An Investigation of Cross-Language Information Retrieval for User-Generated Internet Video

Ahmad Khwileh(✉), Debasis Ganguly, and Gareth J.F. Jones

ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
ahmad.khwileh2@mail.dcu.ie, {dganguly,gjones}@computing.dcu.ie

Abstract. Increasing amounts of user-generated video content are being uploaded to online repositories. This 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. We investigate CLIR effectiveness for the blip10000 archive of user-generated Internet video content. We examine the retrieval effectiveness using the title and free-text metadata provided by the uploader and automatic speech recognition (ASR) generated transcripts. Retrieval is carried out using the *Divergence From Randomness* models, and automatic translation using *Google translate*. Our experimental investigation indicates that different sources of evidence have different retrieval effectiveness and in particular differing levels of performance in CLIR. Specifically, we find that the retrieval effectiveness of the ASR source is significantly degraded in CLIR. Our investigation also indicates that for this task the Title source provides the most robust source of evidence for CLIR, and performs best when used in combination with other sources of evidence. We suggest areas for investigation to give most effective and robust CLIR performance for user-generated content.

Keywords: Cross-Language Video Retrieval · User generated content · User generated internet video search

1 Introduction

Recent years have seen a huge rise in the amount and diversity of content stored in online video repositories. In 2015, YouTube[1] the predominant online video sharing site, reported that 300 hours of video content are being uploaded every minute encompassing material in 61 languages [20]. This content comes from a wide variety of sources, with significant amounts created and uploaded privately

[1] www.youtube.com

J. Mothe et al. (Eds.): CLEF 2015, LNCS 9283, pp. 117–129, 2015.
DOI: 10.1007/978-3-319-24027-5_10

with little or no formal editorial control, meaning that the amount and quality of associated metadata is of widely varying quantity and reliability. Further, the amount of content and topical coverage of the content across different languages is very uneven, meaning that satisfying an information need for a user of one language can only be achieved by providing relevant content in another language. One of the challenges for the effective exploitation of this content in this setting is effective multilingual search.

Recent years have seen significant efforts in the area of Cross Language Information Retrieval (CLIR) for text retrieval initially focusing on formally published content and more recently beginning to look at informal social media content. However, while some limited work has been carried out on Cross-Language Video Retrieval (CLVR) for professional videos such as documentaries or TV news broadcasts, there has to date, been no significant evaluation of CLVR for user-generated Internet-based content. A key difference between user-generated Internet content and professionally produced content is the nature and structure of the textual data associated with it. In this setting, retrieval effectiveness may not only suffer from issues arising 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 inconsistencies, and frequently the sparseness of the associated user uploaded metadata for each video. 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 little or no guidance available for the choices that should be made.

In this paper we explore a known-item CLVR task based on a semi-professional Internet video archive constructed from the MediaEval 2012 Search and Hyperlinking [6]. To understand the complexities of the task better, we undertake a detailed performance analysis examining the impact of different source metadata information on CLIR behaviour. The video collection used for this investigation is the blip10000 dataset collected from the Internet video sharing platform Blip.tv [18]. We investigate the CLIR effectiveness of metadata based on ASR, Title and description fields for both short and long queries defined for the MediaEval 2012 task.

The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 describes the test set used in our experiments and the evaluation metric, Section 4 describes initial experiments examining CLIR robustness for each information source, Section 5 describes our approach to improving CLIR effectiveness, and Section 6 concludes and provides directions for further work.

2 Related Work

While we are not aware of a comparable study of CLIR for user-generated Internet video content, there is much related existing work. The most closely related work to that examined in this paper was carried out in tasks within the CLEF

evaluation campaigns[2]. 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 (SDR) with manually translated queries [7,8]. Their tasks involved the retrieval of American English news broadcasts of both unsegmented and segmented transcripts taken from radio and TV news. A more ambitious Cross-Language Speech Retrieval (CL-SR) task ran within CLEF 2005-2007 [19,15,17]. This examined CLIR for a spontaneous conversational speech collection with content in English and Czech content consisting of oral history interviews. The task provided ASR transcripts, automatically and manually generated metadata for the interviews. The goal was to design systems to help searchers to identify sections of an interview that would be most relevant to their information need. The reported results of these tracks showed that the use of manual metadata yielded substantial improvement on the retrieval effectiveness, compared to using ASR transcripts and automatically created metadata.

The VideoCLEF track was then run at CLEF 2008 and CLEF 2009. This task provided Dutch TV content featuring English-speaking experts and studio guests. VideoCLEF piloted tasks involving performing classification, translation and keyword extraction on dual language video using either machine learning or information retrieval techniques. Participants were provided with Dutch archival metadata, Dutch speech transcripts, and English speech transcripts [10,11].

The multimedia CLIR tasks at CLEF focused on professionally curated content. Whether it was documentaries, TV shows or interviews, this had high quality metadata provided with it. For example, domain experts following a carefully prescribed format created the manual metadata for CLEF 2005-2007. The CLEF tasks were followed by the establishment of the MediaEval benchmarking campaign in 2010 [14]. Activities at MediaEval have focused on various multimedia search tasks, but have not included any CLIR elements.

Other recent work has explored searching video of user generated content, but this has not included an element of CLIR. The most relevant video search task is the known-item search task which was established by the TREC Video Retrieval Evaluation (TRECVID)[3] in it's known-item search task (KIS) [16]. This was included at TRECVid annually from 2010 to 2012. Results were rather inconsistent from year to year in terms of the retrieval effectiveness of different search approaches, one conclusion being the difficulty of setting up such a task on Internet collections.

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 text. One example of this is the work described in [4], which explored the retrieval of questions posed in formal English across user-generated (informal) documents of Arabic collected from forum posts. Their results showed that the retrieval performance could be enhanced by applying an informal text classifier to help the translation of informal content. The work described in [12]

[2] www.clef-initiative.eu/
[3] http://trecvid.nist.gov

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

also reported a CLIR task for informal Chinese documents. They proposed to use pseudo relevance feedback (PRF) approaches to improve retrieval effectiveness, and showed they can be useful to reduce the impact of translation errors on retrieval effectiveness.

3 Experimental Test Set and Evaluation

To the best of our knowledge, our work is the first to explore the issues of CLIR on video that is collected from a user-contributed source on the Internet. Content creators from varied backgrounds with differing motivations and interests created this content without any central editor 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.

The blip10000 collection used in our experiments is described in detail in [18]. This collection is a crawl of the Internet video sharing platform Blip.tv[4]. It was originally used as the content dataset for the MediaEval 2012 Search and Hyperlinking task [6]. The blip10000 collection contains the crawled videos together with the associated metadata. This metadata is comprised of the Titles and short descriptions for each video that were manually provided by the video uploader. In addition, associated ASR transcripts were also provided. The collection consists of 14,838 videos having with a total running time of ca. 3,288 hours, and a total size of about 862 GB.

Table 1 shows the variations of individual fields between the videos. For example, while one video may have no ASR, another may contain over 20K terms. Of particular relevance to our investigation are the following aspects of the data:

- *The distribution of the document lengths*: since there is no restriction on document lengths and they are found to be highly variable. Such length variability poses a challenge for any retrieval task. A breakdown of the details of the various fields in our blip10000 test collection is shown in Table 1.
- *High variability in automatic speech recognition (ASR) quality of the transcripts of the video*: Even though the same ASR system is used, the variation in the audio quality, speaking styles and speakers leads to significant variability in the accuracy of the transcripts.

[4] http://blip.tv/

– *Inconsistencies and sparseness of the associated user uploaded metadata*: Titles may be very short having only one or two terms, while descriptions can be generic and incomplete, making their utility for retrieval very varied.

For our experiments we indexed the metadata fields separately and in combination, as described in the experiments in Sections 4 and 5.

3.1 Query Construction for the CLIR Task

The MediaEval 2012 Search and Hyperlinking task [6] was a known-item search task, a search for a single previously seen relevant video (the *known-item*). This task provided 60 English queries collected using the Amazon Mechanical Turk[5] (MTurk) crowd-sourcing platform. Each query contains a full query statement providing a detailed described of the required features of the single relevant target video (long query) and a terse web type search query for the same item (short query). To create our CLIR test set, we extended the original monolingual English by giving the queries to Arabic, Italian and French native speakers, and asking them to rewrite them into natural queries in their native language. Both short and long queries were expressed into Arabic. In addition, the short query set was also expressed in Italian, while the long query set was further expressed in French.

In order to explore CLVR for this task, we used the Google translate API[6] to translate these translated topics back into English. As would be expected, for some queries, machine translation (MT) produced a slightly different queries than the monolingual ones. In addition to the expected deletion/insertion edits, there were also some Named Entity Errors (NEEs) for Out-Of-Vocabulary (OOV) items that *Google MT translation* could not translate correctly. These edits and translation errors pose a challenge to the retrieval effectiveness of the CLIR over the monolingual one. For our investigation, we explored both the short and long queries to give a better understanding of the effect of query length on retrieval behaviour for both the monolingual and CLIR tasks. The query sets used in out investigation are labelled as follows:

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

Since the retrieval task is a known-item search for which we are seeking to retrieve the single known relevant item, we evaluate our investigations using the standard metric for this task, the mean reciprocal rank (MRR) metric computed as shown in Equation 1.

[5] http://www.mturk.com/
[6] https://developers.google.com/translate/

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

where $rank_i$ indicates the rank of the ground truth known item that the ith query is intended to find.

4 CLIR Using Single Field Indexes

The first part of our investigation examines the behaviour of the separate information fields in the CLIR framework. We are particularly interested here in the impact of errors in automatic translation or inconsistencies on retrieval effectiveness, given the noise in 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. We report this by observing the significance of change between the CLIR and monolingual performance using the same setting and across all query sets. For running 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

We report the results for both long and short query sets to examine the impact of query length and the natural language form of the long queries.

Our single field CLIR retrieval experiments were carried out using the Terrier retrieval engine[7]. Stop-words were removed based on the standard Terrier list, and stemming performed using the Terrier implementation of Porter stemming. We used the PL2 [2] model, a probabilistic retrieval model from *the Divergence From Randomness (DFR)* framework. The reason we selected this model over other retrieval models, is our data collection and experiments specifications; our Internet based data collection has very large variations in the lengths of the metadata and documents shown in Table 1. Previous studies such as [3] showed that the PL2 model 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. The PL2 document scoring model is defined in Equation 2.

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

where $Score(d,Q)$ is the score for a document d for each query term t of the query Q, λ is the Poisson distribution of F/N; F is the query term frequency

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

Table 2. 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

Table 3. The t-values according to the % MRR reduction for each index

	CL-AR-Sh	CL-AR-Lg	CL-IT-Sh	CL-FR-Lg
Title_index	-1.69	-1.73	-0.05	-1.77
ASR_index	**-1.94***	**-2.50***	-1.58	**-2.04***
Desc_index	-0.829	-0.44	-0.32	-0.47

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

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.

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

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 normalization. To set the parameter c, we followed the empirically standard settings used in [3,9], which are $c = 1$ for short queries and $c = 7$ for long queries.

Our results for each index are shown in Table 2, these show that MRR is *lower* in all cases for the CLIR task. Thus retrieval effectiveness of all fields is negatively impacted for CLIR. This confirms the expected additional retrieval challenge that arises from the imperfect query translation. MRR for the AR queries is reduced to a higher degree than for the French and Italian queries. This is likely to arise due to the relative difficulty of Arabic MT [1]. 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.

Also, looking at the MRR reduction rates for each index indicates they have different responses to the query translation; notable the impact is greatest on the ASR transcript indexes across all languages pairs using both short and long queries. To better understand the significance of these CLIR reductions in MRR, we computed the statistical significance of each drop. We calculated the t-value for the difference at the 95% confidence level after representing all monolingual and CLIR MRRs in pairs on every query level. The significance test results in terms of t-values for the indexes searched for all CLIR constructed queries

are shown in Table 3. 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. Table 3 indicates that when using the one-field per index, for both long and short queries, ASR_index has the least robustness, with a statistically significant negative drop in Arabic and in French with ($p < 0.05$). For the Italian queries, the MRR reduction rates of the ASR index (ASR_index) were not statistically significant, but still had the highest negative impact comparing to searching other fields (Title and description fields).

We conclude from this experiment that even if they are incomplete, informal, short and sometimes unreliable, the user-uploaded Titles and meta descriptions are yet more robust in the CLIR setting than the ASR fields. As noted earlier, the degree of ASR recognition errors may vary from video to another on Internet video due to the huge variation of the audio quality. The interaction between recognition error rate, document length and retrieval behaviour is highly complex, as observed in [5]. 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.

5 CLIR Using Combined Metadata Fields

Having examined the effectiveness of the three separate fields for monolingual retrieval and CLIR, in this section we explore the potential for combining them for improving retrieval effectiveness. For this investigation, we carried out another set of experiments that combined the evidence from the individual fields. We combined the three fields with varied field weighting. For these combined field experiments we use the DFR PL2F model [13] which is a modified version of the PL2 model [2] 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.

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

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_x also needs two parameters w_x, c_x to be set. Hence, for scoring every indexed document we needed to set these parameters: C_x which is the set of per-field length normalization parameters c_x that need to be set for every field as $C_x = \{ c\_asr, c\_title, c\_desc\}$. Also for W_x which is the set of

Table 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

per-field boost factors w_x that need to be set for each field as W_x ={ w_asr, w_title, w_desc}.

For our investigation of the retrieval effectiveness with combination of all three fields, we explore giving higher weight to a specific field over the others by creating a single-weighted retrieval model for each source of evidence (field). To set the parameter values for our proposed single-weighted retrieval models we followed these steps:

- Construct a model based on the PL2F that targets a single field x from each (ASR, title, desc) as PL2ASR, PL2Title, PL2Desc.
- Give 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, Cx ={7,7,7} for long queries. We also follow the empirically standard settings applied in [3,9].
- For W_x, we set the w_x value for the targeted field, and the rest to be fixed at 1, to give a priority for field x over the others as in W_x = {w_x,1,1}. The reason why we chose the fixed weights to be 1 was to allow for the presence of their term frequencies, but with normal (not boosted) weights.

The combination weighting schemes are shown in Table 4, in each case only one field has a weight boost of w_x. To examine retrieval behaviour, we varied the w_x boost parameters for each proposed model from 1 to 60 using increments of 1. The first weighting iteration at the weighting point 1 is the same for all models where they have W_x = {1,1,1}. Figure 1 shows the MRR performance at each weighting point for the long queries (the CL-AR-Lg and the CL-FR-Lg query sets), and the short queries (the CL-AR-Sh and the CL-IT-Sh query sets). As can be seen in Figure 1, fields behave *differently* with the weight boosting. The best CLIR precision performance is always achieved by giving a higher weight towards the Title field across the AR, IT and FR queries[8]. Across all the weighting points and all languages, the PL2Title model shows higher performance than other fields for both short and long query sets. It is also shown in these figures, that we even get lower performance when we give progressively higher weights to the ASR and Desc fields. The strong CLIR performance of the PL2Title indicates the stability and the robustness of Title fields for Internet videos over other fields. Also, the fact that these Titles may have been written by the video uploader with more attention than the descriptions could be attributed to the following reasons:

[8] Also worth mentioning that the MTurk task used to construct all query sets did not expose any associated video metadata to the query creators.

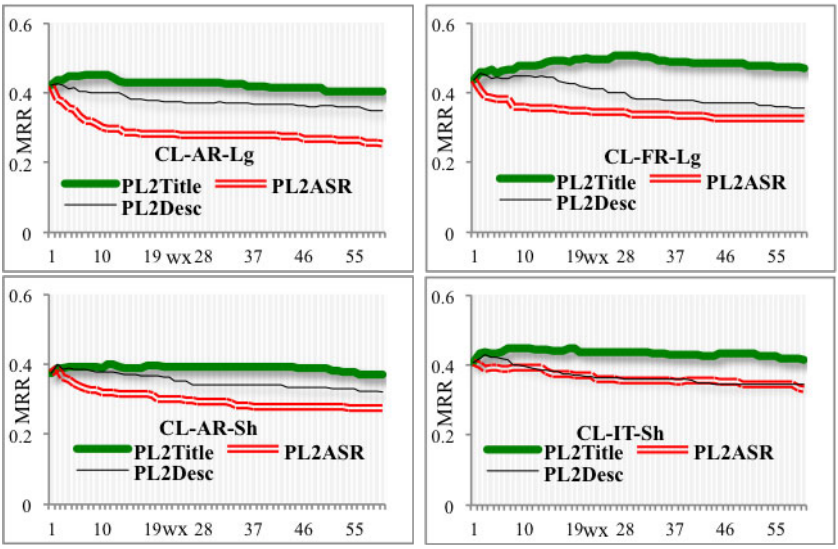

Fig. 1. CLIR peformance (MRR) of the single_weighted models across all weighting points (wx) using both short and long query sets

Table 5. Mono vs. CLIR Recall performance represented by the number of found documents on 100 results cut-off

	Mn-Sh	CL-AR-Sh	CL-IT-Sh	Mn-Lg	CL-AR-Lg	CL-FR-Lg
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

- The uploader thought it is vital to have a meaningful 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 video viewers.
- The quality of textual content of Title field, which is shown to have more CLIR robustness, can be attributed to its shortness; in which it was *only* helpful for limited amount of queries without introducing any noise that would negatively affect the overall retrieval performance.

Comparing the MRR for PL2Title with the values shown in Table 2, it can also be seen that performance for the PL2Title is almost double that of the result for the separate Title field run. 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 order to better understand how the field combination improves retrieval effectiveness, we examined the Recall of the individual fields and the combinations. Table 5 shows the total number of known-items retrieved in the top 100 results for each field set (including pairs-combined fields). It can be seen here that the Title field has low recall in isolation (due to its shortness issue), but it can boost the Recall of the other fields when used in combination. The results in Figure 1 suggest that the Title field brings additional evidence without bringing noise, unlike the Desc and ASR fields which degrade effectiveness when their weight is increased.

6 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 user-generated Internet video 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 videos descriptions which can be informal, 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 significantly drop for CLIR. We then explored field combination and showed that giving higher weight to the Titles over other fields gives improved CLIR performance. In general, our experiments suggest that giving higher weight towards the fields which have a lower CLIR robustness degrades retrieval effectiveness.

Our analysis of these fields effectiveness gives us suggestions for further investigation. 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. Many CLVR search requests have the potential to exploit the use of visual features, we intend to explore the integration of visual features into our retrieval framework in further experiments.

References

1. Alqudsi, A., Omar, N., Shaker, K.: Arabic machine translation: a survey. Artificial Intelligence Review, 1–24 (2012)
2. Amati, G.: Probabilistic Models for Information Retrieval based on Divergence from Randomness. Ph.D. thesis, Department of Computing Science, University of Glasgow (2003)
3. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) **20**(4), 357–389 (2002)

4. Bagdouri, M., Oard, D.W., Castelli, V.: 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 (2014)
5. Eskevich, M., Jones, G.J.F.: Exploring speech retrieval from meetings using the AMI corpus. Computer Speech & Language (2014)
6. Eskevich, M., Jones, G.J.F., Chen, S., Aly, R., Ordelman, R., Larson, M.: Search and hyperlinking task at MediaEval 2012 (2012)
7. Federico, M., Bertoldi, N., Levow, G.-A., Jones, G.J.F.: CLEF 2004 cross-language spoken document retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 816–820. Springer, Heidelberg (2005)
8. Federico, M., Jones, G.J.F.: The CLEF 2003 cross-language spoken document retrieval track. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 646–652. Springer, Heidelberg (2004)
9. He, B., Ounis, I.: On setting the hyper-parameters of term frequency normalization for information retrieval. ACM Transactions on Information Systems (TOIS) **25**(3), 13 (2007)
10. Larson, M., Newman, E., Jones, G.J.F.: Overview of VideoCLEF 2008: automatic generation of topic-based feeds for dual language audio-visual content. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 906–917. Springer, Heidelberg (2009)
11. Larson, M., Newman, E., Jones, G.J.F.: Overview of VideoCLEF 2009: new perspectives on speech-based multimedia content enrichment. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsikrika, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 354–368. Springer, Heidelberg (2010)
12. Lee, C.-J., Croft, W.B.: Cross-language pseudo-relevance feedback techniques for informal text. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C.X., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 260–272. Springer, Heidelberg (2014)
13. Macdonald, C., Plachouras, V., He, B., Lioma, C., Ounis, I.: University of Glasgow at WebCLEF 2005: experiments in per-field normalisation and language specific stemming. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 898–907. Springer, Heidelberg (2006)
14. MediaEval: MediaEval Benchmarking Initiative for Multimedia Evaluation (2014). http://www.multimediaeval.org/ (retrieved September 30, 2014)
15. Oard, D.W., Wang, J., Jones, G.J.F., White, R.W., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF-2006 cross-language speech retrieval track. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 744–758. Springer, Heidelberg (2007)
16. Over, P., Awad, G., Fiscus, J., Antonishek, B., Michel, M., Smeaton, A.F., Kraaij, W., Quénot, G., et al.: TRECVID 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: TRECVID 2011-TREC Video Retrieval Evaluation Online (2011)
17. Pecina, P., Hoffmannová, P., Jones, G.J.F., Zhang, Y., Oard, D.W.: Overview of the CLEF-2007 cross-language speech retrieval track. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 674–686. Springer, Heidelberg (2008)

18. Schmiedeke, S., Xu, P., Ferné, I., Eskevich, M., Kofler, C., Larson, M.A., Estève, Y., Lamel, L., Jones, G.J.F., Sikora, T.: 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 (2013)
19. White, R.W., Oard, D.W., Jones, G.J.F., Soergel, D., Huang, X.: Overview of the CLEF-2005 cross-language speech retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 744–759. Springer, Heidelberg (2006)
20. YouTube Press: Statistics - YouTube (2015). http://www.youtube.com/yt/press/statistics.html (retrieved April 1, 2015)