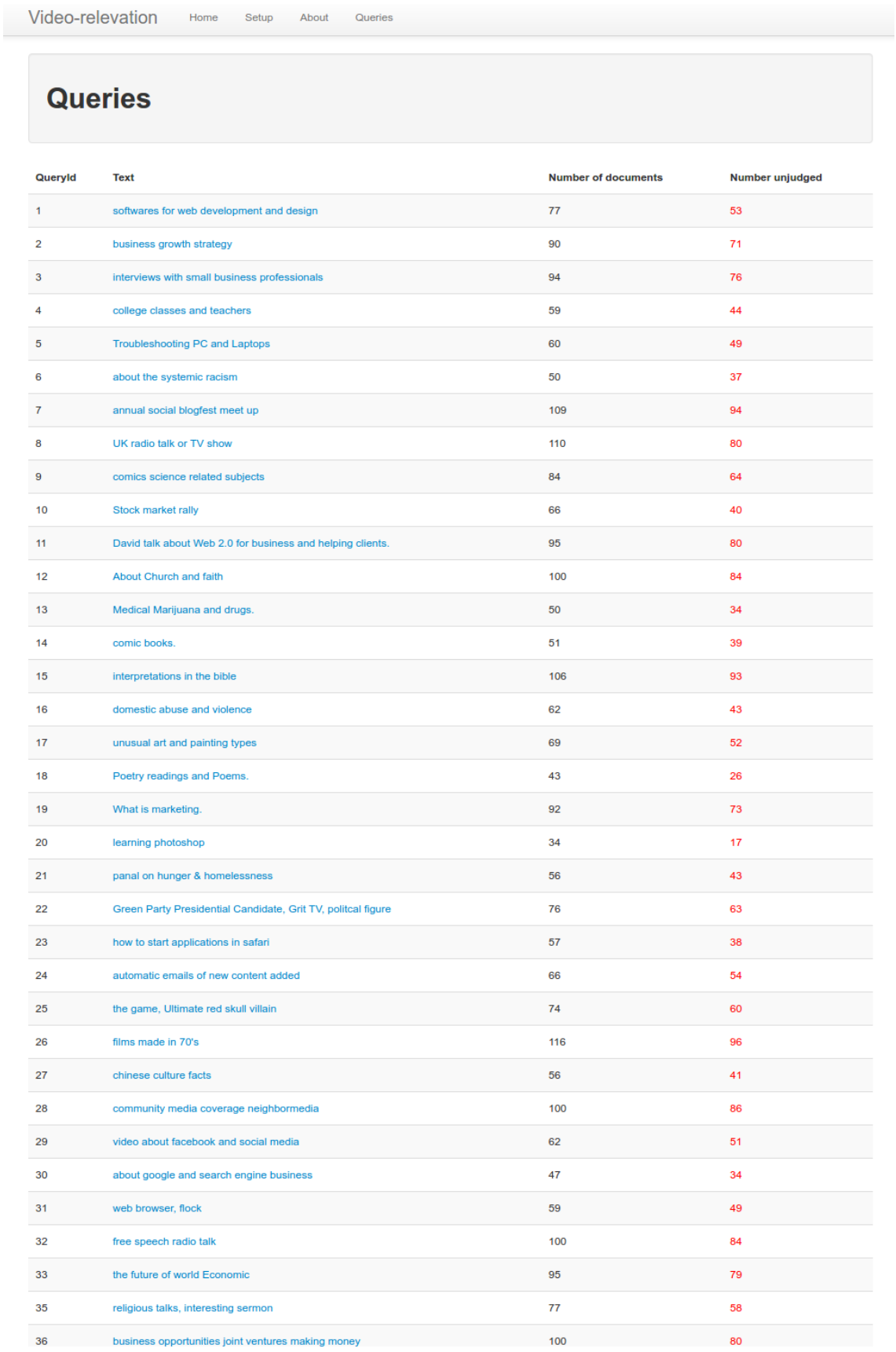

Figure 4.9: Video Relevation - Queries pages.

In the next section, we explain how these MT systems were set up and integrated into our IR evaluation framework.

4.5.4 Machine Translation for CL-UGS Topic sets

Our experimental investigation of CL-UGS retrieval used two translation tools. An off-the-shelf MT approach using the online Google Translate API, and an open-box MT approach using the Moses statistical MT toolkit. These two systems are explained as follows.

Translation us Google Translate API

We utilised Google translate api [19] to translate each of the queries, we built a simple application in Java that connects to the translate API using a JSON snippet. This JSON snippet contains the query to translate, the query source language ("FR or AR"), and the target language ("EN").

Moses MT System

Our MT system is a phrase-based (Koehn et al., 2003), that is developed using the Moses SMT toolkit[20]. Moses provides an implementation of different tools that can be used for MT. Moses has two main components as follows.

- *Training*: This component takes the raw data (parallel and monolingual) build a machine translation model out of it. Raw data are being tokenzied and parallel sentences from each language are then word-aligned typically using GIZA++ [21] tool (Gao and Vogel, 2008). GIZA++ provides an implementation to a set of statistical models developed at IBM in the 80s. These word alignments are used to extract phrase-phrase translations. A language model is then built using monolingual data in the target language, the decoder uses

[19]https://developers.google.com/translate
[20]http://www.statmt.org/moses/
[21]Available at `http://www.cs.cmu.edu/~qing/`

this to adjust the fluency of the output. Tuning is the final stage of MT creation, where different models are weighted against each other to generate the best possible translations.

- *Decoding* : This component is responsible for performing the translation. Given a trained machine translation model and a source sentence, it will translate the source sentence into the target language. Moses decoder seeks to find the highest scoring sentence in the target language (based on the the trained translation model) corresponding to a given source sentence. The decoder can output the single best candidate translation, or it is also possible for the decoder to output n-best candidates as ranked list of the translation candidates.

For our task, we created two MT systems using Moses which are AR-to-EN and FR-to-EN to translate the Arabic and French queries to English. The AR-to-En MT system was trained using the bilingual training corpora listed in Table 4.7. All training datasets were provided by Linguistic Data Consortium (LDC) [22] that included Modern Standard Arabic and other most popular dialects of Arabic. For the FR-to-EN MT, we used News-Commentary (Nc7)[23] (Tiedemann, 2012), and Europarl (Eparl7)[24] (Koehn, 2005) corpora, as shown in Table 4.8.

Arabic data was tokenised using MADA-ARZ version 0.4 (Habash et al., 2013). For French and English we used the default NLTK tokensier implemented by Moses toolkit [25].

For both systems, word alignments in both directions were calculated using a multi-threaded version of the GIZA++ [26] tool (Gao and Vogel, 2008). The parameters of our MT system were tuned using Minimum Error Rate Training (Och, 2003). Also for both system, we utilised the monolingual English data of the training data to build 4-gram back-off language model using Moses SRILM Toolkit [27] for improv-

[22]https://catalog.ldc.upenn.edu/
[23]http://opus.nlpl.eu/News-Commentary.php
[24]http://www.statmt.org/europarl/
[25]http://www.nltk.org/
[26]Available at `http://www.cs.cmu.edu/~qing/`
[27]http://www.statmt.org/moses/?n=FactoredTraining.BuildingLanguageModelntoc3

ing the fluency of the output. In the next section, we show how these MT systems are evaluated for our cross-lingual retrieval task.

LDC Corpus	dialect	AR tokens	EN tokens
bolt	Egyptian	1.70M	2.05M
thy		282k	362k
bbnturk		1.52M	1.58M
bbnegy		514k	588k
gale	Moderen Standard Arabic (MSA)	4.28M	5.01 M
fouo		717 k	791k
ummah		3.61M	3.72M
iraqi	Iraqi	1M	1.14M
bbnlev	Levantine	1.59M	1.81M
Total		15.2M	17M

Table 4.7: The sizes and the dialect of bilingual LDC training corpora for the Arabic-to-English Moses MT.

Corpus	FR tokens	EN tokens
Eparl7	2.3M	2.2M
Nc7	1.0M	1.2M
Total	3.3M	3.4M

Table 4.8: The sizes of bilingual training corpora used for the French-to-English Moses MT.

4.6 Designing a Retrieval Framework for CL-UGS

As explained before, UGS content contains several fields of varying quality and format, therefore setting up a retrieval framework for UGS content involves making choice between multiple experimental design options. In this section, we design preliminary experiments to design a CL-UGS framework and develop the baseline for our investigation. The aim of these experiments is mainly to answer two major questions as follows.

First, in Section 4.6.1, we study how UGS fields should be processed, and best represented for retrieval. Then, in Section 4.6.2, we conduct an experimental investigation to study how MT should be implemented for CL-UGS retrieval.

4.6.1 Document Representation for UGS retrieval

In this experiment, we conduct our initial investigation to report the baseline performance of UGS retrieval and compare the robustness of both using structured and unstructured document representations for our task. Our experiments in this section are similar to those explored previously by Jones et al. (2007). In this experiment, alternative metadata fields and transcripts are utilised and combined as source-of-evidence in both structured and unstructured document format for retrieval. The aim of this investigation is to answer the question of whether considering field-based and structured representation of UGS content is beneficial for UGS retrieval or not. The difference between both document structures in UGS settings is explained as follows.

- In *unstructured representation*, all fields (including ASR transcripts and metadata) are combined together and treated as one source-of-evidence during indexing and retrieval.

- In *structured representation*, indexing and retrieval are carried out on the field-level. In this settings, each field in UGS (title, description or transcript) is indexed and weighted differently for retrieval.

The UGS search tasks we study in this experiment are explained as follows.

- Known-item search using the English monolingual topic set MN-Kn for long queries.

- Adhoc search using the English monolingual topic set MN-Ad for short queries.

For the structured experiment, we indexed the fields separately on each document as previously shown in Figure 4.10, where each field is tuned and weighted separately for retrieval. As described previously, the structured document representation has three fields which are *ASR, Title and Desc.* for retrieval in this experiment, we used the DFR PL2F model[28] (Macdonald et al., 2005). This is a modified version

[28]Terrier implementation of this model can be found in `http://terrier.org/docs/v4.0/javadoc/org/terrier/matching/models/PL2F.html`

of the PL2 model explained earlier. The PL2F model is designed to adopt per-field weighting when combining multiple evidence fields into a single index for search. The term frequencies from document fields are normalised separately and then combined in a weighted sum.

PL2F uses the same document scoring function as PL2, as explained previously in Equation 2.3, but here tfn is the weighted sum of the normalised term frequencies in the normalised term frequencies tf_X for each field x, in our case $x \in (ASR, title, desc)$ as indicated by Equation 4.6. Where l_x is the length of the field x in document d. $avgl_x$ is the average length of the field x across all documents, and c_x, w_x are the per-field normalisation parameters. This per-field normalisation feature in PL2 modifies the standard PL2 document scoring function to include the weighted sum of the normalised term frequencies tf_x.

$$tf_n = \sum_x (w_x.tf_x.\log_2(1 + c_x.\frac{avgl_x}{l_x})), (c_x > 0) \tag{4.6}$$

tf_x also needs two parameters w_x, c_x to be set. Hence, for scoring each indexed document we need to set these parameters: C_x is the set of per-field length normalisation parameters c_x that need to be set for every field as C_x =\{ c_asr, c_title, c_desc\}, and W_x is the set of per-field boost factors w_x that need to be set for each field as W_x =\{ w_asr, w_title, w_desc\}.

For the unstructured document experiment, we use the PL2 retrieval model, with no field-weighting involved as shown in Figure 4.11 where each one document has single field which is *the UGS field*. Unlike the document structure shown in Figure 4.10, for this experiment, we combined the content from all fields without differentiating between them. Indexing and retrieval using Terrier Engine as explained in Section 4.2. The parameters tuning method that is used for both PL2 and PL2F is explained in the next section.

Figure 4.10: Example of a combined-field structured document that contains three fields (Title, Desc and ASR).

```
<DOC>
<DOCNO>
EconomyInCrisis-AFutureOfCleanEnergy384.
flv.ogv
</DOCNO>
<TITLE>
A Future of Clean Energy
</TITLE>
<DESC>
To move forard the U.S. must use
clean energy.
</DESC>
<ASR>
Hello and welcome to daily news and information up
update . Today's topic if future of clean energy
after just passing in the House of Representatives
bible vote Bible of two to 19 to two 12 . The
newly minted Waxman Markey clean energy and Security
security Act act could possibly be . In a it a
landmark piece of legislation for the United States
the intent of the bill is to increase protections
for American workers VA BA climate context border
tax provisions provision . This provision provisional .
Place plays a tariff on goods produced in countries
which do not uphold the same environmental health
and safety regulations as the United States the White
House is currently not keen on the idea of a border
tax provision provisions as ...
</ASR>
</DOC>
```

Figure 4.11: Example of a unstructured UGS document which contains one field (UGS field)

```
<DOC>
<DOCNO>
EconomyInCrisis-AFutureOfCleanEnergy384.
flv.ogv
</DOCNO>
<UGS>
A Future of Clean Energy
To move forard the U.S. must use
clean energy.
Hello and welcome to daily news and information up
update . Today's topic if future of clean energy
after just passing in the House of Representatives
bible vote Bible of two to 19 to two 12 . The
newly minted Waxman Markey clean energy and Security
security Act act could possibly be . In a it a
landmark piece of legislation for the United States
the intent of the bill is to increase protections
for American workers VA BA climate context border tax
provisions provision . This provision provisional .
Place plays a tariff on goods produced in countries
which do not uphold the same environmental health
and safety regulations as the United States the White
House is currently not keen on the idea of a border
tax provision provisions as ...
</UGS>
</DOC>
```

Parameter Tuning For PL2 and PL2F

The PL2 model has single hyper-parameter c that has to be tuned for UGS field of the unstructured document, while the PL2F model has two sets of hyper-parameters for the three fields which are C_x =\{ c_asr, c_title, c_desc\}, and W_x is the set of per-field boost factors w_x that need to be set for each field as W_x =\{ w_asr, w_title, w_desc\}.

In order to find the optimal PL2's c parameters for the combined UGS content, we use *a 2-fold cross-validation evaluation paradigm* (Shtok et al., 2012). The Mn-Ad queries were used for training and testing. The Mn-Ad topic set was randomly split into training and testing sets (30 queries each). During the training process, we performed data sweeping through the range of [0.1, 2] with an interval of 0.1, and through the range of [0.5, 20] with an interval of 0.5. We performed 20 different splits to switch the roles between the training and the testing sets, and reported the best performing c parameter in terms of MAP, by taking the average between the 20 different runs.

For tuning the PL2F weighting in this experiment, we fixed the W_x weights \{ w_asr, w_title, w_desc\} as \{ 1, 1, 1\} to give equal weight to all fields for this baseline experiment assuming that all of these fields are equally valuable for UGS retrieval[29].

For tuning the C_x weights \{ c_asr, c_title, c_desc\} of PL2F, we also used a 2-fold cross-validation using the Mn-Ad queries with the same tuning methods explained for the PL2. The PL2F C_x tuning process is explained as follows.

- For each parameter of the C_x weights, we fixed other parameters at 1 value (i.e. to tune c_asr, we fix c_title = 1, c_desc = 1).
- We ran data sweeping through the range of [0.1, 2] with an interval of 0.1, and through the range of [0.5, 20] with an interval of 0.5.
- We performed 20 different splits to switch the roles between the training and

[29]Note that in the Chapter, as part of Section 5.3, we try to adjust the weights of these fields in order to assess their effectiveness

Table 4.9: Obtained optimal PL2 parameters for each UGS field.

Field	Optimal parameter
c_UGS (Unstructured field)	1.2
c_ASR	1.3
c_Title	5
c_Desc	6

the testing sets and reported the best performing c parameter in terms of MAP, by taking the average between the 20 different runs.

The optimal c parameters found for each field using the cross-validation paradigm are reported in Table 4.9.

Finally, in order to assess the impact of tuning the c parameter for the PL2/PL2F models, we demonstrate its sensitivity on the MAP performance for each field (UGS, ASR, Title and Desc). We use the *optimal paradigm* proposed by Shtok et al. (2012) to find the optimal c hyper-parameter for each field evaluated using the Mn-Ad topic set.

Figure 4.12 shows the impact of tuning the c for each field on the MAP performance. It can be seen from Figure 4.12 that the optimal MAP performance for the UGS and ASR fields are found by setting the c parameter between 1 and 1.5, while the optimal parameter for Title/Desc to set c between 5 and 6.

It should be noted that these parameters did not show a statistically significant improvement over the one (which is setting c at 1) that is recommended in (Amati, 2003; Amati and Van Rijsbergen, 2002). The reason for this is that the high variation in length for each field and each document, which means that it is not possible to tune these parameters to work optimally for all UGS fields, documents and queries.

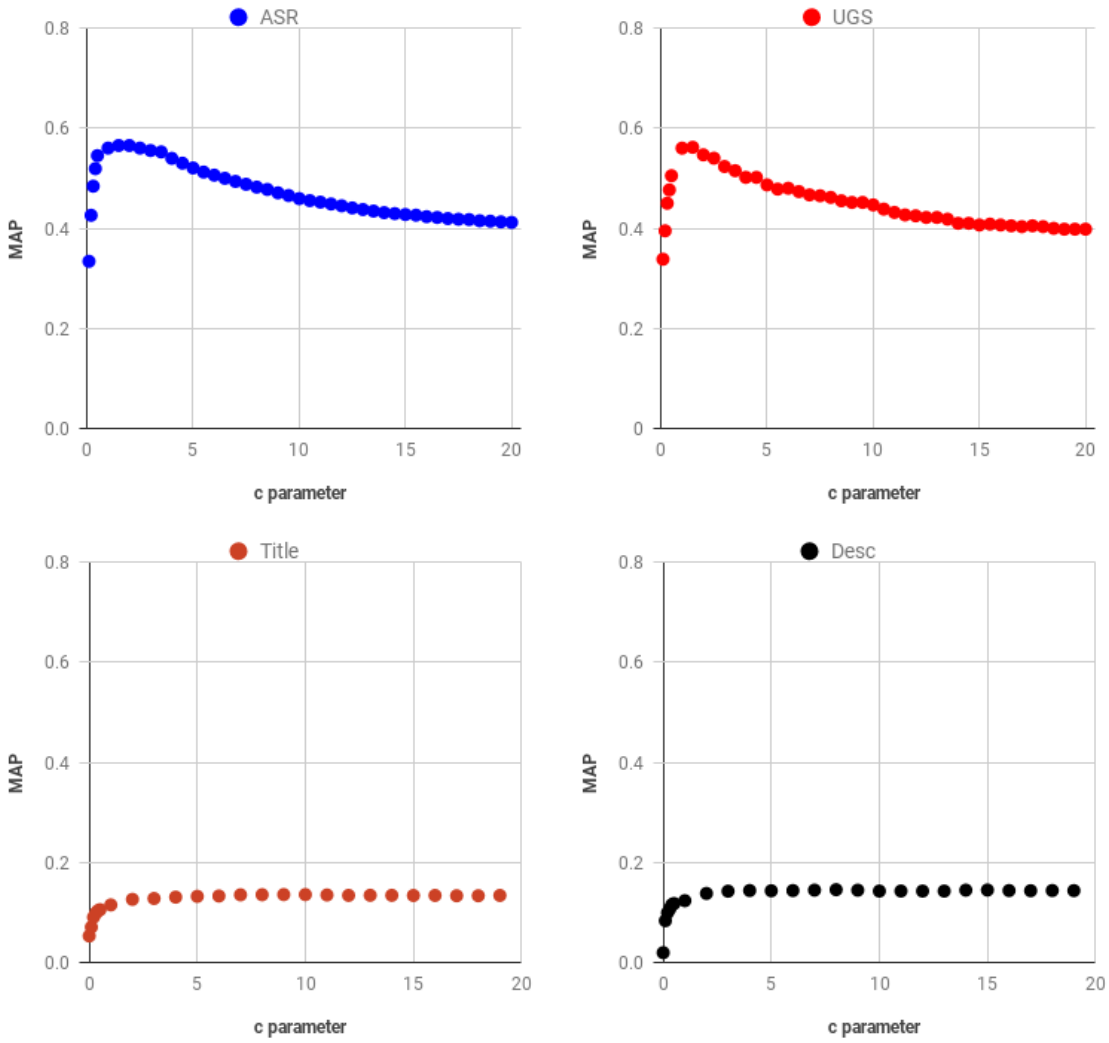

Figure 4.12: PL2 c hyper-parameter sensitivity for UGS fields using the Mn-Ad topic set.

Table 4.10: MRR performance for Mn-Kn topic set using both structured and unstructured document representation (Doc.Rep). ** indicates Statistically significant values with p-value < 0.05.*

	MRR	Recall	Doc.Rep	Model
Mn-Kn	0.3450	0.7508	Unstructured	Pl2
Mn-Kn	0.4634*	0.8667*	Structured	Pl2F

Table 4.11: MAP performance for Mn-Ad topic set using both structured and unstructured document representation (Doc.Rep). ** indicates Statistically significant values with p-value < 0.05.*

	MAP	Recall	Doc.Rep	Model
Mn-Ad	0.5313	0.8429	Unstructured	Pl2
Mn-Ad	0.5833*	0.9002*	Structured	Pl2F

Experimental results

Table 4.10 shows the performance (in terms of MRR) for the Known-item search using the Mn-Kn topic set, while Table 4.11 shows the retrieval performance (in terms of MAP) for the adhoc UGS search using the Mn-Ad topic set.

Results in both tables indicate the superiority of using the structured representation of three fields (ASR, Title and Desc) over the unstructured one. This result demonstrates the benefit of using PL2F model over PL2 for UGS content as it offers a per-field weighting/tuning for this task.

These results indicate that improved retrieval performance is achieved when greater significance is given to the metadata fields such as the Title and description, rather than merging them with the ASR field which may not actually provide a reliable and meaningful description of the video content,

4.6.2 Selecting MT for Query Translation in CL-UGS retrieval

In the previous section, we studied the effectiveness of using structured document representation over unstructured representation for UGS retrieval. The objective of this section is look into another feature of the experimental framework related to the CL-UGS. In particular, we seek to study which translation tool is more effective

for CL-UGS. We evaluate the effectiveness of using an open-box state-of-the-art MT system (developed using Moses MT system) over off-the-shelf MT tool (using Google Translate API) for query translation in CL-UGS. Details and implementation of these tools were provided previously in Section 4.5.4. The next sections describe the experimental setting we use to conduct this evaluation.

Experimental Setup

We setup the following CLIR search tasks over our UGS collections (Blip10000) for this investigation as follows.

- CLIR search based on Off-the-shelf MT using the translated topic sets using the Google API tool (*Cl-Ar and Cl-Fr*).
- CLIR search based on open-box MT machine translated topic sets using Moses system (*Cl-Ar-Moses and Cl-Fr-Moses*).

We use the same experimental settings explained in the previous section. Both PL2 and PL2F are again utilised for for both unstructured/structured UGS document representation as described in the previous section.

Experimental Results

Table 4.12 shows retrieval performance for CLIR search using PL2 model with unstructured document representation, while Table 4.13 shows the retrieval performance for CLIR search using PL2F model with the unstructured document representation. In order to evaluate the performance of the CLIR topic sets, we also show the respective monolingual performance using the Mn-Ad and the percentage of difference for each CLIR topic set against the monolingual performance (*% over Mono*).

Results from these tables reveal some preliminary insights for CL-UGS, which we summarise as follows.

- *Structured vs unstructured representation*: comparing the results between both Table 4.12 and Table 4.13 shows the superiority of using structured retrieval using PL2F for CL-UGS. This confirms the conclusion from the previous section where considering a per-field tuning during retrieval and indexing can help in both monolingual and cross-lingual settings.

- *Monolingual vs cross-Lingual UGS retrieval performance*: The obtained *% over Mono* percentages in both Tables show indicate that the performance is significantly lower for all CLIR queries across different translations. The lower performance of CLIR topics can be attributed to impact of translation edits in the CL-UGS settings. The best performing CLIR query was using CL-Fr (translated using the Google translate API) was between 22% and 25% lower than the monolingual one. This result is consistent with the result of previous work, which was reported in a similar French-to-English cross-lingual speech retrieval (Pecina et al., 2007), that the best CLIR runs using manually-created metadata with professional speech data was at 20% percent lower than the monolingual performance.

- *Machine Translation for CL-UGS*: comparing the performance results for CLIR queries indicate that using off-the-shelf for both Arabic-To-English and French-to-English CL-UGS is more effective. To better understand improvement gained by using the black-box MT tool, we computed the t-value (at the 95% confidence-level) for the AP performance difference between using both Google MT tool and Moses MT for each query. The test results in terms of t-values are shown in Table 5.2. This significance test shows that there is an 18% improvement in using Google MT for French (by comparing CL-Fr to CL-Fr-Moses topic sets), and 40% for Arabic (by comparing CL-Ar to CL-Ar-Moses topic sets). This results confirm that using Google MT API which is trained using data from the World Wide Web (WWW) information, is more suitable for this task over the use of Moses MT which is trained using much

Table 4.12: CL-UGS retrieval performance using PL2 Model with unstructured representation. ** indicates Statistically significant values with p-value < 0.05.*

	MAP	*Recall*	*% over Mono*
Mn-Ad (baseline)	0.5513	0.9167	-
Cl-Fr	0.4307	0.7833	**-21.88%*
Cl-Fr-Moses	0.3261	0.7667	**-40.85%*
Cl-Ar	0.2884	0.6667	**-47.69%*
Cl-Ar-Moses	0.2046	0.5667	** -62.89%*

Table 4.13: CL-UGS retrieval performance using PL2F model with structured representation. ** indicates Statistically significant values with p-value < 0.05.*

topic set	*MAP*	*Recall*	*% over Mono*
Mn-Ad (baseline)	0.5833	0.9333	-
Cl-Fr	0.4420	0.7833	**-24.22%*
Cl-Fr-Moses	0.3737	0.8011	** -35.93%*
Cl-Ar	0.3277	0.6833	** -43.82%*
Cl-Ar-Moses	0.2333	0.5833	** -60.00%*

more limited data sources. This results also match the previous performance evaluation of using Google MT tool versus training an MT system for CLIR (Leveling et al., 2009; Zhou et al., 2012) that using Google MT can be more effective.

4.6.3 Analysing the CL-UGS Performance

Our experiments from the previous section showed that the performance for the translated queries (CLIR) was significantly lower than the monolingual queries. In this section, we perform a query-level performance analysis of our CLIR queries to understand the underlying reasons behind this performance degrade. We calculate the difference in performance between each monolingual English query from the Mn-

Table 4.14: Comparison between the CL-UGS performance obtained by Off-the-shelf MT tool vs Moses MT translation according to the % AP reduction for each query. **Statistically significant values with p-value < 0.05.*

	CL-Ar vs CL-Ar-Moses	Cl-Fr vs CL-Fr-Moses
t-value	-3.88 *	* -2.22
Percentage	- 40% *	* -18%

Ad and its corresponding translated query from Arabic and French CLIR topic sets (CL-Ar and CL-Fr). Figure 4.13 shows the query-by-query performance compar-

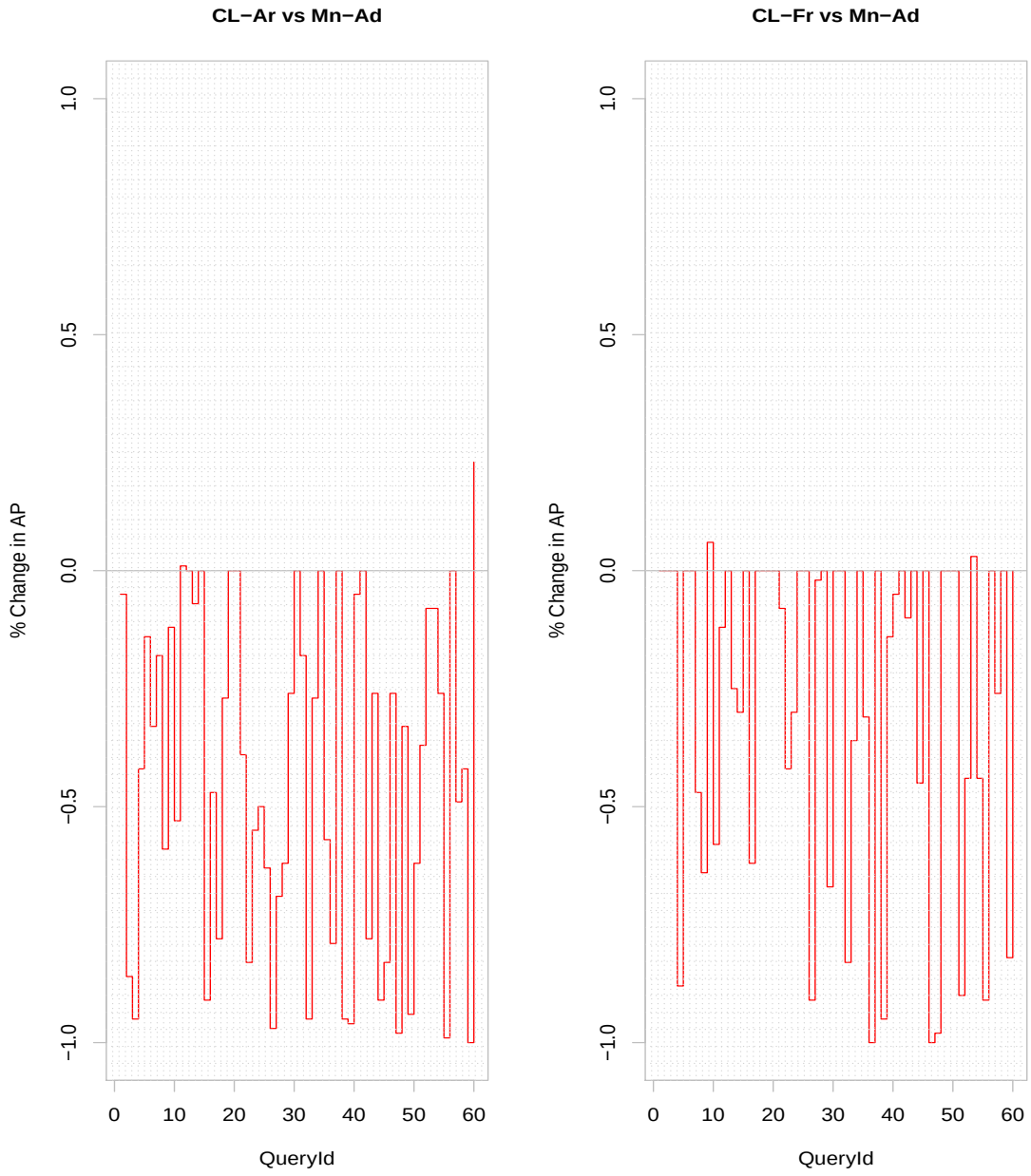

Figure 4.13: Query-level performance measured using the percentage of performance change between the CLIR version and the monolingual one(% Change in AP) for both Cl-Ar and Cl-Fr queries.

isons measured using the percentage of change between the CLIR version and the monolingual one (% change in Average Precision (AP)) for the CL-AR and CL-FR topic sets. It can be seen that the performance is significantly lower for most of CLIR queries.

To better understand the cause of these changes, we extracted some of the most impacted queries from both CLIR query sets (CL-AR and CL-FR) to analyse the translation noise that led to this performance degrade. Table 4.15 shows examples of queries which were negatively impacted from CL-Ar topic set (Arabic queries), while Table 4.16 shows examples of queries which were negatively impacted from CL-Fr topic set (French queries). To analyse the translation impact on the performance of these queries, we study the translation edits of the CLIR queries in terms of word choice which includes addition, deletion or substitutions by comparing them to the monolingual version (Mn-Ad). We analysed the retrieved results obtained by the following queries from Table 4.15 and 4.16 to understand the translation impact.

- CL-AR Query (14) from Table 4.15 : This query had only one translation edit that is the substitution of words "*bible*" to "*Gospel*". As the word "*Gospel*" is an Out-Of-Vocabulary word for the blip10000 collection, this change resulted in the retrieval of many documents that are irrelevant (non religious).

- CL-AR Query (48) from Table 4.15 : This query had two translation edits that are the substitution of words "*digital*" to "*Electronic*", and the reordering of the words "*today*" and "*media*". As most of the relevant documents for this query is about digital media topic, the substitution of words "*digital*" to "*Electronic*" resulted in dropping the rank of many relevant documents (about digital media) in retrieved list for this query.

- CL-FR Query (9) from Table 4.16 : The substitution of "*rally*" to "*recovery*" and the deletion of the word "*stock*" were the translation edits for this query. The deletion of the word "*stock*" had a major effect in as it shifted the query focus away from the initial intention and resulted in missing the retrieval of many relevant documents about the "*stock market*".

- CL-FR Query (13) from Table 4.16 : This query had only one translation edit that which is the deletion of the word "*book*". This deletion resulted in low-

Table 4.15: Examples of queries from the Mn-Ad and CL-Ar sets which have been negatively impacted in the CLIR experiment.

Query id	Mn-Ad	CL-Ar
14	interpretations in the bible	Interpretations in the Gospel
37	creating web sites	Action Websites
46	Economic outlook	The future of the economy
48	todays digital media	Electronic Media today
58	female worker	The realization of women in society

Table 4.16: Examples of queries from Mn-Ad and CL-Fr set which have been negatively impacted in the CLIR experiment.

Query id	Mn-Ad	CL-Fr
9	Stock market rally	Market recovery
15	domestic abuse and violence	Family Violence
13	comic books.	Comics
46	american jobs factories	Jobs US plants
48	college classes and teachers	University courses and professors

ering the rank of many relevant documents in the retrieval list, these relevant documents are specifically discussing and reviewing "*books about comics*".

Overall, Figure 4.13 shows that majority of the queries have lower performance in the CLIR settings, while the rest have either equal or small increase on the performance after translation. This indicates that the word choice and the translation edits conducted by the MT systems to translate the queries can have a significant impact on the retrieval effectiveness for CLIR. In order to deal with these issues in our CL-UGS task, in this work, we utilise query expansion (QE) to add helpful terms to the translated queries to improve the retrieval effectiveness. Furthermore, we also utilise Query Performance Prediction (QPP) to assist the MT system in selecting translations that are predicted to be more effective for our task.

4.6.4 Experimental conclusions

In this section, we conducted preliminary investigation to help us design a retrieval framework for UGS content. The aim of these experiments was to understand how to represent the UGS documents for retrieval, and how translation should be conducted

for CL-UGS using currently available tools and training data.

Our experimental results show superiority for using structured representation with per-field tuning/weighting for this task. We also found that the use of black-box MT using Google API is more effective for query translation for our CL-UGS retrieval task.

Overall, the experiment presented in this section provides guideline on how a monolingual and cross-lingual retrieval framework should be designed in order to carry out an experimental investigation of UGS data. Therefore, for our upcoming investigations, we use these findings where we setup our experiment to use both structured representation with per-field weighting, and a black-box MT to translate the topic sets for CLIR, in particular, we will use the PL2F retrieval model and the CL-Fr and CL-Ar topics which were translated using Google Api. In the next chapter, we take a deeper look into UGS retrieval by studying the retrieval effectiveness of each of the UGS fields and their robustness in CLIR settings.

4.7 Summary

In this chapter, we provided an overview of the component used to conduct the experimental investigation described in this thesis. We described the components of a standard IR text collections, and then outlined the features of the blip1000 UGS collection used for our investigation. This was followed by by describing the details of the MediaEval S&H known item topic sets, and our extension of this topic set to enabling monolingual adhoc search for UGS content. We also explained how we extended these new adhoc topic sets into CLIR topics using MT systems developed by our work to enable CL-UGS retrieval for this task.

Finally, we presented an experimental investigation for designing a retrieval framework for UGS content. Our preliminary experiments indicate that using structured representation is beneficial for UGS retrieval. In terms of query translation for CL-UGS, we found that using black-box MT tool is more effective for this task,

although it still performs significantly lower comparing to monolingual search.

We are now ready to move into our first investigation of this thesis. In the next chapter, we describe our experiments that attempt to answer the first research question of this thesis (RQ1) which aims at understanding the challenges of UGS retrieval by analysing the retrieval effectiveness of each field in UGS content .

Chapter 5

Investigating User-Generated Speech Retrieval

This chapter presents our experimental investigation to address the first research question (RQ1) introduced in Chapter 1 of this thesis. This question is targeted toward understanding the challenges of monolingual and cross-lingual UGS retrieval. To the best of our knowledge, this is the first attempt to formally study monolingual and cross-lingual IR for UGS content. Therefore, before trying to develop or propose any new approach, it is important to understand the main challenges of UGS retrieval task, and how these differ from those posed by other types of spoken content in previous studies of SCR tasks (e.g. Federico and Jones, 2004; Federico et al., 2005; Jones et al., 2007; Pecina et al., 2008; Larson et al., 2010).

5.1 Motivation

The state-of-the-art work on spoken content indicates the superiority of using manual metadata over ASR transcripts for maintaining the retrieval effectiveness (e.g. Jones et al., 2007; Pecina et al., 2007). Interesting follow-up questions to these studies include:

- *Do the findings from previous studies still stand for similar setting over user-*

generated spoken content, where the metadata is also user-generated, highly varied, very subjective and sometimes unreliable?

- *How do UGS fields behave in monolingual and cross-lingual retrieval settings?.*

Our experiments in this chapter investigate search effectiveness over an archive of user-generated Internet video content for both monolingual and cross-lingual settings, with known-item and adhoc search tasks.

To understand the task better, we undertake a detailed performance analysis to examine the impact of different sources of metadata information on search behaviour. The investigation in this chapter is limited to the application of standard Information Retrieval (IR) methods with current MT tools for this task in order to establish the basis for further investigations in the upcoming chapters.

The document and topic sets used in our experiments were all explained in Chapter 4 (Sections 4.3, 4.5.1 and 4.5.2) and are based on the blip10000 archive and topic introduced in the Mediaeval benchmark (Eskevich et al., 2012b). We examine retrieval effectiveness of the *Title* and *Description* metadata provided by the video uploader and ASR transcripts of the content.

We report our experimental investigation in this chapter as follows.

- We analyse the retrieval effectiveness and robustness for each source-of-evidence in our UGS retrieval tasks.

- We examine the effectiveness of combining individual fields and adjusting their weighting to study their interaction.

The remainder of this chapter is structured as follows. Section 5.2 describes initial retrieval experiments the relative retrieval effectiveness of each source of evidence (ASR, Title and Description metadata), Section 5.3 describes our approach to improving retrieval effectiveness for UGS content using careful adjustment of the retrieval algorithm setting, *Section 5.4* concludes the chapter and provides directions for the following chapters.

5.2 Single Field Retrieval

In this section, we examine the behaviour of the separate document information fields in our cross-lingual framework. We are particularly interested here in the impact of translation errors or inconsistencies on retrieval effectiveness given the noise in the ASR transcripts, the shortness of the Title field, and the inconsistencies of the description field. This is examined by evaluating the robustness of each field to measure how the retrieval effectiveness behaves in both monolingual and cross-lingual framework.

Throughout the investigation in this chapter, we define *CLIR Robustness* as how well a field (or source-of-evidence) performs in a cross-lingual setting. We observe this by measuring the significance of the change between the monolingual and cross-lingual performance using the same setting.

5.2.1 Experimental Setup

To run our single-field evaluation experiment, we constructed three indexes for our investigation as follows.

- **ASR_index** which contains only the ASR transcripts of the UGS content.
- **Title_index** contains only the title fields of the UGS content .
- **Desc_index** contains only the description fields of the UGS content .

To conduct our single-field retrieval on each of these indexes, we use the PL2 model with the same unstructured settings [1] and parameters explained in the previously in Section 4.6.1. We used the monolingual adhoc *(Mn-Ad)* and known-item *(Mn-Kn)* queries that were described in Sections 4.5.1, 4.5.2, together with their cross-lingual French and Arabic versions translated using Google MT API[2] *(Cl-Ar)*

[1]Note that although Section 4.6.1 results reported that the structured setting is more effective, the PL2 with unstructured representation is still used since this single-field retrieval and having per-field weighting is not possible.

[2]Google Translation is used since it has shown to outperform the Moses system in Section 4.6.2

Table 5.1: Mono vs. cross-lingual performance per index. Results are reported in terms of MAP except for the known-item queries (MN-Kn) which are reported in terms of MRR.

Field	Mn-Kn	Mn-Ad	Cl-Ar	Cl-Fr
Title_index	0.2827	0.1254	0.0705	0.1022
ASR_index	**0.4513**	**0.5887**	**0.2627**	**0.4618**
Desc_index	0.2432	0.1316	0.0630	0.1117

and *Cl-Fr*. Each query set contains 60 topics.

5.2.2 Experimental Results and Discussion

Retrieval performance for each index across each topic set is shown in Table 5.1. Comparing the performance across the fields, it is clear that the ASR evidence is the most effective and significantly outperforms the two other fields. By contrast, ASR is considered to be the longest modality, and within the UGS content it contains the main content of the uploaded video, while title/description are *optional* shorter fields and contain less information that may not be helpful in addressing the user's information need.

As would be expected, the results in Table 5.1 also show that performance are *lower* in all cases for the cross-lingual task. Thus retrieval effectiveness of all fields is negatively impacted for cross-lingual. This confirms the expected additional retrieval challenge that arises from the imperfect query translation. Performance for the Arabic queries is reduced to a higher degree than for the French. This is most likely due to the relative difficulty of Arabic MT (Alqudsi et al., 2012). One significant challenge for Arabic to English MT relates to named entities. For instance, a query including the word 'dreamweaver' (the proprietary web development tool) was expressed as 'dreamweaver' for both FR and IT, while for AR, it was represented by "الدريموفر" which resulted in it being an OOV term for *Google Translate* and being transliterated into a completely different word 'Aldirimovr' which was not useful for retrieval using the English language metadata.

Further, looking at reduction in MAP for each index indicates they have different

Table 5.2: AR/FR cross-lingual - the t-values according to the % MAP reduction for each index. **Statistically significant values with p-value < 0.05.*

	CL-Ar	Cl-Fr
Title_index	-2.29*	-0.53
ASR_index	**-3.04***	**-2.42***
Desc_index	-2.41*	-0.62

responses to the query translation; notably the impact is greatest on the index of the ASR transcript field across all languages using both known-item and adhoc queries.

To better understand the significance of these cross-lingual reductions in MAP, we computed the statistical significance of each reduction. We calculated the t-value for the difference at the 95% confidence level after representing all monolingual and cross-lingual performance in pairs at every query level. The significance test results in terms of t-values for the indexes searched using the cross-lingual queries, ***(Cl-Ar)*** and ***(Cl-Fr)***, are shown in Table 5.2.

Looking at the t-values, we can observe that French queries are less challenging than the others since the performance was not significantly different from monolingual. Furthermore, Table 5.2 indicates that the ASR transcripts do indeed have the lowest CLIR robustness. On searching the single-field indexes, for both CL-Ar and Cl-Fr queries, ASR_index had the least robustness with a statistically significant negative reduction with ($p<0.05$).

We conclude from this experiment that even if they are incomplete, short and sometimes unreliable, the user-uploaded titles and meta descriptions are more robust in the cross-lingual setting than the ASR fields. As noted earlier, the degree of ASR recognition errors may vary from one video to another with the UGS settings due to the wide variation in the audio quality. The interaction between recognition error rate, document length and retrieval behaviour is highly complex, as observed by (Eskevich and Jones, 2014), in Chapters 6 and 7 we explore this effect in more detail with a view to improving retrieval of the ASR transcript field using both query expansion and text segmentation.

Table 5.3: Mono vs. cross-lingual performance with field pair combinations

Two Fields	MN-Kn	MN-Ad	Cl-Ar	Cl-Fr
TitleDesc_index	0.3020	0.1752	0.0849	0.1490
ASRDesc_index	0.5245	0.5593	0.2503	0.4364
ASRTitle_index	0.4527	0.5601	0.2667	0.4658

5.3 Retrieval with Combined Metadata Fields

Having examined the effectiveness of the three separate fields for monolingual retrieval and cross-lingual, in this section we explore the potential of combining them for improving retrieval effectiveness. For this, we first combine the fields in pairs, and then as it was shown in Figure 4.10, we integrate the three fields but with varied field weighting.

5.3.1 Experimental settings

For all the combined-field experiments, we use the DFR PL2F model (Macdonald et al., 2005) with structured settings and the same tuning hyper-parameters, previously developed in Section 4.6.1.

5.3.2 Experimental Results for Two Field Combinations

Table 5.3 shows retrieval performance for fields combined into pairs which were indexed using the PL2F retrieval model with equal weights. We are interested here in the potential for improved retrieval by using fields in combination. Comparing the results in Table 5.3 and the earlier results shown in Table 5.1, we can see that field combination is more effective for both monolingual and cross-lingual tasks. Further improvement could be potentially obtained by weighting fields differently, which we study in more detail the next section.

Table 5.4: Weighting scheme W_x for the single-weighted retrieval models

	ASR	Title	Desc
PL2ASR	w_x	1	1
PL2Title	1	w_x	1
PL2Desc	1	1	w_x

5.3.3 Experimental Results for Three Field Combinations

In this section we describe our investigation of the retrieval effectiveness with combination of all three fields. To compare the robustness of each field, we investigate giving higher weight W_x to a specific field over the others. We refer to this as the *single-weighted model* where one single field has a higher weight than the other fields.

To set the values for our proposed single-weighted retrieval models we adopted the following steps:

- Construct a model based using the PL2F document scoring that targets a single-field x from each (ASR, title, desc) which we refer to as *PL2FASR, PL2FTitle, PL2FDesc*
- Assign c_x value to each fields to allow length normalisation for the term frequency of each field, we followed the empirically learned parameters for each field explained in Section 4.6.1
- For W_x, we set the w_x value for the targeted field, and fixed the rest at 1, to give priority to field x over the others, as in $W_x = \{w_x,1,1\}$. The reason why we chose set them to be 1 was to allow for the presence of their term frequencies, but with normal (i.e. not boosted) weights.

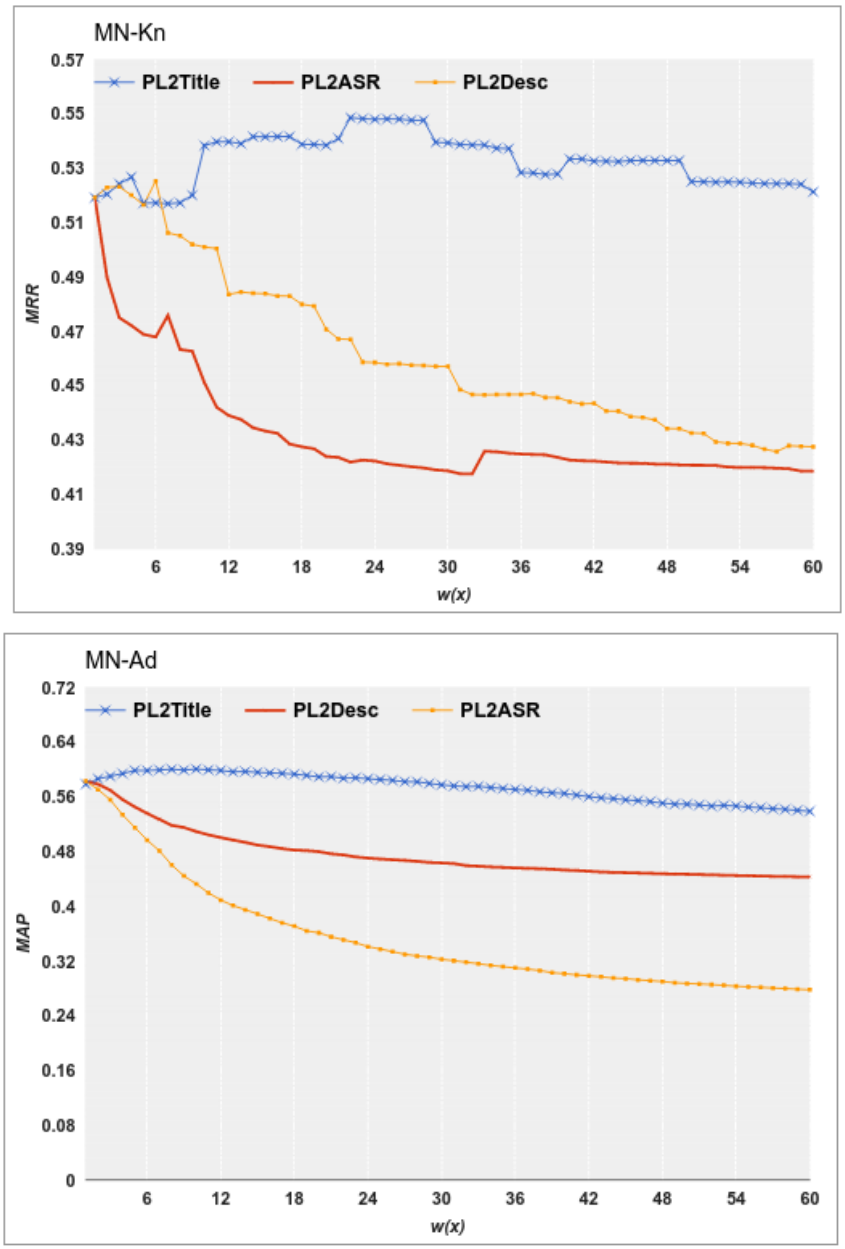

Figure 5.1: Monolingual performance (in terms of MRR and MAP) for the single_weighted models across all weighting points (wx) using both known-item and adhoc topic sets.

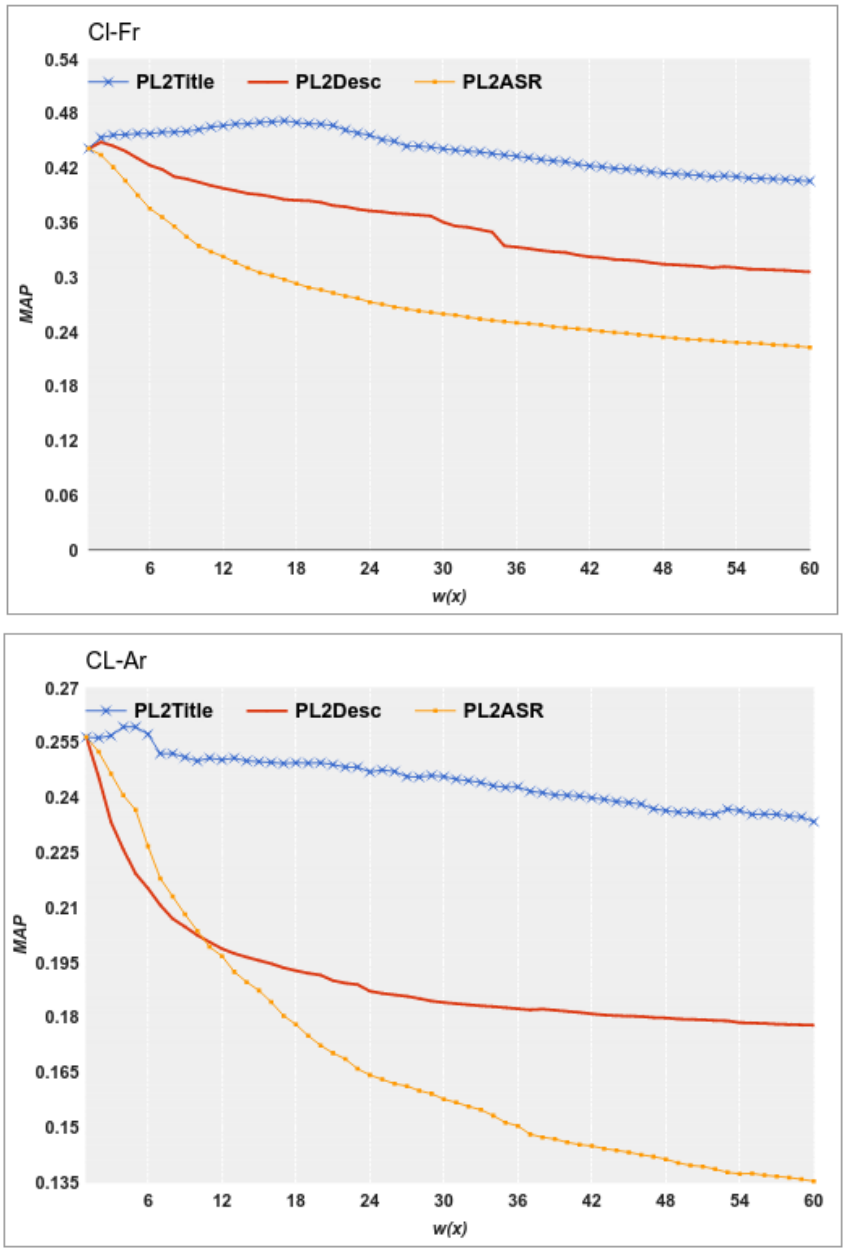

Figure 5.2: cross-lingual performance (in terms of MAP) for the single_weighted models across all weighting points (wx) using both French and Arabic topic sets.

The combination weighting schemes are shown in Table 5.4, in each case one field has a weight boost w_x. To examine retrieval behaviour, we vary the w_x boost parameters for each model in the range 1 to 60 using increments of 1. The first weighting iteration at the weighting point $wx = 1$ is the same for all models where they have $W_x = \{1,1,1\}$.

Experimental Results and Discussion

Figure 5.1 shows the performance at each weighting point for monolingual queries (Mn-Kn and Mn-Ad), while Figure 5.1 shows the performance for the cross-lingual queries (Cl-Ar and Cl-Fr) .

The best cross-lingual/monolingual performance is always achieved by giving a higher weight to the title field for all topic sets. Across all the weighting points and languages pairs, the PL2Title model shows higher performance than other fields for both known-item and adhoc topic sets. [3]

Moreover, it can also be seen from these figures, that we get lower performance when we progressively give higher weights to the ASR and Desc fields. The strong performance of the PL2Title model indicates the stability of title fields for our Internet videos over the other fields. Also, the fact that the titles have been written by the video uploader with more attention than the descriptions could be referred to the following reasons:

- The uploaders thought it is important to have a high quality, meaningful title for their video since it would help in promoting it on the video-sharing site.
- The uploaders believed that it has more importance since it is shown at the header of their video, while the description is generally shown below the video and may not be examined at all by the viewer.

It could also be the case that for the known-item queries, the users who wrote

[3]Note that PL2title performance has more fluctuation. As its measured by the MRR performance on Mn-Kn topics which are the long queries (as explained in Section 4.4.1), increasing the weights on titles can have some major effect on the rank of the single known-item.

the queries have viewed the videos and might have been more likely to include the titles of the videos in their query to find the intended video, because they believed that it would be easier to find them using the title of the video. However, it should be noted that the MTurk task that was used to create the queries for the Search and Hyperlinking MediaEval task did not display the video title while the user was writing the query which was created with the intention of being suitable to re-find the known-item video.

Table 5.5: Mono vs. cross-lingual Recall performance for each field combination.

	Mn-Kn	Mn-Ad	Cl-Ar	Cl-Fr
Title_index	0.5333	0.1533	0.1021	0.1333
ASR_index	0.7833	0.9001	0.5833	0.7667
Desc_index	0.5667	0.2667	0.1214	0.1761
TitleDesc_index	0.7000	0.2921	0.2033	0.2833
ASRDesc_index	0.8033	0.9008	0.5597	0.7567
ASRTitle_index	0.8167	0.9267	0.5933	0.7819
All_Index	0.8667	0.9333	0.6098	0.7833

Comparing the performance for PL2Title with the values shown in Table 5.1, it can be also seen that the performance for PL2Title is almost triple the one obtained by the independent Title field (Title_index). As the w_x increases for the Title field, we can see that there is some further improvement, with the optimal weight depending on the query type/length and the language pair used for cross-lingual (whether it is French-to-English using the Cl-Fr queries or Arabic-to-English using the Cl-Ar).

In an attempt to better understand how the field combination improves retrieval effectiveness. The experiments also examined the recall of the individual fields and the combinations. Table 5.5 shows the recall obtained for each field at 1000 documents results cut-off. It can be seen here that the Title field has the lowest recall in isolation, but that it can boost the Recall of the other fields when used in combination. Results in Figure 5.1 suggest that the title field brings additional evidence without bringing noise, which is not the case for Desc and ASR fields that degrade effectiveness when their weight is increased.

Overall, our experimental investigation shows that for the Title single-field retrieval, the performance is poor due to poor recall. This arises from the fact that the Title and Descriptions of most documents are very short. So in many cases the query does not match well with the relevant documents, but when the query does match with the relevant documents, it does so well. In the case of the ASR single field-retrieval, the fields are much longer, so the chance for relevant items to at least partially match with the query is higher, therefore the recall is higher than for the Title and Description indexes.

5.4 Summary

In this chapter, we examined cross-lingual search over user-generated videos for Arabic-English, French-English for known-item and adhoc search tasks based on the blip10000 collection.

We studied the retrieval effectiveness and challenges of three different sources of information, namely, ASR transcripts, which are challenged by recognition errors, video titles, which can be very short and lack content, and video descriptions, which can be generic and incomplete.

In terms of *retrieval effectiveness*, we found that the ASR is the most important field of the UGS content and contributes to higher recall and precision performance over other metadata fields. This in fact contradicts with conclusion from previous SCR work on non UGS content (e.g. Jones et al., 2007; Pecina et al., 2007), which reported that searching the manual metadata over ASR transcripts is more effective.

In terms of *cross-lingual robustness*, we found that the ASR transcript field has the lowest robustness across other fields and its performance can drop significantly for cross-lingual due to the interaction of translation and transcription errors. We found that Titles are most reliable and robust evidence but it suffers from recall problems due to its shortness.

We then explored field combination to evaluate the retrieval performance of all

fields together, and our investigation show that giving higher weight to the titles over other fields gives improved cross-lingual performance. Our field-combination experiments confirm the low robustness and reliability of the ASR evidence. In particular, we show that tuning the retrieval settings to give a higher weight towards the fields which have a lower cross-lingual robustness such as ASR evidence and metadata description can significantly degrade the retrieval effectiveness.

In terms of CL-UGS, although we used the state-of-art MT tools for conducting the query translation, our empirical experiments demonstrate that the *translations edits* can have a significant negative impact on the retrieval effectiveness for UGS. The ASR transcription errors, UGS noise and translation errors can indeed contribute to major problem of vocabulary mismatch in UGS retrieval. In the next section we explain how plan to address this issue.

5.5 Research Directions for Further Studies

To answer our RQ1 (What are the challenges of UGS retrieval), which aims at studying the retrieval challenges of UGS, and draw the directions to the investigation of the following RQs, we summarise the major challenges from our analysis in this chapter together with our proposed response to address them in the following points.

- Overall retrieval effectiveness and robustness of the ASR transcripts: As shown in our experiments ASR is the most important evidence used to answer most of the queries. We aim to address these ASR retrieval effectiveness using two approaches in the upcoming chapters using :

    1. *In Chapter 6, towards RQ2 (Can QE be beneficial for UGS retrieval) investigation*, we aim to improve the query representation to address vocabulary mismatch between the query and the ASR transcripts, we adopt the query expansion approach to improve the retrieval effectiveness.

    2. *In Chapter 7, 8, towards RQ3 (Can speech segmentation be beneficial for*

UGS retrieval) we aim to improve the document representation to improve the robustness of this evidence. Our goal is to remove unnecessary noise in the transcripts that may harm the retrieval effectiveness. We adopt text segmentation of the ASR transcripts to achieve that.

- High translation quality is required to maintain a reasonable UGS retrieval performance : our experiments showed that query translation edits has a major negative effect on the retrieval effectiveness of UGS retrieval. This issue has shown to be more significant for Arabic language in particular. Our results indicate the high sensitivity of the UGS search to the translation quality. The reason for this is the large amount of noise that is already presents in the UGS data and the interaction with additional noise from the translation errors can harm the retrieval effectiveness to a significant level. *In Chapter 8 and towards our investigation of RQ4 (Can we develop an adaptive CLIR method for UGS retrieval)*, we aim to improve the translation quality by building an MT system and use resources from our UGS task to improve it.

Chapter 6

Field-Based Query Expansion For UGS Retrieval

Following our investigation of retrieval using individual and combined UGS fields in the previous chapter, this chapter is concerned with addressing the second research question (RQ2) introduced in Chapter 1 of this thesis that involves the exploration of the query representation within UGS using QE techniques. QE is utilised in this task to address the vocabulary mismatch between queries and UGS content by enriching the original query using new terms from highly ranked documents.

In Sections 5.2 and 5.3 of the previous chapter, we studied the retrieval effectiveness of individual fields (ASR, Titles and descriptions). Our experimental results indicated that the overall performance is more robust when these fields are combined with trained weights. In this chapter, we seek to utilise these UGS fields in different way; where we modify the query itself using information extracted from these fields for QE, in order to improve the overall effectiveness for UGS retrieval.

6.1 Motivation for QE

The underlying hypothesis behind QE, is that by expanding the queries to include important terms will make it easier to find relevant documents by addressing the

vocabulary mismatch of UGS retrieval. Ideally, the QE should reduce the impact of noise in UGS by adding helpful terms from top relevant documents to improve the overall performance. However, if the new terms are not well correlated with the user information need expressed in the query, application of QE can actually reduce retrieval effectiveness. Applying QE to our task requires special consideration to the noise presented in UGS content to avoid any pitfalls.

Previous research has suggested that QE techniques can be useful for improving both monolingual and CLIR effectiveness for many tasks (Bellaachia and Amor-Tijani, 2008; Carpineto and Romano, 2012; Zhou et al., 2012). However, most of this research have focused primarily on collections of professionally written or formal text which has little amount of noise (Wang and Oard, 2005; Lam-Adesina and Jones, 2006; Singhal and Pereira, 1999). There has been some, but very limited, work on building QE techniques for the noisy data, such as the work on OCR Data (Tong et al., 1996), and more recently, in the user-generated informal text (Lee and Croft, 2014).

In this chapter, we are interested in taking the challenge of applying QE in UGS settings with a focus on avoiding problems that may arise from noise of each field in UGS. In particular to our task, we are concerned with the translation errors of the query, the transcription errors of the ASR, as well as the inconsistency issues of the user-generated metadata.

6.2 Baseline Application of QE in UGS Retrieval

Our initial goal of this chapter is to answer the question of whether standard QE techniques can work effectively within UGS retrieval. Therefore, for the next set of experiments, we explore the effectiveness of QE across the different UGS search tasks we studied in the previous chapter.

We investigate running a state-of-the-art QE approach using all the query sets (Mn-Ad, Cl-Ar, Cl-Fr and Mn-Kn) using the structured representation of all UGS

	MAP	QE
Cl-Ar	0.2536	0.2592
CI-Fr	0.4455	0.4609
Mn-Ad	0.5887	0.5932
	MRR	QE
Mn-Kn	0.5122	0.4462*

Table 6.1: Performance of QE runs for alternative query sets. QE values are the MAP and MRR calculated after the QE is applied, respectively. Numbers which are marked * is statistically significant difference at the $0 > 0.05$ confidence level.

fields combined, with PL2F model, which was explained in Section 4.6.1. For all of our QE experiments, we employ Terrier implementation of the DFR BO1 QE mechanism (Amati, 2003), that was presented in Section 2.2.2, to extract the most relevant terms from the top documents. We first test the default QE using the default parameter settings, where we set it up to extract the 10 most informative terms from the top 3 returned documents . These settings were suggested by Amati (2003) for the DFR QE after conducting multiple experiments on several test collections.

6.2.1 Experimental results and Discussion

Results in Table 6.1 show the effect of using QE for our UGS task. Comparing baseline retrieval to those obtained after the application of QE shows that this application of QE is not helpful for any of these tasks as none of these runs obtained a significant improvement over the original MAP. Furthermore, results in Table 6.1 show that the MRR obtained significantly lower performance that is reduced by 15%.

By contrast, QE seeks to select the most-relevant terms from highly ranked documents of the initial run, unlike text retrieval tasks, many issues with UGS can hinder this process. We summarise these challenges into two major factors as follows.

- For UGS content, it is very common that the original query performance is too low, meaning that the top-ranking documents can be non-relevant, and may result into an errorful extraction of the expansion terms. In other words, a lower performance (i.e low AP) in the initial retrieval run indicates lower

quality of the feedback documents used for expansion, and hence lower QE effectiveness as it is also discussed by He and Ounis (2009). The relationship between AP and QE with more details in Chapter 8.

- Even if the performance for a query of the initial retrieval run is high, the top-ranking documents can be all relevant but may still contain a lot of noise. This noise can be presented in the ASR transcript or metadata fields which includes terms that are non-relevant to the topic of the query. Adding these terms to the query can result in harming the retrieval performance due to a poor choice of expansion terms. The problem of performance drift in QE has long been suggested in the literature (Mitra et al., 1998b; Terra and Warren, 2005) as a major challenge for QE. In the context of DFR QE, the drift issue happens when the top-terms are selected as informative but they are not relevant because they are about different topic. This topic drift issue is made worse in the context of our UGS task where the topic and quality of each video has no specific or consistent theme or style and fields have huge variation of length shown in Table 4.2. For example, where ASR transcripts can be up to 20K length, Desc can be as long as 3K length.

To better understand the QE issues in our UGS retrieval, we run QE with a range of different parameters settings other than the default ones. We explore taking the top 3 and top 5 terms from the top ranking 3, 5 and 10 documents.

This produced a total of 6 different QE runs which we compare to the baseline run that does not involve any expansion (no_QE). Since all query sets show similar conclusions in relation to QE performance in Table 6.1, we limit our investigation to the UGS adhoc retrieval task using the Mn-Ad query sets.

The results shown in Table 6.2 show performance in terms of MAP, Recall and P@10. This results indicate that none of the QE runs obtained significant improvement over the baseline. We tested the statistical significance at $p < 0.05$ for each run by computing the difference at the query level between the precision obtained

Table 6.2: Retrieval performance for QE runs. Docs_Terms represents the QE parameters selected for each run ("Docs" is the number of documents used for expansion, while "Terms" is the number of terms using in the QE).

Docs_Terms	MAP	Recall	P@10
No_QE (baseline)	0.5833	0.9002	0.5450
3_3	0.6109	0.9167	0.5517
3_5	0.5753	0.9167	0.5433
3_10	0.5801	0.8833	0.5433
5_3	0.5753	0.9167	0.5183
5_5	0.5763	0.9167	0.5183
5_10	0.5677	0.8833	0.5133
10_3	0.5932	0.9333	0.5400

by baseline (no_QE) and that obtained using QE, and none of these improvements was statistically significant.

Furthermore, it can be seen from Table 6.2 that QE performance is better when using lower numbers of terms and documents, but the improvement achieved is not significant. The reason for this limitation in QE for UGS could be attributed to the fact *that taking less documents for the expansion reduces the chance of introducing irrelevant terms that may appear in documents ranking lower than the top of the list.*

We also carried out further QE runs with additional *top term* parameters of 2,7 and 10 terms, and also extended the *top document* parameters to be explored to the range 2 to 25 with an increment of 2. This, produced *60 different QE runs in total* (in addition to the ones reported in Table 6.2) to analyse the performance of this approach. Figure 6.1 shows the performance achieved using these different QE runs.

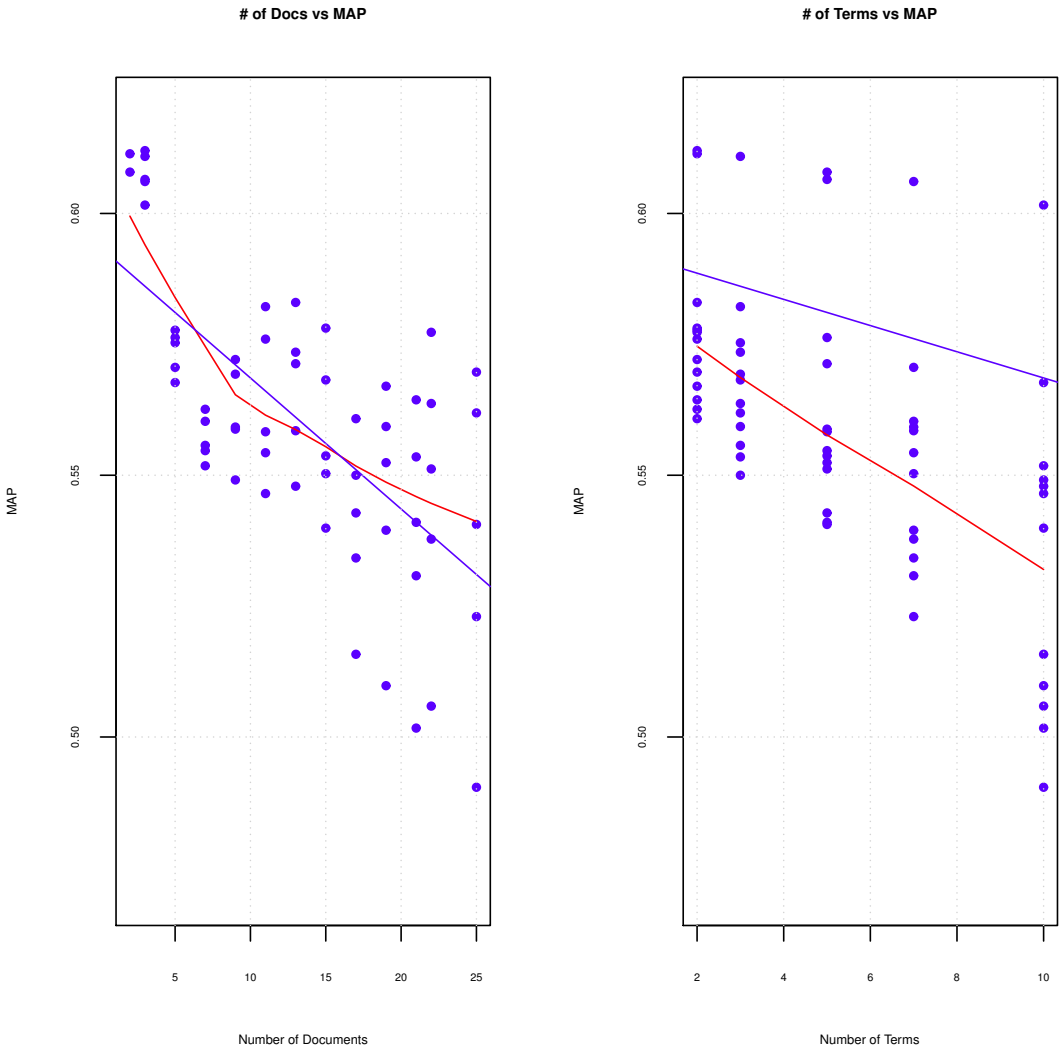

Figure 6.1: MAP performance (blue dots) for alternative term and document parameter values for QE using full ASR evidence. The relationship between the number of documents/terms vs the MAP performance is demonstrated using the linear regression fit line (blue curve), and LOWESS (Locally Weighted Scatter-plot Smoothing) local regression fit line (red curve).

Overall, Figure 6.1 suggests that the chance of adding poor expansion terms after QE increases when more documents and more terms are added. The noise associated within these UGS documents provides a major limitation for QE effectiveness when selecting new terms from retrieved relevant documents.

The following example illustrates an example of query drift issue for the query *"Software Development and Web Design "*, which is looking for a guide on the topic of software development and web design. This query had an initial AP performance of 0.38. The terms (Internet, show, Web, program, tv, develop) were generated from running the DFR Bo1 QE Model (see Section 2.2.2 for details on how this model works) to take the top 5 terms from the top 3 documents. The two expansion terms *Web* and *develop* also appear in the original query, therefore, the weight for each of these terms is boosted to be greater than 1. The new expansion terms (internet, show, program, tv) are added to the original query and their weights are adjusted based on their informativeness. The final expanded and reweighted query is outlined as follows. *Software ×1.000, Development ×1.916, Web × 1.201, Design ×1.000, Internet × 0.332, show × 0.837, program × 0.683, tv × 0.842*

The expanded query with the new terms achieved an AP of 0.21, that is 44% less than the performance of the original query. As can be seen, this new query expanded the focus of the information need from looking to find videos about software development/web design to include internet tv shows and programs (as recognised in the ASR transcripts). Therefore, the scores of a non-relevant documents about this topic have been boosted. As result of this, the new query has resulted in a loss of focus and performance drift and impacted the retrieval effectiveness of the original query.

In general, this experiment shows that applying QE for UGS task requires a further analysis and understanding of the expansion sources, and the quality of the feedback documents (expansion documents) to deal the noise associated with it. Since UGS fields have different relative importance for retrieval as empirically observed in Sections 5.2 and 5.3, we hypothesis that selecting fields which are relevant

can improve the QE effectiveness. In the next section, we introduce our proposed field-based QE which seeks to better understand and analyse the utility of each source for QE.

6.3 QE using Fields for UGS Retrieval

A QE framework using fields was proposed for monolingual text retrieval in (He and Ounis, 2007). This framework proposed an extended term weighting scheme for QE that utilised the fields of the documents to achieve higher performance than traditional document level QE.

In this investigation, we adopt this approach and further adjust it to a single-field QE technique that allows us to assess the effectiveness of each field for QE. In this method, QE is performed as follows.

- Initial retrieval is done similar to our setting for the experiment in Section 5.3 using the structured field combination. We use also the PL2F model (described in Section 5.3) for this retrieval experiment.
- The top-n terms are extracted using the DFR QE model from a certain field combination of the top-m documents, and added to the initial query. We set these parameters to be the default parameters with 10 documents and 3 terms.
- Retrieval is done again using the new query to produce the final ranking of the retrieved results.

To understand the effectiveness of using each field for QE, we conduct several QE runs taking in consideration each field and their combinations. We carried out our field-based QE runs as follows.

- *exp-ASR*: Queries expanded using the ASR field only.
- *exp-Title*: Queries expanded using the Title field only.
- *exp-Desc*: Queries expanded using the Desc field only.

Table 6.3: Retrieval performance for QE runs using different fields combination

	MAP	Recall	P@10
No_QE	0.5887	0.9000	0.545
exp-TitleASR	0.5911	0.9333	0.5350
exp-All	**0.5932**	0.9333	**0.5400**
exp-TitleDesc	0.5787	0.9167	0.5267
exp-ASRMeta	0.5923	0.9333	0.5400
exp-ASR	0.5911	0.9333	0.5350
exp-Title	0.5891	0.9167	0.5350
exp-Meta	0.5701	0.9167	0.5200

- *exp-ASRTitle*: Queries expanded using the ASR and Title fields only.
- *exp-TitleDesc*: Queries expanded using the Title and description fields only.
- *exp-ASRDesc*: Queries expanded expanded using the Desc and ASR field only.
- *exp-All* : Queries expanded using all fields combined together.

6.3.1 Experimental results and Discussion

The field-based QE experiments are designed to explore the retrieval performance of each field in our UGS data set. The retrieval performance for each QE run is shown in Table 6.3. These results suggest the performance of the proposed single field runs (i.e exp-ASR, exp-Title, exp-Desc) do not improve *significantly* over the baseline run (No_QE). Even when field are combined using (exp-ASRMeta, exp-All, etc...), only small improvement is gained over the baseline.

The QE effectiveness is hindered by the fact that we have multiple sources of noise coming from non-relevant fields that appears in the highly ranked documents. This noise can cause a change in the query focus and results harms the initial retrieval performance. To better understand the robustness for each QE run and compare it to the baseline, we study the difference of AP (ΔAP) at each query level for all the proposed run over the baseline run (No_QE), where the ΔAP on a particular query level for a QE run (exp-x) is calculated through ΔAP = *(AP(exp-x) - AP(No_QE))*.

The ΔAP results for each field-based QE runs is shown in Figure 6.2, while Figure 6.3 shows the ΔAP for the exp-AllQE run. Since the DFR QE model uses the informativeness measure to weight the extracted top terms (see Equation 2.10, and Equation 2.11), the decreases and increases of the ΔAP values that are shown in Figure 6.2 and Figure 6.3 can be explained as follows.

- $\Delta AP = 0$, means that the extracted terms were identified as less informative/important, which is the reason why the QE has a minimal effect over the baseline (No_QE). Based on the DFR definition of informativeness (see Equations 2.10,2.11), this indicates that the terms added to this particular query were given *lower weight* because they are not only common in the top-n documents but also in the whole document collection.

- $\Delta AP > 0$ indicates that the extracted terms were identified as highly informative and they were *relevant* to the query, which is the reason why QE shows an increase over the baseline. This in turn suggests that many runs were able to improve over multiple and different queries which indicates that these fields have varying effectiveness for QE.

- $\Delta AP < 0$ means that the extracted terms were identified as highly informative but they were *not relevant* to the query, which is the reason why QE had a negative effect over the baseline run. The ΔAP values show that this incorrect prediction has significantly impacted all runs. Results in Figure 6.2 and Figure 6.3 show that the decrease in performance is common to all QE runs including exp-AllQE which combines all fields.

Overall, the ΔAP numbers from different QE runs suggest that these fields have varying effectiveness and robustness for QE. The results suggest that a careful selection of the evidences and sources for expansion can potentially improve the effectiveness of QE for this task. To test this hypothesis, in the next section, we run an additional experiment using the best field combination for each query.

	MAP	*Recall*	*MAP@10*
no_QE	0.5759	0.9300	0.5267
exp-AllQE	0.5932	0.9333	0.5400
exp-optimal	0.6354*+	0.9731*+	0.5791*+

Table 6.4: Retrieval performance for the optimal QE run. * indicates a statistically significant improvement over the baseline (no_QE). While + indicates a statistically significant improvement over(ex-AllQE)

6.4 Selecting the Best Fields for QE

In this section, we seek to confirm whether careful selection of the fields can actually improve the effectiveness of QE for this UGS retrieval task. We develop an optimal QE, which we refer to as *exp-optimal*, that selects the best combination for QE. The exp-optimal QE is implemented as follows.

- For each query, we run QE using all possible field combinations (exp-TitleASR, exp-All, exp-ASRMeta, exp-ASR, exp-Title, exp-Meta).
- The human relevant-judgement information (groundtruth data) is used to calculate the QE (AP) for running each possible field combination.
- The optimal QE run that has the maximum QE (AP) is selected for running the expansion.

Performance results for exp-optimal QE run are shown in Table 6.4. Figure 6.4 also shows the ΔAP for each query for the exp-optimal QE run. The results in Table 6.4 confirm that QE can be much more effective and indeed provide significant improvement over both baseline and traditional QE runs.

By contrast, comparing Figure 6.4 to Figure 6.3 and Figure 6.2, it can be seen that using the optimal field combination can improve the overall robustness of QE and avoid many instances of negative ΔAP. However, Figure 6.4 shows that even in the case of using the best evidence for QE, performance can be negatively impacted after QE.

As explained before, performance drift in QE can be attributed to the terms introduced from the non-relevant parts of the feedback documents used for expan-

sion. The results in Figure 6.4 suggest that even in the case of removing non-relevant fields, performance drift can happen due to the non-relevant content that appears in the UGS fields themselves. In the next chapter, we explore how to extract relevant segments from the ASR field to address this issue.

6.5 Summary and direction towards the upcoming chapters

In this chapter, we investigated the application of QE for UGS retrieval. To the best of our knowledge, our work is the first to provide an analysis of QE performance on such a real-world task of UGS retrieval.

The work presented in this chapter differs from prior QE SCR investigations by using an internet-based UGS collection; where audio data is highly variable in many aspects, including the audio conditions of the recording, the microphones used, the fluency and informality of the language used by the speaker.

The aim of this chapter has been to answer our research question RQ2 :

1. How do traditional QE work under such a setting of noisy data collected from Internet videos?

2. Can we have an effective QE approach that adaptively utilises these data sources to expand individual queries in order to improve overall retrieval effectiveness?

As for RQ2.1 : The results in Section 6.2 indicate that applying traditional QE such as the one studied (Carpineto and Romano, 2012; Wang and Oard, 2005; Terol et al., 2005; Lam-Adesina and Jones, 2006) in UGS retrieval is not effective due to the noise and structure complexity of the UGS data. We found that UGS sources can have a varying reliability for QE, and even when they are combined together, the retrieval effectiveness can still be negatively impacted due the poor expansion term selection which shifts intended focus of the initial query. This issue arises due to

the non-relevant content present in the expansion documents. To handle this issue, in Section 6.3 we proposed a new field-based QE that only selects certain fields of the expansion documents for QE, and showed how the overall performance can be improved using this approach. However, our analysis suggests that field-based QE can be improved by selecting different field combinations for each query.

To verify this assumption, and work our way towards answering the RQ.2.2, in Section 6.4, we investigated developing an adaptive QE that utilises different field combination for each query and selects the one that maximises the performance after QE. Our results suggest that the performance can be *significantly improved* by picking the right source of evidence for expanding each query. However, our result shows that even when using the optimal field combination for QE, poor term selection can still occur due to the noise and non-relevant content that is presented in the fields themselves. We found that even when fields such as ASR transcripts are relevant, they can have non-relevant topics that present noise into the QE term selection process and harm the retrieval effectiveness.

In an attempt to address this issue, in the next chapter, we investigate adjusting the representation UGS document using text segmentation in order to find relevant content that can be utilised for QE in UGS.

Furthermore, RQ.2.2 (Can we have an effective QE approach that adaptively utilises UGS data sources to expand individual queries in order to improve overall retrieval effectiveness?), is not answered yet as the optimal adaptive approach proposed in Section 6.3 is artificial since it is not possible to have the ground-truth data to evaluate the performance of each query. Instead, in Chapter 8, we show how Query Performance Prediction (QPP) method can be utilised in adaptive QE to select the content that maximise QE effectiveness for particular query .

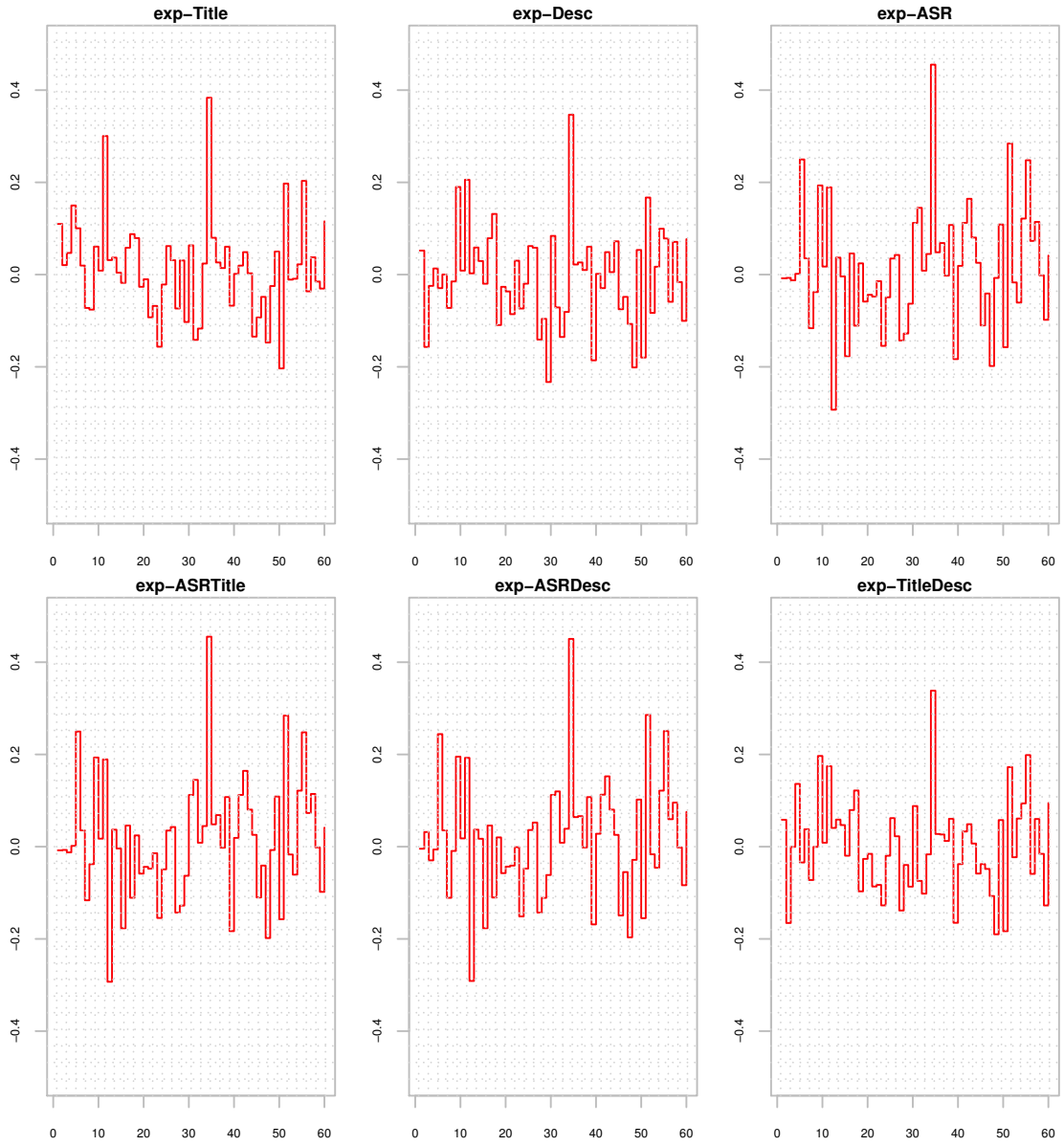

Figure 6.2: Obtained ΔAP per each query for all QE runs. (Numbers (1-60) on the x-axis represent the Query IDs).

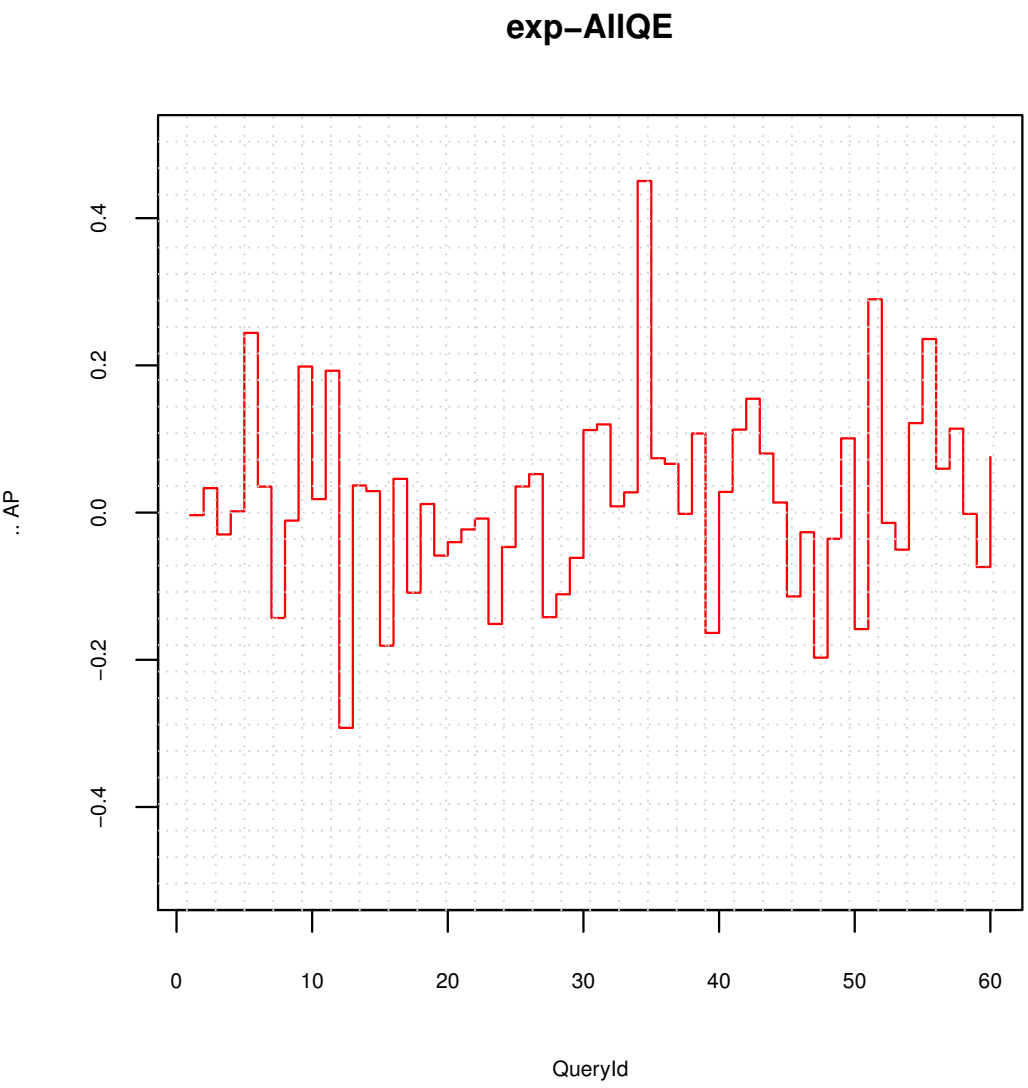

Figure 6.3: Obtained ΔAP per each query for the exp-AllQE.

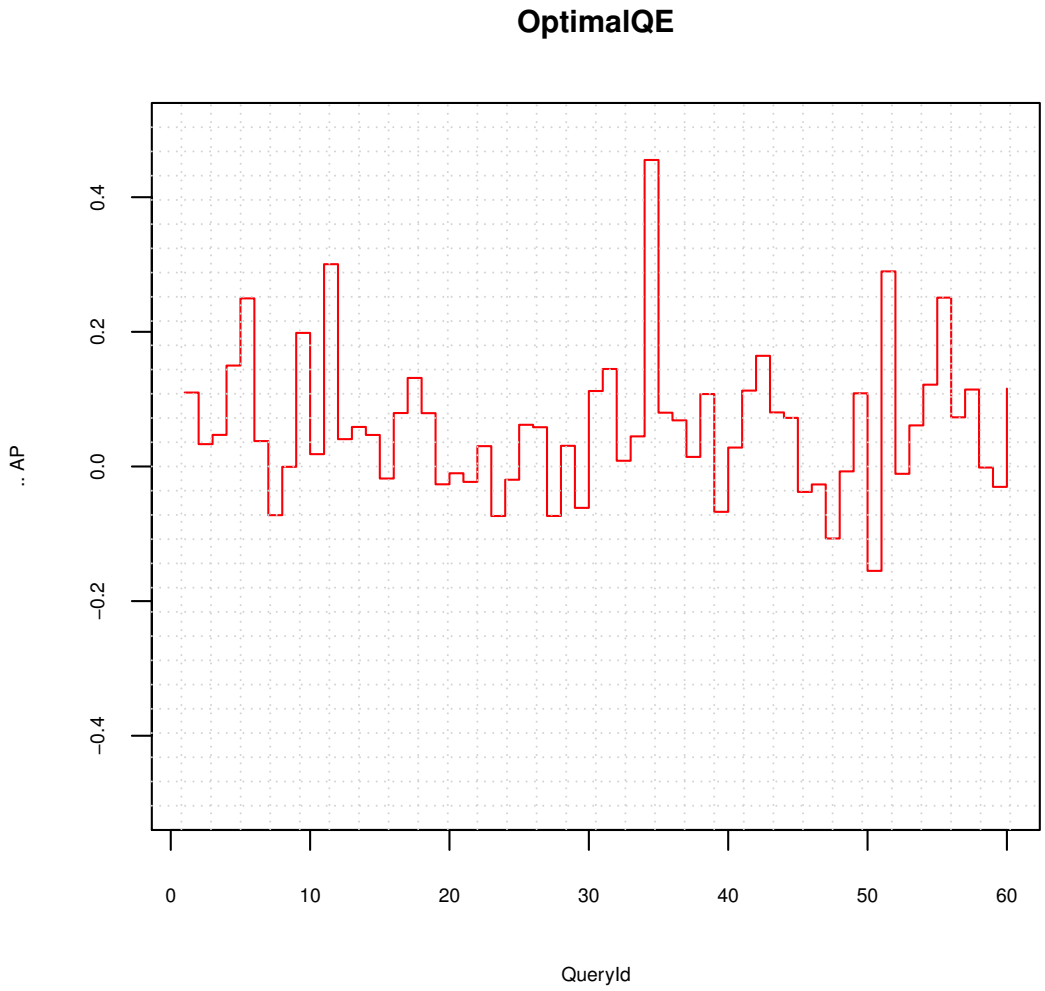

Figure 6.4: Obtained ΔAP per each query for the exp-optimal.

Chapter 7

Segment-Based Query Expansion for UGS Retrieval

The preceding chapters explored RQ1 and RQ2 of this thesis, namely towards understanding the challenges of cross-lingual/monolingual retrieval and QE for UGS content. The aim of this chapter is to address the RQ3 of this thesis as follows.

1. Can automatic speech segmentation be beneficial in improving QE effectiveness for UGS content?

2. What are the characteristics of the most effective speech evidence (segment or document-level) for QE in UGS retrieval?

3. Can we develop a technique to predict the most effective speech segmentation's for each query?

In Chapter 6, we showed that the use of field combination proves effective for improving QE in UGS. However, our experimental analysis revealed that the term selection process in QE can still be impacted to the non-relevant terms presented in the expansion documents. In this chapter, we attempt to address QE issues by utilising speech segmentation. The motivation for our experimental investigation in this chapter is outlined in the next section.

7.1 Motivation

Our empirical analysis in Chapter 5 indicated that a fundamental challenge for UGS retrieval is the vocabulary mismatch between user queries and relevant UGS documents. This mismatch problem, as reported by previous research (Mitra et al., 1998a; Sanderson, 1994), often occurs when queries are vague, short or imperfect or when relevant documents have a complex topical structure.

QE is a popular technique used to bridge this vocabulary gap between the query and its relevant documents (Wang and Oard, 2005; Terol et al., 2005; Lam-Adesina and Jones, 2006; Carpineto and Romano, 2012). Our experimental investigation in the previous chapter reported that QE issues can result in drifting away from the intended focus of the initial query and impact the overall retrieval effectiveness. This drift results from the extraction of poorly chosen expansion terms, where these terms often belong to a topic different to that of the original query . We observed that documents in our social UGS collection have a greatly varying lengths and topical structures [1] that pose a major challenge with respect to QE issues. We reported that that a careful selection of expansion evidence is required to maintain a robust QE for UGS retrieval.

In an attempt to address to improve QE effectiveness, previous work in text retrieval suggested that in some cases using passages or subtopics of the whole document is more relevant than using the whole document for QE (Wilkinson, 1994), and that incorporating *segment (also called* passage*) evidence* can be helpful in avoiding topic drift issues (Mitra et al., 1998b; Xu and Croft, 1996; Allan, 1995; Callan, 1994).

In the context of SCR, several techniques have been explored to address QE issues. For example, prior work on the CLEF collections utilised manually created summaries or segmented spoken content to improve the effectiveness of QE (Lam-

[1]Topic drift is common and major problem within our QE task since, as explained in Section 4.3, documents in our test collection are based on videos which were uploaded to the social video sharing site by 2,237 different uploaders and covering a 25 different topics with varying recording quality and differing lengths.

Adesina and Jones, 2006; Wang and Oard, 2005). These summaries and segments were created *manually by professional indexers* and provided by the task organisers (Pecina et al., 2007).

Unfortunately, having these manually generated summaries or segments within large-scale UGS content is very unlikely due to the cost required to create them. Instead, we propose to apply automatic text segmentation techniques to the ASR transcripts, such as the ones studied in (Wartena, 2012) for passage retrieval, and investigate their effectiveness for QE in SCR for UGS. In the next sections we explain how the ASR transcripts were segmented and prepared for our segment-based QE task.

7.2 Segmenting Speech Transcripts for QE

Previous research on passage retrieval (Callan, 1994; Eskevich, 2014) grouped text segmentation into three classes as follows.

- *Discourse* segmentation based on textual units such as sentences, paragraphs and sections.
- *Semantic* segmentation based on the content and the topic of the text itself (e.g. the C99 technique (Choi, 2000)).
- *Window-based* methods where text is segmented based on a number of words or textual units.

These approaches to text segmentation have been studied extensively within the task of *passage retrieval* which is concerned with retrieval of portions of documents that are relevant to a user query, and thus preventing, or at least reducing the amount of non-relevant material presented to the user (Eskevich, 2014; Eskevich et al., 2012a; Wartena, 2012).

However, our work is *the first* to investigate their utility for QE in adhoc and cross-lingual UGS retrieval. We investigate the application of QE based on segments

in user-generated spoken content retrieval; where we separate speech transcripts into shorter units. This seeks to reduce the impact of noise that may arise from ASR errors, and irrelevant subtopics appearing the top ranking documents. Our hypothesis is that *incorporating speech segments will allow us to detect the relevant parts of a document and prevent QE from selecting harmful expansion terms from non-relevant content.*

We explain how we utilise each of semantic, discourse and window-based segmentation for our QE task in the following sections.

7.2.1 Semantic segmentation

This type of segmentation is based on the lexical cohesion within the ASR transcript. The most well-known topic segmentation algorithms are: TextTiling (Hearst, 1997) and C99 (Choi, 2000). TextTiling, developed by Hearst (Hearst, 1997), is an unsupervised linear topic segmentation algorithm that uses cosine similarity to estimate similarity between blocks of words and assumes that similar words belong to the same topic. The calculation is accomplished using two vectors containing the number of terms occurring in each block. C99 was introduced by Choi (Choi, 2000), and uses a matrix-based ranking and a clustering approach, and assumes the most similar words belong to the same topic.

We used the standard UIMA[2] Text Segmentor[3] for the implementations of the segmentation algorithms TextTiling and C99 which were originally written by Choi (2000).

Similar to the work of (Eskevich et al., 2012a; Wartena, 2012), the punctuation inserted by the ASR system was used as the sentence boundaries for segmentation. Since we do not have the ground truth segments of this dataset, it was not possible to fully optimise the hyper-parameters for these algorithms. Instead we took a sample ASR document with a length of 752 words and manually evaluated the quality of

[2]http://uima.apache.org/UIMA
[3]https://code.google.com/p/uima-text-segmenter/

the segments based on 4 different variations of parameters (including the default ones that are suggested by Choi (Choi, 2000) in his implementation), and picked the one that produced the best segments in terms of detecting manually identified topic boundaries.

7.2.2 Discourse segmentation

This segmentation is based on dividing the ASR transcripts into consecutive silence bounded utterances from the same speaker (Aly et al., 2011). We hypothesise that silence points can be useful for detecting topic boundaries, where each point is considered a segment and can be used as QE evidence. We used those created by the ASR system based on detecting silence, where a new segment is produced whenever a silence point is detected.

7.2.3 Window-based segmentation

This type of segmentation has long been suggested to be the most effective for multiple tasks (Allan, 1995; Mitra et al., 1998a; Xu and Croft, 1996) in text retrieval and spoken passage retrieval (Wartena, 2012; Eskevich et al., 2012a). Using this technique, in our UGS retrieval task, ASR transcripts were segmented in a fixed-length sequence (window) of words For the window-based segmentation. We segmented the transcripts into 50, 100 and 500 words fixed lengths.

We also employ window-based approaches studied in (Liu and Croft, 2002) such as the *Half-overlapped fixed-length passages* where the first window starts at the first term in a document, and subsequent windows start at the middle of the previous one.

Stop word removal was done for all segmentation techniques based on the standard Terrier list[4], and stemming performed using the Terrier implementation of Porter stemming [5]. In the next section we show how we setup our segment-based

[4]http://terrier.org/docs/v2.2.1/javadoc/uk/ac/gla/terrier/terms/Stopwords.html
[5]http://terrier.org/docs/v4.0/javadoc/org/terrier/terms/PorterStemmer.html

QE for this task.

7.2.4 Setting up the Speech Segments for QE

For our segment-based QE investigation, we built different retrieval indexes for the ASR transcripts as follows.

- *ASR* indexes each ASR transcript as one document.
- *fix50* indexes each 50 words length window, *fix100* indexes each 100 words length window, *fix500* indexes each 500 words length window.
- *over50, over100, over500* indexes are for the fixed length half-overlapping window segments with length of 50, 100 and 500 words.
- The *C99* index considers the segments created using C99 as documents, the *TextTile* index uses the TextTile segments and *SP* index which uses the speaker segments of the collection transcripts.

Field metadata (Titles and Descriptions) were added for each document/segment with no modification.

The statistics for each of these segment indexes are shown in Table 7.1. Comparing these segment indexes to the ASR index shown in Table 4.2, it can be seen that all of them produce larger indexes with less length variation. The SP, TextTile, over50 indexes appear to be the largest with an average of 100 segments per ASR document, but also produced the shortest average segment length. Fixed-length segments are obviously much more consistent in length.

7.3 Segment-Based QE for UGS Retrieval

Previous work studied multiple approaches for the utilisation of segmentation in QE for text-based IR tasks (Allan, 1995; Xu and Croft, 1996; Mitra et al., 1998a). However, no such study has been reported on UGS content. Furthermore, prior work

Table 7.1: Statistics for segment indexes: number of indexed documents (docs), average segment length (Avg.len), standard deviation of document length (St.len) and average number of generated segments per document (Segs-doc)

Segments	docs	Avg.len	St.len	Segs-doc
SP	902,209	22.83	43.8	102.5
C99	57,104	400.3	405.5	6.3
TextTile	795,770	28.2	25.1	105
fix50	457,347	49.5	4.0	47.6
fix100	231,244	98.1	11.1	24.1
fix500	50,508	463.7	105.8	5.3
over50	909,514	49.52	3.9	94.6
over100	457,051	98.11	11.0	47.5
over500	94,294	464.1	103.8	9.8

has been limited only to the application of window-based and discourse-segments types for QE.

The most common to utilise passages in QE is using the Local Context Analysis (LCA) technique. Proposed by Xu and Croft (1996), LCA works by retrieving the top-ranking window-passages from the collection and use them for QE. An improved version of the LCA approach, proposed by Liu and Croft (2002), considers re-ranking the documents based on their *best-scoring segment* from *each document* in the collection for QE.

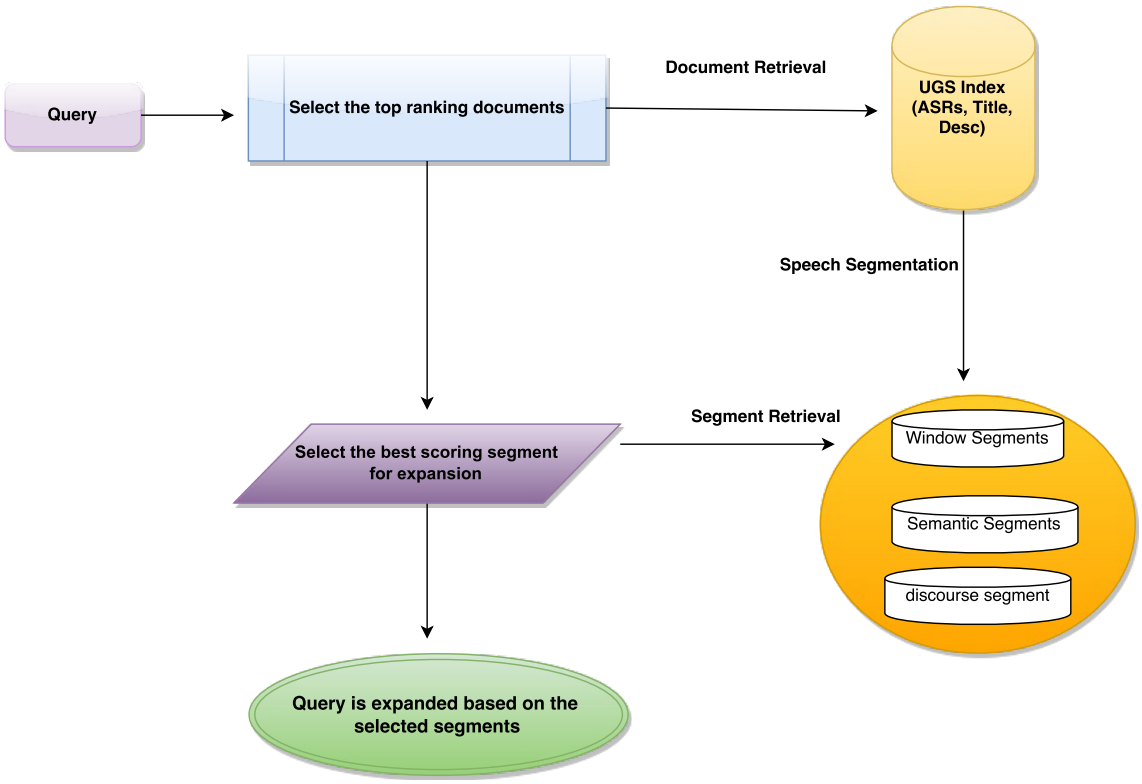

Figure 7.1: Overview of the proposed segment-based QE.

To implement our segment-based QE, we follow a similar approach to that proposed by Liu and Croft (2002), where the *best-scoring segment* from each document is utilised to rank the documents, and then the top ranked segments are used for QE. Figure 7.1 presents an overview of the segment-based QE. The proposed QE algorithm is implemented as follows.

1. After initial retrieval, retrieved documents are re-ranked based on the score of their highest scoring segment.

2. The best-scoring segments for each of the top-ranked documents are then used for expansion, where the top ranking segments are taken from segment indexes shown in Table 7.1.

3. DFR QE Bo1 model is used to extract the top terms form the set of top-scoring segments.

We analysed the performance of different Segment-based QE based on C99, SP,

Table 7.2: Performance for using QE with optimal parameters (Docs_Terms) for full-document evidence (ASR), and each of the studied segmentation schemes (SP, C99, Fix50, fix100, fix500, over50)

QE run	Docs_Terms	MAP	Recall	P@10
ASR	3-3	0.6109	0.9167	0.5517
SP	9-2	0.6043	0.9167	0.5600
C99	5-5	**0.6151**	0.9167	0.5850
fix50	25-10	**0.6140**	0.9167	0.5633
fix100	25-5	**0.6167**	0.9200	0.5633
fix500	25-3	**0.6134**	0.9000	0.5650
over50	25-2	0.6070	0.9167	0.5617
over100	21-10	**0.6158**	0.9000	0.5700
over500	17-5	0.6087	0.9000	0.5617
TextTile	22-2	0.6098	0.8833	0.5517

TextTile, over50, over100, over500, fix50, fix100 and fix500 segmentation of the ASR transcript.

Parameter settings for these QE runs were tuned similarly to those described in Section 6.2, where we generated 60 different runs using each of the segmentation schemes.

Figure 7.2 shows the accumulative retrieval performance for each QE type calculated by summing the MAP performance obtained for the 60 different QE runs generated through all tuning parameters explained in Section 6.2.

From Figure 7.2, it can be seen that all segment-based QE runs were less affected by poor expansion term selection, even when more terms and documents were included. This demonstrates the effectiveness of the segment-based QE approach in terms of sensitivity to the associated noise in the top-ranking documents compared to the full-ASR QE approach (which is the first box to the left in Figure7.2).

The *effectiveness* of these segment-based QE approaches can be attributed to the fact that segments are more likely to avoid unnecessary noise in the feedback documents, as well as allowing relevant terms to be selected from those segments that appear within long ASR transcripts that may contain multiple non-relevant topics.

Table 7.2 shows the *best* retrieval performance obtained for each QE run. Al-

though they are generated automatically and contains less information, it can be seen that most segment-based QE runs achieved similar performance to that obtained by the full ASR.

Some segmentation methods (highlighted in **bold**) such as based on C99 and fixed-length segmentation achieved sigificantly better performance than the using the ASR evidence. While other segmentation types produced lower performance such as TextTile and SP.

The effectiveness of these approaches over others can be attributed to their ability to more reliably detect the topic boundaries in this noisy collections and dealing with the document lengths, and topic inconsistency for this task. Better detection of the topic boundaries provides a better ranking of the segments used as feedback evidence for QE, and hence improved the retrieval performance.

By contrast, comparing results obtained for each QE in Table 7.2 to the statistics of each segmentation in Table 7.1, it can be seen that for some QE runs which rely on *higher* average number of segments per document (Segs-doc)), obtained *lower* MAP values. TextTile and discourse segmentation techniques which tend to produce more segments and detect all possible segments (i.e. the Segs-doc value is more than 100 in Table 7.1), obtained lower retrieval performance than the rest. This results show that tuning segmentation algorithms to have a larger number of discovered segments over precision (quality of produced segments) may produce many unsatisfactory segments that can rather harm the ranking of good relevant segments, and hence negatively impact the robustness of the segment-based QE. However, the previous conclusion is not true for the overlapping segmentation, in which the over100 produced better segmentation than the over50 and over500, so the combination of both precision and recall is needed for an effective segmentation.

To summarise, our result in Figure 7.2 show that all segment-based QE were more *robust* in dealing with poor expansion term selection than the full-ASR QE approach. However, the results in Table 7.2 indicate that the relationship between

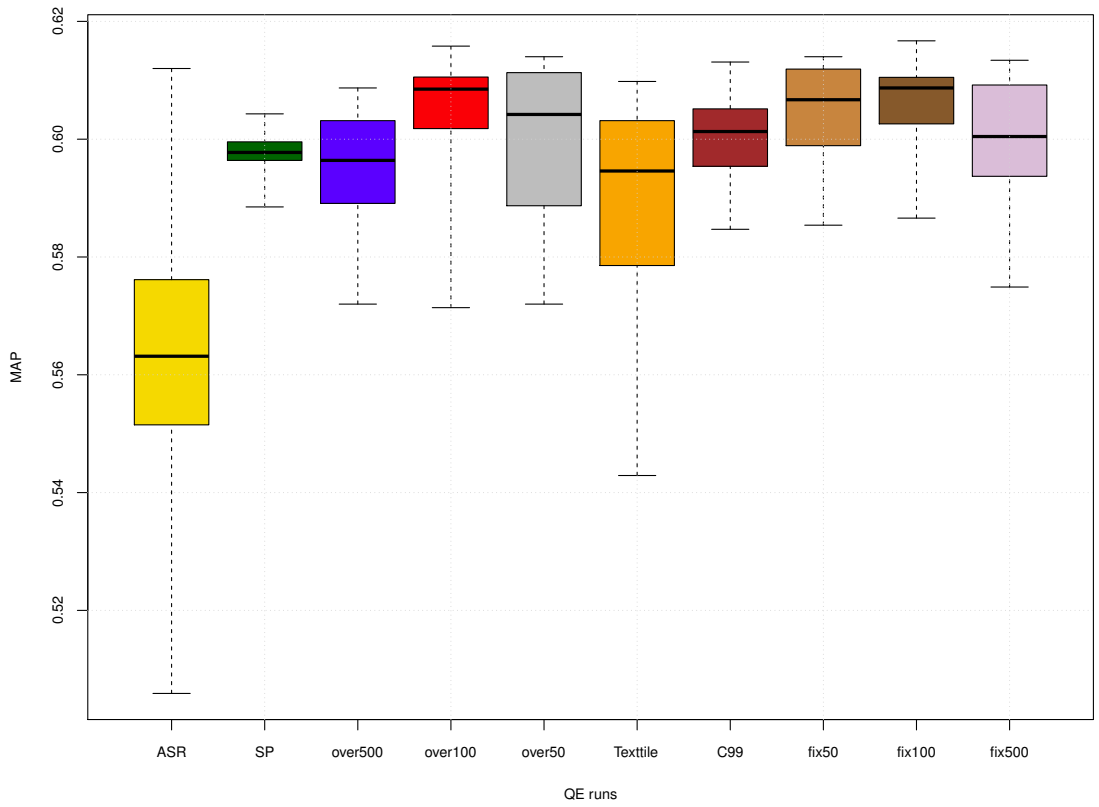

Figure 7.2: Accumulative retrieval performance for each QE type (including full and segment evidence) calculated by summing the obtained MAP performance from the 60 different QE runs generated through all tuning parameters explained in Section 6.2.

the length of the feedback documents, the type of speech segmentation used, and the obtained QE effectiveness is not straightforward and rather more complicated to interpret. This can be due to multiple factors which we discuss more in detai in the next section in studying which segmentation is better for QE.

7.4 Which Segmentation scheme is better for QE in UGS Retrieval

Our investigation so far has analysed the first question of RQ3 which is RQ3.1 (Can automatic speech segmentation be beneficial in improving QE effectiveness for UGS content?); our experiments in the previous section showed that using segmentation

for QE can improve the term selection robustness in UGS retrieval. The next question is RQ3.2 where we seek to understand the characteristics of the most effective speech evidence for QE in UGS retrieval. When we sought to compare different segmentation in the previous section, our experiments showed that no segmentation or speech evidence is consistently the best for QE in UGS data.

In this section we aim to answer RQ3.2 of this thesis, reproduced as follows. *What are the characteristics of the most effective speech evidence for UGS retrieval?*

In order to identify which evidence is more effective for QE, in this section, we provide a further analysis of the performance of each segmentation scheme for QE.

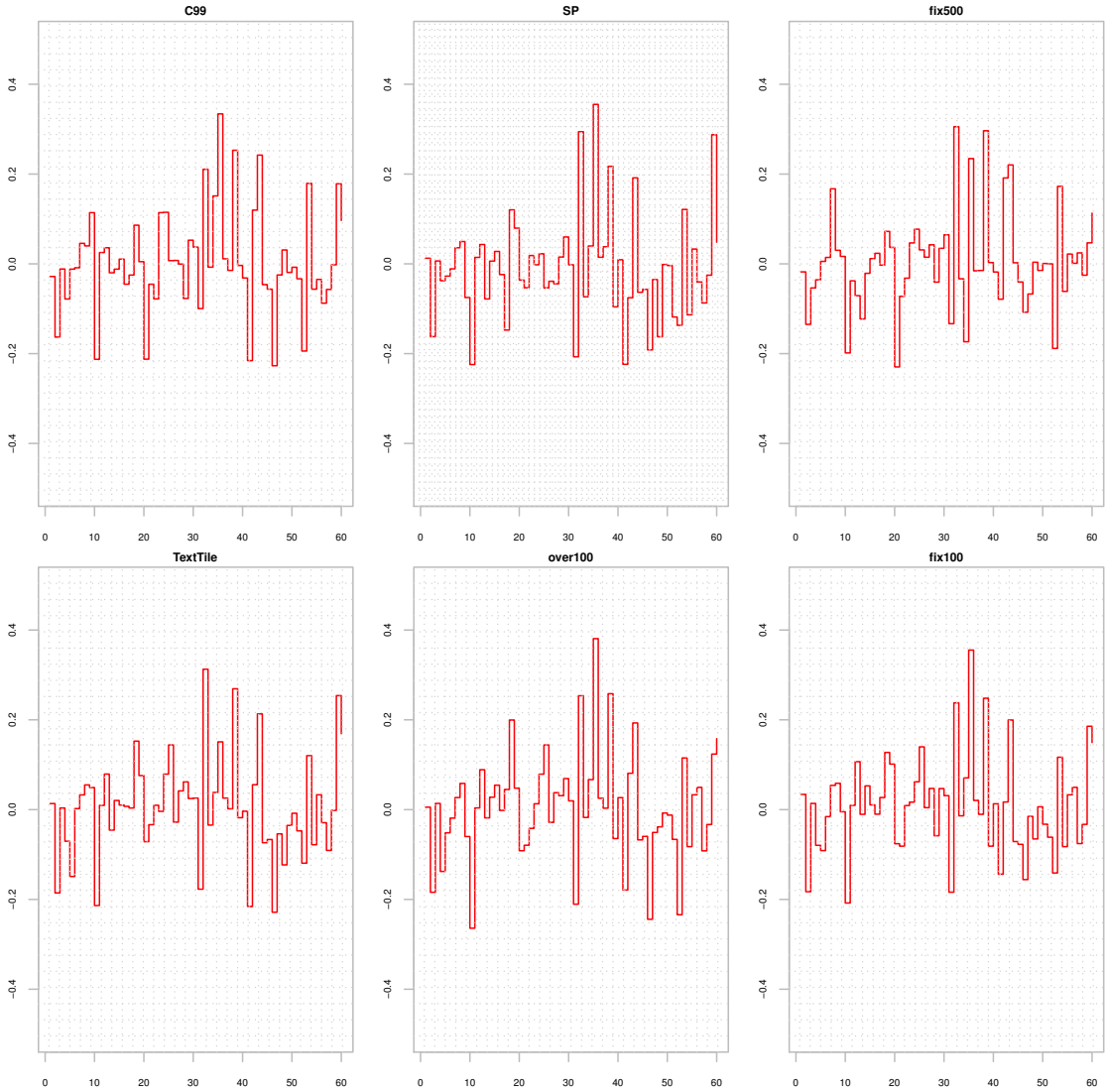

Figure 7.3: ΔAP between full document and segment-based QE for every query

To analyse the effectiveness of the different QE runs shown in Table 7.2, we show the changes in AP value (ΔAP) for each segment-based QE run relative to the baseline run for each query in Figure 7.3.

The results in Figure 7.3 indicate that each QE run has different performance for each query; some have increased performance over the baseline where segment-based QE was more effective, while others show decreased performance over the baseline where document-based QE was better.

The ineffectiveness of segment-based QE can be attributed to the fact that the segmentation process is sometimes ineffective where poor detection of topic boundaries can also harm the rank of good documents and lead to a reduced score of possibly good terms for QE. It can also be the case that segmentation is not actually needed for some queries, as discussed in previous work by (Gu and Luo, 2004), since some of the relevant documents only contain a single-topic and *thus using full-ASR document evidence is often better for expansion.*

To study the relationship between QE performance and the length of the evidence used for QE, we provide an analysis between the average length of feedback documents and the obtained in QE. We calculated this for the QE experiments in Section 6.2 that uses the full-ASR documents as feedback. Figure 7.4 shows the effect of documents-length on the performance of QE for the 60 different runs. As can be seen, some QE runs obtained better ΔMAP even though they are too long. This indicates that there is no evidence which length is more suitable length for QE.

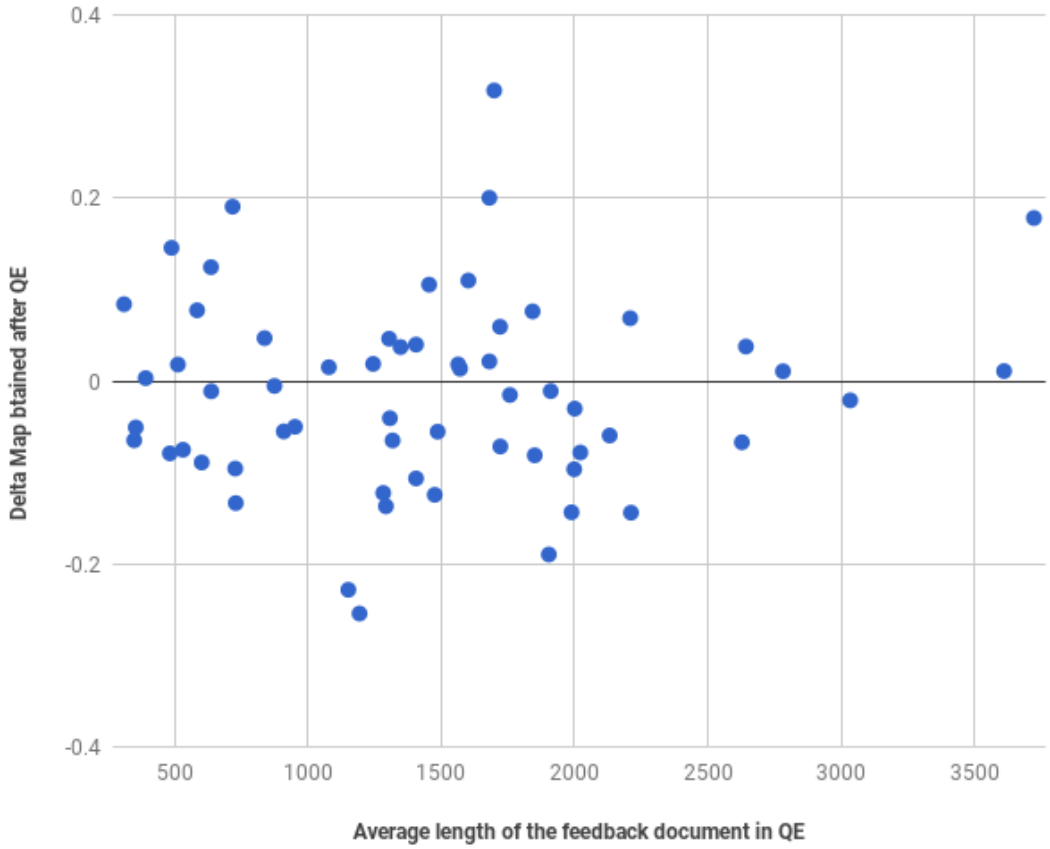

Figure 7.4: Relationship between the average length of feedback document used for QE and the obtained ΔMAP after QE (represented by the blue dots).

Overall, Figure 7.4 shows that no correlation between both the length of feedback document used for expansion and the obtained better ΔMAP after expansion. As previously explained, the ΔMAP depends on the structure of relevant document that can be useful for QE, which varies from query to another. Some useful documents are single topic and long, in which segmentation is not needed, while other documents are short but have a multi-topic and segmentation is needed.

Our answer to **RQ3.2**, is that no particular segmentation scheme is always effective for all queries. This suggests that a hybrid approach that combines both full-document and segment evidences can be more effective for UGS retrieval.

From the analysis of Figure 7.3, we hypothesise that if we can predict which QE scheme (whether on document or segment-level) to apply for an individual query,

then we could automatically select the one which will give the best performance and improve the overall effectiveness. In the next chapter, we present our approach to test this hypothesis, and to answer **RQ3.3** of this thesis as follows " *Can we develop a technique to predict the most effective speech segmentation for each query?*".

7.5 Summary and Research Direction for the Upcoming Chapters

This chapter has examined the effectiveness of segment-based QE methods for UGS retrieval. Our experiments show that the segment-based approach improves reliability for QE and shows lower sensitivity to QE parameters (number of documents and terms used for the expansion). However, our investigation shows that the segment-based approach can also harm the effectiveness of the QE for the following drawbacks.

- Since it is an automatic process, segmentation can be ineffective in detecting the useful topic boundaries for QE in UGS retrieval.
- For some queries, where relevant and helpful documents have a single-topic structure, topic segmentation may not be needed at all.

Overall, our investigation revealed that none of the evaluated segment-based or document-based QE approaches is always better for all queries in UGS retrieval. Our experiments indicate that in the UGS settings, the variation in the length, quality and structure of the relevant documents can harm the effectiveness of both techniques across different queries.

Further, the results show that the overall performance can benefit from an adaptive hybrid approach that combine both the segment and document evidence for QE. In the next chapter , we attempt to utilise Query Performance Prediction (QPP) approaches to develop an adaptive QE approach that predicts, for each query, which

evidence is more effective for this query and selects the one with objective of maximising the retrieval effectiveness.

Chapter 8

Query Performance Prediction for Query Expansion

The previous two chapters studied the effectiveness of Query Expansion (QE) for User-Generated Speech (UGS) retrieval as follows. *Chapter 6* investigated the performance of QE approaches based on different fields combination, whilst *Chapter 7* studied the effectiveness of using segmentation for QE. Our experimental efforts reveal that for the automatically generated transcripts of the highly variable UGS content, there is no ideal use of evidences or field combination that can be robust and always effective for QE in UGS retrieval.

Our investigation of QE in UGS retrieval reported the need for an adaptive QE technique that can improve the overall performance by predicting which setting or evidence is more suitable for QE for individual queries. In this chapter, we investigate the development of a prediction framework to select the right setting and evidence to be used in QE for UGS retrieval.

8.1 Motivation

Recent advances in Query Performance Prediction (QPP) methods provide effective techniques to predict the retrieval performance in the absence of user relevance

judgement information. As previously explained in Chapter 2 (Section 2.3), previously proposed QPP approaches focused on estimating the effectiveness of a search performed in response to a query. In this chapter, we describe our proposed prediction framework that is designed for the prediction of QE performance. This chapter also demonstrates how we develop different adaptive QE approaches that utilise this prediction framework to improve the effectiveness of UGS retrieval.

Our investigation from the previous chapters showed that performance of QE depends on the suitability of the feedback documents or segments in terms of their relevancey to the information need, and the robustness of the QE model in identifying useful expansion terms in these documents. Our experiments showed that feedback documents in UGS retrieval can be wholly or partially non-relevant, and may lead to selecting terms for QE that are not well-correlated with relevance to the information need (Mitra et al., 1998a).

Furthermore, in the previous chapter, we demonstrated how these issues can be potentially mitigated by selecting the right evidence for QE. In this chapter, we explore the use of QPP methods to select the most effective evidence for QE.

The experiments presented in chapter are designed towards answering the last part of the proposed RQ3 (RQ3.3), "*Can we develop a technique to predict the most effective speech segmentation's for each query?*" . Therefore, our experiments in this chapter are designed to examine the prediction of the QE outcome using QPP methods, such as the ones reported in (Shtok et al., 2012; Hauff, 2010; Zhou and Croft, 2007), and utilise them to develop an adaptive QE method for UGS.

8.2 Query Performance Prediction

In this section, we give an overview of existing pre-retrieval and post-retrieval QPP approaches from the literature. A more extensive review of QPP approaches can be found in (Carmel and Yom-Tov, 2010; Hauff, 2010).

8.2.1 Pre-retrieval QPP

Pre-retrieval QPP methods rely on measuring the statistics/characteristics of the query terms calculated over the index collection. Pre-retrieval methods can be implemented at indexing time, and are more efficient than the post-retrieval causing less overhead to the retrieval system since no retrieval is required in the prediction process.

Several pre-retrieval methods have been studied and experimented by the literature, the next section explains the most representative pre-retrieval methods.

IDF-based Methods

The most widely used QPP methods, which have proven to be very effective, rely on the Inverse Document Frequency (IDF) of query terms, are referred to as IDF-based QPP (Cronen-Townsend et al., 2002). The IDF value for a term is usually calculated using the INQUERY formula (Allan et al., 1995; He and Ounis, 2006), as shown in Equation 8.1, where Nt is the number of documents that contain the query term t, and N is the number of documents in the whole collection.

$$IDF(t) = \frac{\frac{log2(N+0.5)}{Nt}}{log2(N+1)} \tag{8.1}$$

IDF-based QPP approaches are implemented by taking an aggregation of the idf values across the query terms as follows.

- ***AvIDF*** : takes the average of the IDF values obtained by the query terms as shown in Equation **??**, where ql is query length that is the number of terms in the query.
- ***SUMIDF*** : takes the resultant sum of the IDF values of the query terms as shown in Equation 8.3.
- ***MAXIDF*** : takes the maximum obtained IDF value from the query terms as shown in Equation 8.4.

$$AvIDF = \frac{1}{ql} \sum_{t}^{Q} IDF(t) \tag{8.2}$$

$$SUMIDF = \sum_{t}^{Q} IDF(t) \tag{8.3}$$

$$MAXIDF = \max_{t}^{Q}[IDF(t)] \tag{8.4}$$

Another common IDF-based QPP is the averaged inverse collection term frequency (avICTF) of the query terms (Plachouras et al., 2004; He and Ounis, 2006). The AvICTF is defined as shown in Equation 8.5, where Q is the query which has set of terms t, $token_{coll}$ is the number of tokens in the whole collection, ql is the query length, and $F(t)$ is the term frequency of t across the whole collection.

$$AvICTF = \frac{log_2 \prod_Q (\frac{token_{coll}}{F(t)})}{ql} \tag{8.5}$$

Linguistics Methods

This type of QPP is based on analysing the linguistics features of the query terms such as the the query length (AvQL) which is based on the average number of content words (non stop-words) in a query. AvQL is shown in 8.6, where nt_c is the number of content terms t_c in the query Q (He and Ounis, 2004; Mothe and Tanguy, 2005; He and Ounis, 2006).

$$AvQL = nt_c/ql \tag{8.6}$$

In addition to AvQL, Mothe and Tanguy (2005) studied the effectiveness of multiple query-based linguistic features as follows.

- Morphological features : such as the average number of morphemes per query, average number of suffixed tokens and average number of proper nouns and acronyms.

- Syntactic features : such as the average number of conjunctions, prepositions and personal pronouns as detected by the POS tagger.

Mothe and Tanguy (2005) reported that the only effective feature that showed some positive correlation with the retrieval performance is the number of nouns in the query.

Statistical Methods

Statistical QPP methods are basical on extracting statistics from the query and document collection to infer the retrieval performance.

Query Scope (QS) is popular effective Statistical-based QPP which makes use of the document frequencies (DF) of the terms (He and Ounis, 2004, 2006). In QS, a higher DF of the query terms indicates that they are very common, probably not helpful for finding relevant documents and result into a lower effectiveness of the query. The query scope calculated as shown in the Equation 8.7, where n_q is the number of documents containing at least one of the query terms, and N is the number of documents in the whole collection.

$$QS(q) = -log(n_q/N) \tag{8.7}$$

More advanced pre-retrieval techniques were proposed by Zhao et al. (2008), called the Summed Collection Query similarity (SCQ). SCQ approaches utilise both document frequency and IDF of the query term to predict the query performance. The SumSCQ score is defined as shown in Equation 8.8, where N is the total number of documents in the collection, F is the frequency of query term t in the collection, and df is document frequency of the query term t.

$$SumSCQ = \sum_{t \epsilon Q} (1 + ln(F) \times ln(1 + \frac{N}{df}) \tag{8.8}$$

Similar to the IDF-based QPP, in addition to the SumSQC, there are also two different aggregation methods of the SQC across the query terms as follows.

- AvSQC, takes the average across of all resultant similarities for each query term as shown in Equation 8.9 where ql is the query length.

- MaxSQC, takes the Maximum value among all resultant similarities for each query term in Equation 8.10.

$$AvSQC = \frac{1}{ql}\sum_{t \epsilon Q}(1 + ln(F) \times ln(1 + \frac{N}{df}) \tag{8.9}$$

$$MaxSQC = max[t \forall Q(1 + ln(F) \times ln(1 + \frac{N}{df})] \tag{8.10}$$

Zhao et al. (2008) also proposed another type of QPP that is computationally more expensive than all previously explained approaches called VarTFIDF. VarTFIDF approaches rely on the distribution of the TF.IDF weights(Zobel and Moffat, 2006) of the query terms across the documents. The TF.IDF score of each query term for each document (t_d) is calculated as follows.

$$w_{t,d} = 1 + ln(tf_d) \times ln(1 + \frac{N}{df}) \tag{8.11}$$

Based on the recommendation of Zhao et al. (2008), TF.IDF weights for query terms that are not present in a document are assigned zero values. Similar to the SQC and IDF QPP approaches, this also has three versions of aggregations: SUM, MAX and Avg as follows.

- SUM.VarTFIDF :takes the sum of the query term TFIDF weight deviations. For example the SUM for VarTF.IDF for a query q is defined as showing in Equation 8.12, where the $\overline{w_t}$ is defined as showing in Equation 8.15, where Dt is the set of documents that contain query term t.

- Avg.VarTFIDF : takes the average of the query term TFIDF weight deviations as shown in Equation 8.13.

- MAX.VarTFIDF : takes the maximum deviation obtained across all query term TFIDF weight deviations as shown in Equation 8.14.

$$SUM.VarTFIDF(q) = \sum_{t \epsilon Q} \sqrt{\frac{1}{df} \sum_{d \epsilon D} w(t) - \overline{w_t}} \tag{8.12}$$

$$Avg.VarTFIDF(q) = \frac{1}{ql} \sum_{t \epsilon Q} \sqrt{\frac{1}{df} \sum_{d \epsilon D} w(t) - \overline{w_t}} \tag{8.13}$$

$$Max.VarTFIDF(q) = max[t \forall Q \sqrt{\frac{1}{df} \sum_{d \epsilon D} w(t) - \overline{w_t}}] \tag{8.14}$$

$$\overline{w_t} = \frac{\sum_{t \epsilon q} w_{t,d}}{D_t} \tag{8.15}$$

8.2.2 Post-Retrieval QPP

The idea behind post-retrieval QPP is to utilise the retrieved results to estimate the retrieval performance. In comparison with to the pre-retrieval methods, post-retrieval methods have shown to be more effective and provide higher accuracy of prediction (Hauff, 2010; Kurland et al., 2012). However, this higher accuracy comes at cost of efficiency as post-retrieval methods require an actual retrieval run to be performed in order to make the prediction.

In this section, we introduce the key methods for post-retrieval QPP as follows.

Query Clarity Score

Cronen-Townsend et al. (2002) proposed the *query Clarity Score* QPP. This approach is based on measuring the clarity of the retrieved results with respect to the collection or the corpus. The basic idea of query clarity is that a query language model constructed from the retrieved results should be different to that constructed from that constructed from the corpus as whole. The divergence between the language models of both the retrieved results and that of corpus has been evaluated in multiple forms within the clarity framework. In their work, Cronen-Townsend et al.

(2002), initially used the KL divergence method to estimate the degree of query ambiguity.

Amati and Van Rijsbergen (2002) used the information gained of the retrieved results within the DFR framework (InfoDFR) to estimate the query difficulty. InfoDFR is measured by capturing the divergence between the frequency of query terms in the retrieved result and that in the collection. The idea of InfoDFR showed a significant correlation with query performance but did not show any correlation between this predictor and the effectiveness of QE. A detailed list of the clarity prediction forms is given in (Hauff, 2010).

Robustness-based QPP

Another family of post-retrieval QPP methods is based on evaluating the *Robustness* of the initial retrieved results. *Robustness* QPP indicates that the more coherent the results are, the better performance is assumed to be.

Robustness-based QPP typically measures the degree of query perturbations in the retrieved results. Query perturbation is measured using the overlap between the list of documents retrieved in response to the entire query, and the documents retrieved in response to each query term (Yom-Tov et al., 2005; Vinay et al., 2006; Zhou and Croft, 2006; Cronen-Townsend et al., 2006).

Query Feedback (QF) prediction, proposed by Zhou (2008), has been reported to be the most effective robustness-based approach. To predict the retrieval performance, QF compares the results list of the original query to the result list of an artificial query that is generated from the original list. Higher similarity between the two lists indicates a better retrieval performance.

QPP using Score Analysis

The state-of-the-art post-retrieval QPP techniques use information induced from analysing the retrieval scores $Score(d)$ of the results set ${D_q}^{[res]}$ produced by retrieval method M. ${D_q}^{[res]}$ represents the scores obtained by the retrieved list of documents

d for a query q..

In probabilistic terms, the resultant score $Score(d)$ of a document d represents the estimated relevance probability r of a document d with respect to q $Score(d) \equiv \mathcal{P}(d|q,r)$.

Score-based QPP methods (Zhou and Croft, 2007; Shtok et al., 2012), are based on analysing the performance of the top k ranked documents The followings are the two main score-based QPP.

- **Weighted Information Gain (WIG)** (Zhou and Croft, 2007), is a well-established and effective QPP technique based on the weighted entropy of the top k ranked documents. WIG works by comparing the scores of the top-k documents $\forall d \epsilon {D_q}^{[k]} Score(d)$ to that obtained by the corpus $Score(D)$. Score (D) is the average score of all the result list as shown in Equation 8.16.

  WIG is often combined with a query length normalisation $\frac{1}{\sqrt{|q|}}$ to make the scores comparable over different queries, defined in equation (8.17).

$$Score(D) = \frac{\sum score(d)}{{D_q}^{[res]}} \tag{8.16}$$

$$WIG(q,M) = \frac{1}{k} \sum_{d \epsilon D_k} \frac{1}{\sqrt{|q|}} (Score(d) - Score(D)) \tag{8.17}$$

- **Normalised Query Commitment (NQC)** :Another similar post-retrieval QPP technique that has also shown to be very effective for many search tasks is the *Normalised Query Commitment (NQC)* (Shtok et al., 2012). NQC is based on estimating the potential amount of query drift in the list of top k documents by measuring the *standard deviation* of their retrieval scores. A high standard deviation indicates reduced topic drift and hence improved

retrieval defined in equation (8.18) where $\overline{\mu} = \frac{1}{k}\sum_{d \epsilon D_q{}^{[k]}} Score(d)$.

$$NQC(q, M) = \frac{1}{Score(D)} \sqrt{\sum_{d \epsilon D_q{}^{[k]}} \frac{1}{k}(Score(d) - Score(\overline{\mu})} \tag{8.18}$$

Using the standard deviation as an approach to QPP was also studied by Pérez-Iglesias and Araujo (2010), and Carmel and Yom-Tov (2010). The authors of these papers suggested that the standard deviation does not need to be computed for all results, instead a cut-off points of for set of results is sufficient for making an effective prediction .

More recent work by Cummins (2012) utilised Monte Carlo simulations to prove the effectiveness of using the standard deviation for QPP. Both WIG and NQC are tuned to have a strong linear relationship with the AP of the query in which the only variable that needs to be decided is the top-k documents that are required for the comparison/prediction.

In the next section we show we utilise these QPP methods in QE for our UGS retrieval task.

8.3 Probabilistic Prediction Framework for QE

As previously explained, the goal of QPP methods in IR to predict the retrieval performance an IR system. Such prediction can be used to tune the retrieval settings to maximise the overall system effectiveness. QPP enables estimation of IR effectiveness for the current query. None of the previously reported work in QPP has focused on prediction of QE performance.

In this section, we propose a prediction framework to estimate the effectiveness of QE denoted as follows.

For a query q, and $D_q{}^{[prf]}$ is the list of pseudo feedback documents D_{prf} in response to q, S_{prf} is the list of pseudo feedback documents (segments) in response to q. Assuming that $\mathcal{Q}$ is the event of being an effective feedback list for expansion,

the goal of the prediction task is to estimate $\mathcal{P}(D_q^{[prf]}|q, \mathcal{Q})$, which seeks to answer the following question:

What is the probability $\mathcal{P}(.)$ ***that this feedback list*** (prf) ***being effective*** $\mathcal{Q}$ ***to expand this query*** (q)***?***

Our proposed framework seeks to identify the best feedback list[1] $\overline{prf}$ for each query q from prf_{list} where $prf_{list} \in \{D_{prf}, S_{prf}\}$ by selecting the one that gives the highest probability as shown in equation (8.19).

$$\overline{prf} = \underset{prf \in prf_{lists}}{\operatorname{argmax}} \mathcal{P}(prf|q, \mathcal{Q}) \tag{8.19}$$

The next section explains how this probability function is defined for our QE task.

8.3.1 The Weighted Expansion Gain Approach

We rely on the previously proposed QPP methods to devise the probability function $\mathcal{P}(prf|q, \mathcal{Q})$ that estimates the effectiveness of QE.

For example, both WIG and NQC, which were explained in Section 8.2.2, are designed to have a strong linear relationship with the AP of the query in which the only variable that needs to be decided is the top-k documents.

For our QE task, we introduce a new element to the prediction, the number of document prf which influence the effectiveness of QE. This introduces a modified version of the WIG predictor, which we refer to as the *Weighted Expansion Gain (WEG)*, for predicting the QE performance.

WEG assumes that the top-k documents $D_q^{[k]}$ for each query are composed of two subsets $D_q^{[prf]}$ and $D_q^{[nprf]}$ defined as follows.

- $D_q^{[prf]}$ is the set of prf feedback documents that are assumed relevant and

[1]The best feedback list is defined as the list of feedback documents that are most robust and effective for QE.

effective for expansion $\forall d \epsilon D_q^{[res]}$ ($d_{\mathcal{R}i}$ where $0\mathcal{R}iprf < k$).

- $D_q^{[nprf]}$ is the set of documents that are assumed non-relevant and ineffective for expansion.

Figure 8.1 shows an example of how a retrieved list is analysed into three features for prediction in WEG, namely, top-k documents, $D_q^{[prf]}$ and $D_q^{[nprf]}$. Note that both $D_q^{[prf]}$ and $D_q^{[nprf]}$ lists are ranked among the top-k documents and right after the prf documents ($prf < nprf < k$).

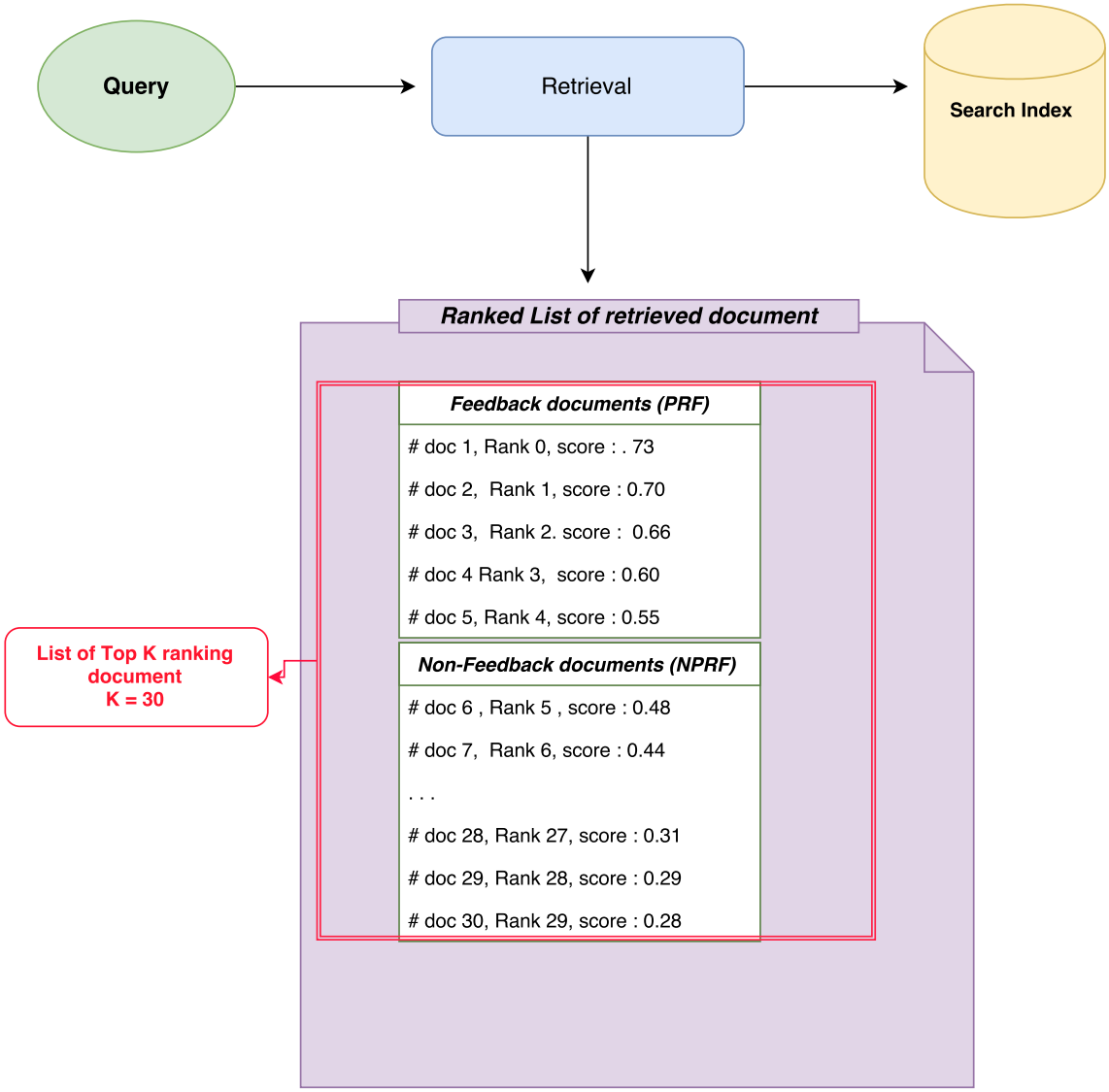

Figure 8.1: Example of the features used by the proposed WEG for the prediction of QE, where top-k documents = 30, prf= 5 and $nprf$ = 25.

Although the goal is different[2], WEG relies on a hypothesis that is closely related to that proposed in (Shtok et al., 2012) by assuming that topic drift is caused by **misleading** documents that are highly ranked in the results because they are *similar but not relevant* to the query. The WEG predictor aims to analyse the quality of the prf documents by measuring the likelihood that they have any misleading documents in the list that would otherwise harm the QE performance. This is estimated by measuring the weighted entropy of the prf documents against the top-ranked yet non-pseudo $nprf$ set of documents.

Unlike WIG, which uses the *centroid* of general non-relevant document $Score(D)$, the proposed WEG uses the *centroid* of the $nprf$ documents scores as showing in Equation 8.20.

$$Cnprf \equiv Cent(Score({D_q}^{[nprf]})) \equiv \frac{1}{nprf} \sum_{d \epsilon {D_q}^{[nprf]}} Score(d) \tag{8.20}$$

$Cnprf$ is used as a reference point for estimating of the effectiveness of the prf documents in QE, as shown in Equation 8.21.

$$WEG(q, D_{prf}) = \frac{1}{prf} \sum_{d \epsilon D_{prf}} \frac{1}{\sqrt{|q|}} (Score(d) - Cnprf) \tag{8.21}$$

The proposed WEG predictor requires three parameters to be tuned as follows.

- prf : the number of feedback document used for the expansion in QE.
- *top-k* : similar to WIG and NQC (explained in Section 8.2.2), WEG also requires the top-k, which is the set of top ranking documents used for prediction.

  The number of $nprf$ documents does not need to be tuned or decided since it is automatically set according to prf and k parameters as $nprf = k - prf$.

The final formal probabilistic representation of our proposed WEG predictor is

[2]The approach of Shtok et al. (2012) was proposed to predict the query performance while our WEG is proposed to predict the QE performance.

shown in Equation 8.22.

$$\mathcal{P}(D_q{}^{[prf]}|q,\mathcal{Q}) \equiv WEG(q,D_{prf}) = \frac{1}{prf}\sum_{d\epsilon D_{prf}} \mathcal{P}(d|q,r) - Cent(\mathcal{P}(D_q{}^{[nprf]})|q,r)) \tag{8.22}$$

In the next section, we evaluate the prediction quality for different QPP methods (including the proposed WEG predictor). Then, in Section 8.5, we present our proposed utilisation of this QPP framework to develop an adaptive QE approach for UGS retrieval.

8.4 Evaluating Prediction Quality

We evaluate alternative QPP methods in terms of their effectiveness for predicting QE performance in our UGS task. We measure the Pearson correlation coefficient ρ between the actual AP performance after expansion QE(AP) for queries in a given test set, and the predicted values assigned to these queries by each prediction method (WIG, NQC, WEG and others). As explained previously in Section 2.3.1, *higher correlation value indicates increased prediction performance.*

8.4.1 Experimental settings

We set up three different QE tasks using different collections (text, Web and our UGS collection) to evaluate the robustness of the proposed QPP approaches. In addition to the proposed predictor WEG, we study several pre- and post-retrieval QPP methods that have shown to be the most effective in (Zhao et al., 2008; Hauff, 2010; Shtok et al., 2012).

For each task, QE is performed by utilising the top ranked documents returned by running the initial query as feedback documents. QE then extracts the most common terms (top-terms) from the feedback documents (top-docs), adds them to the query and runs the retrieval again to produce the final output.

In order to further test and evaluate the proposed framework, in addition to the

Blip10000 UGS collection, we conduct our experiments over other text-based TREC collections used in previous QPP studies (Zhou and Croft, 2007; Hauff et al., 2008; Hauff, 2010; Shtok et al., 2012; Kurland et al., 2012) namely:

- *WT10G* (topics 451-550) data collection that contains 1,692,096 web documents.
- *ROBUST* collection (disks 4&5-CR, topics 301-450, 601-700), which contains 524,929 news text documents.

We use the titles of TREC topics as the main queries. These collections were previously explained in Chapter 4, Section 4.4.

For retrieval and QE, the PL2 and BO1 models are used. The c parameters for the blip10000 were set as empirically determined in Section4.6.1. While for ROBUST and WT10G, we set the c parameters at at $c = 9$ which were empirically determined for these collections in (Lv and Zhai, 2011).

To setup the QE parameters, we explore 5 different combinations of {*top-terms,top-docs*[3]} : {(10,30),(3,10),(3-3),(5-5),(10-3)} to assess the relationship between the QE(AP) and the predictors. Similar to the work reported in (Shtok et al., 2012; Kurland et al., 2012), to calculate the the centriod mean ($Cnprf$) and the standard deviation, we assume that all scores follow uniform normal distributions. In the next section, we explain how tune the parameters for post-retrieval QPP.

8.4.2 Post-retrieval parameters

As explained in Section 8.2.2, post-retrieval QPP methods require some parameters to be tuned prior to the prediction. In this experiment, we examine three post-retrieval methods, WIG, NQC and our proposed WEG method. To set the k parameter for these predictors, we use the *cross-validation paradigm* explained in Section 4.6.1 since it was proposed for evaluation of QPP in (Shtok et al., 2012). The tuning method is explained as follows.

[3] *Note that the terms :* top-docs, feedback document and prf list refer to the same list, *which is the number of documents used for expansion.*

Term-docs	10-30	3 -10	3-3	5-5	10 -3
Robust					
WEG	**0.52***	**0.51***	**0.51***	**0.55***	**0.55***
WIG	0.39*	0.42*	0.42*	0.44*	0.46*
NQC	0.36*	0.34	0.30	0.33	0.34
Avg.VarTFIDF	0.21	0.18	0.24	0.14	0.22
MaxSCQ	0.09	0.12	0.21	0.13	0.11
MaxIDF	0.04	0.11	0.14	0.12	0.03
AvICTF	0.18	0.29	0.26	0.22	0.21
UGS data (Blip10000)					
WEG	**0.40***	**0.55***	**0.53***	**0.53***	**0.53***
WIG	0.23	0.45*	0.43*	0.42*	0.45*
NQC	0.17	0.38*	0.38*	0.38*	0.38*
Avg.VarTFIDF	0.15	0.23	0.19	0.14	0.22
MaxSCQ	0.13	0.19	0.18	0.11	0.12
MaxIDF	0.11	0.16	0.19	0.13	0.09
AvICTF	0.28	0.29	0.21	0.26	0.27
WT10g					
WEG	**0.50***	**0.56***	**0.52***	**0.52***	**0.56***
WIG	0.38*	0.45*	0.44*	0.44*	0.42*
NQC	0.36*	0.46*	0.36*	0.45*	0.43*
Avg.VarTFIDF	0.22	0.13	0.19	0.14	0.17
MaxSCQ	0.13	0.18	0.23	0.15	0.09
MaxIDF	0.24	0.25	0.28	0.16	0.21
AvICTF	0.26	0.31	0.27	0.24	0.19

Table 8.1: Obtained correlation coefficients between QE(AP) for each QE run vs each QPP (Avg.VarTFIDF MaxSCQ, MaxIDF and AvICTF as pre-retrieval methods, WEG, WIG and NQC as post-retrieval methods) on three different collections. Correlations which are significant at the 0.05 confidence level are those marked with *.

The Mn-Ad query set was randomly spilt into training and testing sets (30 queries each). During the training process, the k parameter for WEG, NQC and WIG for each corpus was tuned using manual data sweeping through the range of [5, 100] with an interval of 5, and through the range of [100,500] with an interval of 20. We performed 20 different splits to switch the roles between the training and testing sets, and reported the best performing k parameter that yields optimal prediction performance (as measured by Pearsons correlation [4]), by taking the average over the 20 different runs. The obtained optimal k parameters for post-retrieval predictors across the three collections are shown in Table 8.2.

[4]Note that the main difference between the tuning method used in Section 4.6.1, is the scoring function used which is here in this section is the Pearsons correlation

WEG optimal parameters					
Term-docs	10-30	3 -10	3-3	5-5	10 -3
Robust	90	150	75	200	150
UGS data (Blip10000)	95	125	55	175	135
WT10g	105	200	65	195	200
WIG optimal parameters					
Term-docs	10-30	3 -10	3-3	5-5	10 -3
Robust	10	10	5	10	15
UGS data (Blip10000)	15	5	5	10	25
WT10g	10	15	15	5	15
NQC optimal parameters					
Term-docs	10-30	3 -10	3-3	5-5	10 -3
Robust	150	150	140	100	125
UGS data (Blip10000)	135	120	130	125	155
WT10g	125	145	150	140	150

Table 8.2: Optimal k parameters for post-retrieval predictors across the three collections

The WEG predictor also requires the prf parameter to be set, this is set automatically based on the number of feedback documents used for each QE. $nprf$ is set as the difference between both the prf and k values for each run.

8.4.3 Experimental Results and Discussion

Table 8.1 shows prediction results for pre-retrieval and post-retrieval QPP which was measured using the optimal parameters obtained using the cross-validation evaluation paradigm explained in the previous section.

Table 8.1 shows that the post-retrieval QPP methods are more effective than the pre-retrieval methods overall. This is can be attributed to the fact that post-retrieval methods generally rely on stronger evidence than pre-retrieval methods which is the obtained scores of each prf document. Furthermore, the fact that post-retrieval methods require exhaustive parameter tuning for QE, as explained in Section 8.4.2 make them more suitable for this task. Post-retrieval QPP adapts to each query by using the obtained scores of the retrieved documents as signal of it is estimated relevancy. This score combine multiple query-related signals of term frequency, inverse document frequency and length normalisation as explained

Section 2.1.1.

The results in Table 8.1 reveal that all post-retrieval predictors have strong ability to estimate the QE performance for most runs. WEG is shown to significantly outperform other predictors across all QE runs, and over the three data collections used in this investigation. This is due to the fact that WEG is tuned to focus on the actual prf and $nprf$ documents, which in this case have the highest influence over the QE performance. Furthermore,

Compared to other post-retrieval methods, the effectiveness of WEG over NQC and WIG indicates the benefit of analysing the set top-ranking, yet non pseudo-relevant documents (the $nprf$ list), as a reference to predict the QE performance.

Furthermore, although they are designed to predict the original query performance (AP), WIG and NQC still provided good estimation of QE(AP). This in fact confirms the relationship between the original query performance AP and QE(AP). In other words, we found that if the AP is too low, the QE process does not have a good prf list to extract useful expansion terms.

The same finding about this relationship was previously discussed in detail by He and Ounis (2009). Furthermore, this relationship is also confirmed in terms of the optimal k parameters (as shown in Table 8.2 for WIG and NQC values) that maximise the correlation for both predictors with QE(AP); where the optimal parameters for both WIG and NQC were similar to that reported by Shtok et al. (2012) for predicting the AP. While the optimal k values for WEG vary across different QE runs and depend on the number of prf documents considered for expansion. In general, in terms of the optimal parameters, we found that for better WEG prediction, the size of the $nprf$ set must be at least 3 times the size of the prf set.

In the next section we examine the utilisation this prediction framework in QE for our UGS task.

8.5 Adaptive Segment-based QE for UGS Retrieval

In this section, we show how the proposed QPP framework from Section 8.3 can be used to implement an adaptive QE designed specifically for UGS retrieval. The key idea of our adaptive QE utilises QPP to select the best prf list of documents for expansion for each query.

Experimental results in the previous three chapters demonstrated that QE can be optimised if the right source-of-evidence is selected for expansion. The adaptive QE is implemented to identify the best performing evidence between the alternative segments and meta-data fields which were explored in the previous chapters. In particular, the adaptive QE technique is designed to automatically select the field and speech segment that is expected to maximise QE effectiveness. Our approach is designed to answer the following questions.

- How to identify the metadata field that is best for QE? This is achieved by calculating the prediction for QE over two fields (Title or Desc) or their combination, and selecting the one that maximises expected performance.

- How to identify the speech evidence which is best for QE in UGS retrieval? this is achieved by calculating the prediction for QE over between alternative segmentation types and the full-document evidence.

In the next section we show how we utilise the WEG predictor introduced in Section 8.3.1 to answer these research questions.

8.5.1 Implementation of the adaptive QE algorithm

Our proposed adaptive QE approach applies Equation 8.19 by calculating the WEG value in Equation 8.22 for every candidate prf list extracted from different evidences (segments/documents or metadata fields) for each query. Then, it selects the $\overline{prf}$ that achieves the highest prediction value for expansion. The proposed QE approach is designed to run in two stages as follows.

Select between three meta-data fields combination by calculating the WEG value from the following.

- $\mathcal{P}(desc|q, \mathcal{Q})$ where *desc* is the list of description fields extracted from top-ranking documents. Therefore, since we use the WEG QPP in our task, we assume that $\mathcal{P}(desc|q, \mathcal{Q})$ ≡WEG(desc).

- $\mathcal{P}(title|q, \mathcal{Q})$ where *title* is the list of titles extracted from top-ranking documents. $\mathcal{P}(title|q, \mathcal{Q})$ ≡WEG(title).

- $\mathcal{P}(titledesc|q, \mathcal{Q})$ where *titledesc* is the list of title and desc fields (combined) extracted from top-ranking documents.

  $\mathcal{P}(titledesc|q, \mathcal{Q})$ ≡WEG(titledesc).

After deciding which meta-data field is better for expansion, the QE selects which speech evidence between the three different speech segmentation (fix100, C99, over100)[5] and the full ASR document as follows.

- $\mathcal{P}(fix|q, \mathcal{Q})$ where fix is the list of fix100 segments from top-ranking documents. We assume that $\mathcal{P}(fix|q, \mathcal{Q})$ ≡WEG(fix).

- $\mathcal{P}(over|q, \mathcal{Q})$ where *over* is the list of over100 segments extracted from top-ranking documents. $\mathcal{P}(over|q, \mathcal{Q})$ ≡WEG(over).

- $\mathcal{P}(C99|q, \mathcal{Q})$ where $C99$ is the list of C99 segments extracted from top-ranking documents. $\mathcal{P}(C99|q, \mathcal{Q})$ ≡WEG(C99).

- $\mathcal{P}(ASR|q, \mathcal{Q})$ where ASR is the list of ASR transcripts extracted from top-ranking documents. $\mathcal{P}(ASR|q, \mathcal{Q})$ ≡WEG(ASR).

[5]note that the definition for each segmentation is similar to that explained in the Section 7.2.4. We limited our exploration to these three options for simplicity, as they have shown to be the most effective for our task (see Section 7.3.

Therefore, the WEG predictor is calculated seven times for each query before the selection is made. Note that both segments and field indexes are all generated during indexing [6] to speedup the QE process and reduce the running time.

The overall process of the proposed adaptive QE is shown in Figure 8.2. In the next section we explain the experimental settings we use to examine the effectiveness of our approach.

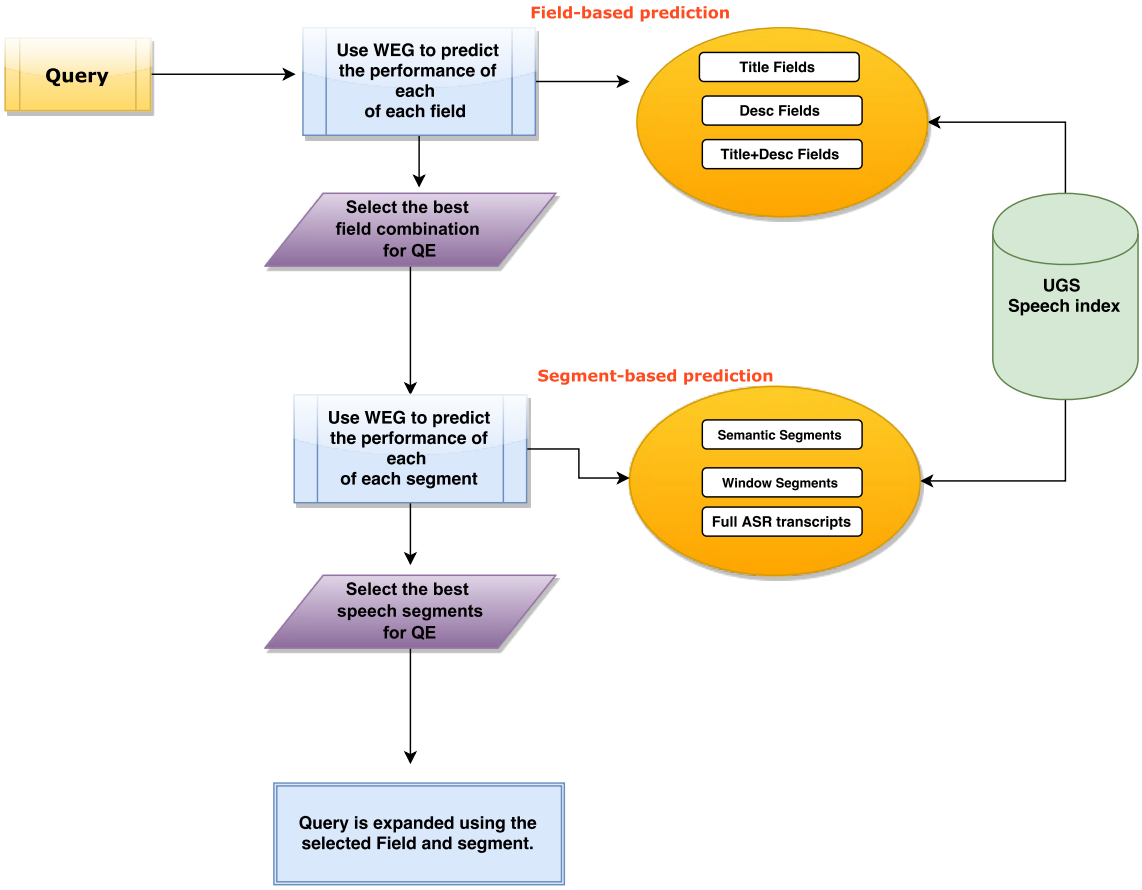

Figure 8.2: Adaptive QE technique using WEG predictor.

8.5.2 Experimental Setting

We run our adaptive QE using all query sets that studied in Chapter 5, namely : Mn-Ad, Mn-Kn, Cl-Fr, Cl-Ar, in order to investigate the effectiveness of this approach in both monolingual and cross-lingual settings. We again used the PL2F

[6]Segment indexes are generated as explained in previous chapters; Section 6.3 for metadata fields, and Section 7.2.4 for the segmentation

	Cl-Ar	Cl-Fr	Mn-Kn	Mn-Ad
Adaptive-QE	0.2921	0.5312	0.6138	0.724
Optimal-QE	0.4233	0.6305	0.6949	0.7724
Baseline-QE	0.2592	0.4609	0.4462	0.5932
no_QE	0.2536	0.4455	0.5122	0.5887
% difference over baseline	**12.69%**	**15.25%**	**37.56%**	**22.05%**
% difference over no_QE	**15.18%**	**19.24%**	**19.84%**	**22.98%**
Acc	43%	52%	70%	72%

Table 8.3: MaP performance of the proposed adaptive QE runs compared to baseline and no_QE runs. Statistically significant differences are highlighted in **bold**

retrieval model with the same structured settings as explained in Section 4.6.1.

We also use the QE parameters value (10 terms 3 documents) to be able to compare all results. The WEG predictor is used to perform the actual prediction of the adaptive QE. We use the obtained optimal of $k = 135$ shown in Table 8.2.

To make them comparable, we also normalise the final scores of each extracted feedback list using the standardisation technique $\frac{Score(d)-\mu}{\sigma}$ before calculating the actual prediction value of each prf list, where μ is the average score of the retrieved results list and σ is the overall standard deviation. In the next section, we explain the metrics used for evaluating our adaptive QE method.

8.5.3 Evaluation of the Adaptive QE Method

In this section we explain how we evaluate our proposed approach. For comparison, we report retrieval effectiveness for a baseline no QE condition *no_QE*, and using standard QE *the baseline-QE*, which the standard QE with no prediction involved similar to the one applied in Chapter 6.

For each adaptive run, we also report the performance of its *optimal QE run* which is obtained using the ideal selection of the best prf list based on relevance-judgement information (ground-truth data) for each query.

Finally, for each QE run, we report the prediction accuracy (*Acc*) of the adaptive selection. This is calculated by how many times the QPP makes optimal prediction.

The results for the different adaptive QE runs are shown in Table 8.3. It can

be seen that adaptive QE runs for all query sets obtained an improved performance over the Baseline-QE. Across the different query sets, the best performance is always obtained by the adaptive QE with a statistically significant (tested at the 0.05 confidence level) improvement over the baseline-QE and the no_QE runs.

The obtained improvement in retrieval performance for the adaptive QE confirms the effectiveness of the WEG predictor in identifying the setting that is most robust for QE. Table 8.4 shows the selection statistics performed by the adaptive QE for each query set, which demonstrates the adaptive QE were able to examine each metadata field and speech segments to improve the retrieval performance. However, this improvement varies between different tasks, in particular, the improvement for the Cl-Ar was significantly lower than others. As discussed in Chapter 5 CL-Ar is the most challenging query set which leads into lower prediction quality of their performance.

Overall, the results shown in Table 8.3 indicate that prediction accuracy of WEG (*Acc*) was between 43% and 72%, showing that there is still room for improvement. By contrast, our proposed adaptive QE is a simple modification of the QE process uses a single engineered feature, the document retrieval score, to rank each source-of-evidence and select the one that is predicted to perform better. The aim of this adaptive QE approach is to demonstrate the effectiveness of our proposed prediction framework and suggest it as possible solution to improve the QE performance for UGS retrieval. Ideally, the accuracy of this prediction approach could be possibly improved by training a machine learning model with more document and query related features/signals involved with QPP signals. We leave this potential improvement of this adaptive approach for future work.

8.5.4 Efficiency of the adaptive QE Method

As previously explained, the adaptive QE approach applies the prediction framework across individual fields and selects the one that is predicted to improve the performance.

Field/segment combination	**Mn-Ad**	**Mn-Kn**	**Cl-Fr**	**Cl-ar**
Fix100-title	2	1	3	12
Fix100-Desc	0	0	2	0
Fix100-titleDesc	7	12	12	13
Over100-title	0	0	0	0
Over100-Desc	2	0	0	0
Over100-titleDesc	6	7	7	8
C99-title	1	12	8	5
C99-Desc	0	0	0	1
C99-titleDesc	4	4	4	4
ASR-title	8	7	7	2
ASR-Desc	7	3	3	2
ASR-titleDesc	23	14	14	13

Table 8.4: Selection statistics for each field/segment combinations (described in Section 8.5.1) as performed by the adaptive QE for each query set

In our demonstration of the QPP framework in Section 8.5.1, the proposed adaptive QE calculates WEG for 7 possible alternatives index files of each field and speech segment. A more comprehensive approach would be to expand this number to more UGS fields. For example, more social meta-data evidences could be integrated in these indexes, such as tweets or user comments, relating to each UGS document. As the retrieval run over each index must be performed at run time, the efficiency of this adaptive method become a concern.

In this section, we evaluate efficiency of our proposed approach by simply using the wall-clock time model (Gysel et al., 2016; Wurzer et al., 2016). We measure the wall clock by looking at the actual running time of the algorithm in seconds, averaged over 5 runs on a regular desktop machine [7]. The overall average time for traditional standard QE (Baseline QE), is 6.7 secs, while the one for the adaptive QE is at 9.2 secs. This indicates that around 37% increase in the overall time to expand the each query using our proposed algorithm. At the same time, the average time required to calculate the WEG for each index file (field or speech segment combination) accounted for less than a 6% increase over the Baseline QE, which is still reasonable given the statistically significant increase in retrieval effectiveness.

[7]Machine specs : Linux machine with a processor of 6-core Intel Xeon E5-1650 V4; 3.5GHz, 15MB cache, and RAM of 64GB 2400MHz DDR4

In terms of efficiency, the advantage of our QPP approach is that it does not to require any further extraction of any document or query related features, and the only signal it relies on is the retrieval scores of each document. Furthermore, as explained in the previous section, the segment and field indexes are generated offline to reduce the overall run time.

Overall, although it is not perfect as it obtained 40-70% accuracy as shown in Table 8.3, the proposed prediction method still saves significant amount of time that would otherwise need to be evaluated manually to estimate the effectiveness of each field/segment with help of human-judgement.

8.6 Summary

This chapter presented a novel QE prediction framework that utilises the retrieval scores of the feedback documents to estimate QE effectiveness. We proposed the WEG predictor, which is a post-retrieval QPP method that uses the information entropy of the feedback document to predict the QE effectiveness. WEG predicts the performance of QE based on comparing the scores of the feedback documents that are used for QE against that are obtained by the non-feedback document in the top ranking list.

In Section 8.4, we evaluated effectiveness of our QPP approach in terms of predicting the QE effectiveness. Our experimental evaluation in Section 8.4 demonstrated the effectiveness of our proposed predictor WEG over other state-of-art QPP techniques for QE not only for our UGS task but also for other collections, such as ROBUST and WT10G (Voorhees, 2003b).

Finally, Section 8.5 showed the use of our proposed prediction framework in developing a new adaptive QE method for UGS retrieval.

Our proposed adaptive QE works by predicting the most effective speech and metadata source-of-evidence for QE. Our experimental evaluation demonstrated the effectiveness of the proposed QE approach in both mono-lingual and cross-lingual

UGS retrieval. However, the results obtained indicate that the adaptive approach is less effective for cross-lingual due to translation noise. In an attempt to address this translation issue, in the next chapter, Chapter 9, we investigate how to improve the translation quality for this task using QPP.

Chapter 9

Adaptive CL-UGS Retrieval

The experiments presented in Chapter 5 showed the high sensitivity of the CL-UGS effectiveness to the translation quality of the search queries. Even when a state-of-the-art translation tool is used, our experimental investigation from suggested that significantly higher translation quality is required in order to maintain an effective cross-lingual search for UGS (CL-UGS) content.

In this chapter, we aim to improve translation quality for UGS retrieval. In particular, we try to address our RQ4 which comprises the following questions.

1. Can we develop a prediction technique to estimate the translation effectiveness for CL-UGS retrieval?

2. Can we implement an adaptive CLIR technique that is able to select the most-effective translations for UGS retrieval?

To answer these questions, we examine the potential of improving CLIR effectiveness by predicting the translation effectiveness using QPP, and selecting the most effective translation to perform in UGS retrieval. We propose a novel QPP framework that relies on pre- and post-retrieval predictors to estimate the quality of translation for our CL-UGS retrieval task.

Our contribution to this chapter are two fold. *Firstly*, we describe our proposed prediction framework to evaluate translation effectiveness in CL-UGS, and

empirically evaluate it on alternative translation outputs extracted from an Arabic-to-English and French-to-English MT systems. **Secondly**, we show how this framework can be integrated in CL-UGS retrieval to improve translation quality. In the next section, we describe the motivation behind our proposed method.

9.1 Motivation

As discussed in Chapter 1, the growing archives of UGS content available online are widely diverse in style, media and the language used. Within the scale of this content, the balance between use of languages is very uneven. For example, for Arabic UGS content, the amount of content available is relatively very small compared to other languages, which results in a significant demand from bilingual Arabic speakers to access information in other languages, most notably English. CLIR comes as an effective tool to bridge the language barrier between user search queries in one language and the target documents in another language (Oard and Diekema, 1998) (see section 2.4.1 for more background on MT and CLIR). The simplest and most commonly adopted approach in CLIR is to use MT to translate the user's query. As explained before, in most cases, MT is used as a black-box to an otherwise unchanged monolingual IR. Many different MT systems have been studied in CLIR research for different tasks, e.g. (Oard and Hackett, 1997; Magdy and Jones, 2014). However, no single MT system has been reported to be effective for all CLIR tasks.

Chapter 5 reported that beyond the challenges in monolingual UGS retrieval, a further challenge is raised from the translation errors in a cross-lingual search setting. We investigated the use of Moses MT as an open-box system, and Google Translate as a black-box off-the-self MT system for our CL-UGS task. Our experimental evaluation in Section 4.6.2 reported that black-box Google MT outperformed open-box Moses MT systems for both Arabic-to-English and French-to-English CLIR. The same conclusion has been observed by CLIR researchers for many language pairs (Zhou et al., 2012). For instance, during the CLEF 2009 workshop (Leveling

et al., 2009; Zhou et al., 2012), the best performing non-off-the-shelf MT achieved just 70% of the performance achieved by Google Translate.

However, later experiments in Section 5.3, we found black-box MT for Arabic/French is still ineffective and results in a significant decrease in retrieval effectiveness compared to the mono-lingual one. Our investigation on task of CL-UGS from previous chapters concluded that having an effective translation quality is vital to cope with the challenges that arise from the uncontrolled amount of noise in the social metadata together with errorful speech transcription.

At the same time, from our results for the us of open-box Moses MT tool to this task in Section 4.6.2, showed that the "best" translation suggested by the open-box MT does not always produce the most effective translation for optimal CL-UGS performance, and better translations are often produced with lower translation confidence by the MT system.

To address word translation ambiguity in CLIR , Qu et al. (2002) utilised co-occurrence statistics in a reference corpus consisting of documents from World Wide Web and external corpus in the target language to choose the best target translation for each source query word. They demonstrated the effectiveness of their approach in Spanish-to-English and Chinese-to-English CLIR tasks in the bilingual track of CLEF 2002.

In this chapter, we propose to use QPP methods to choose relevant translations based on the search index to improve CLIR effectiveness for UGS retrieval using an open-box MT system. We compare the effectiveness of this approach against a standard online black-box MT (Google translate tool) for our CL-UGS search task. In particular, we investigate the use of QPP to select from an *N-best* list of alternative translations for q query generated by a statistical MT systems.

In the next section, we present a prediction framework that utilises recent advances in QPP methods to estimate expected retrieval performance for specific query translation based on both the translated query itself and the output of the translation process.

9.2 Query Performance Prediction Methods For CL-UGS

In Section 2.3, we provided a background on the existing QPP methods in IR. In this chapter, we are interested in studying the application of QPP to predict the translation quality of queries in CLIR. In this section, we describe the pre-retrieval and post-retrieval QPP approaches we use in this task.

9.2.1 Pre-retrieval QPP for CL-UGS

We evaluate the effectiveness of multiple pre-retrieval QPP methods previously explained before in Section 8.2.1 to predict the translation quality for our CL-UGS task. However, specifically with regards to pre-retrieval methods, we argue that having the IDF-based approaches is *not* a good predictor for this task. This argument is supported by the following hypothesis.

By definition, IDF gives a higher weight to unique terms across the search collection. While this might be useful for a retrieval model to rank documents, using IDF is not reliable for QPP since it also gives high values for translation candidates which are *misleading* terms. We define *misleaders as terms that are rare across the collection (hence having high IDF values), but not relevant to the topic of the current query.* These misleaders can result in query topic drift (Mitra et al., 1998a) and thus negatively impact on retrieval effectiveness.

Another source of misleader terms is words which are *Out-of-Vocabulary (OOV)* with respect to the MT. In this situation, the MT system produces incorrect translations of terms which the MT system cannot by definition translate correctly. To cope with the previously-stated issues, we propose a new QPP technique as follows.

Average Term Fluency (AvgFL)

To deal with misleaders arising from IDF prediction issues, we propose a new simple prediction technique which we refer to as *the Average Term Fluency (AvgFL).* Term

fluency indicates whether a query contains the same terms that appear in documents collection or not. *Higher fluency is assumed to lead to better query-document matching, and hence an improved QPP effectiveness.* We rely on the collection frequency (cf) of each term to indicate its fluency in the given collection D. The cf is normalised by the DF to penalise non-helpful terms which appear in all documents in collection.

The proposed AvgFL is calculated as shown in Equation 9.1; where k is the number of t terms in query Q, cf_t is the cf which is the number of times t appears in the collection D, and df_t indicates the number of documents containing the term t.

$$AvgFL(Q) = \frac{1}{k}\sum_{t \epsilon Q}^{k}(log(cf_t + 1)/(log(df_t + 1) + 1)) \tag{9.1}$$

9.2.2 Post-Retrieval QPP For CL-UGS

As explained in Section 8.2.2, state-of-the-art post-retrieval QPP techniques utilise information induced from analysing the retrieval scores $Score(d)$ of the results set ${D_q}^{[res]}$ produced by retrieval method M, where ${D_q}^{[res]}$ represents the list of document ids retrieved for a query together with their ranks $\mathcal{R}i$, and scores $Score(d)$ sorted according to their relevancy to a query q (Zhou and Croft, 2007; Shtok et al., 2012).

We use both WIG and NQC for this task, as previously explained these QPP are designed to have a strong linear relationship with the performance of the query in which the only variable that needs to be decided is the top-k documents.

For our task, we also introduce a modified version of the WEG method introduced in the Section 8.3.1. We call it **Weighted Relevancy Gain (WRG)**. WRG focuses on the scores of the top-ranked assumed *relevant* documents vs other top-ranked but assumed *non-relevant* documents. Unlike other post-retrieval predictors, this approach assumes that the top-k documents ${D_q}^{[k]}$ for each query are composed of two subsets ${D_q}^{[rel]}$ and ${D_q}^{[nrel]}$ defined as follows.

- ${D_q}^{[rel]}$, set of rel relevant documents that are assumed relevant for query q.

- $D_q^{[nrel]}$, set of documents that are assumed non-relevant.

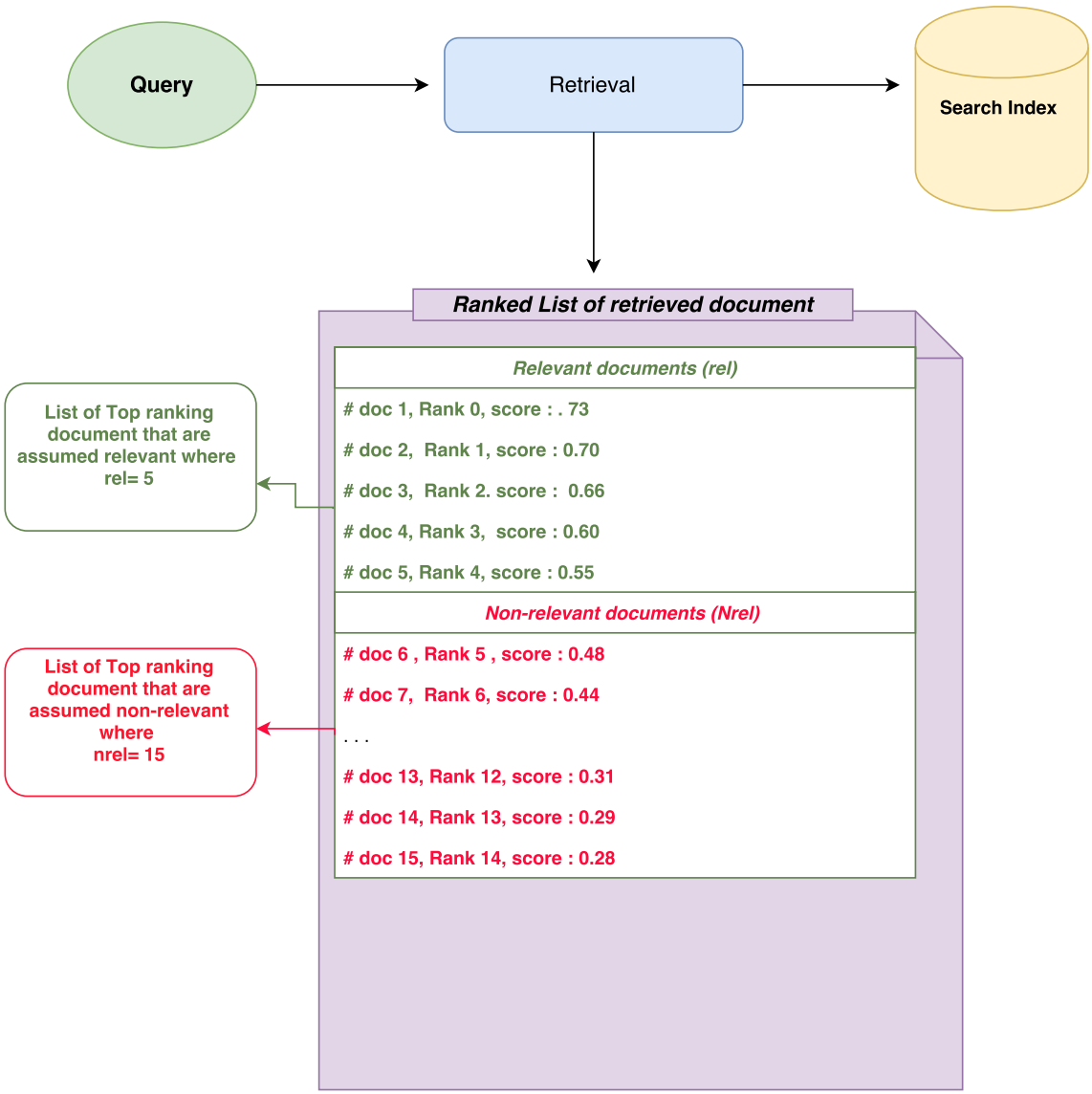

Figure 9.1: Example of the prediction elements used by the proposed WRG predictor, where the assumed relevant documents $rel = 5$, and the non-relevant documents $nrel = 15$.

The WRG predictor aims to analyse the quality of the rel documents by measuring the likelihood that they contain non-relevant documents. This is estimated by measuring information carried by the assumed rel documents against other top-ranked yet non-relevant $nrel$ set of documents. Figure 9.1 shows an example how retrieved list is analysed into the 2 predictions elements in WRG.

WRG uses $Cnrel$ that is the *centroid* of the $nrel$ documents scores ($Score(D)$) as defined in Equation 9.2. $Cnrel$ is then used as a reference point for estimating the

effectiveness. The proposed WRG predictor is implemented as shown in Equation 9.3.

$$Cnrel \equiv Cent(Score({D_q}^{[nrel]})) \equiv \frac{1}{nrel} \sum_{d \epsilon {D_q}^{[nrel]}} Score(d) \tag{9.2}$$

$$WRG(q, D_{rel}) = \frac{1}{rel} \sum_{d \epsilon D_{rel}} \frac{1}{\sqrt{|q|}} (\frac{Score(d)}{Cnrel}) \tag{9.3}$$

Note that the WRG requires 2 parameters: the number of *rel* documents and the number of *nrel* documents to perform the actual estimation.

9.3 Implementing QPP in CL-UGS retrieval

We propose to utilise QPP for CL-UGS as follows.

Assume T_q is an MT translated version of q and ${T_q}^{[n]}$ is the list of n-best translations generated by an MT translation system T. Assuming $\mathcal{Q}$ is the event of being an effective translation of q for getting relevant content in CL-UGS, the goal of this prediction task is to estimate $\mathcal{P}(T_q|q, Q)$ (the likelihood of the translation T_q given that a relevance event happens for q), which seeks to answer the following question.

What is the probability $\mathcal{P}(.)$ for each translation candidate T_q from the top n-list generated by translation system T being an effective translation $\mathcal{Q}$ of a query q for CL-UGS?

Our proposed framework relies on QPP to rank the best translations ${T_q}^{[n]}$ generated by MT system T based on the probability function $\mathcal{P}(T_q|q, Q)$. We the QPP methods previously explained in section 2.3 to predict the retrieval effectiveness of each translation candidate T_q. For example, we assume that AvICTF (He and Ounis, 2006) can be taken as prediction function $\mathcal{F}$ to indicate the effectiveness of translation candidates T_q as $\mathcal{P}(T_q|q, Q) \equiv \mathcal{F}(T_q) \equiv AvICTF(T_q)$.

In the next section we explain how QPP methods are evaluated for our CL-UGS

task.

9.4 Evaluating QPP in CL-UGS Retrieval

We evaluate state-of-the-art pre-retrieval and post-retrieval QPP techniques explained in Section 2.3 for this prediction task. In the next section, we explain the experimental settings we used for this evaluation.

9.4.1 Experimental Setup

In order to evaluate QPP methods for our CL-UGS task we configured the following modules.

- CL-UGS task using the Cl-Fr and Cl-Ar query sets that were translated using Google translate.
- CL-UGS task using the Cl-Fr-Moses and Cl-Ar-Moses query sets that were translated by taking the single best (*SingleBest*) translation recommended by the open-box Moses system.
- N-best translations in Arabic and French for each query of the English monolingual queries (Mn-Ad) by Moses system.
- QPP module to score each of the n-best translations and re-rank them accordingly.

The generation of the *CL-UGS* for Cl-Fr and Cl-Ar, Cl-Fr-Moses and Cl-Ar-Moses query sets is as explained previously in Section 4.5.3. We used the PL2F retrieval model with the same structured settings as described in Section 4.6.1

To generate the *n-best translations*, we used the open-box Moses (Koehn et al., 2003), we developed both Arabic-To-English and French-to-English explained in Section 4.5.4. After setting up the MT systems, we generated the top 100 translations

Table 9.1: Average number of candidate translations generated for each query (*nbest/query*), and total number of candidate translations generated per each MT system (*Total nbest*).

	nbest/query	**Total nbest**
French-to-English	97	5820
Arabic-to-English	94	5640

list for each query. Table 9.1 shows the statistics for the number of generated candidate translations per each language. For Arabic-to-English system, the overall number of translation candidates was 5,640 instances, and 5,820 for French-to-English system with an average of over 90 different translations per query for both system.

To set up the *QPP module*, we implemented several methods as follows.

- For pre-retrieval QPP, We test IDF as predictor including MaxIDF, AvgIDF and SumIDF Methods. We also test the SCQ, VARTFIDF, AvQL and QS predictors (See Section 8.2.1 for more detail about these methods).

- For post-retrieval QPP, we test the three post-retrieval QPP methods explained in the Section 9.2.2 and Section 8.2.2, namely, WRG, WIG and NQC. In the next section, we explain how we tune parameters for these QPP methods for our task.

9.4.2 Parameters Tuning for the Post-retrieval QPP

As explained in Section 8.2.2, post-retrieval QPP methods require some parameters to be tuned. In our task, four different parameters are required to be tuned as follows.

- k parameter for NQC QPP.

- k parameter for WIG QPP.

- *rel* parameter for WRG QPP. [1].

[1]Note that *nrel* parameter for WRG QPP does not need to be tuned since it is decided based on the scale of *rel*

Table 9.2: The optimal k parameters obtained for post-retrieval predictors

QPP	k parameter
WIG	10
NQC	150

For our experiments, we used the following approach to tune NQC, WIG. The k parameter for both NQC and WIG parameter was tuned using the *cross-validation paradigm* that gives parameters that yield optimal prediction performance for each predictor on a set of queries (Shtok et al., 2012). We used the Mn-Ad query set which has 60 monolingual EN queries to tune the WIG and NQC parameters.

The Mn-Ad query set was randomly split into training and testing sets (30 queries each). During the training, parameters k was tuned using manual data sweeping through the range of [5, 100] with an interval of 5, and through the range of [100,500] with an interval of 20. We performed 20 different splits to switch the roles between the training and testing sets and reported the best performing k parameter that yield optimal prediction performance in terms of Pearsons correlation.

The WRG has two parameters, rel that is the number of documents that are assumed relevant, and $nrel$, the number of documents that are assumed non-relevant. As previously explained, to achieve better estimation, the size of rel must be higher that $nerl$. In order to be able to tune the rel parameter, we fix the size $nrel$ with regards to the rel. We explore 4 different versions of WRG with different combination between rel and $nrel$ as follows.

- **WRG-2** : where we fixed $nrel$ as $rel \equiv 2 * nrel$
- **WRG-3** : where we fixed $nrel$ as $rel \equiv 3 * nrel$
- **WRG-5** : where we fixed $nrel$ as $rel \equiv 5 * nrel$
- **WRG-10** : where we fixed $nrel$ as $rel \equiv 10 * nrel$

After fixing $nrel$, the rel was then tuned similar to the k parameter in WIG and NQC using the *cross-validation paradigm* on the Mn-Ad query set.

Table 9.3: The optimal parameters obtained for WRG Predictor

QPP	*rel parameter*	*nrel parameter*
WRG-2	50	WRG-2
WRG-3	30	WRG-3
WRG-5	20	WRG-5
WRG-10	10	WRG-10

The optimal parameters obtained for the post-retrieval predictors are shown in Table 9.3, and the optimal parameters obtained for WRG predictor is shown in Table 9.2. The optimal k parameter obtained for WIG was 10, while for NQC it was 150, these are similar to those recommended in (Shtok et al., 2012).

Table 9.4: Correlation Coefficients vs AP for each query translation from Ar-to-En MT system against each QPP. Correlation that are significant at the 0.05 confidence level are marked in **bold**.

	Pearson	**Kendall's tau**	**Spearman's**
Pre-retrieval Predictors			
VarTFIDF (Zhou and Croft, 2007)	-0.20	-0.165	-0.194
SCQ (Zhou and Croft, 2007)	0.248	0.137	0.201
Qs (He and Ounis, 2004)	**-0.319**	-0.221	**-0.29**
AvQL (Mothe and Tanguy, 2005)	-0.193	-0.126	-0.208
SumIDF (Cronen-Townsend et al., 2002)	0.069	0.110	0.163
AvgIDF (Cronen-Townsend et al., 2002)	0.030	0.086	0.128
MaxIDF (Scholer et al., 2004)	-0.044	0.019	0.035
AvICTF (He and Ounis, 2006)	0.118	0.162	0.210
AvgFL (Equation 9.1)	**0.446**	**0.313**	**0.395**
Post-retrieval Predictors			
WRG-2 (Equation 9.3)	**0.440**	**0.301**	**0.359**
WRG-3 (Equation 9.3)	**0.463**	**0.321**	**0.384**
WRG-5 (Equation 9.3)	**0.452**	**0.333**	**0.363**
WRG-10 (Equation 9.3)	**0.412**	**0.291**	**0.349**
WIG (Equation 8.17)	**0.405**	0.260	**0.333**
NQC (Equation 8.18)	**0.385**	0.22	**0.321**

9.5 Evaluating Prediction Quality

As previously explained in Section 2.3.1, the effectiveness of QPP methods is usually evaluated by measuring correlation between values assigned by the QPP method and the actual retrieval performance, in terms of the *AP (Average Precision)* of each query.

The quality of each predictor is evaluated in our CL-UGS task by measuring the Pearson linear correlation coefficient ρ between both the MAP , which is measured using human relevant assessment for each candidate translation of the 100-best, and the values assigned to these queries by each prediction method.

For evaluating each QPP, we measure the correlation by utilising the MAP obtained using each of the translations in the 100-best list for both the Arabic-to-English and French-to-English systems. For the SQC and VarTFIDF, we examine the best performing aggregation, as reported by (Zhao et al., 2008; Hauff, 2010), which are MaxSQC and SUM.VarTFIDF.

In addition to Pearson's correlation, we also evaluate Kendalls tau and Spearman correlations to report the nonlinear relationship between these predictors and the retrieval performance.

Results in terms of correlations for the French-to-English MT system is shown in Table 9.4, while Table 9.5 shows the QPP quality using the Arabic-to-English MT system. These results are analysed in the following sections.

Table 9.5: Correlation Coefficients vs AP for each query translation from French-to-English MT system against each QPP. Correlation that are significant at the 0.05 confidence level are marked in **bold**.

	Pearson	**Kendall's tau**	**Spearman's**
Pre-retrieval Predictors			
VarTFIDF (Zhao et al., 2008)	-0.153	-0.116	-0.142
SCQ (Zhao et al., 2008)	0.174	0.114	0.171
Qs (He and Ounis, 2004)	-0.179	-0.103	- 0.132
AvQL (Mothe and Tanguy, 2005)	-0.112	-0.069	-0.084
SumIDF (Zhao et al., 2008)	0.115	0.084	0.120
AvgIDF (Zhao et al., 2008)	0.033	0.018	0.026
MaxIDF (Zhao et al., 2008)	-0.039	-0.015	-0.024
AvICTF (He and Ounis, 2006)	0.033	0.022	0.029
AvgFL (Equation 9.1)	**0.429**	**0.293**	**0.378**
Post-retrieval Predictors			
WRG-2 (Equation 9.3)	**0.420**	0.294	**0.371**
WRG-3 (Equation 9.3)	**0.443**	**0.333**	**0.382**
WRG-5 (Equation 9.3)	**0.403**	0.283	**0.377**
WRG-10 (Equation 9.3)	**0.391**	0.271	**0.375**
WIG (Equation 8.17)	**0.386**	0.251	**0.361**
NQC (Equation 8.18)	0.220	0.129	0.190

Table 9.6: Example of candidate translations for an Arabic Query

Example of an Arabic Query	سوفتويرات لتصميم و برمجة مواقع الويبْ.
Candidate Translation (T1)	“ سوفتويرات ***for the development and web design***”
Candidate Translation (T2)	“ سوفتويرات ***for the development and design internet”.***

9.5.1 Pre-retrieval Quality

In terms of pre-retrieval QPP, as can be seen from the results shown in both Table 9.4 and Table 9.5, IDF-based predictors are found to have the least robustness of the predictors examined. The reliability issue regarding misleading terms (as discussed in Section 9.2.1) significantly impacted the prediction quality of these predictors. To further illustrate this issue, consider the Arabic example query

سوفتويرات لتصميم و برمجة مواقع الويبْ.

This query has two candidate EN translations (T1 and T2) as shown in Table 9.6.

The main difference between these two translations is the word “web” vs “internet”. While the word “internet” is rare term with a higher IDF value, it is considered as a misleader in this query since it shifts the original topic of the query “web design”. Thus, this has resulted in a query topic-drift, and hence a false prediction of its performance.

In contrast, Table 9.4, and Table 9.5 show that prediction quality is improved for all QPP methods which are less focused on the uniqueness of the terms and do not rely *solely* on the IDF in its calculation (i.e. Qs, SQC).

The proposed AvgFL measure is shown to have the highest quality over all tested pre-retrieval QPP methods, showing a consistent statistically significant prediction across different correlation measures. This arises as result of its robustness in utilising the fluency measure to identify relevant translations, and penalising the ones that contain OOV or very unique words in the collection.

9.5.2 Post-retrieval Quality

Results from Table 9.4 and Table 9.5 show that post-retrieval QPP methods are more robust and perform better than the pre-retrieval methods overall for both French-to-English and Arabic-to-English systems.

This is due to the fact that post-retrieval methods are based on the actual scores of the translations in which at least one retrieval run has been used for the prediction. Unlike, pre-retrieval methods, post-retrieval requires exhaustive parameter tuning, as previously explained in Section 9.4.2. Both parameter tuning and the time required to generate the post-retrieval QPP is a major efficiency issue by comparison to pre-retrieval QPR (the average time to generate the pre-retrieval QPPs was around 10% to that of the post-retrieval QPPs).

WRG has the highest prediction quality across all predictors. The robustness of WRG is due to the fact that it relies on stronger evidence in the form of the scores of the relevant and non-relevant documents. While NQC and WIG rely only on one parameter, i.e. the top k ranked documents, WRG relies on further tuning of the top k parameter into both the *rel* and *nrel* documents to provide better estimation.

9.6 Using QPP to Find Relevant Translations in CL-UGS

In this section, we investigate the potential for these QPP techniques to be used in an adaptive CLIR method that is able to automatically identify the most relevant translations for CL-UGS. The process of this adaptive CLIR method is shown in Figure 9.2.

The main idea is to use the translation candidate that is predicted to have the highest retrieval effectiveness for each query. Using the same settings explained in Section 9.4.1, we implement the adaptive CLIR algorithm as follows.

1. For each query, the MT system is used to generate up to the 100-best trans-

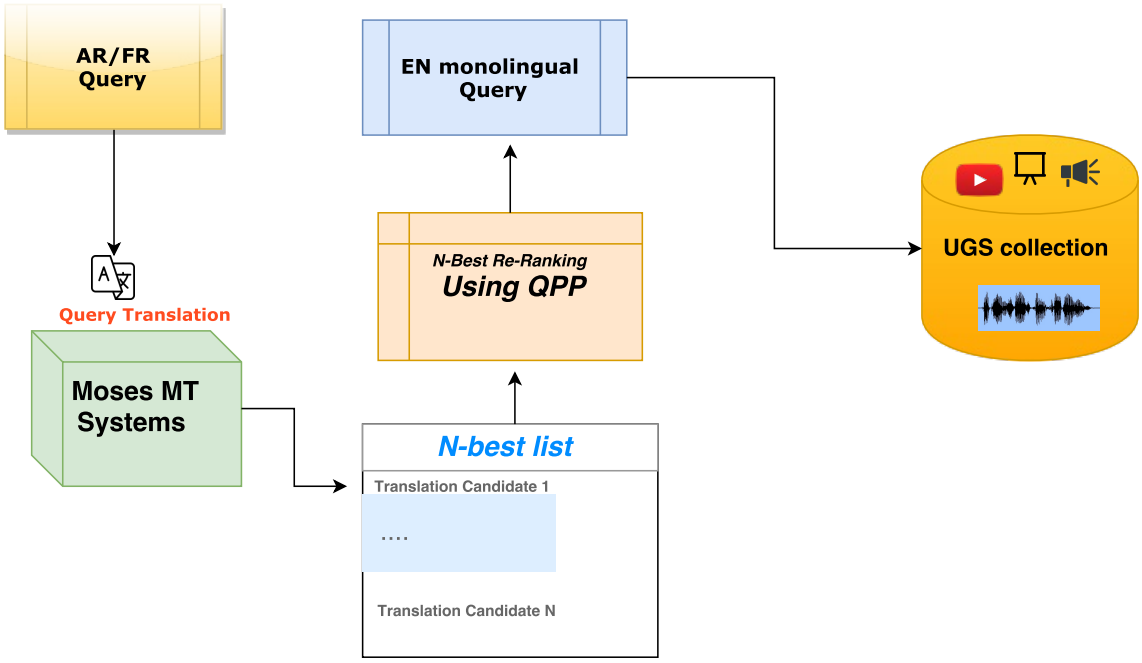

Figure 9.2: Adaptive CLIR method.

lations to form a selection pool.

2. QPP is used to score each translation candidate from the selection pool based on its estimated retrieval performance.

3. CL-UGS Retrieval is then performed using the translation which has the highest score and is predicted to be most effective for each query.

We evaluate our adaptive CLIR technique against three baselines as follows.

- *black-box MT*: this is performed using the the Cl-Fr and CL-Ar query sets which are translated using Google translate as an example of an off-the-shelf black-box MT tool.

- *SingleBest*: this is performed using CL-Fr-Moses and CL-Ar-Moses, which are the 1-best translation output generated by our Moses MT system.

- *100BestAP*: this baseline uses the ground truth data to get the best performing translation in terms of AP from the 100-best translations generated by Moses MT.

In the next section, we report our results using the proposed adaptive CLIR for Arabic CL-UGS.

9.6.1 Experimental Results - for Arabic CL-UGS

The adaptive and baseline retrieval performance results for the Arabic-to-English CL-UGS experiments are shown in Table 9.7.

For clarity, we also report the percentage of improvement over each of these baselines as an additional columns as follows.

- *% blackbox MT* column indicates the improvement in MAP over the BlackBox MT baseline.
- *% SingleBest* column indicates the improvement in MAP over the SingleBest baseline.
- *% 100BestAP* column indicates the improvement in MAP over the 100BestAP baseline.

	MAP	% blackbox MT	% SingleBest	% 100BestAP
Baseline CLIR				
Off-shelf black-box	0.3277	-	*40.46%	*-35.31%
100BestAP	0.5066	* 54.59%	*117.15%	-
SingleBest	0.2333	*-28.81%	-	*-53.95%
Adaptive CLIR using Pre-Retrieval				
MAXIDF	0.2082	-36.47%	-10.76%	*-58.90%
QL	0.1827	-44.25%	*-21.69%	*-63.94%
SumSQC	0.2215	-32.41%	-5.06%	*-56.28%
AvgFL	**0.3107**	**-5.19%**	***33.18%**	**-38.67%**
avgICTF	0.2319	-29.23%	-0.60%	*-54.22%
SumVarTFIDF	0.2419	-26.18%	3.69%	*-52.25%
Qs	0.2003	-38.88%	-14.14%	*-60.46%
Adaptive CLIR using Post-Retrieval				
WRG-3	**0.3899**	**18.98%**	***67.12%**	**-23.04%**
NQC	0.3379	3.11%	*44.83%	*-33.30%
WIG	0.3423	4.46%	*46.72%	*-32.43%

Table 9.7: ***Arabic-to-English CL-UGS*** Baseline and adaptive CLIR results using both pre-retrieval and post-retrieval QPP. Percentages % with * indicate *statistically significant* at 95% confidence level

The *Baseline CLIR* results from Table 9.7 show that black-box Google MT introduced a 40% performance improvement to the SingleBest output from the open-box Moses, which confirms the previously reported results in (Leveling et al., 2009) that using off-the-shelf MT tools can be easier and more effective over the Singlebest.

On the other hand, looking at **% 100BestAP** results from the open-box with ideal AP performance, confirms that the open-box MT can indeed be improved by looking at other translations candidates that are more relevant for CL-UGS.

The pre-retrieval QPP performance from the *Adaptive CLIR using Pre-Retrieval* block of Table 9.7 shows that pre-retrieval QPP can be used to find the best translation from the 100-best extracted. Comparing each of the pre-retrieval QPP tested in this task confirms the conclusion obtained from Table 9.4 and Table 9.5 where the AvgFL is indeed the most effective for this task.

The *Adaptive CLIR using Post-Retrieval* block of Table 9.7 shows how the post-retrieval QPP methods are more effective, generally, for finding the most effective translation in CLIR. This confirms previously reported conclusions on comparing pre-retrieval and post-retrieval QPP, i.e. that post-retrieval QPP is consistently more effective (Hauff, 2010).

Overall, the WRG predictor is the most effective with significant improvement of 67% over the SingleBest and 19% over the black-box tool. This also confirms that the correlation results reported in Table 9.4 where WRG has the highest correlation to AP in terms of predicting the translation quality.

Results from Table 9.7 indicate that QPP techniques can indeed help in re-ranking translation candidates of an open-box MT system, and hence improve its translation quality for CLIR purposes. Both AvgFL and WRG predictors, which were designed specifically for this task, served as an adequate reference to find the most effective translations and improve over the SingleBest output that is suggested originally by the MT system. However, none of the reported adaptive Arabic CL-UGS results were able to match or even come close to the ideal performance baseline (100BestAP). This suggests that there is still scope for further improvement in terms

of prediction quality. To verify these conclusions we also test our proposed adaptive CLIR method for the French CL-UGS in the next section.

	MAP	% blackbox MT	% SingleBest	% 100BestAP
Baseline CLIR				
blackbox MT	0.4307	-	*32.08%	*-19.22%
100BestAP	0.5332	*23.80%	*63.51%	-
SingleBest	0.3261	*-24.29%	-	*-38.84%
Adaptive CLIR using Pre-Retrieval				
MAXIDF	0.1689	* -60.78%	* -48.21%	*-68.32%
QL	0.1692	*-60.72%	*-48.11%	*-68.27%
SumSQC	0.2571	*-40.31%	*-21.16%	*-51.78%
AvgFL	**0.3507**	**-18.57%**	**7.54%**	**-34.23%**
avgICTF	0.2012	*-53.29%	*-38.30%	*-62.27%
SumVarTFIDF	0.2793	*-35.15%	-14.35%	*-47.62%
Qs	0.2909	*-32.46%	-10.79%	*-45.44%
Adaptive CLIR using Post-Retrieval				
WRG^2	**0.4323**	**0.37%**	***32.57%**	**-18.92%**
NQC	0.3309	*-23.17%	1.47%	*-37.94%
WIG	0.3464	*-19.57%	6.23%	*-35.03%

Table 9.8: *French-to-English CL-UGS* -Baseline and adaptive CLIR results using both pre-retrieval and post-retrieval QPP. Percentages % with * indicate *statistically significant* change at 95% confidence level

9.6.2 Experimental Results - for French CL-UGS

The adaptive and baseline retrieval performance results for the French-to-English CL-UGS experiments are shown in Table 9.8.

The *Baseline CLIR* results from Table 9.8 show the effectiveness of Google MT for French comparing to the Arabic ones observed in the previous section. This results indicate that Google MT significantly out-performs the SingleBest output from Moses MT by 32%. However, the results suggest that Moses open-box can in fact be better than using Google translate by using the 100BestAP.

Furthermore, the results from Table 9.8 confirm our initial hypothesis in this chapter that open-box MT could be improved by adjusting the translation quality for CL-UGS purposes, where the 100BestAP obtained a significant improvement of 63% over the SingleBest. WRG and AVGFL were also the most effective QPP

approaches for the French CL-UGS task. Using the WRG, the adaptive approach was able to gain significant improvement of 33% over the SingleBest performance. However, the WRG improvement over the Google MT was at less than 1%.

Overall, comparing the results obtained by Google MT and the ideal 100BestAP from both Arabic and French CL-UGS in Table 9.7 and Table 9.8 , it is clear that both are equally effective and our adaptive approach can benefit from combining both translations. In the next section, we explore the use of our proposed WEG QPP in predicting the best translations from both Google MT and Moses MT.

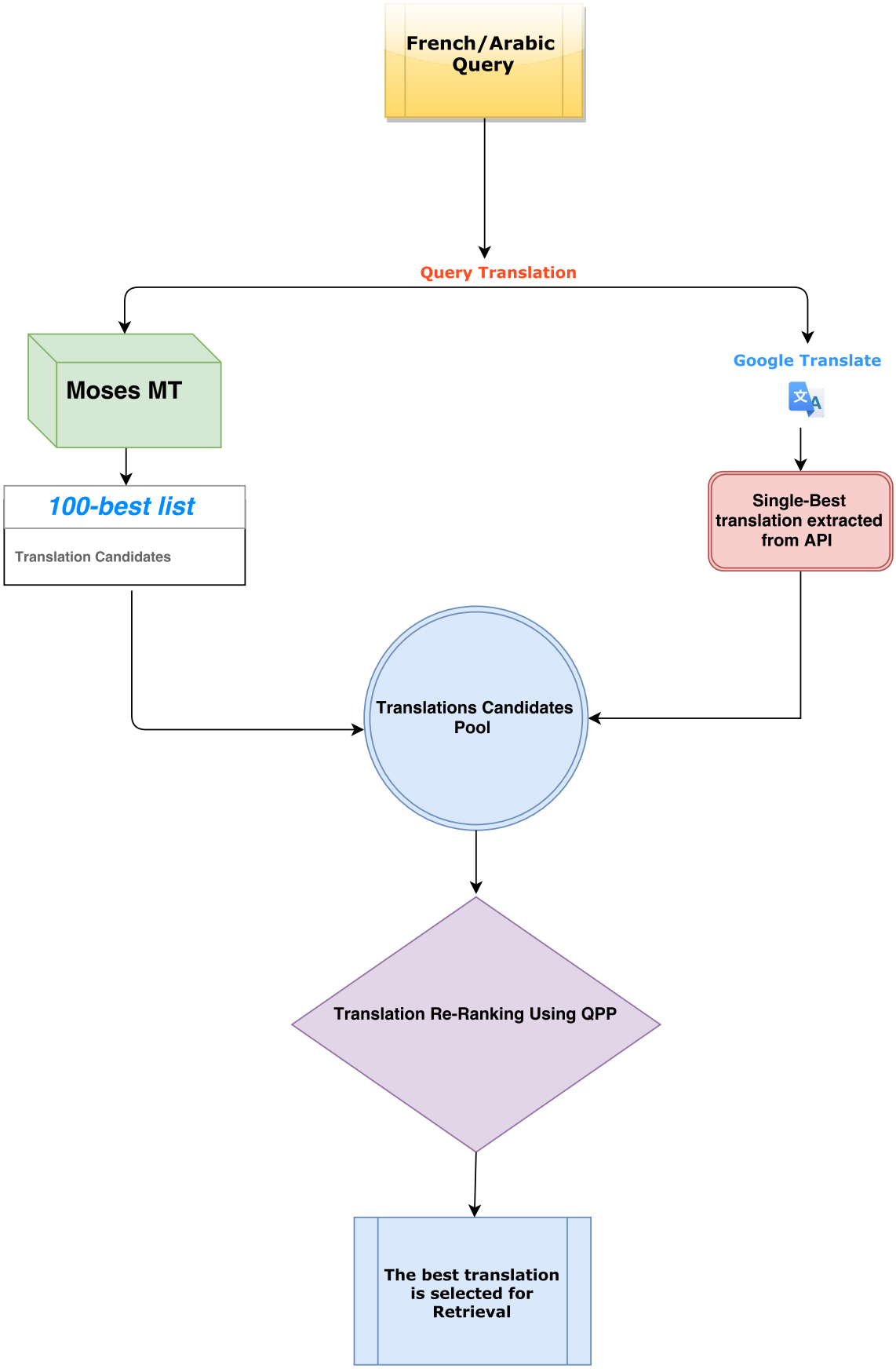

Figure 9.3: Combining different translations for CL-UGS using QPP

9.7 Utilising multiple MT systems for Adaptive CL-UGS

In this section we aim to combine translations from both Google MT and Moses MT to improve translation quality for our CL-UGS task.

We utilise the proposed WRG approach to predict the best translation output between both Google and Moses MT. This new adaptive approach is shown in Figure 9.3 and implemented as follows.

- For each query, the translation output from both Google and Moses is added to a *Translations candidates pool*
- WRG is used to score each translation from the pool based on their estimated relevancy.
- Finally, the best-scoring translation is used for CL-UGS.

9.7.1 Experimental Settings

We use the same experimental settings as Section 9.4.1). For our investigation in this section, we conduct our proposed adaptive CLIR approach using two runs as follows.

- ***Gl-Single*** : in this run, for each query, the adaptive approach selects the most effective translations between two options; the single-best (1-best) translation between Google translation and 1-best translation from Moses MT.
- ***Gl-Best100***: in this run, the adaptive approach utilises the WRG to select the most effective translations between multiple options; the 1-best translation from Google translation API, and each of the 100-best translations from Moses MT.

In addition to the MAP, we also report percentages of difference over three baselines; using % blackbox MT, % SingleBest and % 100BestAP that were explained in the previous section.

We also report the *Accuracy* of each technique, by measuring how many times our approach made the right selection (the one that maximises IR effectiveness) between the available selections. This is calculated by comparing the selection made by the WRG predictor and the selection based on the best AP using the groundtruth data.

9.7.2 Experimental Results and Discussions

The Adaptive CL-UGS performance for Arabic is shown in Table 9.9, and while Table 9.10 shows the obtained performance for French.

As expected, combining different systems brings a significant improvement for both French and Arabic translation in CL-UGS. The Gl-100Best runs were significantly more effective than the GL-Single for both Arabic and French. This indicates the importance of adding the 100-best translations from the open box MT in order to have significant improvement.

Comparing the results in Table 9.9, and Table 9.10 to those in Table 9.7, and Table 9.8, we can see that the performance is always better when combining both Google MT and Moses MT.

For Arabic CL-UGS, the proposed approach using **Gl-100Best** run in Table 9.9 was able to obtain a significant improvement of 34% over the black-box Google MT tool, and up to 88% significant improvement over using the open-box Moses MT with SingleBest. For French CL-UGS, the improvement was at 27% over Google MT tool, and 68% over the performance of Moses MT.

Overall, the reported result in Table 9.9, and Table 9.10 indicate the benefit of combining MT output from different MT systems improve translation quality for CL-UGS retrieval.

	Gl-SingleBest	**Gl-100Best**
MAP	0.3409	0.4379
% blackbox MT	4.03%	*33.63%
% SingleBest	*46.12%	*87.70%
% 100BestAP	-0.75%	0.93%
Accuracy	83%	49%

Table 9.9: The Adaptive CL-UGS performance for both Arabic using combined translations from Google MT and Moses MT.

	Gl-SingleBest	**Gl-100Best**
MAP	0.4655	0.5479
% blackbox MT	8.08%	*27.21%
% SingleBest	*42.75%	*68.02%
% 100BestAP	-12.70%	2.76%
Accuracy	76%	53%

Table 9.10: The Adaptive CL-UGS performance for both French using combined translations generated by Google MT and Moses MT.

9.8 Summary

In this chapter we presented a framework for predicting translation quality for our CL-UGS retrieval task. We proposed novel QPP approaches to estimate the effectiveness of a translation when there is no human evaluation of retrieval available. The contributions and conclusions of this chapter can be summarised under the following points.

- We evaluated the effectiveness of using different state-of-the-art QPP approaches to predict translation effectiveness for CL-UGS. Our experimental investigation reveals that IDF-based prediction is not effective for this task because of the misleading very unique terms which can result in unreliable prediction.

- We proposed a novel pre-retrieval QPP technique for this task called AvgFL (Average Fluency) that is designed to detect misleading very unique and OOV words. AvgFL relies on capturing the translation fluency measured over the search collection to estimate the effectiveness for CL-UGS.

- For post-retrieval QPP, we also proposed the WRG (Weighted Relevancy Gain) approach that is modified version of the well-established WIG predictor (Zhou and Croft, 2007). WRG is based on measuring the information carried by the relevant documents retrieved by a certain translation.

- Our experimental evaluation reports the robustness of these proposed approaches in predicting the translation effectiveness for an Arabic-to-English and French-to-English CL-UGS tasks over other state-of-art QPP methods.

- Finally, we demonstrated how QPPs could be utilised by an adaptive CLIR model that is able to find the most-relevant translations for CL-UGS. Our work in this chapter presented two prototypes for implementing the proposed QPP framework into the MT module in CL-UGS. The first is to use it in open-box MT to re-rank the n-best translations based on their predicted retrieval performance in CL-UGS. The second combines translations from different sys(tems and the select the one that is predicted to perform better for CL-UGS. Both demonstrations obtained a significant improvement in the CL-UGS performance.

Following the investigation of our RQ4 (Can we implement an adaptive CLIR technique that is able select the most-effective translations in UGS retrieval?) which is the last research question of this thesis. The next chapter summarises our findings and contribution, and provides suggestions for further investigations in UGS retrieval.

Chapter 10

Conclusions

The work of this thesis has explored the topic of effective cross-lingual retrieval for online user-generated speech (UGS), we studied the challenges of both monolingual and cross-lingual UGS retrieval. This research demonstrated how techniques such as query expansion, speech segmentation and query performance prediction can be beneficial in coping with the UGS challenges and improving the overall effectiveness for UGS retrieval.

In this chapter, we summarise the contributions of this study and outline potential directions for future work.

10.1 Research Questions Revisited

In this section, we revisit the research questions which were posed in Chapter 1, and state how each has been addressed in this thesis.

10.1.1 RQ1 : Understanding UGS search as a retrieval task

The work of this thesis was motivated by the demand for a robust information retrieval to enable effective access of UGS content.

In the first experimental chapter, Chapter 5, we sought to understand the main challenges of UGS content by answering the following questions from our **Research**

Question 1 (RQ1) :

1. What are the main challenges for UGS mono-lingual and cross-lingual retrieval and how different are they from other SCR tasks?

2. How do Internet-collected UGS data sources behave in monolingual and cross-lingual retrieval? How do these behave when they are combined/weighted together using state-of-the-art retrieval frameworks?

To answer these questions, we conducted experiments in three stages as follows.

- First, we proposed a retrieval framework for monolingual and cross-lingual search of UGS content and investigated multiple retrieval settings to find the most suitable ones for UGS content.

- We then examined the retrieval effectiveness and robustness for each source-of-evidence in UGS data.

- Finally, we studied the effectiveness in combing each field and adjusting its weight to examine their interaction.

We used the results from these experiments to answer our RQ1, summarised as follows.

- Overall retrieval effectiveness challenges : The variations of quality, length and structure across UGS content makes it challenging to tune or train a retrieval system with single optimal settings that work for all queries. In other words, relevant content for each query can be presented with different structure and quality across UGS data which makes it very challenging for retrieval systems.

- Field reliability and robustness in CL-UGS retrieval: We studied the retrieval effectiveness of three UGS fields in both monolingual and cross-lingual settings, namely, ASR transcripts, which are challenged by recognition errors, video titles, which can be very short and lack content, and video descriptions, which can be generic and incomplete. The results obtained suggest that

ASR is the most important field in CL-UGS and contributes to higher recall and precision performance. However, the ASR field has the lowest robustness across other fields and its performance can drop significantly for CLIR due to the interaction of translation and transcription errors. We found that Titles are the most reliable and robust source-of-evidence but they suffer from recall problems due to their shortness.

- Translation quality in CL-UGS: The experimental investigation showed that translation errors can have a significant negative impact on the retrieval effectiveness for CL-UGS. The ASR transcription errors, together with the UGS noise and translation errors, contribute to major problem of vocabulary mismatch in UGS retrieval. In summary, our experimental results in Chapter 5 indicate that a high translation quality is required to maintain a reasonable cross-lingual performance.

10.1.2 RQ2: Query Expansion for UGS retrieval

The initial investigation of UGS retrieval in this thesis revealed that the retrieval effectiveness can suffer from a major problem of vocabulary mismatch. Our first step towards addressing this issue was by improving the query to match the relevant content. The aim of RQ2 was to investigate the effectiveness of applying Query Expansion (QE) for UGS retrieval. In Chapter 6, we examined the utility of using a traditional QE method to address the vocabulary mismatch problem in UGS retrieval.

Our experiments in Chapter 6 were designed to investigate the following questions of RQ2.

1. How does traditional query expansion approaches work under such a setting of noisy data collected from Internet videos?

2. Can we have an effective QE approach that adaptively utilises these data

sources to expand individual queries in order to improve overall retrieval effectiveness?

Our experimental results in Chapter 6 indicated that applying traditional QE approaches is *not* effective for UGS retrieval. This can be attributed to the noise and complexity of the UGS data used during the expansion process of QE. Our result suggests that UGS fields have a varying and inconsistent reliability for QE, and even when all fields are combined together, the retrieval effectiveness is still be negatively impacted due the topic drift issue.

QE problems arise due to irrelevant content presented in the expansion documents in UGS. Therefore, our initial proposal to improve QE for this task was to discard the fields which were irrelevant from the expansion documents. We proposed an adaptive field-based QE approach that utilises only relevant fields from the expansion documents QE. Our empirical results demonstrate how the retrieval performance can be *significantly improved* by picking the right field-combination for expanding each query.

However, this proposed approach has two main limitations as follows.

1. The proposed approach relies on a ground-truth data to estimate the field relevance. However, having the ground-truth data for each query is not possible due to the scale of UGS content. Therefore, in Chapter 8 we proposed a prediction approach to estimate field effectiveness for QE.

2. Our results also suggest that even when fields such as ASR transcripts are relevant and effective for QE, it can have some irrelevant content that add noise into the QE process, and rather harm the effectiveness. Therefore, in Chapter 7, we proposed to use speech segmentation techniques to identify relevant segments and for QE.

10.1.3 RQ3 : Segment-based Query Expansion

As explained in the previous section, our RQ2 investigation indicated that QE is not reliable for UGS retrieval due to the topic drift problem. In Chapter 7, we studied the use of the segmentation methods to improve the reliability of QE in UGS retrieval. Our experiments were designed to answer the following questions of RQ3.

1. Can having the segment evidence of the ASR transcripts be beneficial in improving the QE effectiveness for UGS?

2. What are the characteristics of the most effective speech evidence (i.e. speaker-based speech segments, full ASR document) for UGS retrieval?

3. Can we develop a technique to predict the most effective speech segmentation for each query?

Several novel QE techniques were presented based on three different speech segmentation methods, namely, semantic, discourse and window-based . To answer RQ3.1, we compared these QE approaches to traditional document-based QE. We found that, in general, all segment-based QE techniques have more reliability and robustness than traditional QE in UGS retrieval. In particular, our results demonstrate how segment-based QE approaches are more effective in dealing with term selection process for QE in UGS retrieval.

However, when we tried to find out which of these methods are the most effective to answer RQ3.2, we found that *none* of them is consistently effective for all queries in UGS retrieval. This can be attributed to the diverse structure and style of UGS documents, in which it is not clear how documents should be segmented in order to maximise retrieval effectiveness for each query.

For example, there might be a relevant document that does not need any segmentation because it contains a single topic, and segment-based QE can actually harm the retrieval effectiveness. Our experimental results suggest the need of an adap-

tive QE technique that takes the advantage of both segment and document-based approaches for each query to improve the overall effectiveness.

RQ3.3 : Query Performance Prediction For Query Expansion

Our analysis of QE for this task in the previous chapters indicated that the effectiveness is improved by predicting the right settings for expansion. Our experiments in Chapter 6, and Chapter 7 showed that selecting the most suitable field and segment evidence for expanding each query can lead to a significant improvement of retrieval effectiveness. The last question of RQ3, which is RQ3.3, specifically investigates the possibility of developing such a prediction approach for UGS retrieval. In Chapter 8, we utilised query performance prediction (QPP) methods to develop a probabilistic prediction framework for QE in UGS retrieval. Thus, our proposal was to use this framework to predict the best settings for QE in UGS retrieval.

We presented Weighted Expansion Gain (WEG) as a novel post-retrieval QPP method that utilises the information entropy of the feedback document to predict the QE performance. WEG predicts the performance of QE based on comparing the scores of the feedback documents that used for QE to that obtained by the non-feedback document in the top ranking list. Our experiments demonstrated the effectiveness of our proposed approach over other state-of-the-art QPP methods for QE not only on our UGS task but also on other retrieval test collections.

Finally, we demonstrated how our proposed QPP method can be utilised to improve QE effectiveness in UGS retrieval.

10.1.4 RQ4 : Adaptive CLIR for UGS retrieval

Successful exploration of the previous research questions motivated us to study the complementary problem of CL-UGS. Our initial CL-UGS experiments in Chapter 5 reported that higher translation quality is required to maintain effective cross-lingual performance. We investigated the uses of Moses MT as open-box system, and Google Translate as a black-box off-the-self MT system for this task. Our experimental

evaluations from the previous chapters revealed that both are ineffective and cause significant decrease in performance comparing to the mono-lingual one.

In Chapter 9, we performed an investigation towards answering RQ4 questions of this thesis, reproduced as follows.

1. Can we develop a prediction technique to estimate the translation effectiveness for cross-lingual UGS retrieval?

2. Can we implement an adaptive CLIR technique that is able select the most-effective translations in UGS retrieval?

To answer RQ4.1, we evaluated the effectiveness of using several state-of-the-art QPP approaches to predict translation effectiveness for CL-UGS. We presented a novel pre-retrieval QPP method for this task called AvgFL (Average Fluency) that is designed to detect misleading very unique and OOV words. For post-retrieval, we also presented the WRG (Weighted Relevancy Gain) QPP that is a modified version of the well-established WIG predictor (Zhou and Croft, 2007). WRG is based on estimating the information entropy of the relevant documents retrieved by a certain translation. Our experimental evaluation demonstrated the effectiveness of the proposed approaches in estimating translation performance for CL-UGS retrieval. Finally, towards answering RQ4.2, we presented adaptive CLIR methods that utilise our proposed QPP framework to select the most relevant translations in UGS retrieval.

10.2 Future Work

While this thesis has applied techniques of Query Expansion, Text Segmentation and Query Performance Prediction in UGS settings to improve the performance of UGS retrieval, there remain a number of avenues which we believe deserve further research. We summarise the areas of future directions into the following points.

- **UGS Retrieval Challenges**: In Chapter 5, we presented an initial investigation of the retrieval challenges UGS content. Our study presented an analysis of three main fields, namely, UGS ASR transcripts, Title and description meta data. Future direction in this area is to study the retrieval effectiveness and challenges of other evidence such as social tweets and comments, as well as the visual evidence of the UGS content.

  Furthermore, our analysis of the field weighting and combinations suggests that a *query-based weighting* of each field can potentially be beneficial for UGS retrieval. Potential venue for future work is to determine how to set up the weighting parameters for each field automatically using QPP, or using other technique such as Expectation Maximisation (EM techniques) (Moon, 1996) and Reinforcement Learning (RL) (van Hasselt et al., 2016).

- **Query Expansion For UGS retrieval** In this work, we investigated the application of standard QE methods in UGS retrieval. We presented adaptive QE methods using Query Performance Predictions (QPP) that is designed to predict the optimal setting to perform QE for each query. However, other QE methods worth investigating in UGS retrieval as follows.

  - Negative relevance feedback (Wang et al., 2008) : The QE methods presented in this work utilised the top ranked documents as relevant documents to improve the initial query in UGS retrieval. However, UGS retrieval may also benefit from negative relevance feedback which utilises non-relevant documents to improve the ranking of initial query. For example, Wang et al. (2008) performed QE based assigning more weight to a term with more occurrences in relevant documents and less weight or negative weight to a term with more occurrences in non-relevant documents. Future work on QE for UGS retrieval can benefit from investigating this approach to improve the retrieval effectiveness.

  - Word embedding similarities : Word embedding similarities can be utilised

to improve or even replace QE in UGS retrieval. For example, recent work by El Mahdaouy et al. (2018) showed that the effectiveness of Arabic information retrieval can be improved by incorporating word embedding semantic similarities scores into existing probabilistic IR models. A similar approach should be considered as potential further work related to our UGS retrieval task. This could be achieved by building a distributed word representation of UGS content using the Continuous Bag of Words (CBOW) and the skip gram models proposed in (Mikolov et al., 2013) to identify semantically similar terms from the collection and use them to expand the query in UGS retrieval.

  - Word embedding segmentation: Our work examined the utility of semantic, discourse and window speech segmentation for QE in UGS retrieval. Future work should study the utility of semantic segmentation using word embedding which has proven to be more effective than C99 and other semantic algorithms. Examining speech segmentation using word-embedding approaches, such as the one studied in (Alemi and Ginsparg, 2015), for QE in UGS retrieval would be an interesting area for further exploration of the segment-based QE.
  - Improved QPP prediction for QE : The adaptive QE approach presented in Chapter 8 is implemented using only the WEG QPP to preform the actual prediction. A potential work direction in this area is to improve the prediction quality for the adaptive QE. This could be done by implementing a modified version of this adaptive approach that relies on combining multiple QPP signals (i.e. both post- and pre-retrieval) to conduct the prediction.

- **Cross-Lingual UGS retrieval** Our investigation of CL-UGS retrieval in this thesis examined CLIR robustness of the ASR transcripts and UGS meta data. We also presented an approach to tune the quality of an MT system

for IR purposes and provide directions for measuring MT quality from an IR perspective using QPP approaches. Using the proposed QPP framework we presented an alternative CLIR approach for Query Translation (QT) in UGS settings.

Further investigation could be to study the Document Translation approach (DT) using QPP. DT CLIR can provide more contextual and native translation than QT (Oard and Hackett, 1997). However, DT is very costly since we need to translate all indexed documents/fields within corpus. This would be even more challenging in the context of video search where the translation of speech, metadata translation will result in noisy and incomplete data that would harm the retrieval effectiveness. However, as we reported in this thesis, not all UGS evidence is equally important for retrieval systems; usually some fields of information can be more important for maintaining higher IR effectiveness. Could QPP methods guide the MT system on what piece of information should be translated in high quality standard? and potentially the MT system has to put an additional efforts into maintaining the translation quality of these fields in order to maintain higher retrieval effectiveness.

Another area that is yet to be studied in cross-lingual UGS retrieval is how the final results should be presented to non-native users. Translating the search results of UGS includes translating noisy metadata and transcripts, which would be very expensive from resource perspective (Parton, 2012). Using QPP to find out which fields of the results are interesting to the user and should be translated would be an interesting research question for further work in this area.

10.3 Closing Remarks

The work presented in this thesis has opened potential new research directions for exploiting the novel problem of UGS retrieval. We believe the contribution of this

work would be beneficial for multiple audiences as follows.

- *Research and Knowledge* : This research is the first effort to explore a novel, and state-of-the-art problem of UGS retrieval. The research presented detailed analysis and insights into the problems of UGS retrieval that can guide upcoming research in SCR and IR in general. Further, the proposed retrieval framework of UGS content can make it easier for research to focus into deeper and more complex UGS problems. We believe that the efforts presented in this thesis will act as a baseline for further research not only in similar IR areas but also other related tasks for processing UGS content.

- *Online users* : The main motivation behind this work was to enable a better user-experience for online users to locate relevant UGS content on the Internet. Furthermore, the proposed cross-lingual retrieval framework is designed to enable content to be found no matter what language it is presented in. Our framework would also benefit content creators and producers by enabling their content to be accessible and available to a wider audience.

- *Industry* : Social media platforms and search engines are required to effectively process, and analyse uploaded UGS content in order to make them accessible to end users (i.e. consumer, viewer). The work of this thesis presented a prototype for the development of UGS retrieval application. We also demonstrated how techniques such as query expansion, and query performance prediction can be integrated to the system to improve user satisfaction. We believe that a successful implementation and integration of the proposed techniques would benefit current social media platforms not only within their search application but also within other modules which involves processing and retrieving UGS content. Finally, as ever more and richer UGS content is made available on social media platforms, the future of UGS search looks to be more promising. UGS search applications could be utilised not only to serve user's information needs, but also to serve business needs by enabling more commercialisation of

UGS content. We believe that the contributions presented in this thesis will guide upcoming advances in UGS-related technologies.

Appendices

Appendix A

Publications

The work presented in this thesis has formed the basis of research papers included in the proceedings of several peer-reviewed conference and referred journals. The work on understanding the retrieval challenges of UGS in **Chapter 5,6** appeared as long conference paper (Khwileh et al., 2015), and was later extended as full journal article (Khwileh et al., 2016).

The proposed segment-based QE method introduced in **Chapter 7** appeared in a long peer-reviewed workshop paper (Khwileh and Jones, 2016). The QPP approach described in **Chapter 8** also appeared as long conference paper (Khwileh et al., 2017b).

Finally, the adaptive Cl-UGS approach presented in **Chapter 9** appeared as long peer-reviewed workshop paper (Khwileh et al., 2017a).

1. **Khwileh, A.**, Ganguly, D., Jones, G. J.F (2015). An investigation of cross-language information retrieval for user-generated internet video. The Proceedings of International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 117-129). Springer.

2. **Khwileh, A.**, Ganguly, D., Jones, G. J. (2016). Utilisation of metadata fields and query expansion in cross-lingual search of user-generated internet video. Journal of Artificial Intelligence Research, 55, 249-281.

3. **Khwileh, A.**, Jones, G. J.F (2016). Investigating segment-based query expansion for user-generated spoken content retrieval. Proceeding of the 14th International Workshop on Content-Based Multimedia Indexing (CBMI), (pp. 1-6). IEEE.

4. **Khwileh, A.**, Afli, H., Jones, G.J.F., Way, A. (2017). Identifying Effective Translations for Cross-lingual Arabic-to-English User-generated Speech Search. In Proceedings of the Third Arabic Natural Language Processing Workshop (pp. 100-109).

5. **Khwileh, A.**, Way, A., Jones, G. J.F. (2017). Improving the Reliability of Query Expansion for User-Generated Speech Retrieval Using Query Performance Prediction. Proceedings of International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 43-56). Springer, Cham.

Appendix B

UGS query sets

In this appendix, we present the query sets (Mn-Ad, Mn-Kn, Cl-Fr, Cl-Fr-Moses, Cl-Ar, Cl-Ar-Moses), which were utilised for our UGS retrieval task throughout this thesis.

Details about each of these query sets can be found in Chapter 3, Section 4.5.

```
0  softwares for web development and design
1  business growth strategy
2  interviews with small business professionals
3  college classes and teachers
4  Troubleshooting PC and Laptops
5  about the systemic racism
6  annual social blogfest meet up
7  UK radio talk or TV show
8  comics science related subjects
9  Stock market rally
10  David talk about Web 2.0 for business and helping clients.
11  About Church and faith
12  Medical Marijuana and drugs.
13  comic books.
14  interpretations in the bible
15  domestic abuse and violence
16  unusual art and painting types
17  Poetry readings and Poems.
18  What is marketing.
19  learning photoshop
20  panel on hunger & homelessness
21  Green Party Presidential Candidate, Grit TV, politcal figure
22  how to start applications in safari
23  automatic emails of new content added
24  the game, Ultimate red skull villain
25  films made in 70's
26  chinese culture facts
27  community media coverage neighbormedia
28  video about facebook and social media
29  about google and search engine business
```

Figure B.1: Adhoc monolingual queries (Mn-Ad) : 0 - 30

```
30  web browser, flock
31  free speech radio talk
32  the future of world Economic
33  Local police calls
34  religious talks, interesting sermon
35  business opportunities joint ventures making money
36  clean and green energy
37  creating web sites
38  The Use of Technology in Our Daily Life
39  candidate representative of Constitution Party interview
40  about the food industries
41  state representative candidate talk
42  open source business concerns.
43  speech of US politician
44  Obama's campaign review
45  comedians discussing politics
46  Economic outlook
47  Tax situation in America..
48  todays digital media
49  games guide on how to play
50  american jobs factories
51  Religious talk about jesus
52  global warming, news, politics
53  oil and global warming issues
54  economics revenue business
55  Human trafficking
56  global economy and US politics
57  Unemployment Crisis in the US
58  female worker
59  Comical world news
```

Figure B.2: Adhoc monolingual queries (Mn-Ad) : 30 - 59.

1	Logiciels pour le développement et design web
2	Stratégie de croissance d'affaires
3	Interviews avec les professionnels des petites entreprises
4	Cours d'université et professeurs
5	Dépannage PC et ordinateurs portables
6	À propos du racisme systémique
7	Rencontre sociale annuelle blogfest
8	Émission TV ou radio au Royaume-Uni
9	Bandes dessinées scientifiques
10	Reprise boursière
11	David parle du Web 2.0 pour les entreprises et le service client
12	À propos de l'église et de la foi
13	Marijuana médicale et médicaments
14	Bandes dessinées
15	Interprétations dans la bible
16	Violence familiale
17	Types d'art et de peintures inhabituels
18	Lecture de poésie et poèmes
19	Qu'est ce que le marketing
20	Apprendre photoshop
21	Débat sur la faim et le problème des sans-abris
22	Candidat du parti vert aux présidentielles, sur Grit TV, figure politique
23	Comment lancer des programmes sur Safari
24	Emails automatiques du nouveau contenu ajouté
25	le jeu, ultime vilain à crâne rouge
26	Films des années 70
27	Données sur la culture chinoise
28	Couverture médiatique de la communauté par neighbormedia
29	Vidéo sur facebook et les réseaux sociaux
30	À propos de google et les affaires des moteurs de recherche

Figure B.3: French version of (Mn-Ad) : 1 - 30.

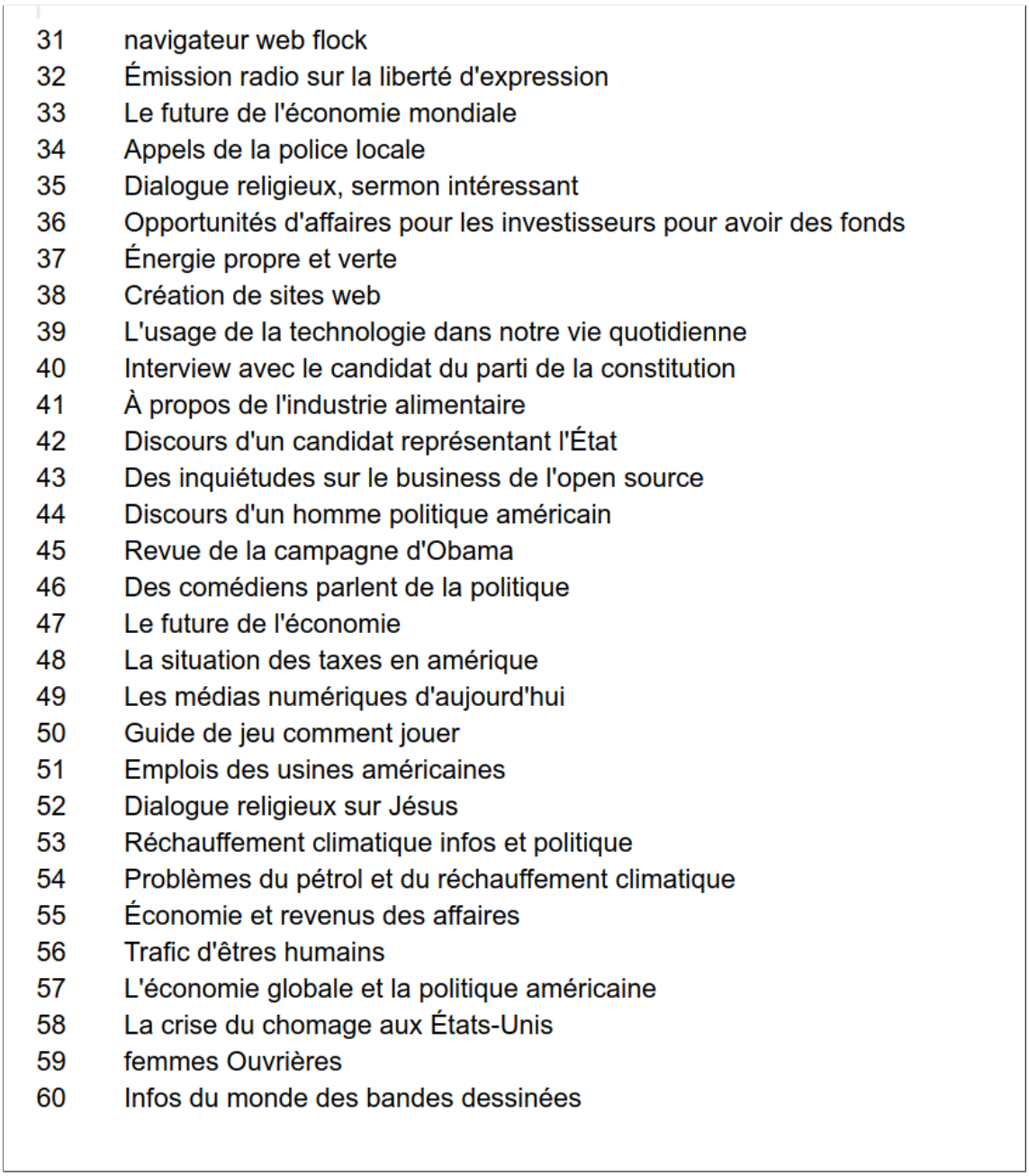

Figure B.4: French version of (Mn-Ad) : 31 - 60.

0	سوفتويرات لتصميم و برمجة مواقع الويب
1	استراتيجية تطوير الاعمال
2	مقابله مع رجال اعمال
3	محاضرات كليه و معلمين
4	حل مشاكل اللابتوبات و الدسكتوب
5	عن العنصريه المنتظمه
6	اجتماع اجتماعي للمدونين
7	محادثه برطانيه على الراديو او على التلفاز
8	كاريكتورات علميه
9	التسارع في اسواق الاسهم
10	ديفيد يتكلم عن اهميه الويب ٢.٠ للاعمال وخدمه الزبائن
11	عن الايمان و الكنائس
12	المريجوانا الطبيه و الادويه
13	كتب الكاريكتير
14	التفسيرات في الانجيل
15	العنف الاسري
16	رسم و فن غير عادي
17	شعر و شعراء
18	ماهو التسويق
19	تعلم الفوتشوب
20	مناظرة عن التسول
21	مرشح ا الحزب الاخضر للرئاسه، علا قرنيفي، شخضيه سياسه
22	كيفيه بدء برامج على سفاري
23	بعث ايميلات للمحتوى الجديد
24	لعبه الريد سكلل فيلان
25	افلام السبعينات
26	عن الحضاره الصينية
27	تغطيه اعمال جمعيه من خلال ال نيبورميديا
28	عن الفيسبوك و وسائل التواصل الاجتماعي
29	عن قوقل و اعمال محركات البحث

Figure B.5: Arabic version of (Mn-Ad) : 0 - 29.

30	المتصفح فلوك
31	برنامج عن حريه التعبير
32	مستقبل الاقتصاد العالمي
33	مكالمات الشرطه المحليه
34	حديث ديني و سيرمون ممتع
35	فرص اعمال للمستثمرمين لجلب الاموال
36	الطاقه البديله و النظيفه
37	عمل مواقع الكترونيه
38	استخدام التكنولوجيا في حياتنا اليوميه
39	مقابله مع مرشح الحزب الدستوري
40	عن صناعات الطعام
41	حديث احد المرشحين للتمثيل الولايه
42	بعض التحفظات على عمل الكود المفتوح
43	حديث احد السياسين الامريكيين
44	تقرير عن حملة اوباما
45	كوميدين يتكلمون عن السياسه
46	مستقبل الاقتصاد
47	وضع الضرائب في امريكيا
48	الاعلام الالكتروني اليوم
49	ارشادات لعبه كيفه اللعب
50	وظائف المصانع الامريكيه
51	حديث ديني عن النبي المسيح
52	الاحتباس الحراري اخبار و سياسه
53	مشاكل الاحتباس الحراري و النفط
54	الاقتصاد و ارباح الاعمال
55	تهريب البشر
56	الاقتصاد العالمي و السياسه الامريكيه
57	مصيبه البطاله في امريكا
58	اعمال المراه في المجتمع
59	اخبار و كاركتورات عالميه

Figure B.6: Arabic version of (Mn-Ad) : 30 - 59.

```
0 Its about Dreamweaver, an application for
building and improving websites.
1 Profit Partner program talks about growing business faster.
2 Curtis Baylor of Allstate gives a small piece of planning
advice for small business using his basic three factors.
3 Jeff Parker knows how to find a company to work for....
Monolith is great.
4 One of the biggest problems with the EEE PC laptop
and how to solve it.
5 IGE members talk about racism.
6 Its about an annual Brooklyn Blogfest
where bloggers and fans meet each other and have fun.
7 Hey guys, I thought this was pretty interesting to listen to.
Minus the fact it should be Judaism,and not Judism
(sounded like Druidism HAH)
I thought his reaction to the news of conversion
was pretty funny.
8 Its of serious comics on science related subjects.
9 Stock market rally (sudden rise in stock prices)
10 Listen to David Leeking tell the benefits of Web .
and how it can help connect all types of businesses
with their client base.
11 Experience of looking for a church
12 Medical Marijuana clinics in California.
13 This is the process a comic book goes through before
it's released.
14 Its about wrong impressions created by artists on Angels
and clarifies the authentic interpretation as per the Bible.
15 California to pass law intended to put an end to
domestic violence by outing the abusers in public.
16 What an unusual painting interview
17 This is a video that includes two different poets, both doing
readings of their work.
18 Marketing skills.
19 I found this clip simple but very helpful. I couldn't
remember how to create a new new pattern, but the steps were
pretty simple and easy to follow.
Hope it can help you guys out too! Enjoy.
20 Too Big to Fail composed by Austin Lounge Lizards.
21 Its a Grit TV presentation
on Green Party Presidential Candidate.
22 one easy step to start an application in safari
23 Sending automatic emails whenever
you add new content to blogs or web sites.
24 The launch of a new villain for the Avengers to fight with.
25 The Dark a film from
26 Brooklyn China Town Community Leader.
27 Neighbormedia at work covering community activities.
28 great talk about facebook
29 What Would Google Do By Jeff Jarvis
```

Figure B.7: Known-item monolingual queries (Mn-Kn) : 0 - 29.

```
30 This is some of the feature of flock, please see
the whole video to see the others.
31 A "free speech" radio commenting on the handling
and cover-ups.
32 Is a second great depression on its way?
33 Local police calls
34 Interesting Sermon
35 Profit Partner Show on Joint Ventures quickest way to
earn money. Fraizer O'lerry a real
estate business expert says he is going to
describe how to earn money through
joint ventures in a short period of time.
36 The video features a recent USA sanctioned clean energy Act.
37 This video helps you to make over your website
for more sales & recognition
38 The video invites interested persons to share
their knowledge on wood working technology,
tools, ideas etc.
39 An interview with Robert Owens Chairman
of the Ohio Constitution Party and
Candidate for Attorney General.
40 do you believe in the food industries
41 Bruce Dammeier is running for state representative.
42 The thing most disrupted by open source.
43 Hillary Clintons speech overview.
44 How much Obama spend on his election campaign?
45 George Carlin shares his thoughts
on politics and voting!
46 Economic wishful thinking
47 Current tax situation in united state...
48 Dancers have a chance to make new art and
money very easily in todays industry.
49 Scary board game video...dragon strike
50 How China is racing ahead of the U.S.
51 American pastor talks about
the relationship between Jesus and children.
52 Randy Hansen shares insight from others about
global topics, such as global warming
53 This is why I hate big oil companies.
54 Nevada economics could be impacted by
Thompson Grass Valley sales.
55 Human Trafficking in the US
56 Economic integration and it's
possible effects on United States politics.
57 the stupid bus trip
58 David Weber's Honor Harrington is
about the first ever female naval command.
59 World events juxtaposed with funny images.
```

Figure B.8: Known-item monolingual queries (Mn-Kn) : 30 - 59.

```
0  Software Development and Web Design
1  Business Growth Strategy
2  Interviews with professionals from small businesses
3  University courses and professors
4  PC Troubleshooting and laptops
5 About Systemic Racism
6 Annual social gathering blogfest
7 TV show or radio in the United Ai
8 Scientists Comics
9 Market recovery
10 David talks about Web 2.0 for companies and customer service
11 About the church and faith
12 Medical Marijuana and medicine
13 Comics
14 Interpretations in the Bible
15 Family Violence
16 Types of art and unusual paintings
17 Reading poetry and poems
18 What marketing
19 Learn photoshop
20 Debate on the issue of hunger and homelessness
21 Green Party candidate for president
on Grit TV, political figure
22 How to run programs on Safari
23 Automatic email of new content added
24 the game, the ultimate villain Red Skull
25 Movies of the 70s
26 Chinese culture Data
27 Media coverage of the community neighbormedia
28 Video on Facebook and social networks
29 About google and business search engine
```

Figure B.9: Google translated French CLIR queries (Cl-Fr) : 0 - 29.

30	Flock web browser
31	Radio program on freedom of expression
32	The future of the world economy
33	Calls to local police
34	Religious dialogue, interesting sermon
35	Business Opportunities for Investors to Have Funds"
36	Clean and green energy
37	Website creation
38	The use of technology in our daily lives
39	Interview with the candidate of the party constitution
40	About the food industry
41	Speech representing the candidate State
42	Concerns about the business of open source
43	Speech American politician
44	Review of the Obama campaign
45	Actors talk about politics
46	The future of the economy
47	The situation in America taxes
48	Digital media today
49	Guide how to play the game
50	Jobs US plants
51	Religious Dialogue on Jesus
52	Global warming and political information
53	Problems Crude Oil and global warming
54	Economic and Revenue Business
55	Human trafficking
56	The global economy and US policy
57	The crisis of unemployment in the US
58	Workers
59	Comics World News

Figure B.10: Google translated French CLIR queries (Cl-Fr) : 30 - 59.

```
0  Logiciels for the development and design web
1  Strategy of growth of business
2  Interviews with professionals from small businesses
3  courts of university and teachers
4  D pannage PC and laptops
5  the racism systemic
6  annual congress blogfest social
7   mission TV and radio in the United Kingdom
8  Bandes drawn scientists
9  stock market Reprise
10  David talking about the Web 2.0
for businesses and the service customer
11  speaking of church and faith
12  medical and medicines Marijuana
13  Bandes drawn
14  Interpr tations in the Bible
15  family Violence
16  Types of art and paints unusual
17  Lecture poetry and poems
18  what is the marketing
19  Learning photoshop
20  Debate on hunger and the problem of homeless
21  Candidat of the Green party in the presidential elections ,
on Grit TV , political figure
22  How launch programmes on Safari
23  automatic Emails new content added
24  the game , the final dirty to skull red
25  Films 1970s
26  Donn es on Chinese culture
27  media Couverture community by neighbormedia
28  Vid o on facebook and social networks
29  speaking of google and business search engines
```

Figure B.11: Moses translated French CLIR queries (Cl-Fr-Moses) : 0 - 29.

```
30  navigator web flock
31   mission on freedom of expression
32  the future of the world economy
33  Appels of local police
34  Dialogue religious , harangue interesting
35  for opportunities of cases for investors for funds
36  Energy clean and green
37  Establishment of websites
38  The use of technology in our everyday lives
39  Interview with the candidate of the constitution
40  speaking of the food industry
41  Mid-term address of a candidate representing the rule
42  there are concerns about the business of the open source
43  Mid-term address a US politician
44  Revue the campaign Obama
45  Plans com diens speak of politics
46  The future of the economy
47  The situation of taxes in am rique
48  the digital media today
49  Guide field how play
50  Emplois American factories
51  religious dialogue on Jesus
52  R chauffement infos and political climate
53  Problems of p trol and climate change
54  Economy and income business
55  the illicit traffic in human beings
56  The global economy and US policy
57  The crisis of chomage in the United States
58  women Ouvri res
59  Infos of the world gangs drawn
```

Figure B.12: Moses translated French CLIR queries (Cl-Fr-Moses) : 30 - 59.

0 سوفتويرات to design and his رمج web sites .
1 the strategy for developing business
2 for him with businessmen .
3 the lectures and teachers .
4 solving the problems of the laptop and الدسكتوب
5 the regular racism
6 a meeting of the bloggers social
7 speaking to برطانيه on the radio or on tv .
8 كاريكتورات scientific .
9 rushing in the stock markets
10 david talks about the importance of web @@2.0@@ for business and serving customers
11 the faith and the churches .
12 medical المريجوانا and medicine
13 comic books
14 in the bible explanations
15 violence prisoners .
16 drawing and art .
17 hair and poets .
18 is marketing .
19 know الفوتشوب
20 the debate about begging
21 , the green party candidate for the presidency , شخضيه , قرتيفي policy .
22 how to start the safari programs .
23 sent e-mails to the new content
24 الريد سكال vilan game
25 the seventies movies
26 the chinese civilization .
27 the acts of the association through the نيبورميديا
28 from facebook and the means of social networking
29 the google and acts of search engines .

Figure B.13: Moses translated Arabic CLIR queries (Cl-Ar-Moses) : 0 - 29.

30 look the browser .
31 program about the freedom of expression .
32 the future of the world economy
33 the local police phone calls
34 a religious and would throw interesting .
35 للمستثمرين business opportunities to get money .
36 clean alternative energy and
37 work electronic websites
38 use technology in the daily حياتنا
39 for him with the constitutional party 's candidate .
40 the food industries .
41 interview with one of the candidates for the state 's representation .
42 some reservations about the work of code open
43 interview with one of the american politicians ,
44 a report about the obama campaign .
45 كوميدين talk about politics .
46 the future of the economy
47 put taxes in america .
48 the electronic media today
49 guidelines on how to play a game .
50 american factories jobs .
51 a religious talk about the prophet messiah
52 and the policy of global warming .
53 the problems of global warming and oil .
54 the economy and the business profits .
55 smuggling human .
56 the world economy , and the american policy .
57 is a disaster unemployment in america .
58 of women in the society .
59 international news , and كاركتورات

Figure B.14: Moses translated Arabic CLIR queries (Cl-Ar-Moses) : 30 - 59.

Bibliography

Abdul-Rauf, S. and Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25(4):341–375.

Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y., and Itou, K. (2008). Test collections for spoken document retrieval from lecture audio data. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T., and Matsui, T. (2011). Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-9, National Center of Sciences, Tokyo, Japan, December 6-9, 2011*.

Alemi, A. A. and Ginsparg, P. (2015). Text segmentation based on semantic word embeddings. *arXiv preprint arXiv:1503.05543*.

Allan, J. (1995). Relevance feedback with too much data. In *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*, pages 337–343.

Allan, J., Ballesteros, L., Callan, J. P., Croft, W. B., and Lu, Z. (1995). Recent

experiments with INQUERY. In *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*.

Alqudsi, A., Omar, N., and Shaker, K. (2012). Arabic machine translation: a survey. *Artificial Intelligence Review*, pages 1–24.

Alqudsi, A., Omar, N., and Shaker, K. (2014). Arabic machine translation: a survey. *Artificial Intelligence Review*, 42(4):549–572.

Aly, R., Verschoor, T., and Ordelman, R. (2011). Utwente does rich speech retrieval at mediaeval 2011. In *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*.

Alzghool, M. and Inkpen, D. (2008). Cluster-based model fusion for spontaneous speech retrieval. In *Proceedings of the ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, pages 4–10. Citeseer.

Amati, G. (2003). *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Department of Computing Science, University of Glasgow.

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

Bagdouri, M., Oard, D. W., and Castelli, V. (2014). CLIR for informal content in arabic forum posts. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1811–1814.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum*, volume 31, pages 84–91. ACM.

Ballesteros, L. and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71. ACM.

Bellaachia, A. and Amor-Tijani, G. (2008). Enhanced query expansion in english-arabic clir. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, pages 61–66. IEEE.

Bendersky, M., Garcia-Pueyo, L., Harmsen, J., Josifovski, V., and Lepikhin, D. (2014). Up next: retrieval methods for large scale related video suggestion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1769–1778. ACM.

Bisazza, A. and Federico, M. (2016). A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Computational Linguistics*, 42(2):163–205.

BlipTV (2017). Bliptv. `https://web.archive.org/web/20120331073050/http://blip.tv/`. Retrieved: 2017-01-30.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Buckley, C., Allan, J., and Salton, G. (1993). Automatic routing and ad-hoc retrieval using SMART: TREC 2. In *Proceedings of The Second Text REtrieval Conference, TREC 1993, Gaithersburg, Maryland, USA, August 31 - September 2, 1993*, pages 45–56.

Buckley, C., Dimmick, D., Soboroff, I., and Voorhees, E. (2006). Bias and the limits of pooling. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 619–620. ACM.

Buckley, C. and Voorhees, E. M. (2000). Evaluating Evaluation Measure Stability. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece.*, pages 33–40, New York, NY, USA. ACM.

Büttcher, S., Clarke, C. L., and Cormack, G. V. (2010). *Information retrieval: Implementing and evaluating search engines*. MIT press.

Callan, J. P. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310. Springer-Verlag New York, Inc.

Carmel, D. and Yom-Tov, E. (2010). Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1.

Chelba, C., Bikel, D., Shugrina, M., Nguyen, P., and Kumar, S. (2012). Large scale language modeling in automatic speech recognition. *arXiv preprint arXiv:1210.8440*.

Chen, H.-H., Hueng, S.-J., Ding, Y.-W., and Tsai, S.-C. (1998). Proper name translation in cross-language information retrieval. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 232–236. Association for Computational Linguistics.

Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Schuster, M., Chen, Z., et al. (2018). The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 26–33.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2006). Precision prediction based on ranked list coherence. *Information Retrieval*, 9(6):723–755.

Cummins, R. (2012). Investigating performance predictors using monte carlo simulation and score distribution models. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1097–1098. ACM.

Darwish, K., Magdy, W., et al. (2014). Arabic information retrieval. *Foundations and Trends® in Information Retrieval*, 7(4):239–342.

Darwish, K. and Oard, D. W. (2003). Probabilistic structured query methods. In *Proceedings of SIGIR '03*, pages 338–344.

Eickhoff, C., Li, W., and de Vries, A. P. (2013). Exploiting user comments for audio-visual content indexing and retrieval. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, pages 38–49.

El Mahdaouy, A., El Alaoui, S. O., and Gaussier, E. (2018). Improving arabic information retrieval using word embedding similarities. *International Journal of Speech Technology*, 21(1):121–136.

Eskevich, M. (2014). *Towards effective retrieval of spontaneous conversational spoken content*. PhD thesis, Dublin City University.

Eskevich, M., Jones, G. J., Aly, R., Ordelman, R. J., Chen, S., Nadeem, D., Guinaudeau, C., Gravier, G., Sébillot, P., De Nies, T., et al. (2013). Multimedia information seeking through search and hyperlinking. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 287–294. ACM.

Eskevich, M., Jones, G. J., Wartena, C., Larson, M., Aly, R., Verschoor, T., and Ordelman, R. (2012a). Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–6. IEEE.

Eskevich, M. and Jones, G. J. F. (2014). Exploring speech retrieval from meetings using the AMI corpus. *Computer Speech & Language*, 28(5):1021–1044.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., Ordelman, R., and Larson, M. A. (2012b). Search and hyperlinking task at mediaeval 2012. In *Working Notes Proceedings of the MediaEval 2012 Workshop, Santa Croce in Fossabanda, Pisa, Italy, October 4-5, 2012*.

Eskevich, M., Magdy, W., and Jones, G. J. (2012c). New metrics for meaningful evaluation of informally structured speech retrieval. In *European Conference on Information Retrieval*, pages 170–181. Springer.

Facebook video (2017). Facebook. https://www.facebook.com/facebook/videos. Retrieved: 2017-01-30.

Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 167–174. ACM.

Federico, M., Bertoldi, N., Levow, G.-A., and Jones, G. J. F. (2005). Clef 2004 cross-language spoken document retrieval track. In *Multilingual Information Access for Text, Speech and Images*, pages 816–820. Springer.

Federico, M. and Jones, G. J. F. (2004). The clef 2003 cross-language spoken document retrieval track. In *Comparative Evaluation of Multilingual Information Access Systems*, pages 646–652. Springer.

Filippova, K. and Hall, K. B. (2011). Improved video categorization from text metadata and user comments. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 835–842. ACM.

Ganguly, D., Leveling, J., and Jones, G. J. (2013). An lda-smoothed relevance model for document expansion: a case study for spoken document retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1057–1060. ACM.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., and Huang, C. (2001). 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*, pages 96–104. ACM.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57.

Garofolo, J. S., Auzanne, C. G. P., and Voorhees, E. M. (2000). The TREC spoken document retrieval track: A success story. In *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 2000, 6th International Conference, College de France, France, April 12-14, 2000. Proceedings*, pages 1–20.

Garofolo, J. S., Voorhees, E. M., Auzanne, C. G., and Stanford, V. M. (1999a). Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio*.

Garofolo, J. S., Voorhees, E. M., Auzanne, C. G., Stanford, V. M., and Lund, B. A. (1999b). 1998 trec-7 spoken document retrieval track overview and results. In *Broadcast News Workshop*, volume 99, page 215.

Garofolo, J. S., Voorhees, E. M., Stanford, V. M., and Jones, K. S. (1997). TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of The Sixth Text REtrieval Conference, TREC 1997, Gaithersburg, Maryland, USA, November 19-21, 1997*, pages 83–91.

Glass, J., Hazen, T. J., Hetherington, L., and Wang, C. (2004). Analysis and processing of lecture audio data: Preliminary investigations. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, pages 9–12. Association for Computational Linguistics.

Glass, J. R., Hazen, T. J., Cyphers, D. S., Malioutov, I., Huynh, D., and Barzilay, R. (2007). Recent progress in the mit spoken lecture processing project. In *Interspeech*, pages 2553–2556.

Google (2017). Google translate. `https://translate.google.com/`. Retrieved: 2017-01-30.

Grefenstette, G. (1998). The problem of cross-language information retrieval. In *Cross-Language Information Retrieval*, pages 1–9. Springer.

Gu, Z. and Luo, M. (2004). Comparison of using passages and documents for blind relevance feedback in information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 482–483. ACM.

Gysel, C. V., de Rijke, M., and Worring, M. (2016). Unsupervised, efficient and semantic expertise retrieval. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 1069–1079.

Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. (2013). Morphological analysis and disambiguation for dialectal arabic. In *HLT-NAACL*, pages 426–432.

Harman, D. (1993). Overview of the second text retrieval conference (TREC-2). In *Proceedings of The Second Text REtrieval Conference, TREC 1993, Gaithersburg, Maryland, USA, August 31 - September 2, 1993*, pages 1–20.

Harman, D. (1994). Overview of the third text retrieval conference (TREC-3). In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, pages 1–20.

Hauff, C. (2010). *Predicting the Effectiveness of Queries and Retrieval Systems*. PhD thesis, University of Twente, Enschede, Netherlands.

Hauff, C., Hiemstra, D., and de Jong, F. (2008). A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 1419–1420.

He, B. and Ounis, I. (2004). Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval, 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004, Proceedings*, pages 43–54.

He, B. and Ounis, I. (2006). Query performance prediction. *Information Systems*, 31(7):585–594.

He, B. and Ounis, I. (2007). Combining fields for query expansion and adaptive query expansion. *Inf. Process. Manage.*, 43(5):1294–1307.

He, B. and Ounis, I. (2009). Studying query expansion effectiveness. In *Advances in Information Retrieval*, pages 611–619. Springer.

He, D. and Wu, D. (2008). Translation enhancement: a new relevance feedback method for cross-language information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 729–738.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Herbert, B., Szarvas, G., and Gurevych, I. (2011). Combining query translation techniques to improve cross-language information retrieval. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 712–715.

Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, Netherlands.

Hull, D. A. (1993). 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. Pittsburgh, PA, USA, June 27 - July 1, 1993*, pages 329–338.

Inkpen, D., Alzghool, M., Jones, G. J. F., and Oard, D. W. (2006). Investigating cross-language speech retrieval for a spontaneous conversational speech collection. In *Proceedings of the Human Language Technology Conference of the NAACL,*

Companion Volume: Short Papers, pages 61–64. Association for Computational Linguistics.

Internetworldstats.com (2017). Internet world users by language top 10 languages. `http://www.internetworldstats.com/stats7.htm`. Retrieved: 2017-01-04.

James, D. A. (1995). *The application of classical information retrieval techniques to spoken documents*. PhD thesis, Citeseer.

Jones, G. J., Zhang, K., Newman, E., and Lam-Adesina, A. M. (2007). Examining the contributions of automatic speech transcriptions and metadata sources for searching spontaneous conversational speech.

Jones, G. J. F. and Edens, R. J. (2002). Automated alignment and annotation of audio-visual presentations. In *Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002, Proceedings*, pages 276–291.

Jones, G. J. F., Foote, J. T., Jones, K. S., and Young, S. J. (1996). Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 30–38.

Jones, K. S. (1973). Index term weighting. *Information Storage and Retrieval*, 9(11):619–633.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models.

In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Khwileh, A., Afli, H., Jones, G. J., and Way, A. (2017a). Identifying effective translations for cross-lingual arabic-to-english user-generated speech search. *WANLP 2017 (co-located with EACL 2017)*, page 100.

Khwileh, A., Ganguly, D., and Jones, G. J. (2015). An investigation of cross-language information retrieval for user-generated internet video. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 117–129. Springer.

Khwileh, A., Ganguly, D., and Jones, G. J. (2016). Utilisation of metadata fields and query expansion in cross-lingual search of user-generated internet video. *Journal of Artificial Intelligence Research*, 55:249–281.

Khwileh, A. and Jones, G. J. F. (2016). Investigating segment-based query expansion for user-generated spoken content retrieval. In *14th International Workshop on Content-Based Multimedia Indexing, CBMI 2016, Bucharest, Romania, June 15-17, 2016*, pages 1–6.

Khwileh, A., Way, A., and Jones, G. J. F. (2017b). Improving the reliability of query expansion for user-generated speech retrieval using query performance prediction. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*, pages 43–56.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Kishida, K. and Kando, N. (2006). *A hybrid approach to query and document translation using a pivot language for cross-language information retrieval*. Springer.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007a). Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007b). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*.

Koopman, B. and Zuccon, G. (2014). Relevation: an open source system for information retrieval relevance assessment. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 1243–1244.

Kurland, O., Shtok, A., Carmel, D., and Hummel, S. (2011). A unified framework for post-retrieval query-performance prediction. In *Conference on the Theory of Information Retrieval*, pages 15–26. Springer.

Kurland, O., Shtok, A., Hummel, S., Raiber, F., Carmel, D., and Rom, O. (2012). Back to the roots: a probabilistic framework for query-performance prediction. In

21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012, pages 823–832.

Lam-Adesina, A. M. and Jones, G. J. (2006). *Dublin city university at CLEF 2005: cross-language speech retrieval (CL-SR) experiments*. Springer.

Lamel, L. and Gauvain, J. (2008). Speech processing for audio indexing. In *Advances in Natural Language Processing, 6th International Conference, GoTAL 2008, Gothenburg, Sweden, August 25-27, 2008, Proceedings*, pages 4–15.

Langlois, T., Chambel, T., Oliveira, E., Carvalho, P., Marques, G., and Falcão, A. (2010). Virus: video information retrieval using subtitles. In *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 197–200. ACM.

Larson, M. and Jones, G. J. (2012). Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(45):235–422.

Larson, M., Newman, E., and Jones, G. J. F. (2009). Overview of videoclef 2008: Automatic generation of topic-based feeds for dual language audio-visual content. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 906–917. Springer.

Larson, M., Newman, E., and Jones, G. J. F. (2010). Overview of videoclef 2009: New perspectives on speech-based multimedia content enrichment. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 354–368. Springer.

Le, V. B., Lamel, L., and Gauvain, J. (2010). Multi-style MLP features for BN transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, pages 4866–4869.

Lee, C.-J., Chen, C.-H., Kao, S.-H., and Cheng, P.-J. (2010). To translate or not to translate? In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 651–658. ACM.

Lee, C.-J. and Croft, W. B. (2014). Cross-language pseudo-relevance feedback techniques for informal text. In *Advances in Information Retrieval*, pages 260–272. Springer.

Lee, D. and Lee, G. G. (2008). A korean spoken document retrieval system for lecture search. *CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG*, page 73.

Lee, H. and Lee, L. (2014). Improved semantic retrieval of spoken content by document/query expansion with random walk over acoustic similarity graphs. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 22(1):80–94.

Leveling, J., Zhou, D., Jones, G. J. F., and Wade, V. (2009). Document expansion, query translation and language modeling for ad-hoc IR. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, pages 58–61.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 22(4):745–777.

Liu, X. and Croft, W. B. (2002). Passage retrieval based on language models. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 375–382.

Liu, Y., Liu, Q., and Lin, S. (2006). Tree-to-string alignment template for statistical machine translation. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*.

Lo, W. K., Li, Y., Levow, G., Wang, H., and Meng, H. M. (2003). Multi-scale document expansion in english-mandarin cross-language spoken document retrieval. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*.

Locke, W. N. and Booth, A. D. (1955). *Machine translation of languages: fourteen essays*. Published jointly by Technology Press of the Massachusetts Institute of Technology and Wiley, New York.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8.

Luong, M.-T. (2016). *NEURAL MACHINE TRANSLATION*. PhD thesis, Department of Computer Science, Stanford University.

Lv, Y. and Zhai, C. (2011). Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 7–16.

Macdonald, C., Plachouras, V., He, B., Lioma, C., and Ounis, I. (2005). University of glasgow at webclef 2005: Experiments in per-field normalisation and language specific stemming. In *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, pages 898–907.

Magdy, W. (2011). *Toward Higher Effectiveness for Recall-Oriented Information Retrieval: A Patent Retrieval Case Study*. PhD thesis, School of Computing, Dublin City University.

Magdy, W. and Jones, G. J. (2014). Studying machine translation technologies for large-data clir tasks: a patent prior-art search case study. *Information Retrieval*, 17(5-6):492–519.

Masumura, R., Hahm, S., and Ito, A. (2011). Language model expansion using webdata for spoken document retrieval. In *INTERSPEECH*, pages 2133–2136.

McCarley, J. S. (1999). Should we translate the documents or the queries in cross-language information retrieval? In *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999.*

MediaEval (2017). MediaEval Benchmarking Initiative for Multimedia Evaluation. `http://www.multimediaeval.org/`. Retrieved: 2017-01-03.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitra, M., Singhal, A., and Buckley, C. (1998a). Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–214. ACM.

Mitra, M., Singhal, A., and Buckley, C. (1998b). Improving automatic query expansion. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 206–214.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). Meetings about meetings: research at ICSI on speech in multiparty conversations. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, pages 740–743.

Mothe, J. and Tanguy, L. (2005). Linguistic features to predict query difficulty. In *ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop*, pages 7–10.

Naaman, M. (2012). Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. *Multimedia Tools and Applications*, 56(1):9–34.

Neubig, G. (2017). Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*.

Nikoulina, V., Kovachev, B., Lagos, N., and Monz, C. (2012). Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 109–119. Association for Computational Linguistics.

Oard, D. W. and Diekema, A. R. (1998). Cross-language information retrieval. *Annual review of information science and technology*, 33:223–256.

Oard, D. W. and Hackett, P. G. (1997). Document translation for cross-language text retrieval at the university of maryland. In *Proceedings of The Sixth Text REtrieval Conference, TREC 1997, Gaithersburg, Maryland, USA, November 19-21, 1997*, pages 687–696.

Oard, D. W., Wang, J., Jones, G. J. F., White, R. W., Pecina, P., Soergel, D., Huang, X., and Shafran, I. (2006). Overview of the CLEF-2006 cross-language speech retrieval track. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, pages 744–758.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational*

Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan., pages 160–167.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Over, P., Awad, G. M., Fiscus, J., Antonishek, B., Michel, M., Smeaton, A. F., Kraaij, W., and Quénot, G. (2011). Trecvid 2010–an overview of the goals, tasks, data, evaluation mechanisms, and metrics.

Over, P., Ianeva, T., Kraaij, W., and Smeaton, A. F. (2005). Trecvid 2005-an overview,trecvid 2005 workshop notebook papers, gaithersburg, md, usa.

Parton, K., McKeown, K., Allan, J., and Henestroza, E. (2008). Simultaneous multilingual search for translingual information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 719–728.

Parton, K. P. (2012). *Lost and Found in Translation: Cross-Lingual Question Answering with Result Translation*. Columbia University.

Pecina, P., Hoffmannová, P., Jones, G. J., Zhang, Y., and Oard, D. W. (2008). Overview of the CLEF 2007 Cross-Language Speech Retrieval Track. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 674–686.

Pecina, P., Hoffmannová, P., Jones, G. J. F., Zhang, Y., and Oard, D. W. (2007). Overview of the CLEF-2007 cross language speech retrieval track. In *Working Notes for CLEF 2007 Workshop co-located with the 11th European Conference on Digital Libraries (ECDL 2007), Budapest, Hungary, September 19-21, 2007*.

Pérez-Iglesias, J. and Araujo, L. (2010). Standard deviation as a query hardness estimator. In *SPIRE*, volume 10, pages 207–212. Springer.

Peters, C., Braschler, M., and Clough, P. D. (2012). *Multilingual Information Retrieval - From Research To Practice*. Springer.

Peters, C., Gonzalo, J., Braschler, M., and Kluck, M., editors (2004). *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers*, volume 3237 of *Lecture Notes in Computer Science*. Springer.

Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Inf. Retr.*, 4(3-4):209–230.

Plachouras, V., He, B., and Ounis, I. (2004). University of glasgow at TREC 2004: Experiments in web, robust, and terabyte tracks with terrier. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Qu, Y., Grefenstette, G., and Evans, D. A. (2002). Resolving translation ambiguity using monolingual corpora. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 223–241. Springer.

Renals, S., Hain, T., and Bourlard, H. (2008). Interpretation of multiparty meetings the ami and amida projects. In *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, pages 115–118. IEEE.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.

Robertson, S. E. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1998). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In *Proceedings of The Seventh*

Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998, pages 199–210.

Robertson, S. E., Walker, S., Jones, S., and Hancock-Beaulieu, M. (1994). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*. NIST.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In *The SMART retrieval system – Experiments in automatic document processing*. Prentice Hall.

Rogati, M. and Yang, Y. (2001). Cross-lingual pseudo-relevance feedback using a comparable corpus. In *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers*, pages 151–157.

Rose, R. (1991). Techniques for information retrieval from speech messages. *Lincoln Laboratory Journal*, 4(1):45–60.

Rudinac, S., Larson, M., and Hanjalic, A. (2009). Exploiting visual reranking to improve pseudo-relevance feedback for spoken-content-based video retrieval. In *10th Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2009, London, United Kingdom, May 6-8, 2009*, pages 17–20.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sanderson, M. (1994). Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 142–151.

Santos, R. L. T., McCreadie, R., and Plachouras, V. (2011). Large-scale information retrieval experimentation with terrier. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 2601–2602.

Schmiedeke, S., Kofler, C., and Ferrané, I. (2012). Overview of the mediaeval 2012 tagging task. In *Working Notes Proceedings of the MediaEval 2012 Workshop, Santa Croce in Fossabanda, Pisa, Italy, October 4-5, 2012*.

Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, M. A., Estève, Y., Lamel, L., Jones, G. J. F., and Sikora, T. (2013). Blip10000: a social video dataset containing SPUG content for tagging and retrieval. In *Multimedia Systems Conference 2013, MMSys '13, Oslo, Norway, February 27 - March 01, 2013*, pages 96–101.

Scholer, F., Williams, H. E., and Turpin, A. (2004). Query association surrogates for web search. *JASIST*, 55(7):637–650.

Shakery, A. and Zhai, C. (2013). Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs. *Inf. Retr.*, 16(1):1–29.

Shannon, C. E. (2001). A mathematical theory of communication. *Mobile Computing and Communications Review*, 5(1):3–55.

Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 58–65.

Shtok, A., Kurland, O., Carmel, D., Raiber, F., and Markovits, G. (2012). Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.*, 30(2):11:1–11:35.

Singhal, A. (1997). *Term Weighting Revisited*. PhD thesis, Cornell University, Department of computer science.

Singhal, A. and Pereira, F. C. N. (1999). Document expansion for speech retrieval. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 34–41.

Singhal, A., Salton, G., Mitra, M., and Buckley, C. (1996). Document length normalization. *Inf. Process. Manage.*, 32(5):619–633.

Smeaton, A. F., Over, P., and Kraaij, W. (2006). Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2006, October 26-27, 2006, Santa Barbara, California, USA*, pages 321–330.

Sokolov, A., Hieber, F., and Riezler, S. (2014). Learning to translate queries for CLIR. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 1179–1182.

Sparck-Jones, K., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6):779–840.

Stark, L. A., Whittaker, S., and Hirschberg, J. (2000). ASR satisficing: the effects of ASR accuracy on speech retrieval. In *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, pages 1069–1072.

Stroppa, N. and Way, A. (2006). MATREX: DCU machine translation system for IWSLT 2006. In *2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006*, pages 31–36.

Terol, R. M., Palomar, M., Martínez-Barco, P., Llopis, F., Muñoz, R., and Noguera, E. (2005). The university of alicante at CL-SR track. In *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, pages 769–772.

Terra, E. L. and Warren, R. (2005). Poison pills: harmful relevant documents in feedback. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 319–320.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218.

Toderici, G., Aradhye, H. B., Pasca, M., Sbaiz, L., and Yagnik, J. (2010). Finding meaning on youtube: Tag recommendation and category discovery. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3447–3454.

Tong, X., Zhai, C., Milic-Frayling, N., and Evans, D. A. (1996). OCR correction and query expansion for retrieval on OCR data – CLARIT TREC-5 confusion track report. In *Proceedings of The Fifth Text REtrieval Conference, TREC 1996, Gaithersburg, Maryland, USA, November 20-22, 1996*.

Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016). Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*.

Ture, F. (2013). *Searching to translate and translating to search: When information retrieval meets machine translation*. PhD thesis, University of Maryland, College Park.

van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with

double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2094–2100.

Varshney, S. and Bajpai, J. (2014). Improving performance of english-hindi cross language information retrieval using transliteration of query terms. *CoRR*, abs/1401.3510.

Vinay, V., Cox, I. J., Milic-Frayling, N., and Wood, K. R. (2006). On ranking the effectiveness of searches. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 398–404.

Volkmer, T. and Natsev, A. (2006). Exploring automatic query refinement for text-based video retrieval. In *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006, July 9-12 2006, Toronto, Ontario, Canada*, pages 765–768.

Voorhees, E. M. (2003a). Overview of the TREC 2003 question answering track. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, pages 54–68.

Voorhees, E. M. (2003b). Overview of the TREC 2003 question answering track. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, pages 54–68.

Voorhees, E. M. (2004). Overview of the TREC 2004 robust track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M. (1996). Intelligent access to digital video: Informedia project. *IEEE Computer*, 29(5):46–52.

Wang, J. and Oard, D. W. (2005). CLEF-2005 CL-SR at maryland: Document and query expansion using side collections and thesauri. In *Accessing Multilingual*

Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers, pages 800–809.

Wang, X., Fang, H., and Zhai, C. (2008). A study of methods for negative relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 219–226. ACM.

Wartena, C. (2012). Comparing segmentation strategies for efficient video passage retrieval. In *10th International Workshop on Content-Based Multimedia Indexing, CBMI 2012, Annecy, France, June 27-29, 2012*, pages 1–6.

White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., and Huang, X. (2005). Overview of the CLEF-2005 cross-language speech retrieval track. In *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, pages 744–759.

Wilkinson, R. (1994). Effective retrieval of structured documents. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 311–317.

Willett, P. (2006). The porter stemming algorithm: then and now. *Program*, 40(3):219–223.

Woodland, P. C., Johnson, S. E., Jourlin, P., and Jones, K. S. (2000). Effects of out of vocabulary words in spoken document retrieval. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, pages 372–374.

Wrede, B. and Shriberg, E. (2003). Relationship between dialogue acts and hot spots in meetings. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 180–185. IEEE.

Wurzer, D., Osborne, M., and Lavrenko, V. (2016). Randomised relevance model. *CoRR*, abs/1607.02641.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM.

Yom-Tov, E., Fine, S., Carmel, D., and Darlow, A. (2005). Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 512–519.

Youtube (2017). Youtube. `http://www.youtube.com/`. Retrieved: 2017-01-30.

YouTube Press (2016). Statistics - YouTube. `http://www.youtube.com/yt/press/statistics.html`. Retrieved: 2016-11-30.

Zhang, J. (2017). *Domain adaptation for statistical machine translation and neural machine translation*. PhD thesis, Dublin City University.

Zhao, Y., Scholer, F., and Tsegay, Y. (2008). Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, pages 52–64.

Zhou, D., Truran, M., Brailsford, T. J., Wade, V., and Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Comput. Surv.*, 45(1):1:1–1:44.

Zhou, Y. (2008). *Retrieval performance prediction and document quality*. University of Massachusetts Amherst.

Zhou, Y. and Croft, W. B. (2006). Ranking robustness: a novel framework to predict query performance. In *Proceedings of the 2006 ACM CIKM International Con-*

ference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006, pages 567–574.

Zhou, Y. and Croft, W. B. (2007). Query performance prediction in web search environments. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 543–550.

Zobel, J. and Moffat, A. (2006). Inverted files for text search engines. *ACM Comput. Surv.*, 38(2):6.