Investigating Segment-Based Query Expansion for User-Generated Spoken Content Retrieval

Ahmad Khwileh
ADAPT Centre, School of Computing
Dublin City University
Dublin 9, Ireland
Email: ahmad.khwileh2@mail.dcu.ie

Gareth J. F. Jones
ADAPT Centre, School of Computing
Dublin City University
Dublin 9, Ireland
Email: gjones@computing.dcu.ie

***Abstract*—The very rapid growth in user-generated social multimedia content on online platforms is creating new challenges for search technologies. A significant issue for search of this type of content is its highly variable form and quality. This is compounded by the standard information retrieval (IR) problem of mismatch between search queries and target items. Query Expansion (QE) has been shown to be an effect technique to improve IR effectiveness for multiple search tasks. In QE, words from a number of relevant or assumed relevant top ranked documents from an initial search are added to the initial search query to enrich it before carrying out a further search operation. In this work, we investigate the application of QE methods for searching social multimedia content. In particular we focus on social multimedia content where the information is primarily in the audio stream. To address the challenge of content variability, we introduce three speech segment-based methods for QE using: *Semantic segmentation*, *Discourse segmentation* and *Window-Based*. Our experimental investigation illustrates the superiority of these segment-based methods in comparison to a standard full document QE method for a version of the MediaEval 2012 Search task newly extended as an adhoc search task.**

I. Introduction

An ever increasing amount of user-generated multimedia content is being uploaded to online repositories. To facilitate access to this content, it is becoming increasingly important to develop advanced Information Retrieval (IR) techniques focused on the search challenges posed by this content. In this paper we explore the use of Query Expansion (QE) methods for IR for user-generated content where the information is primarily in the spoken data stream, for which search relies on Spoken Content Retrieval (SCR) techniques. Research on SCR initially investigated IR for planned speech content such as news broadcasts and documentaries [1], [2]. The focus then shifted towards spoken content that is produced spontaneously such as interviews, lectures and TV shows [3]. However, most existing work on SCR has been conducted on formally produced and curated content, and despite the upsurge in user generated spoken content on social media platforms, there has been little research on SCR for this content. In common with other IR tasks, a fundamental challenge for IR with social media content is the vocabulary mismatch between user queries and the relevant documents in the collection. This mismatch problem often occurs when queries are vague, short or imperfect or when documents are noisy, long or have a complex structure [4]. QE [5] is a popular technique suggested to bridge this vocabulary gap between the query and its relevant documents. In QE, an initial query is enriched using the top ranking documents returned by the retrieval system for an initial search using the original query. While QE techniques have been shown to improve retrieval in multiple SCR tasks [6], [7], they have not been studied for noisy and inconsistent User-Generated Content (UGC). In this case, QE effectiveness is hindered by errors in automated transcripts of the content, and also by topical, quality and length variations of the content. These issues can result in *query drift*, where QE changes the focus of the initial query. Query drift results from the extraction of poorly chosen expansion terms, where these terms often belong to a topic different to that of the original query [4].

The key novelty of the work reported here addresses QE for retrieval from a UGC multimedia archive where the information is primarily in the spoken media stream. Our work focuses on the blip10000 Internet video archive [8]. These videos were uploaded to the social video sharing site blip.tv by 2,237 different uploaders and covering a 25 different topics[1] with varying recording quality and differing lengths. The statistics of Automatic Speech Recognition (ASR) transcripts extracted from these videos are shown in Table I. Some of these videos, such as news broadcasts and TV shows are carefully authored, edited and quality controlled, while others such as videoblogs and personal recordings are not. The inconsistency in document lengths, content quality, together with the noise in the automatically generated transcripts presents challenges for IR systems. In this paper, we investigate the application of QE methods for this highly variable UGC content with errorful transcripts, and propose new QE techniques that utilise well-established text segmentation algorithms.

[1]Art, Autos and Vehicles, Business, Citizen Journalism, Comedy, Conferences and Other Events, Documentary, Educational, Food and Drink, Gaming, Health, Literature, Movies and Television, Music and Entertainment, Personal or Auto-biographical, Politics, Religion, School and Education, Sports, Technology, The Environment, The Mainstream Media, Travel, Videoblogging, Web Development/Sites

978-1-4673-8695-1/16/$31.00

TABLE I
WORD-LEVEL LENGTH STATISTICS FOR BLIP10000 ASR TRANSCRIPTS.

Stan.Dev	2399.5
Avg.Length	703.0
Median	1674.8
Max	20451.0
Min	10.0

The remainder of this paper is structured as follows: Section II provides some background and related work on SCR and QE, we then describe the speech segmentation techniques used in our experiments in Section III. Section IV introoduces our experimental settings. Sections V and VI describe our investigation of standard QE and our new segment-based QE method. Section VII concludes and highlights directions for further research in this area.

II. SPOKEN CONTENT RETRIEVAL

Speech retrieval systems generally make use of automatic speech recognition (ASR) systems to derive time-coded transcripts of the speech associated with videos. In most cases, these ASR transcripts serve as the primary basis and core component of the retrieval process. However, since ASR is an imperfect process, these transcripts generally contain insertion, deletion, and substitution errors. Such errors can adversely affect the SCR process. Researchers in SCR have studied the impact of these transcription errors on SCR behaviour, and have observed that interaction between transcription errors and SCR is rather complex and highly dependent on the retrieval task and the transcript quality [9].

SCR research has also studied methods for improving the errorful ASR transcripts for retrieval purposes [6], [7], [10]. For example, Singhal et al. [10] examined the use of Document Expansion (DE) to alleviate the effect of transcription mistakes on the retrieval effectiveness. Their work sought to recover those words that might have been in the original video, but had been mis-recognized by enriching document ASR transcripts with terms drawn from highly ranked documents that share the same topic. Previous work has also investigated document re-ranking using QE techniques for SCR. For example the work of [6], [7] investigated the effectiveness of document re-ranking approaches on professional video collections with consistent quality, length, style and topical structures. These videos were provided by the CLEF[2] speech retrieval tracks [3]. The main issue for these document re-ranking approaches (whether using QE or DE) is that they can be affected by the problem of query drift introduced in the previous section. which can impact negatively on SCR effectiveness. This is most commonly a problem when the documents are long and a single document may contain multiple topics, as discussed in [11]. In SCR, query drift can happen due to the relatively long ASR transcripts (such the ones shown in Table I) where a single spoken document may caontain multiple sub-topics. Several techniques have been explored to address this issue. For example, previous work on the CLEF collections [7], [6] utilised manually created summaries or manually segmented spoken content to improve the effectiveness of QE expansion. These summaries and segments were created manually by professional indexers and provided by the task organisers [3].

Unfortunately, having these manually generated summaries or segments within large-scale UGC video content is very unlikely due to the cost required to create them. Instead, we propose to use text segmentation techniques of ASR transcripts, such as the ones studied in [12] for passage retrieval and investigate their robustness/effectiveness for QE in SCR for UGC. Our work differs from prior QE SCR investigations by using an internet-based UGC multimedia collection; where audio data is highly variable in many aspects, including the audio conditions of the recording, the microphones used, the fluency and informality of the language used by the speaker. The most closely related work to that examined in this paper is our previous study [13], [14] which used the same blip10000 UGC data collection to evaluate the CLIR/monolingual SCR/QE effectiveness for different UGC modality such as UGC metadata and ASR transcripts in a known-item search task. This work indicated that the retrieval effectiveness of the ASR transcript can drop significantly when combined with translation at the CLIR. In this paper, we focus on improving the retrieval effectiveness of the ASR transcripts specifically by tuning their representation in semantic, discourse and window based segments, and propose a QE technique that utilises these segments to improve overall effectiveness.

III. SEGMENTING SPEECH TRANSCRIPTS FOR QE

Previous research on passage retrieval [15] grouped text segmentation into three classes: *Discourse* segmentation based on textual units (sentences, paragraphs, sections), *Semantic* segmentation based on the content and the topic of the text itself (e.g Textiling, C99), and *Window-based* methods where text is segmented based on a number of words or textual units. The utility of text segmentation has been studied extensively within the task of passage spoken retrieval which is concerned with retrieval of portions of documents that are relevant to a user query, and thus preventing, or at least reducing, the amount of non-relevant material presented to the user [16], [12]. However, our work is the first to investigate their utility for QE in SCR. We investigate the application of QE based on segments in user-generated spoken content retrieval; where we separate speech transcripts into shorter units. This seeks to reduce the impact of noise that may arise from ASR errors, and irrelevant subtopics appearing the top ranking documents. Our hypothesis is that *incorporating speech segments allows us to detect the relevant parts of a document and prevent QE from assigning a high score non-relevant content*.

The segmentation techniques we utilize for topic-based segmentation for use in QE are: *Semantic segmentation* based on the lexical cohesion within the ASR transcript using two

[2]http://clef2015.clef-initiative.eu/

well-established topic segmentation algorithms: TextTiling [17] and C99 [18]. *Discourse segmentation* into consecutive silence bounded utterances from the same speaker . We hypothesize that silence points can be useful for detecting topic boundaries where each point is considered a segment and can be used as QE evidence. *Window-based segmentation* in a fixed-length sequence (window) of words. This type of segmentation has long been suggested to be the most effective for multiple for QE [11], [4], [19] in text retrieval and spoken passage retrieval [12], [16].

TextTiling, developed by Hearst [17], is an unsupervised linear topic segmentation algorithm that uses cosine similarity to estimate similarity between blocks of words and assumes that similar words belong to the same topic. The calculation is accomplished using two vectors containing the number of terms occurring in each block. C99 was introduced by Choi [18], and uses a matrix-based ranking and a clustering approach, and assumes the most similar words belong to the same topic. To perform C99 and TextTiling segmentation, we used implementation in the UIMA[3] Text Segmentor[4] Similar to the work of [16], [12], the punctuation inserted by the ASR system was used as the sentence boundaries for segmentation. Since we do not have the ground truth segments of this dataset, it was not possible to fully optimise the hyper-parameters for these two algorithms. Instead we took a sample ASR document with a length of 752 words and manually evaluated the quality of the segments based on 4 different variations of parameters (included the default ones that are suggested by by Choi [18] in his implementation), and picked the one that produced the best segments in terms of detecting topic boundaries.

For speaker-turn segmentation, we used the ones produced by the ASR system based on detecting both silence points and the speakers' turn taking, where a new segment is produced whenever a silence point is detected. For the window-based segmentation, we segmented the transcripts into 20, 50, 100, 200, 300, 500 words fixed length, however we only report the 50, 100 and 500 segmentation in this paper due to space limitations. Stop word removal was done for all segmentation based on the standard Terrier list [5], and stemming performed using the Terrier implementation of Porter stemming [6].

IV. Experimental Settings

The blip10000 collection used in our experiments is described in detail in [8]. This collection is a crawl of the Internet video sharing platform Blip.tv[7]. It was originally used as the content dataset for the MediaEval 2012 Search and Hyperlinking task [**?**]. The blip10000 collection contains the crawled videos together with the associated ASR transcripts. The collection consists of 14,838 videos with a total running time of ca. 3,288 hours, and a total size of about 862 GB. For our investigation we remove all videos that do not contain any spoken content. We also filter out all non-English videos as well as those containing less than 10 words, This results in a final collection of 9,615 ASR transcripts. The statistics of these transcripts are shown in Table I which indicates the significant variations in the length between the videos. Of particular relevance to our QE investigation are the following challenging aspects of the data:

Structure and topical variations : As explained before, videos were uploaded by 2,237 different individuals. They have differing recording styles and covering 25 topical areas. Such variation poses challenges for term selection in QE. *Distribution of the document lengths*: Variations in document length pose challenges for any retrieval task. *High variability in ASR quality*: Even though the same ASR system is used, the variation in the audio quality, speaking styles and speakers leads to significant variability in the quality of the transcripts.

The MediaEval 2012 Search and Hyperlinking task [**?**] 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[8] (MTurk) crowd-sourcing platform. Each query contains a full query statement providing a detailed description 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).

For our work, we created an adhoc version of these queries by manually generalising the known-item ones. This was done by removing specific terms that are related to one relevant item and re-writing the whole query into more natural adhoc form; for example the query "Troubleshooting the EEE PC 900 Laptop" was changed to "Troubleshooting PC and Laptops". To create the relevance judgements for these adhoc queries, a pooling method was used by combining different result lists created using different IR methods . We ran each query with 6 different retrieval models (TFIDF, Okapi, PL2, language modeling (LM) Himestra, DLH13) in different indexes. These indexes included combinations between the textual metadata associated with each video (title and descriptions) and the ASR transcripts. Retrieval runs were produced using the Terrier retrieval platform[9]. Stop words were removed based on the standard Terrier list, and stemming performed using the Terrier implementation of Porter stemming. We then used the NTCIR pooling script[10] to generate the top 30 results for each query. We adjusted *Relevation*[11], an open source IR relevance judging system introduced in [20], to embed videos together with their metadata and assigned two fluent bilingual English reviewers to evaluate the results of each query. The agreement level between the reviewers was 93%. We produced a relevance file containing each query together with the list of videos selected as relevant by both reviewers. Each query has between 7-13 relevant videos with an average of 9 relevant items per query, this number depends mostly on the difficulty of each query and the availability of relevant documents in the collections.

[3]http://uima.apache.org/UIMA
[4]https://code.google.com/p/uima-text-segmenter/
[5]http://terrier.org/docs/v2.2.1/javadoc/uk/ac/gla/terrier/terms/Stopwords.html
[6]http://terrier.org/docs/v4.0/javadoc/org/terrier/terms/PorterStemmer.html
[7]http://blip.tv/
[8]http://www.mturk.com/
[9]http://www.terrier.org/
[10]http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html
[11]https://github.com/ielab/relevation

TABLE II
INDEX STATISTICS: NUMBER OF INDEXED DOCUMENTS (DOCS), AVERAGE SEGMENT LENGTH (AVG.LEN), LENGTH STANDARD DEVIATION (ST.LEN) AND AVERAGE NUMBER OF SEGMENTS ASR TRANSCRIPT (SEGS-DOC)

	docs	Avg.len	St.len	Segs-doc
SP	902209	22.83	43.8	102.5
C99	57104	400.3	405.5	6.287
Textile	795770	28.2	25.1	105
fix50	457347	49.5	4.0	47.6
fix500	50508	463.7	105.8	5.3
fix100	231244	98.1	11.1	24.1

TABLE III
QE RUNS USING FULL ASR DOCUMENTS

Docs_Terms	MAP	Recall	P@10
No QE	0.5887	0.9	0.545
3_3	0.6109	0.9167	0.5517
3_5	0.5753	0.9167	0.5433
3_10	0.5801	0.8833	0.5433
5_3	0.5753	0.9167	0.5183
5_5	0.5763	0.9167	0.5183
5_10	0.5677	0.8833	0.5133

For our QE investigation, we built 7 different indexes for the ASR transcripts as follows: *ASR* indexes each ASR transcripts as one document, *fix50* indexes each 50 words length window, *fix100* indexes each 100 words length window, *fix500* indexes each 500 words length window, *C99* index which considers the C99 produced segments as a document, the *TextTile* index uses the Textile segments and *SP* index which uses the speaker segments of the collection transcripts. The statistics for each of these segment indexes are shown in Table II. Comparing these segment indexes to the ASR index shown in Table I, it can be seen that all of them produced larger indexes with less length variation. The SP and TextTiling indexes appear to be the largest with an average of over 100 segments per ASR document, but also produced the shortest average segment length. Fixed-length segments are obviously much more consistent in length.

We used the PL2 IR model, a probabilistic retrieval model from *the Divergence From Randomness (DFR)* framework [21]. We selected this model over other available retrieval models after preliminary experiments showed that this model is more suitable for our task and data collection. Previous studies such as [22] have shown that PL2 has less sensitivity to length distribution compared to other retrieval models. PL2 is thus suitable since our Internet-based data collection has huge variation in document lengths as shown in Table I. The PL2 document scoring model is defined as shown in Equation 1, 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 2, 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 in [21], [22], which is $c = 1$.

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

$$tf_n = \sum_d (tf. \log_2 (1 + c. \frac{avg_l}{l})), (c > 0) \quad (2)$$

For QE, we use the *Divergence From Randomness* (DFR) QE mechanism [21] to weight the top extracted terms from the top documents. DFR QE generalizes Rocchios method to implement several term weighting models that measure the informativeness of each term in a relevant or pseudo relevant set of documents. DFR first applies a term weighting model to measure the informativeness of the top ranking terms (top-terms) in the top ranking documents (top-doc). The main concept of the DFR term weighting model is to infer the informativeness/importance 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, which is 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 the most effective [21], [22]. The weight w of a term t in the top-ranked documents using the DFR Bo1 model is shown in Equation 3, 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.

$$w(t) = tf_x . \log_2 (\frac{1 + P_n}{P_n}) + \log_2 (1 + P_n) \quad (3)$$

The query term weight qt_w, which was obtained from initial retrieval is further adjusted according the newly obtained weighting values of $w(t)$ for both the newly extracted terms and the original ones as using Equation 4, where the $w_{max}(t)$ is indicated by the maximum obtained $w(t)$ of the expanded query terms.

$$qt_w = qt_w + \frac{w(t)}{w_{max}(t)} \quad (4)$$

Some of the original query terms may *not* appear in the top-terms, in which the above formula would only give them the same weight they would have in single-pass retrieval settings.

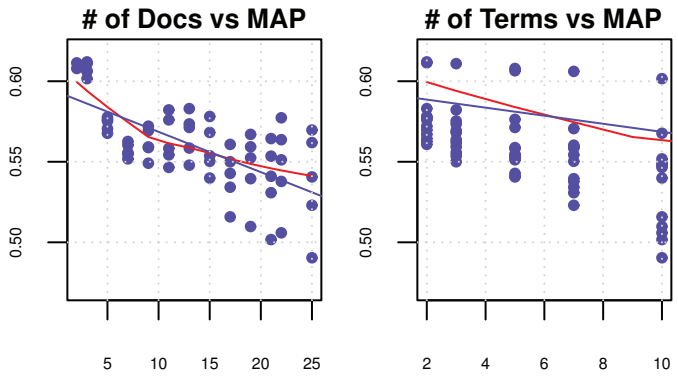

Fig. 1. MAP performance for alternative terms and documents values.

V. Query Expansion For User-Generated Speech Content

In the first part of our investigation, we evaluate the effectiveness of traditional QE for internet-based speech document retrieval. We run QE by taking full ASR documents as feedback evidence. For our initial runs, we explore taking the top 3 and top 5 terms from the top ranking 3, 5 and 10 documents produced by running the initial query. This produced a total of 6 different QE runs which we compare to the baseline run which does not involve any expansion (no QE). We used the retrieval model PL2 and QE model BO1 described in the previous section (see Equations 1,3) to calculate the results shown in Table III, which shows performance in terms of Mean Average Precision (MAP), Recall and Precision for top 10 documents (P@10). The results in Table III indicate that none of the QE runs actually improve performance significantly over the baseline. We tested the statistical significance at $p < 0.05$ for each run by computing the difference at the query level between the precision obtained by baseline (no QE) and that obtained using QE, and none were significant.

As can be seen from these results, QE performance is better when using lower numbers of terms and documents, but the improvement achieved is not significant. The reason for this is that taking less documents for feedback reduces the chance of dealing with noise that may appear in the top ranking documents. This noise can be attributed to the non-relevant segments of these top ranking documents which introduce query drift for some queries and impact on the overall improvement in MAP. To verify this assumption, we carried out further QE runs with additional *top term* parameters of 2,7 and 10 terms. We also extended the *top document* parameters to be explored to the range 2 to 25 with an increment of 2. This produced a total of 60 QE runs (in addition to the ones reported in Table III) to evaluate the performance of this approach. Figure 1 shows how the chance of query drift after QE increases when more documents and more terms are explored. The noise associated within these ASR documents provides a limitation for QE effectiveness when selecting new terms from relevant speech transcripts. In the next section we seek to avoid this noise by using the relevant segments as evidence for QE.

TABLE IV
Retrieval performance results for optimal QE runs

	Doc-Terms	MAP	Recall	P@10
ASR	3-3	0.6109	0.9167	0.5517
SP	9-2	0.6043	0.9167	0.56
Tt	22-5	0.612	0.9	0.5633
C99	5-5	**0.6429**	0.9167	**0.585**
fix50	25-10	**0.6452**	**0.9333**	**0.585**
fix100	25-5	**0.6468**	**0.9333**	**0.59**
fix500	25-3	**0.6403**	0.9167	**0.585**

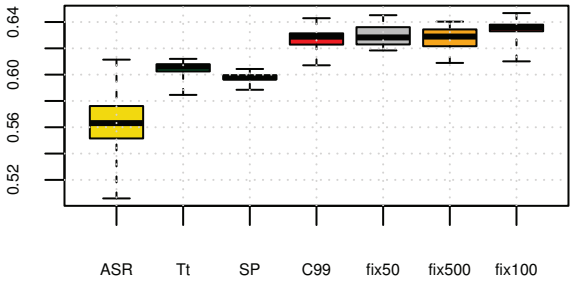

Fig. 2. MAP performance for each QE runs using all tuning parameters.

VI. Investigating Segment-Based Query Expansion

Previous work has explored multiple approaches for utilising segments in QE for text IR tasks [11], [4], [19]. However, no such study has been reported on spoken content. Also, prior work focused only on window-based and discourse-segments for QE. For example, Allan et al.[11] suggested using only long passages from feedback documents. Local Context Analysis (LCA) is by far the most well known approach to utilisation of segments for QE. Proposed by Xu and Croft [19], LCA works by retrieving the top-ranking window-passages from the documents and using them for QE. We implement a similar approach to the LCA technique to investigate the robustness of different segments evidence for QE, where the top ranking segments are taken from segment indexes shown in Table II. The segment-based QE used in this experiment was were carried out as follows:

- The highest scoring terms were extracted from the top ranking documents retrieved in response to executing the query on each separate indexes in Table II. This allows us to employ a passage-retrieval system during the expansion phase.
- Query retrieval was then done using the full ASR index shown in Table I.

We analysed the performance of 6 different QE runs based on C99, SP, Tt, fix50, fix100 and fix500 segmentation of the ASR transcript. Parameter settings for these QE runs were tuned similarly those described in the previous section, where we generated 60 runs for each of the segmentation schemes.

Figure 2 shows the MAP performance obtained by each QE run, including for comparison, the full ASR run obtained in the previous section. It can be seen that generally segment-based QE runs were less affected by query drift, even when more terms and documents were included. This demonstrates the robustness of the segmentation QE approach in terms of sensitivity to noise in the top ranking documents compared to the full ASR approach. The *robustness* of these segment-based QE approaches can be attributed to the fact that segments are more likely to avoid unnecessary noise in the feedback documents, as well as allowing relevant terms to be selected from those segments which appear within long ASR transcripts which may contains multiple non relevant topics.

Table IV shows the best retrieval performance obtained for each QE run. It can be seen that some segment-based QE runs based on C99 and fixed-length segmentation yielded statistically significant performance (highlighted in **bold**) improvement over the full ASR QE and over the baseline; while others (TextTile and SP) did not. This indicates that these C99 and fixed-length segments produced much better segmentation for QE than the others. The effectiveness of these approaches over others can be attributed to their ability to more reliably detect the topic boundaries for this data collections, and hence provide a better ranking of the segments used as feedback documents in QE. As shown in Table II, Textiling and SP have the highest average number of segments per document (Segs-doc), this may produce many unsatisfactory segments that can rather harm the ranking of relevant segments.

VII. Conclusion and Future work

In this work we studied the issues and challenges of applying QE approaches for a collection of user-generated short videos where the information is primarily in the spoken data stream collected from the Internet. We investigated tuning the representation of the spoken source into three type of segments for QE (Semantic, Discourse and Window-based), to improve the effectiveness of QE in this setting. Our experiments show that using segment evidence for QE is more robust than using full document evidence in terms of dealing with query draft issues. We found that QE using C99 and fixed-length segmentation produces the most effective segmentation for the utility of QE for this task.

This work opens directions for multiple potential further investigations for improving document re-ranking techniques for similar tasks and dealing with noise and topic drift issues. QE techniques that involve the combination of multiple segmentation evidence for each query would be an interesting idea to investigate in future work. For example, we plan to develop a QE approach that automatically predicts the quality of each segment based on different heuristics and select the one that is predicted to be most robust and effective for the current query.

Acknowledgements

This research was partially supported by Science Foundation Ireland as part of ADAPT Centre (Grant No. 13/RC/2106).

References

[1] M. Federico and G. J. F. Jones, "The clef 2003 cross-language spoken document retrieval track," in *Comparative Evaluation of Multilingual Information Access Systems*. Springer, 2004, pp. 646–652.

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

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

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

[5] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys (CSUR)*, vol. 44, no. 1, p. 1, 2012.

[6] J. Wang and D. W. Oard, *Clef-2005 cl-sr at maryland: Document and query expansion using side collections and thesauri*. Springer, 2006.

[7] A. M. Lam-Adesina and G. J. Jones, *Dublin city university at CLEF 2005: cross-language speech retrieval (CL-SR) experiments*. Springer, 2006.

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

[9] M. Eskevich and G. J. F. Jones, "Exploring speech retrieval from meetings using the ami corpus," *Computer Speech & Language*, 2014.

[10] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 34–41.

[11] J. Allan, "Relevance feedback with too much data," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995, pp. 337–343.

[12] C. Wartena, "Comparing segmentation strategies for efficient video passage retrieval," in *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*. IEEE, 2012, pp. 1–6.

[13] A. Khwileh, D. Ganguly, and G. J. F. Jones, "An investigation of cross-language information retrieval for user-generated internet video," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, 2015, pp. 117–129.

[14] A. Khwileh, D. Ganguly, and G. J. Jones, "Utilisation of metadata fields and query expansion in cross-lingual search of user-generated internet video," *Journal of Artificial Intelligence Research*, vol. 55, pp. 249–281, 2016.

[15] J. P. Callan, "Passage-level evidence in document retrieval," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 1994, pp. 302–310.

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

[17] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[18] F. Y. Y. Choi, "Advances in domain independent linear text segmentation," in *Proceedings of NAACL'00*, 2000.

[19] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 4–11.

[20] B. Koopman and G. Zuccon, "Relevation!: an open source system for information retrieval relevance assessment''

[21] A. Gianni, "Probabilistic models for information retrieval based on divergence from randomness," Ph.D. dissertation, Department of Computing Science, University of Glasgow, 2003.

[22] G. Amati and C. J. Van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 357–389, 2002.