Journal of Artificial Intelligence Research 55 (2016) 249-281 Submitted 03/15; published 01/16

Utilisation of Metadata Fields and Query Expansion in Cross-Lingual Search of User-Generated Internet Video

Ahmad Khwileh AHMAD.KHWILEH2@MAIL.DCU.IE
Debasis Ganguly DGANGULY@COMPUTING.DCU.IE
Gareth J. F. Jones GJONES@COMPUTING.DCU.IE
ADAPT Centre, School of Computing
Dublin City University
Dublin 9, Ireland

Abstract

Recent years have seen significant efforts in the area of Cross Language Information Retrieval (CLIR) for text retrieval. This work initially focused on formally published content, but more recently research has begun to concentrate on CLIR for informal social media content. However, despite the current expansion in online multimedia archives, there has been little work on CLIR for this content. While there has been some limited work on Cross-Language Video Retrieval (CLVR) for professional videos, such as documentaries or TV news broadcasts, there has to date, been no significant investigation of CLVR for the rapidly growing archives of informal user generated (UGC) content. Key differences between such UGC and professionally produced content are the nature and structure of the textual UGC metadata associated with it, as well as the form and quality of the content itself. In this setting, retrieval effectiveness may not only suffer from translation errors common to all CLIR tasks, but also recognition errors associated with the automatic speech recognition (ASR) systems used to transcribe the spoken content of the video and with the informality and inconsistency of the associated user-created metadata for each video. This work proposes and evaluates techniques to improve CLIR effectiveness of such noisy UGC content. Our experimental investigation shows that different sources of evidence, e.g. the content from different fields of the structured metadata, significantly affect CLIR effectiveness. Results from our experiments also show that each metadata field has a varying robustness to query expansion (QE) and hence can have a negative impact on the CLIR effectiveness. Our work proposes a novel adaptive QE technique that predicts the most reliable source for expansion and shows how this technique can be effective for improving CLIR effectiveness for UGC content.

1. Introduction

Increasing amounts of user generated multilingual video content (UGC) are being uploaded to social video-sharing websites such as Youtube (2015), Facebook (Facebook video, 2015), BlipTv (2015) and many others. In 2015, YouTube, the predominant online video sharing site, reported that 300 hours of video content were being uploaded every minute in 61 different languages (YouTube Press, 2015). The ease and the flexibility of video content production, coupled with the low cost of publishing and wide potential reach, are resulting in an exponential growth in the number of videos available on the Web. At the same time, due to increasing user demands for accessing and viewing this content, it is very important to manage it in such a way so as to facilitate effective and efficient access to it. The very large amounts of this content are creating the need for the development of sophisticated video retrieval systems, presenting new challenges and exciting opportunities for Information Retrieval (IR) research (Bendersky, Garcia-Pueyo, Harmsen, Josifovski, & Lepikhin, 2014; Naaman, 2012).

One of the key challenges for the effective exploitation of UGC content in a multilingual setting is effective search between the languages of the user queries and the content metadata. From the multilingual perspective, the quality of UGC depends solely on the characteristics of the individuals who actually produce or upload videos in each language. This lack of formal editorial control means that the uploaded videos are typically very varied across languages in terms of audio, visual and metadata quality. Moreover, *the quantity and topical coverage* of this content across different languages is very uneven, often meaning 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, this is even more likely for video material where little UGC Arabic content is currently available in certain topics such as cultural or historical topics. This in fact is a classical use-case for Cross-Language Information Retrieval (CLIR) which seeks to enable users to enter search queries in one language to retrieve relevant content in another one. Translation technologies are key to successfully bridging the language gap between a user's query and the relevant content (Oard & Diekema, 1998; Herbert, Szarvas, & Gurevych, 2011).

The quality of monolingual search over UGC content is dependent on the effective utilization of the available metadata. CLIR search effectiveness will further depend on translation quality between query and content languages. There are many potential choices for how to design a robust CLIR framework for an Internet video search task, but the current lack of detailed investigation means that there is a lack of understanding of the specific challenges it represents and thus little or no guidance available for the choices that should be made in developing such a framework.

In this paper, we investigate CLIR search effectiveness over an archive of user-generated Internet video content originally used for the MediaEval 2012 Search and Hyperlinking task (Eskevich, Jones, Chen, Aly, Ordelman, & Larson, 2012a), which we extend to a CLIR task. We examine retrieval effectiveness using the *title* and the *description* metadata provided by the video uploader and automatic speech recognition (ASR) transcripts of the content. We further investigate the application of automatic query expansion on each source for improving the CLIR retrieval performance. Retrieval and query expansion are carried out using the *Divergence From Randomness* (DFR) IR model, and automatic translation is carried out using *Google Translate* (2015). To understand the task better, we undertake a detailed performance analysis examining the impact of different source metadata information on CLIR behaviour. However, our current investigation is limited to the application of state-of-the-art Machine Translation (MT) and information retrieval (IR) methods to this task, in order to establish the basis for further investigations.

The remainder of this paper is structured as follows: **Section 2** gives some general background on CLIR, **Section 3** reviews related work, **Section 4** describes the test set used in our experiments and the evaluation metric, **Section 5** describes initial retrieval experiments examining the relative CLIR effectiveness of each source of evidence (ASR, Title and Description), **Section 6** describes our approach to improving CLIR effectiveness using careful adjustment of the retrieval algorithm setting, **Section7** describes our approach to improving CLIR effectiveness using automatic query expansion techniques, **Section 8** concludes the paper and provides directions for further work.

2. Cross Language Information Retrieval

As stated previously, the goal of CLIR is to satisfy a user information need expressed as a query in one language using a content from another language. CLIR techniques use translation to bridge this language barrier between the query and the indexed content. These techniques differ mainly as to where the translation module is to be placed, either in the query processing or the document indexing stage. Figure 1 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

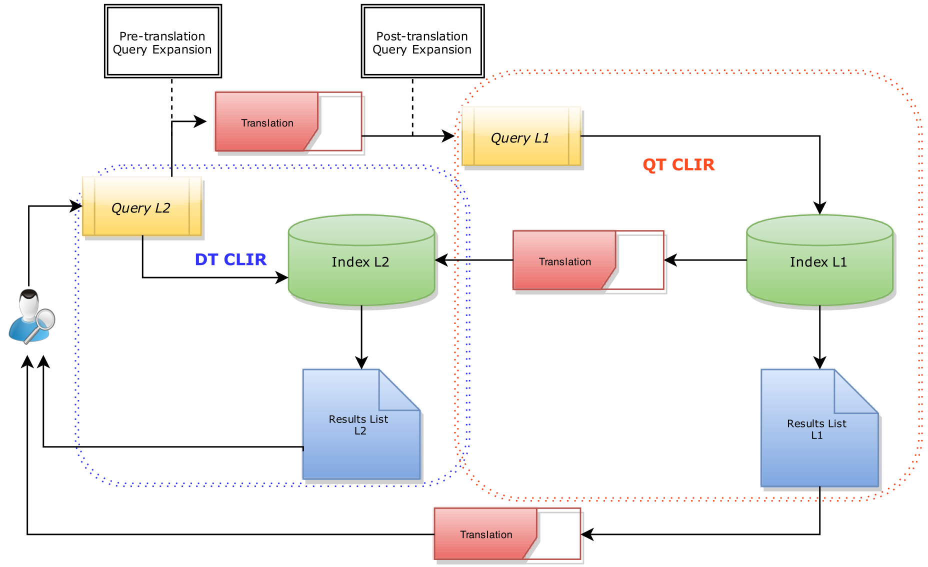


Figure 1: Document and Query based CLIR Techniques.

CLIR technique (Oard & Diekema, 1998; Herbert et al., 2011; Sokolov, Hieber, & Riezler, 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. Yet 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. Previous literature has explored multiple techniques to overcome these issues either by improving the translation quality using various translation techniques (Chen, Hueng, Ding, & Tsai, 1998; Gao, Nie, Xun, Zhang, Zhou, & Huang, 2001; Varshney & Bajpai, 2014; Lee, Chen, Kao, & Cheng, 2010) or by improving the query itself using query reformulation techniques such as Query Expansion (QE) in a Relevance Feedback (RF) process (Carpineto & Romano, 2012).

QE in RF operates by selecting terms from documents which have been marked as relevant by the user and adding them to the original query. In the absence of user created relevance data, the top ranked documents can be assumed relevant in a process often referred to as pseudo relevance feedback (PRF). Although noisy, PRF has shown to be effective for improving the overall retrieval effectiveness (Bellaachia & Amor-Tijani, 2008). In cross-lingual settings, the query can be expanded before translation to provide a more effective query for translation using a pre-translation QE technique (Ballesteros & 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 1, in order to alleviate the impact on IR effectiveness arising from translation problems (Ballesteros & Croft, 1998; Rogati & Yang, 2002).

The alternative to QT CLIR is Document Translation (DT CLIR) where all documents in the collection are translated to the query language (Oard & Hackett, 1998; Lee & Croft, 2014). Several arguments suggest that DT should be competitive or superior to QT CLIR for some tasks, due to the fact that it is less sensitive to translation errors. DT has the advantage that all translation is carried out offline prior to 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 1, since documents are already translated at index time. However, while DT CLIR has been shown to be effective for several tasks, its application in CLIR settings is often 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 CLIR technique which has shown to be effective, 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 (McCarley, 1999; Kishida & Kando, 2006; Parton, McKeown, Allan, & Henestroza, 2008).

Several approaches have been proposed to carry the translation process in the CLIR framework. The most commonly used ones are bilingual dictionaries and machine translation (MT) (Zhou, Truran, Brailsford, Wade, & Ashman, 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, Hedlund, Keskustalo, & Järvelin, 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 automatic translation of free-text from one natural language to another (Nikoulina, Kovachev, Lagos, & Monz, 2012; Magdy & Jones, 2014). While MT can also have similar dictionary coverage problems, the creation of a 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 off-the-shelf MT systems. Most CLIR research has dealt with the translation module as a black-box without any control over the translation process, and rather used one of the freely available online translation tools such as Google Translate (2015), Bing translate (2015) and others, which have proven to be effective. For example, in the CLEF evaluation campaigns 2009 (CLEF, 2015a), the best performing non-Google MT system achieved just 70% of the performance achieved by Google Translate tool (Leveling, Zhou, Jones, & Wade, 2010; Zhou et al., 2012).

For this experimental investigation, we choose the default and most common CLIR settings, which is a QT CLIR technique that utilises the Google MT translate tool. We nevertheless, plan to explore other CLIR settings (e.g. DT-CLIR with open-box MT system) in our future work.

3. Related Work

Several CLIR tasks have been explored across different domains and document types (Peters, Braschler, & Clough, 2012). The most closely related CLIR work to that examined in this research was carried out in tasks within the CLEF evaluation campaigns on professionally generated video content (2015b).

From 2002-2004 the Cross-Language Spoken Document Retrieval (CL-SDR) task investigated news story document retrieval using data from the NIST TREC 8-9 Spoken Document Retrieval (SR) with manually translated queries (Federico & Jones, 2004; Federico, Bertoldi, Levow, & Jones, 2005). The aim of these tasks was to evaluate 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 & Jones, 2004), while using dictionary term-by-term translation, this reduction increased to between about 40% and 60%, which highlights the challenge for CLIR over video collections (Federico et al., 2005).

A more ambitious Cross-Language Speech Retrieval (CL-SR) task ran within CLEF 2005-2007 (White, Oard, Jones, Soergel, & Huang, 2006; Oard, Wang, Jones, White, Pecina, Soergel, Huang, & Shafran, 2007; Pecina, Hoffmannová, Jones, Zhang, & Oard, 2008). This examined CLIR for a spontaneous conversational speech oral history collection with content in English and Czech. The tasks provided ASR transcripts, automatically and manually generated metadata for the interviews. The goal for both Czech and English tasks was to create systems that could help monolingual and cross lingual searchers identify sections of an interview that they wish to listen to on Czech and English interviews. The reported results of these tasks showed that the use of manual metadata yielded substantial and statistically significant improvement in retrieval effectiveness over ASR transcripts and automatically created metadata. A further investigation carried by Inkpen, Alzghool, Jones, and Oard (2006) on the CL-SR standard collection 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 the CLEF 2007 CL-SR task to present a method for combining results from different retrieval models in order to improve the overall retrieval effectiveness. They also provided a comparison between both ASR and manual metadata, indicating the superiority of the manual metadata for maintaining the retrieval effectiveness. Another interesting follow up study, reported by Jones, Zhang, Newman, and Lam-Adesina (2007), examined and compared the CLIR effectiveness of each source of evidence included in this collection. Results from this work indicate that searching the manually generated metadata gives higher performance in terms of recall and precision then search of noisy ASR transcripts.

The VideoCLEF task was then introduced at CLEF 2008 and CLEF 2009. This task provided Dutch TV content featuring English-speaking experts and studio guests. VideoCLEF piloted tasks involved performing classification, translation and keyword extraction on dual language video using either machine learning techniques or treating it as an IR task. Participants were provided with Dutch archival metadata, Dutch speech transcripts, and English speech transcripts (Larson, Newman, & Jones, 2009, 2010).

This previous work on CLVR focused on running CLIR tasks on a professional video broadcast 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 in the CLEF 2005-2007 with consistent speech quality of word error rate of 25% across the collections used (White et al., 2006; Oard et al., 2007; Pecina et al., 2008).

The CLEF tasks were followed by the establishment of the MediaEval benchmarking campaign in 2010 (MediaEval, 2015). Activities at MediaEval have focused on various multimedia search tasks, but have not included any CLIR elements.

The emergence of user-generated video content on the web has introduced new search opportunities and challenges in exploitation of the user-generated metadata (Eickhoff, Li, & de Vries, 2013; Filippova & Hall, 2011; Toderici, Aradhye, Pasca, Sbaiz, & Yagnik, 2010).

While CLIR for published text has been explored for a wide variety of language pairs for many years, recent research has begun to explore CLIR for user-generated informal text. One example of this work is the one done by Bagdouri, Oard, and Castelli (2014) which explored the retrieval of questions posed in formal English across user generated documents of Arabic collected from a forum posts. They employed a DT CLIR approach where they translated the Arabic informal text into English. Their results show that retrieval precision can be enhanced by applying an informal text classifier to help the translation of informal content. Lee and Croft (2014) also experimented with a CLIR task for informal documents. They developed a CLIR task over a large collection of Chinese forum posts and demonstrated that translation noise is increased by the informal text used in discussion forums. Their retrieval approach proposed to use PRF approach to improve retrieval effectiveness. Their results showed that PRF approaches can be useful in reducing the impact of translation errors on retrieval effectiveness for their tasks.

UGC has begun to attract considerable research interest in video retrieval and indexing in the recent years. While none of this work has so far included an element of 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 retrieval (Eickhoff et al., 2013; Filippova & Hall, 2011; Toderici et al., 2010), other work has focused on the quality of visual/audio features within the scale and the dynamics of UGC content (Bendersky et al., 2014; Chelba, Bikel, Shugrina, Nguyen, & Kumar, 2012; Langlois, Chambel, Oliveira, Carvalho, Marques, & Falcão, 2010). Moreover, from 2010, the TREC Video Retrieval Evaluation (TRECVID) (2015), the main video retrieval benchmark in the multimedia community, provided a collection of Internet videos to be used in several

Table 1: Length statistics for indexed blip10000 fields.

	Title	Desc	ASR
Stan.Dev	3.0	106.9	2399.5
Avg.Length	5.3	47.7	703.0
Median	5.0	24.0	1674.8
Max	22.0	3197.0	20451.0
Min	0.0	1.0	0.0

tasks. However, the design of TRECVID tasks has mainly focused on exploiting visual information for applications at the shot level (concept detection), or short video clips (event detection) and others. One task which is relevant to this work is the known-item search task (KIS) (Over, Awad, Fiscus, Antonishek, Michel, Smeaton, Kraaij, & Quénot, 2011) at TRECVID, the task aimed to explore 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.

In this work, we focus on studying the retrieval challenges of Internet-based UGC multimedia collections 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. These challenges can produce more ASR errors which affect retrieval effectiveness not only in monolingual retrieval, as reported by Eskevich, Jones, Wartena, Larson, Aly, Verschoor, and Ordelman (2012b) and Eskevich (2014), but also in cross-lingual settings when combined with query translation.

To the best of our knowledge, our work is the first effort to explore the issues of CLIR on video collected from a user-contributed source on the Internet. Thus creators from varied backgrounds and differing motivations and interests have generated the content without any central editorial control of style, format or quality. This makes the uploaded videos *very varied* in terms of the amount and quality of manually added metadata descriptions, and thus challenging from multiple retrieval perspectives. Of particular relevance to our investigation are the following aspects of the data:

- *Distribution of document lengths*: There is no restriction on document length which it found to be highly variable. Such length variability poses a challenge for any retrieval task, but it can be particularly significant within CLIR due to the presence of translation errors. A breakdown of the details of the various fields in our blip10000 test collection is shown in Table 1.

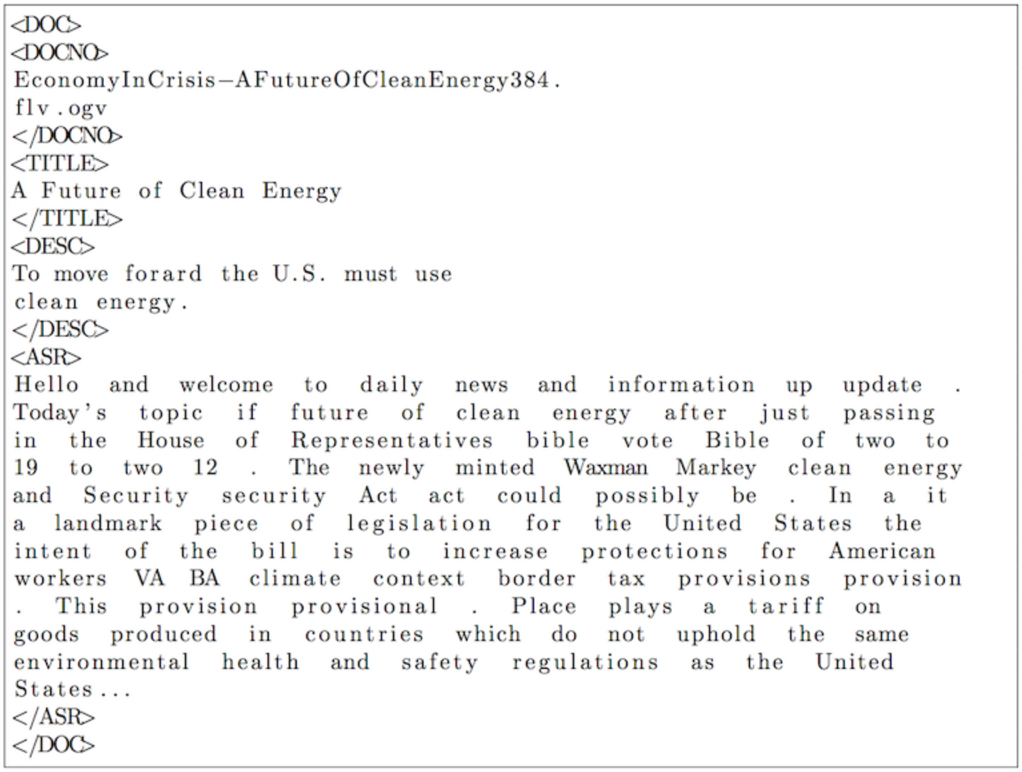
- *High variability in ASR quality of the video transcripts*: Even if the same ASR system is used, the variation in the audio quality, speaking styles and speakers, generally leads to significant variability in the accuracy of the transcripts.

- *Inconsistencies and sparseness of the associated user contributed metadata*: The titles may be very short having only one or two terms, while descriptions can be generic, informal and sometimes incomplete, making their utility for retrieval very varied.

4. Experimental Test Set and Evaluation

The blip10000 collection used in our experiments is a crawl of the Internet video sharing platform Blip.tv (Schmiedeke, Xu, Ferrané, Eskevich, Kofler, Larson, Estève, Lamel, Jones, & Sikora, 2013). It was originally used as the content dataset for the MediaEval 2012 Search and Hyperlinking task (Eskevich et al., 2012a). The blip10000 collection contains the crawled videos together with the associated metadata. This metadata is composed of the titles and descriptions for each video that were provided by the video uploader. In addition, associated ASR transcripts were also provided for most of videos. The collection consists of 14,838 videos having a total running time of ca. 3,288 hours, and a total size of about 862 GB[1].

The length statistics of the fields are shown in Table 1. It can be noted from this that there is a huge variation in the length distributions across different fields. Table 1 also highlights the variations of individual fields between the videos. For example, while one video may have no ASR, another may contain over 20K terms. For our experiments we indexed the metadata fields separately, and in combination as shown on Figure 2.



```
<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...
</ASR>
</DOC>
```

Figure 2: Example of a combined-field document.

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

Table 2: Monolingual English query vs Arabic-English translated query example.

Monolingual (MN) Query :
<top> <num>37 </num> <Mn-Lg>the video features a recent USA sanctioned clean energy Act.</Mn-Lg> <Mn-Sh>clean energy legislation USA</Mn-Sh> </top>
Machine Translated (CL) Query :
<top> <num>37</num> <CL-AR-Lg>Video displays Identify measures the United States toward alternative Energy</CL-AR-Lg> <CL-AR-Sh>Rules United States toward alternative energy</CL-AR-Sh> </top>

4.1 Query Construction for the CLIR Task

The MediaEval 2012 Search and Hyperlinking task (Eskevich et al., 2012a) was a known-item search task, a search for a single previously seen relevant video (the *known-item*), which provided 60 English queries collected using the Amazon Mechanical Turk (MTurk) crowd-sourcing platform (2015). Each query contains a full query statement (long query) and a terse web type search query (short query). For our investigation, we explored both the short and long queries to give a better understanding of the query-length independent retrieval behaviour for both the monolingual and CLIR tasks. To create our CLIR test set, we extended the original monolingual English queries by giving them to Arabic, Italian and French native speakers, and asking them to translate them into natural queries into their native language. Both the short and long queries were translated into Arabic. In order to explore the CLIR effectiveness across multiple language pairs, the short query set was also expressed in Italian, while the long query set was constructed in French. Having both types of queries being expressed in two languages (long queries are expressed in Arabic and French, while short queries are expressed in both Arabic and Italian) allowed us to draw better conclusions about the CLIR performance for this task. We used the Google translate API[2] to translate these query sets back into English. As would be expected. The MT translation produced different version of the original monolingual ones; in addition to the expected deletion/insertion edits as shown in the example of Table 2, there were also Named Entity Errors (NEEs) for Out-Of-Vocabulary (OOV) items that Google translation could not translate correctly. These translation edits and errors pose a challenge to the retrieval effectiveness of the MT translated queries compared to the monolingual ones.

The monolingual English query sets which were originally provided for the MediaEval 2012 Search and Hyperlinking task. For our investigation these query sets are labelled as follows:

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

- **Mn-Sh**: 60 EN short queries (monolingual)
- **Mn-Lg**: 60 EN long queries (monolingual)

The CLIR query sets are labelled as follows:

- **CL-AR-Sh**: 60 AR short queries translated into EN
- **CL-AR-Lg**: 60 AR long queries translated into EN
- **CL-IT-Sh**: 60 IT long queries translated into EN
- **CL-FR-Lg**: 60 FR long queries translated into EN

4.2 Mean Reciprocal Rank (MRR) Evaluation Metric

Since the retrieval problem that we are addressing is a known-item search for which we are seeking to retrieve a single known relevant item, we evaluate our investigations using the standard metric for this task is the Mean Reciprocal Rank (MRR) metric computed as shown in Equation 1 where $rank_i$ indicates the rank of the ground truth known item that the ith query is intended to find.

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

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, Clarke, & Cormack, 2010). Moreover, the recall is reported by default at the standard TREC1000 results cut off, but we also report it at cutoff points of 10, 50, 100 documents for some experiments.

5. Single Field Retrieval

The first part of our investigation examines the behaviour of the separate document information fields in the CLIR 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. We examine this question by evaluating the CLIR robustness of each field to measure how the retrieval effectiveness behaves in the CLIR framework. Throughout our investigation in this paper, we define **CLIR Robustness** as how well a field or source of evidence performs in the CLIR framework. We observe this by computing the significance of change between the CLIR and monolingual performance using the same setting and across all query sets. To run our CLIR robustness evaluation experiment, we compare the CLIR effectiveness of each field against a monolingual baseline:

- **ASR_index** contains only the ASR transcript fields.
- **Title_index** contains only the title fields.
- **Desc_index** contains only description fields.

Table 3: Mono vs. CLIR performance per index

	Mn-Sh	CL-AR-Sh	CL-IT-Sh	Mn-Lg	CL-AR-Lg	CL-FR-Lg
Title_index	0.239	0.2288	0.2383	0.2827	0.2244	0.2239
ASR_index	0.4275	0.2748	0.3873	0.4513	0.3487	0.3833
Desc_index	0.2154	0.1943	0.2102	0.2432	0.2285	0.2316

5.1 Retrieval Model

Our single field retrieval experiments were carried out using the Terrier retrieval engine[3]. Terrier is a standard open source IR toolkit providing many of the best established retrieval algorithms and widely used by the IR research community. Stop-words were removed based on the standard Terrier list, and stemming performed using the Terrier implementation of Porter stemming. We used the PL2 model[4], a probabilistic retrieval model from *the Divergence From Randomness (DFR)* framework (Gianni, 2003). The reason we selected this model over other available retrieval models is the characteristics of our data collection. Previous studies such as(Amati & Van Rijsbergen, 2002) 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 aligns with our known-item experiment. PL2 is thus suitable since our Internet based data collection has huge variation in lengths, whether at the field level or the document level, as shown in Table 1. The PL2 document scoring model is defined as shown in Equation 2, where $Score(d, Q)$ is the retrieval matching score for a document d for query term t and λ is the Poisson distribution of F/N, F is the query term frequency of t 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 normalized term frequency defined in Equation 3, where l is the length of the document d. avg_l is the average length of the documents, and c is a free parameter for the normalization. To set the parameter c, we followed the empirically determined standard settings recommended by Amati and Van Rijsbergen (2002) and He and Ounis (2007b), which are $c = 1$ for short queries and $c = 7$ for long queries.

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

$$tf_n = \sum_{d} (tf. \log_2(1 + c.\frac{avg_l}{l})), (c > 0) \qquad (3)$$

5.2 Experimental Results and Discussion

Our results for each index are shown in Table 3, these show that MRR is *lower* in all cases for the CLIR task. Thus retrieval effectiveness of all fields is negatively impacted

[3] http://www.terrier.org/

[4] Terrier implementation of this model can be found in :
http://terrier.org/docs/v4.0/javadoc/org/terrier/matching/models/PL2.html

Table 4: AR CLIR - the t-values according to the % MRR reduction for each index

	CL-AR-Sh	CL-AR-Lg
Title_index	-1.69	-1.73
ASR_index	**-1.94***	**-2.50***
Desc_index	-0.829	-0.44

**Statistically significant values with p-value < 0.05.*

Table 5: FR and IT CLIR - the t-values according to the % MRR reduction for each index

	CL-IT-Sh	CL-FR-Lg
Title_index	-0.05	-1.77
ASR_index	-1.58	**-2.04***
Desc_index	-0.32	-0.47

**Statistically significant values with p-value < 0.05.*

for CLIR. This confirms the expected additional retrieval challenge that arises from the imperfect query translation. MRR for the Arabic queries is reduced to a higher degree than for the French and Italian queries. This is most likely due to the relative difficulty of Arabic MT (Alqudsi, Omar, & Shaker, 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 MRR for each index indicates they have different responses to the query translation; notably the impact is greatest on the index of the ASR transcript field across all languages using both short and long queries.

To better understand the significance of these CLIR reductions in MRR, 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 CLIR MRRs in pairs at every query level. The significance test results in terms of t-values for the indexes searched for the Arabic CLIR queries are shown in Table 4 and for the French and Italian CLIR queries in Table 5. Looking at the t-values, we can observe that IT queries were less challenging than the others since the performance was not significantly different from monolingual.

Furthermore, both Tables 4 and 5 indicate that the ASR transcripts do indeed have the lowest robustness in the CLIR setting. On searching the single-field indexes, for both long and short queries, ASR_index had the least robustness with a statistically significant negative reduction in Arabic and in French with ($p<0.05$). For the Italian short queries, MRR reduction rates of the ASR index (ASR_index) were not statistically significant, but still had the highest negative impact over other fields.

We conclude from this experiment that even if they are incomplete, short and/or sometimes unreliable, the user-uploaded titles and meta descriptions are more robust in the CLIR setting than the ASR fields. As noted earlier, the degree of ASR recognition errors may vary from one video to another on Internet, 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), and we plan to explore this effect in more detail in future work with a view to improving the CLIR robustness of the ASR transcript field.

6. Retrieval with Combined Metadata Fields

Having examined the effectiveness of the three separate fields for monolingual retrieval and CLIR, in this section we explore the potential of combining them for improving retrieval effectiveness. For this investigation, we carried out another set of experiments that combined the evidence from the individual fields. For this, we first combine the fields in pairs, and then as shown in Figure 1, we integrate the three fields but with varied field weighting.

6.1 Retrieval Model

For the combined field experiments we use the DFR PL2F model[5] (Macdonald, Plachouras, He, Lioma, & Ounis, 2006). This is a modified version of the PL2 model used in the previous section. 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, shown in Equation 2, 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. 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 normalization parameters. This per-field normalization feature in PL2 modifies the standard PL2 document scoring function to include the weighted sum of the normalised term frequencies tf_x.

$$tf_n = \sum_x (w_x . tf_x . \log_2(1 + c_x . \frac{avgl_x}{l_x})), (c_x > 0) \qquad (4)$$

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 normalization 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}.

6.2 Two Field Combinations

Table 6 shows MRR values for fields combined into pairs for which were indexed using the PL2F retrieval model. We are interested here in the potential for improved retrieval using fields in combination. Comparing the results in Table 6 and the earlier results shown in Table 3, we can see that field combination is more effective for both monolingual and CLIR tasks. Further improvement could probably be obtained by weighting fields differently.

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

Table 6: Mono vs. CLIR performance with field pair combinations

	Mn-Sh	CL-AR-Sh	CL-IT-Sh	Mn-Lg	CL-AR-Lg	CL-FR-Lg
TitleDesc_index	0.2503	0.2421	0.2463	0.3020	0.2795	0.2614
ASRDesc_index	0.4394	0.3624	0.3951	0.5245	0.3905	0.4326
ASRTitle_index	0.4295	0.3676	0.3820	0.4527	0.3451	0.3768

Table 7: 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

However, the main goal of our investigation is the potential for combining all three fields, and we explore this in more detail in the next section.

6.3 Three Field Combinations

In this section we describe our investigation of the retrieval effectiveness with combination of all three fields. We explore giving higher weight to a specific field over the others.

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): PL2FASR, PL2Title, PL2Desc.

- Assign an equal c_x value to all fields to allow full-length normalization for the term frequency of each field as in C_x = {1,1,1} for short queries, and Cx ={7,7,7} for long queries. We also followed the empirically standard settings recommended by Amati and Van Rijsbergen (2002), and He and Ounis (2007b).

- For W_x, we set the w_x value for the targeted field, and the rest to be fixed at 1, to give priority for field x over the others, as in W_x = {w_x,1,1}. The reason why we chose them to be 1 was to allow for the presence of their term frequencies, but with normal (is not boosted) weights.

The combination weighting schemes are shown in Table 7, in each case one field has a weight boost w_x. To examine retrieval behaviour, we vary 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}.

6.3.1 Experimental Results and Discussion

Figure 3 shows the MRR performance at each weighting point for the long queries (CL-AR-Lg and CL-FR-Lg query sets), and the short queries (CL-AR-Sh and CL-IT-Sh query sets). As can be seen in Figure 3, fields behave differently with weight boosting. The best CLIR

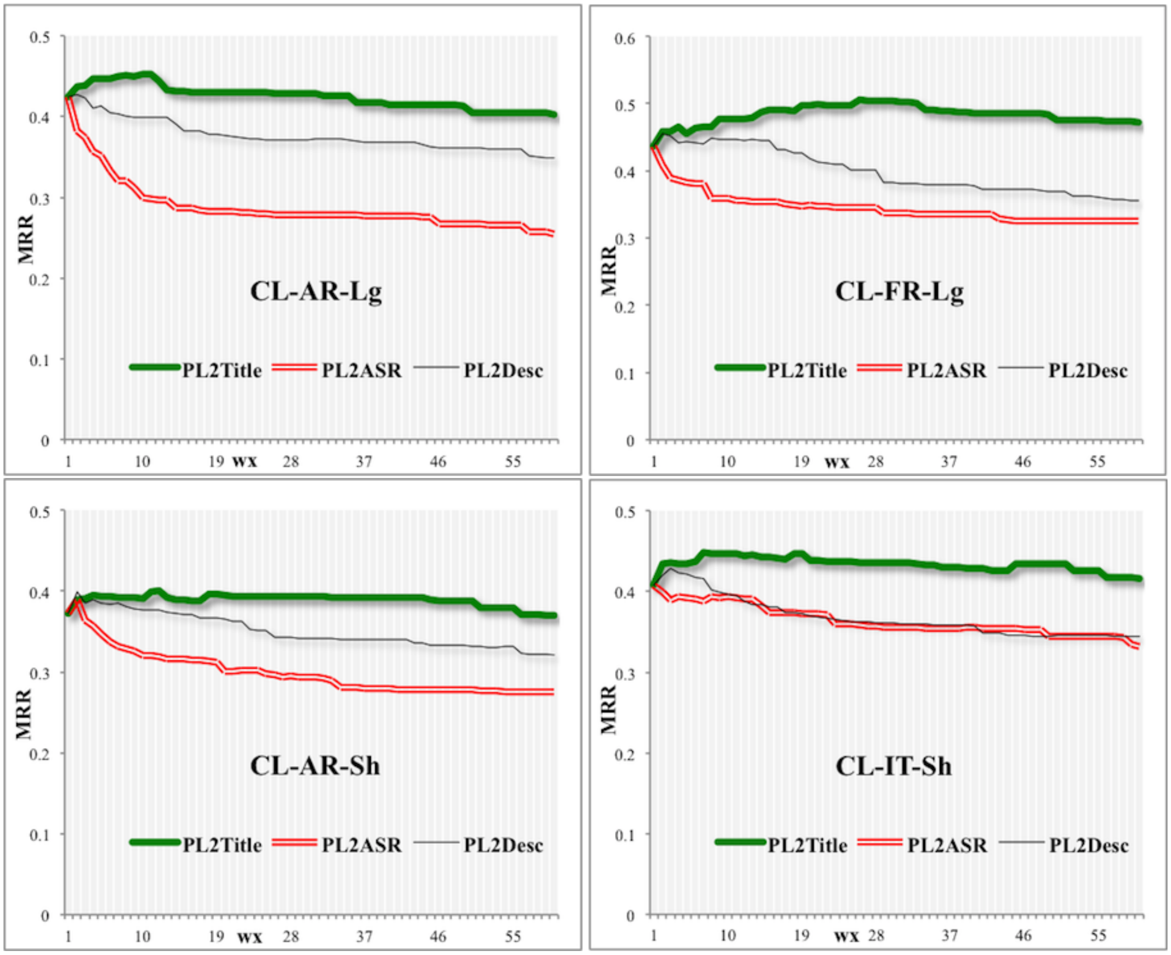


Figure 3: MRR CLIR performance for the single_weighted models across all weighting points (wx) using both short and long query sets.

precision performance is always achieved by giving a higher weight to the title field for all of the AR, IT and FR query sets. Across all the weighting points and all languages pairs, the PL2Title model shows higher performance than the other fields for both short and long query sets.

Moreover, it can also be seen from these figures, that we get lower performance when we give progressively higher weights to the ASR and Desc fields. The strong CLIR 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 may have been written by the video uploader with more attention than the descriptions could be referred to the following reasons:

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

It could be also the case that for the known-item queries, the users who wrote the queries had viewed the videos and might be more likely to include the titles of the videos in their query to find the intended video, because they believe that it would be easier 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 8: Single index Mono vs. CLIR Recall performance represented by the number of found documents with cut-off values of 10, 50, and 100.

10 - cut off						
	Mn-Sh	CL-AR-Sh	CL-IT-Sh	Mn-Lg	CL-AR-Lg	CL-FR-Lg
Title_index	19	16	19	23	18	19
ASR_index	35	31	34	34	29	29
Desc_index	17	15	15	23	21	20
TitleDesc_index	20	17	19	25	21	22
ASRDesc_index	35	30	35	40	32	33
ASRTitle_index	35	31	34	34	28	38
All_Index	37	33	37	40	33	34
50 - cut off						
	Mn-Sh	CL-AR-Sh	CL-IT-Sh	Mn-Lg	CL-AR-Lg	CL-FR-Lg
Title_index	25	23	23	30	28	26
ASR_index	42	38	42	43	37	39
Desc_index	29	23	27	31	27	28
TitleDesc_index	33	28	30	37	29	31
ASRDesc_index	47	41	44	46	40	43
ASRTitle_index	42	38	42	43	36	40
All_Index	47	40	47	48	41	45
100 - cut off						
Title_index	25	23	24	32	29	29
ASR_index	46	41	45	47	40	43
Desc_index	34	27	31	34	32	32
TitleDesc_index	38	32	35	42	35	38
ASRDesc_index	50	47	49	50	43	47
ASRTitle_index	46	41	45	46	39	42
All_Index	50	45	50	52	43	49

Comparing the MRR for PL2Title with the values shown in Table 3, it can be also seen that the performance for PL2Title is almost double the one obtained by the independent Title field (Title_index). While the MRR values for the ASR and Desc fields are similar between the two experiments. 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 length and the language pair. In an attempt to better understand how the field combination improves retrieval effectiveness, we examined the Recall of the individual fields and the combinations.

Table 8 shows the total number of known-items retrieved in the top 10, 50 and 100 for each field set. It can be seen here that the Title field has lowest recall in isolation, but that it can boost the Recall of the other fields when used in combination. The results in Figure 3 suggest that the title field brings additional evidence without bringing noise, which is not the case for Desc and ASR fields which degrade effectiveness when their weight is increased.

7. Query Expansion Using Combined Metadata Fields

In Sections 5 and 6, we analyzed the effectiveness of each data source (ASR, Title and description) in the CLIR framework and showed that overall performance is more robust when these fields are combined together in an optimal way. We also showed that adjusting the retrieval settings to give higher weight to more a reliable data source, i.e. the Title, in comparison to other less reliable fields, can benefit the overall CLIR performance.

In this section, we seek to modify the query itself using the field information to improve the retrieval effectiveness. The query modification strategy we explore is based on the query expansion (QE) techniques (Carpineto & Romano, 2012), where we use different sources of evidence (fields) to enrich the original query. The underlying motivation behind applying QE is that expanding the query to include important terms should make it more effective in identifying relevant items. Ideally, the QE should alleviate the impact of the translation errors by adding more terms from top ranked videos that are either relevant or related to the query. The QE effectiveness relies on the quality and the informativeness of the top ranked documents (Amati, Carpineto, & Romano, 2004). The Top ranked documents can be taken from external resources such as WordNet and Wikipedia (Pal, Mitra, & Datta, 2013) or from the local document collection itself by considering the top ranking documents as relevant to the query. We consider more interesting/challenging approach here, which is the local collection expansion approach since we aim to investigate how these noisy collection of web videos can be utilized to improve the query effectiveness. We, nevertheless, plan to consider the use of the external collections for QE in our future investigation.

Existing research has shown that QE techniques can be useful for improving both monolingual and CLIR effectiveness in many languages (Bellaachia & Amor-Tijani, 2008). However, most of this research has focused primarily on collections of professionally written or formal text which has none or a minimum amount of noise. There has been some, but very limited, work on applying QE techniques for the noisy data, such as the work on OCR Data (Tong, Zhai, Milic-Frayling, & Evans, 1996; Lam-Adesina & Jones, 2006), and more recently, on the user-generated informal text (Lee & Croft, 2014). In this section, we are interested in taking the challenge of applying QE in CLIR settings with a focus on avoiding problems that may arise from the noise presented in each source of evidence; in particular for our task, the translation errors of the query, the transcription errors of the ASR as well as the inconsistency errors of user generated textual metadata. We explore the expansion of queries based on the top ranked documents. Expansion terms are intended to make the query more reliable and robust to find the intended relevant item. Our interest in this section can be summarized with the help of the following research questions:

- If we expand the query using the top ranked documents, how might this affect the overall CLIR effectiveness? How effective will QE under such a setting of noisy data collected from Internet videos be?

- Which of the different UGC information sources is more useful for QE? Since these different sources (fields) have different relative characteristics and behaviour for retrieval, as observed empirically in Sections 5 and 6.

We describe our approach to addressing these questions in the following sections. We investigate the reliability of each single field for QE and their challenges in Section 7.1. We then propose an adaptive approach to improve overall QE robustness[6] by selecting the best source for expansion in Section 7.2.

7.1 Query Expansion on Fields

We employ the *Divergence From Randomness* (DFR) QE mechanism proposed by Gianni (2003). This technique computes a weight to rank the terms from the top ranking documents. The DFR QE generalizes Rocchio's method (Salton & Buckley, 1997) to implement several term weighting models that measure the informativeness of each term in the pseudo relevant set.

DFR QE has two stages. First, 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 documents from a random distribution. We use the DFR weighting model called Bo1, a parameter-free DFR model which uses BoseEinstein statistics to weight each term based on its informativeness. This parameter free model has been widely used and proven to be effective (He & Ounis, 2007a; Plachouras, He, & Ounis, 2004; Gianni, 2003). The weight w of a term t in the top ranked documents using the DFR Bo1 model is shown in Equation 5, where tf_x is the frequency of the term in the pseudo-relevant set (top n ranked documents). P_n is given by F/N ; F is the term frequency of the query term in the whole collection and N is the number of documents in the whole collection.

$$w(t) = tf_x . \log_2(\frac{1+P_n}{P_n}) + \log_2(1+P_n) \tag{5}$$

Secondly, the query term weight qt_w, which was obtained from the single-pass retrieval (as described in Equation 2), 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 6, where $w_{max}(t)$ is indicated by the maximum $w(t)$ values among the expanded query terms.

$$qt_w = qt_w + \frac{w(t)}{w_{max}(t)} \tag{6}$$

To illustrate the QE approach used in our experiments, we provide the QE example of the CLIR-AR 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

[6] QE robustness is interpreted in this context as how likely it is that it will improve retrieval performance over the baseline, where the baseline is a single-pass retrieval that is using the original query.

is boosted to be greater than 1[7]. 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 expanding and reweighing of the query is explained as follows :

eee×1.000*, pc*×1.9211*, 900*×1.0000 *,troubleshoot*×1.0000*, laptop*×1.2988*, mac*×0.2195*, us*×0.0000*, classroom*×0.0000.

Some of the original query terms may *not* appear in the top-terms such as (900, EEE and troubleshoot), in which the formula in Equation 6 would only give them the same weight they have in a single-pass retrieval settings. *Such cases might be common in our CLIR settings* due to the presence of named entity translation errors, such as the example of *"Aldirimovr"* as described in Section 5. Since these NEEs are produced by the translation, they will never appear in any of the top ranked documents since they are not present in the collection, therefore their weight will always remain the same as after the expansions. This will pose an extra challenge for overall QE effectiveness which we try to handle in the following sections by designing a post-translation QE that is tuned for this task. The main reasons for us to pick the post-translation QE approach in this experimental investigation are :

- To study the impact of translation errors/translation quality on the QE effectiveness. This can be investigated by running a post-translation analysis of the query expansion performance.
- To investigate whether adding new informative terms to the query can reduce the impact of the translation errors and improve the retrieval effectiveness.

We use the default QE parameter settings in our experiment where we set it up to extract the 10 most informative terms from the top 3 returned documents. These settings were suggested by Gianni (2003) after conducting extensive experiments on several test collections. However, in our case, the task is much more challenging, where we only have one relevant document to find, we also extend these settings to explore other possible parameter combinations. The parameter settings for our proposed QE runs are tuned to explore the top (3, 5, 10) terms from each of the top (3, 5, 7) documents. Our QE runs parameters combinations are as follows:

- Taking the top 3 terms, this includes QE runs that take the top 3 terms from the top 3 documents, the top 5 and the top 10 documents.
- Taking the top 5 terms, this includes QE runs that take the top 5 terms from the top 3 documents, the top 5 and the top 10 documents.
- Taking the top 10 terms, this includes QE runs that take the top 10 terms from the top 3 documents, the top 5 and the top 10 documents.

[7] 1 is the normal weight for any term that appears on the original query

Table 9: Optimized parameters (top-terms and top-doc) that is selected for each QE run

	Terms	Docs
exp-ASR	5	5
exp-Title	3	3
exp-Desc	3	3
exp-All	3	3

To study the best parameters for each field expansion, we explored all the parameter variations. We then chose the best performing setting for each QE run. These optimized settings are shown in Table 9.

A QE framework using fields for monolingual text retrieval was proposed by He and Ounis (2007a). This suggests an improved term-weighting method based on field statistics to achieve better retrieval performance. In this investigation, we adopt this approach to CLIR QE and further tune 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:

- The top n terms are extracted from the top n documents retrieved in response to executing the query on each separate field type (title, description and ASR) in order to investigate their individual effectiveness for QE.

- Retrieval is carried out similarly to our setting for the experiment in Section 6, where we combine all the fields together (see Figure 2), and give an equal weight of 1 to each field. We use the PL2F model (described in Section 6) for this retrieval experiment. Since both query sets (long/short) led to similar conclusions in the previous Sections regarding field effectiveness, we chose to run QE for the short queries only (the CLIR queries of CL-AR-Sh and CL-IT-Sh are described in Section 4.1).

We conducted several QE runs based using each individual field and their combinations. The reason for having such tuned runs is to assess the effectiveness of each field combination for QE. Our proposed field-based QE runs are as follows.

- **exp-ASR**: Queries were expanded using the ASR field only, by only taking the top ranking terms from the top documents which are retrieved from the ASR index.

- **exp-Title**: Queries were expanded using the Title field only, by taking the top ranking terms from the top documents which are retrieved from the Title index.

- **exp-Desc**: Queries were expanded using the Desc field only, by taking the top ranking terms from the top documents which are retrieved from the Desc index.

- **exp-All** : Queries were expanded using a combination of all fields, i.e. ASR, Title and Desc.

- **exp-Non** : This run skips the expansion of the queries and just does single pass retrieval using the original query. Note that we use this as the baseline since we want to assess how effective QE can be for such a challenging task.

Table 10: MRR performance for each QE run.

	CL-AR-Sh	CL-IT-Sh
exp_ASR	0.3502	0.3735
exp_Title	0.3820	0.4060
exp_Desc	0.3470	0.4090
exp_All	0.3571	0.4069
exp_Non	0.3726	.4081

Table 11: Overall Recall (total found known-items) for each QE run.

	CL-AR-Sh	CL-IT-Sh
exp_ASR	53	57
exp_Title	52	56
exp_Desc	51	56
exp_All	53	56
exp_Non	52	56

We run our field-based QE experiments to explore MRR performance for each field. The MRR performance across different runs is shown in Table 10, while the overall recall results are shown in Table 11. It can be seen that MRR for the proposed runs (exp-All, exp-ASR, exp-Title, exp-Desc) does not improve over the baseline run (exp-Non). The exp-Title and exp-Desc get more or less similar MRR performance, while the exp-ASR achieves significantly lower MRR. The fact that we have multiple sources of noise coming either from the translation or from the fields themselves, together with the fact that this is known-item task, where there is only one relevant item that may not be highly ranked in the initial search, can justify the ineffectiveness of these QE runs.

To better understand the robustness for each QE run and compare it to the baseline, we study the difference of MRR (ΔMRR) at each query level for all the proposed runs over the baseline run (exp-Non). Where the ΔMRR on a particular query level for a QE run (exp-x) is indicated through $\mathbf{\Delta MRR = (MRR(exp\text{-}x)\ \text{-}\ MRR(exp\text{-}Non))}$. The ΔMRR results for each CL-AR-Sh query across all the runs are shown in Figure 4, while Figure 5 shows the ΔMRR results for CL-AR-Sh queries.

Since the DFR QE model (see Equation 5 and 6) uses the informativeness measure to weight the extracted top terms, the decreases and increases of the ΔMRR values that are shown in Figures 4 and 5 can be explained as follows.

- $\mathbf{\Delta MRR = 0}$, means the top ranked terms were predicted as less informative, which is the reason why the QE runs have minimal effect over the baseline (exp-Non). Based on the DFR definition of the informativeness (see Equation 5 and 6), this indicates that the terms added to this particular query were given low weight because they are not only common in the top ranked documents, but also in the whole document collection. In other words, the QE run could not find any helpful terms that might potentially improve the overall performance. The ΔMRR performance of each run suggests that this situation occasionally occurs in all QE runs across several queries, particularly on the exp-Title and exp-Desc. This probably arise due to the recall problem that these two fields have, as shown in Table 8 (see also the single-field retrieval experiments of Section 5). In fact, for most of the queries for the exp-Desc and exp-Title runs, the application of QE has a very low or a zero effect on the MRR compared to the single-pass baseline retrieval (exp-Non).

- **ΔMRR > 0**, means that the top ranked terms were predicted to be highly informative and they were *relevant* to the query, which is the reason why QE shows a positive increase over the baseline. This in turn suggests that the exp-ASR runs were able to improve more queries (more positive ΔMRR points) in terms of ranking the known-item over other runs. This is also expected due to the higher recall performance of these fields.

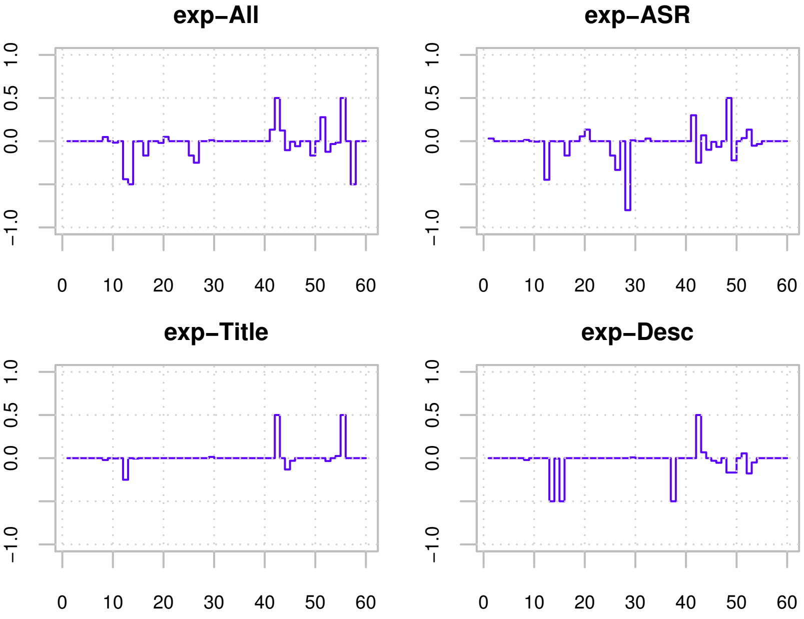


Figure 4: ΔMRR across all QE runs on CL-AR-Sh.

Moreover, all runs have had a positive ΔMRR for some queries. Particularly interesting is the fact that *the combined QE run (exp-All) did not yield the best results.*

- **ΔMRR < 0**, means that the top ranked terms were predicted 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 ΔMRR values show that this incorrect prediction has significantly impacted all runs. In particular, the exp-ASR run is affected the most, as can be seen from the negative ΔMRR values across several queries on both CL-AR-Sh and CL-IT-Sh.

We conclude from these experiments that setting the expansion to be based on the Title and the Desc fields may help to improve the MRR by improving the ranking for some queries. However, this improvement covers only a limited number of queries due to the coverage and

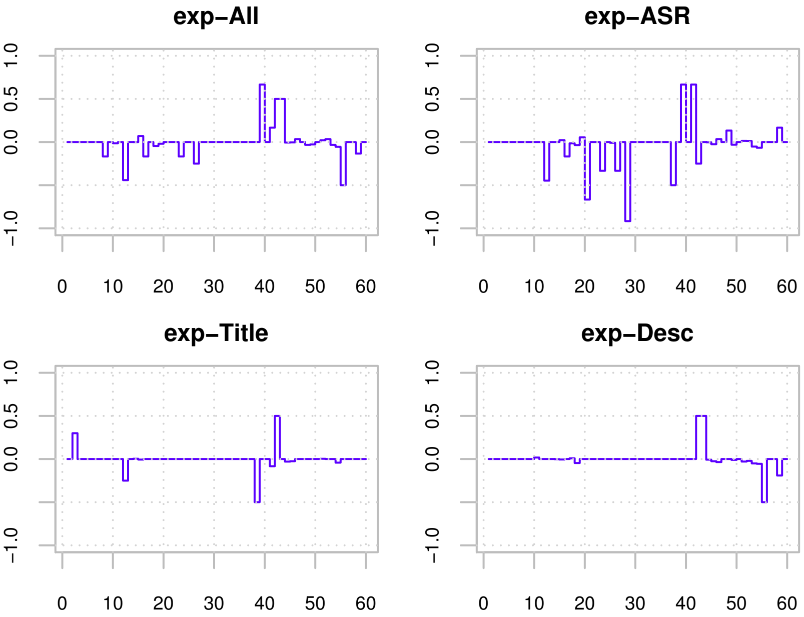


Figure 5: ΔMRR across all QE runs on CL-AR-Sh.

recall issues of these two fields (see Section 5). ASR fields seem to have better coverage in terms of improving the performance for most of the queries. However, this coverage can be negative sometimes when QE model fails to pick the right terms.

Even when the ASR is combined with other fields using the exp-All approach, the overall performance still does *not* improve the MRR with respect to the baseline (exp-Non). As shown in these experiments, each QE run that uses a single field shows positive ΔMRR values for some queries. However, these per-field improvements decrease the average when the fields are combined together (for example, the exp-All MRR values shown in Table 10 are lower than exp-Title for CL-AR-Sh, exp-Desc for CL-IT-Sh). The combination approach is not sufficiently effective because the overall retrieval performance can still be negatively impacted by the noise present in the fields. In the next section, we propose a novel approach for alleviating this issue by selecting only the best source for expansion. This technique adaptively predicts performance of each QE run, and chooses the source that is more likely to achieve a positive impact, as elaborated in the next section.

7.2 Selecting the Best Source for Expansion

In this section we introduce our proposed approach to improve the QE effectiveness for our CLIR task by predicting whether QE is needed or not, and if it is, to select the best source for expansion. In particular, we aim to design a robust QE approach that can prevent or minimize the negative MRR changes that the exp-ASR or other runs can have. We argue that these reductions are caused for two reasons:

- The query performance of the intiail run is just *perfect* for our known-item task. This case can happen if the known relevant document is ranked at first position. So the added new expansion terms can potentially disturb the query performance and negatively impact the retrieval effectiveness.

- As mentioned in the literature (Mitra, Singhal, & Buckley, 1998; Terra & Warren, 2005), QE effectiveness is further challenged by the query drift issue where the expansion terms are informative, but they to do not belong the same topic of the original query. As shown in Table 1, ASR transcripts can be up to 20K length, Desc can be as long as 3K length. These long descriptive fields may cover many different topics and noise, and may not be relevant to the topic of the query. This is common for the dataset used in this task, since we are using user-generated videos where the length, the topic and the ASR quality of each video may have no specific or consistent theme.

To address the above issues we propose a modified QE technique that use a pre-retrieval prediction technique to decide whether QE is likely to be beneficial and which source is more reliable for the expansion. We describe the prediction technique used for this approach in Section 7.2.1. Section 7.2.2 explains our proposed Adaptive QE algorithm. Section 7.2.3 reports the experiments we conduct to investigate the effectiveness of this approach.

7.2.1 Predicting the QE Effectiveness

Predicting QE effectiveness has been proposed within the concept of the selective QE framework by Amati et al. (2004), and has been widely used to improve QE effectiveness. The main concept behind this technique is to disable QE when it is predicted to have a negative impact on the first-pass retrieval performance. This prediction is based on pre-retrieval metrics that assess the application of QE and make a decision on whether to apply it or not. Most of these predictions are based on capturing the query statistics in the collection, such as the query difficulty (Gianni, 2003), the query clarity score (Cronen-Townsend, Zhou, & Croft, 2002), query length (Amati et al., 2004). Full analysis of the effectiveness of every prediction techniques effectiveness is contained in the work of He and Ounis (2006). We chose to use one of the most successful predictors, the Average Inverse Collection Term Frequency (AvICTF) (He & Ounis, 2006). We use this prediction metric in our proposed Adaptive QE technique to predict whether QE is needed or not and which field combination is the most reliable for the expansion.

The AvICTF predictor has previously been tested in the field-based query expansion in several previous studies (He & Ounis, 2007a; Macdonald, He, Plachouras, & Ounis, 2005). The reason to choose this predictor over others is that by definition it has a higher potential to work well for CLIR because it addresses the term frequency aspect of the query. In principle, the term frequency should be a good indicator for predicting translation errors. For example an NEE error like the word "Aldirimovr" would have zero term frequency which can be used to give an indication of overall query performance. AvICTF is also a very low cost metric that uses only local collection statistics to make the prediction, if the AvICTF value is higher than a tuned threshold, then the query is predicted to preform very well using first-pass retrieval and therefore the query expansion is disabled. The AvICTF is defined as follows.

$$AvICTF = \frac{log_2 \prod_Q (\frac{token_{coll}}{F})}{ql} \tag{7}$$

where $token_{coll}$ is the number of tokens in the whole collection. ql is the query length, F is the query term frequency of the whole collection. In the next sections, we describe the adaptive QE algorithm we implemented and our experimental investigation.

7.2.2 Adaptive Field-Based QE Technique

In this section we propose our adaptive QE technique, the adaptivity concept is inspired by the work of He and Ounis (2006). The idea behind this adaptive method is to automatically set the weights of the fields during the query running time. We further redesign this concept to auto-select which field combination is the best source for the expansion. The adaptive field-based QE algorithm we implemented is explained below:

- During indexing, the algorithm processes all possible fields combination into separate indexes, where each index can have one, two or n fields, n is the total number of fields available in the collection.
- During the retrieval of a query Q:
    1. The algorithm calculates the prediction value V across every possible index, then selects the index $Maxindex$ that has the highest prediction value $MaxV$.
    2. The algorithm then makes a decision of whether to apply QE or not by comparing the prediction value $MaxV$ against a trained threshold. If the predicted value is less than the threshold, the algorithm skips QE and moves straight to the retrieval.
    3. If the prediction decision is to proceed with QE, the algorithm expands the query Q by taking the top terms from the top ranked documents in $Maxindex$ then run the retrieval.

Our adaptive QE algorithm predicts the effectiveness of each possible QE field run and chooses the one that is more likely to be useful for each query. In other words, our adaptive QE make the use of (exp-ASR, exp-Title, exp-Desc, exp-Non, exp-All) as well as other possible QE runs that are based on two fields combination such as (exp-ASRDesc, exp-TitleASR, exp-TitleDesc) to produce adaptive QE runs as follows.

- **exp-Adapt-AR**: Expands the AR queries (CL-AR-Sh) using the proposed adaptive QE technique.
- **exp-Adapt-IT**: Expands the IT queries (CL-IT-Sh) using the proposed adaptive QE technique.

For training the prediction and tuning the threshold value, we conduct a two-fold holdout evaluation on our query sets. We divide both the CL-AR-Sh and CL-IT-Sh sets into training and testing query sets as described below:

- AR-train queries set: Contains 30 queries picked randomly from the CL-AR-Sh set.
- IT-train queries set: Contains 30 queries picked randomly from the CL-IT-Sh set.

- Test queries set: Contains the remaining 30 queries from CL-AR-Sh query set and the remaining 30 queries from CL-IT-Sh query set. These two sets are used to evaluate the proposed adaptive QE runs (exp-Adapt-AR and exp-Adapt-IT).

During the training, similar to the work of He and Ounis (2007a), we perform a manual data sweeping through the range of [3, 15] with an interval of 1 for both training queries. The best threshold chosen for exp-adapt-AR on AR-train queries was 6. The best threshold found for exp-adapt-IT on IT-train queries was 9. The difference between these threshold values can be attributed to the distinct level of translation qualities between both query sets. As we discussed previously in Section 5, IT queries have better CLIR performance over the AR ones, and thus QE threshold is different here.

7.2.3 Experimental Results and Discussion

We ran our proposed adaptive QE using the two runs: exp-Adapt-AR and exp-Adapt-IT explained previously. The overall performance results of the two runs for the test queries is shown Table 12.

As we explained before, these adaptive QE runs involve all possible QE runs for one-single run, and use the one that is predicted to have better performance, the selection statistics of each run is indicated in Table 13. As can be seen in Table 12, the overall MRR performance improves when using our proposed adaptive QE technique. Also, looking at the recall performance of Table 12, it appears that this technique successfully improves the

Table 12: Results for adaptive QE runs in terms of MRR and Recall.

MRR		
	Adaptive	(Baseline exp-Non)
exp-Adapt-AR	0.43614	0.3785
exp-Adapt-IT	0.4867	0.4580
Recall		
	Exp-Adapt	Exp-Non (Baseline)
exp-Adapt-AR	22	22
exp-Adapt-IT	22	22

Table 13: The selection statistics for the adaptive QE runs.

	exp-Adapt-AR	exp-Adapt-IT
exp-ASRDesc	4	6
exp-TitleASR	7	7
exp-TitleDesc	4	5
exp-Title	1	2
exp-Desc	2	0
exp-ASR	3	2
exp-Non	5	8
exp-All	4	0

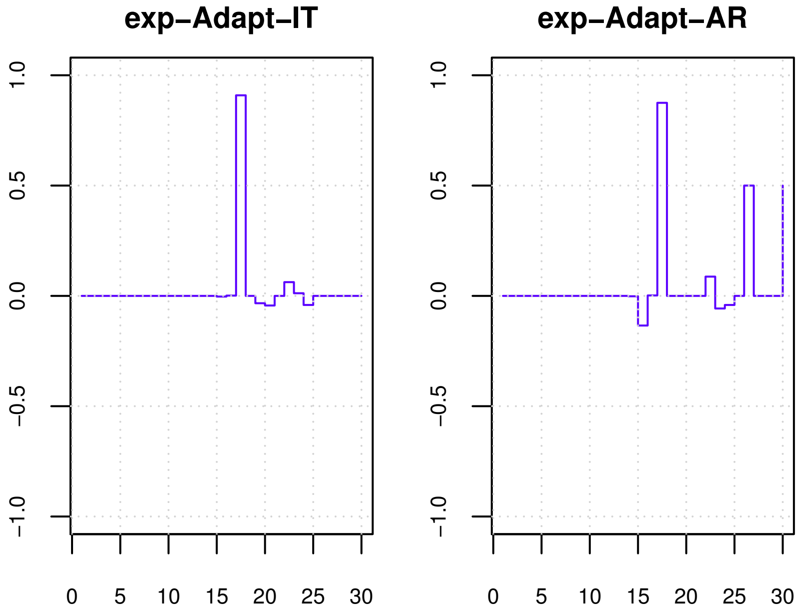


Figure 6: ΔMRR (Baseline Vs Adaptive QE runs) for both languages using the test queries

overall ranking while maintaining the same recall level. This can be attributed to the fact that these runs are able to selectively make use of the best performance from each individual run.

To better understand the improvement over the baseline, similar to our previous experiment in Section 7.1, we also calculate the ΔMRR values for each test query for both AR-train and IT-train query tests in Figure 6. The ΔMRR values in Figure 6 show that this technique reduces the chance of having any significant reduction in MRR over the baseline. Overall, it appears that this QE approach has robustness to improve the overall CLIR performance by selecting the most reliable source of expansion individually for each query.

8. Conclusions and Further Research

This paper has examined CLVR based on text metadata fields for an Arabic-English, French-English and Italian-English known-item search task based on the blip10000 collection. We studied the retrieval effectiveness and challenges of three different sources of information: 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. Our first set of experiments analysed the behaviour of these sources for CLIR by examining their CLIR robustness. We found that the ASR transcript field has the lowest robustness across other fields and its performance can drop significantly for CLIR. We then explored field combination retrieval to explore their performance all together, and our investigation showed that giving higher weight to the titles over other fields gives improved CLIR performance. In general our experiments show that tuning the retrieval settings to give a higher weight towards the fields which have a lower CLIR robustness degrades retrieval effectiveness.

Our work also investigated the effectiveness of automatic query expansion on the CLIR setting for this task. We found that these information sources can have a varying reliability for query expansion, and can have a negative impact on the retrieval effectiveness in the CLIR framework when they are combined together. We proposed an adaptive query expansion technique that automatically selects the most reliable source for expansion based on a well established query performance prediction technique. The results from our experimental investigation show that this technique has better robustness to maintain retrieval performance in the CLIR setting.

In general, we found that in this noisy CLIR setting, between the translation errors, transcription errors, incorrect or incomplete UGC metadata and the very varied document lengths, there might be no single best solution that can be recommended to be used for answering all queries or even expanding/ improving them, but rather an adaptive approach that relies on trained heuristics that are tuned specifically for this task and this collection. The concept of adaptivity will be further studied in our next investigations of this problem.

Our analysis of the CLIR effectiveness for UGC video gives us suggestions for further investigation in many areas. One potential direction for further work is to automatically assess the quality of ASR transcripts and the Description information and assign weights based on quality measures, and also to explore task dependent tuning of the machine translation process. Studying the CLIR effectiveness of other UGC sources of evidence (such as visual information and social interaction data which include tweets, personal profile data) would be an interesting follow up investigation. Another area for future study is to improve the document representation within this framework by developing a document expansion technique that is tuned to improve the overall CLIR robustness and effectiveness of each source of evidence. Also, we plan to expand this work by studying the effectiveness of other available CLIR techniques such as Hybrid and DT CLIR for this task.

References

Alqudsi, A., Omar, N., & Shaker, K. (2012). Arabic machine translation: a survey. *Artificial Intelligence Review*, 1–24.

Alzghool, M., & Inkpen, D. (2008). Cluster-based model fusion for spontaneous speech retrieval. In *Proceedings of the ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, pp. 4–10. Citeseer.

Amati, G., Carpineto, C., & Romano, G. (2004). Query difficulty, robustness, and selective application of query expansion. In *Advances in information retrieval*, pp. 127–137. Springer.

Amati, G., & 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.

Amazon (2015). Amazon mechanical turk - welcome. `https://www.mturk.com/`. Retrieved: 2015-03-30.

Bagdouri, M., Oard, D. W., & 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*, pp. 1811–1814. ACM.

Ballesteros, L., & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum*, Vol. 31, pp. 84–91. ACM.

Ballesteros, L., & 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*, pp. 64–71. ACM.

Bellaachia, A., & Amor-Tijani, G. (2008). Enhanced query expansion in english-arabic clir. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, pp. 61–66. IEEE.

Bendersky, M., Garcia-Pueyo, L., Harmsen, J., Josifovski, V., & 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*, pp. 1769–1778. ACM.

Bing (2015). Bing translator. `http://www.bing.com/translator/`. Retrieved: 2015-03-30.

BlipTV (2015). Bliptv. `https://www.blip.tv`. Retrieved: 2015-03-30.

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2010). *Information retrieval: Implementing and evaluating search engines*. Mit Press.

Carpineto, C., & 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., & 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., & Tsai, S.-C. (1998). Proper name translation in cross-language information retrieval. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp. 232–236. Association for Computational Linguistics.

CLEF (2015a). The clef initiative (conference and labs of the evaluation forum) - clef2009. `http://www.clef-initiative.eu/edition/clef2009`. Retrieved: 2015-03-30.

CLEF (2015b). The clef initiative (conference and labs of the evaluation forum) - homepage. `http://www.clef-initiative.eu/`. Retrieved: 2015-03-30.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 299–306. ACM.

Eickhoff, C., Li, W., & de Vries, A. P. (2013). Exploiting user comments for audio-visual content indexing and retrieval. In *Advances in Information Retrieval*, pp. 38–49. Springer.

Eskevich, M. (2014). *Towards effective retrieval of spontaneous conversational spoken content*. Ph.D. thesis, Dublin City University.

Eskevich, M., & Jones, G. J. F. (2014). Exploring speech retrieval from meetings using the ami corpus. *Computer Speech & Language*.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., Ordelman, R., & Larson, M. (2012a). Search and hyperlinking task at mediaeval 2012..

Eskevich, M., Jones, G. J., Wartena, C., Larson, M., Aly, R., Verschoor, T., & Ordelman, R. (2012b). Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pp. 1–6. IEEE.

Facebook video (2015). Facebook. `https://www.facebook.com/facebook/videos`. Retrieved: 2015-03-30.

Federico, M., Bertoldi, N., Levow, G.-A., & Jones, G. J. F. (2005). Clef 2004 cross-language spoken document retrieval track. In *Multilingual Information Access for Text, Speech and Images*, pp. 816–820. Springer.

Federico, M., & Jones, G. J. F. (2004). The clef 2003 cross-language spoken document retrieval track. In *Comparative Evaluation of Multilingual Information Access Systems*, pp. 646–652. Springer.

Filippova, K., & 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*, pp. 835–842. ACM.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., & 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*, pp. 96–104. ACM.

Gianni, A. (2003). *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. Ph.D. thesis, Department of Computing Science, University of Glasgow.

Google (2015). Google translate. `https://translate.google.com/`. Retrieved: 2015-03-30.

He, B., & Ounis, I. (2006). Query performance prediction. *Information Systems*, *31*(7), 585–594.

He, B., & Ounis, I. (2007a). Combining fields for query expansion and adaptive query expansion. *Information processing & management*, *43*(5), 1294–1307.

He, B., & Ounis, I. (2007b). On setting the hyper-parameters of term frequency normalization for information retrieval. *ACM Transactions on Information Systems (TOIS)*, *25*(3), 13.

Herbert, B., Szarvas, G., & Gurevych, I. (2011). Combining query translation techniques to improve cross-language information retrieval. In *Advances in Information Retrieval*, pp. 712–715. Springer.

Inkpen, D., Alzghool, M., Jones, G. J. F., & 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*, pp. 61–64. Association for Computational Linguistics.

Jones, G. J., Zhang, K., Newman, E., & Lam-Adesina, A. M. (2007). Examining the contributions of automatic speech transcriptions and metadata sources for searching spontaneous conversational speech..

Kishida, K., & Kando, N. (2006). *A hybrid approach to query and document translation using a pivot language for cross-language information retrieval.* Springer.

Lam-Adesina, A. M., & Jones, G. J. (2006). Using string comparison in context for improved relevance feedback in different text media. In *String Processing and Information Retrieval*, pp. 229–241. Springer.

Langlois, T., Chambel, T., Oliveira, E., Carvalho, P., Marques, G., & Falcão, A. (2010). Virus: video information retrieval using subtitles. In *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp. 197–200. ACM.

Larson, M., Newman, E., & 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*, pp. 906–917. Springer.

Larson, M., Newman, E., & 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*, pp. 354–368. Springer.

Lee, C.-J., Chen, C.-H., Kao, S.-H., & 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*, pp. 651–658. ACM.

Lee, C.-J., & Croft, W. B. (2014). Cross-language pseudo-relevance feedback techniques for informal text. In *Advances in Information Retrieval*, pp. 260–272. Springer.

Leveling, J., Zhou, D., Jones, G. J., & Wade, V. (2010). Document expansion, query translation and language modeling for ad-hoc ir. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pp. 58–61. Springer.

Macdonald, C., He, B., Plachouras, V., & Ounis, I. (2005). University of glasgow at trec 2005: Experiments in terabyte and enterprise tracks with terrier.. In *TREC*.

Macdonald, C., Plachouras, V., He, B., Lioma, C., & Ounis, I. (2006). *University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming.* Springer.

Magdy, W., & 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.

McCarley, J. S. (1999). Should we translate the documents or the queries in cross-language information retrieval?. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 208–214. Association for Computational Linguistics.

MediaEval (2015). MediaEval Benchmarking Initiative for Multimedia Evaluation. `http://www.multimediaeval.org/`. Retrieved: 2015-03-30.

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 206–214. ACM.

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.

Nikoulina, V., Kovachev, B., Lagos, N., & 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*, pp. 109–119. Association for Computational Linguistics.

Oard, D. W., & Diekema, A. R. (1998). Cross-language information retrieval. *Annual review of information science and technology*, *33*, 223–256.

Oard, D. W., & Hackett, P. G. (1998). Document translation for cross-language text retrieval at the university of maryland. In *Information Technology: The Sixth Text REtrieval Conference (TREC-6)*, pp. 687–696. US Dept. of Commerce, Technology Administration, National Institute of Standards and Technology.

Oard, D. W., Wang, J., Jones, G. J. F., White, R. W., Pecina, P., Soergel, D., Huang, X., & Shafran, I. (2007). Overview of the clef-2006 cross-language speech retrieval track. In *Evaluation of multilingual and multi-modal information retrieval*, pp. 744–758. Springer.

Over, P., Awad, G. M., Fiscus, J., Antonishek, B., Michel, M., Smeaton, A. F., Kraaij, W., & Quénot, G. (2011). Trecvid 2010–an overview of the goals, tasks, data, evaluation mechanisms, and metrics..

Pal, D., Mitra, M., & Datta, K. (2013). Improving query expansion using wordnet. *CoRR*, *abs/1309.4938*.

Parton, K., McKeown, K. R., Allan, J., & Henestroza, E. (2008). Simultaneous multilingual search for translingual information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 719–728. ACM.

Pecina, P., Hoffmannová, P., Jones, G. J., Zhang, Y., & Oard, D. W. (2008). Overview of the clef-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval*, pp. 674–686. Springer.

Peters, C., Braschler, M., & Clough, P. (2012). *Multilingual information retrieval: From research to practice*. Springer Science & Business Media.

Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information retrieval*, *4*(3-4), 209–230.

Plachouras, V., He, B., & Ounis, I. (2004). University of glasgow at trec 2004: Experiments in web, robust, and terabyte tracks with terrier.. In *TREC*.

Rogati, M., & Yang, Y. (2002). Cross-lingual pseudo-relevance feedback using a comparable corpus. In *Evaluation of Cross-Language Information Retrieval Systems*, pp. 151–157. Springer.

Salton, G., & Buckley, C. (1997). Improving retrieval performance by relevance feedback. *Readings in information retrieval*, *24*(5), 355–363.

Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, M. A., Estève, Y., Lamel, L., Jones, G. J. F., & Sikora, T. (2013). Blip10000: a social video dataset containing spug content for tagging and retrieval. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pp. 96–101. ACM.

Sokolov, A., Hieber, F., & Riezler, S. (2014). Learning to translate queries for clir. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 1179–1182. ACM.

Terra, E., & Warren, R. (2005). Poison pills: harmful relevant documents in feedback. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 319–320. ACM.

Toderici, G., Aradhye, H., Pasca, M., Sbaiz, L., & Yagnik, J. (2010). Finding meaning on youtube: Tag recommendation and category discovery. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3447–3454. IEEE.

Tong, X., Zhai, C., Milic-Frayling, N., & Evans, D. A. (1996). Ocr correction and query expansion for retrieval on ocr data–clarit trec-5 confusion track report.. In *TREC*.

TRECVID (2015). Trec video retrieval evaluation home page. `http://trecvid.nist.gov/`. Retrieved: 2015-03-30.

Varshney, S., & Bajpai, J. (2014). Improving performance of english-hindi cross language information retrieval using transliteration of query terms. *arXiv preprint arXiv:1401.3510*.

White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., & Huang, X. (2006). *Overview of the CLEF-2005 cross-language speech retrieval track*. Springer.

Youtube (2015). Youtube. `http://www.youtube.com/`. Retrieved: 2015-03-30.

YouTube Press (2015). Statistics - YouTube. `http://www.youtube.com/yt/press/statistics.html`. Retrieved: 2015-03-30.

Zhou, D., Truran, M., Brailsford, T., Wade, V., & Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, *45*(1), 1.