DCU at NTCIR-10 SpokenDoc2 Passage Retrieval Task

Maria Eskevich
Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin 9, Ireland
meskevich@computing.dcu.ie

Gareth J. F. Jones
Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

We describe details of our runs and the results obtained for the "2nd round of IR for Spoken Documents (SpokenDoc2)" task. We participated in the passage retrieval from the Corpus of Spoken Document Processing Workshop (SDPWS) task. For our participation in the NTCIR-9 SpokenDoc task, we investigated the use of different content-based segmentation methods that attempt to identify topically coherent units for retrieval. For NTCIR-10 we compare content-based segmentation (the TextTiling algorithm) to division of the content into segments of a fixed number of Inter-Pausal Units (IPUs) using a sliding window, and subsequent combination of overlapping segments into single units in the ranked list of results. Another focus of our submissions to NTCIR-10 is the potential for use of external data for document expansion. For this we used a DBpedia collection for IPU expansion for all segmentation methods.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Experimentation

Keywords

Speech search, passage retrieval, automatic segmentation, document expansion

1. INTRODUCTION

Recorded spoken content varies in quality and potential purpose of use. Therefore the techniques to be developed for its processing and subsequent search and retrieval focus on different use case scenarios. The NTCIR-10 SpokenDoc2 Task continues the exploration into methods for retrieval of spoken terms, passages and whole documents [2] introduced in the NTCIR-9 SpokenDoc Task [1]. DCU participated in the subtask that targeted adhoc spoken passage retrieval from within the Corpus of Spoken Document Processing Workshop (SDPWS) that consists of recordings of oral presentations (26.8 hours). Three official evaluation metrics were used: utterance-based measure (uMAP), passage-based measures: pointwise MAP (pwMAP) and fraction MAP (fMAP).

Previous experiments on passage retrieval have shown that the use of overlapping fixed length segments created using a sliding window combined with removal of overlapping segments in the retrieved results list [1] [7] produces higher mean average precision based scores than the use of lexical coherence based segmentation [4]. This can be explained by the fact that lexical coherence based methods produce segments of varying length, and even with length normalisation perform less consistently in retrieval than fixed length segments. However segments of fixed length do not correspond to the real situation where the length of the topically coherent passages varies and the regions of relevant content within them can also vary depending on the query. Moreover comparison of different segmentation techniques has demonstrated that segments with boundaries that are closer to the actual relevant content boundaries perform better [5]. Therefore we believe that we should continue to investigate the application of lexically coherent segmentation approaches that target relevant passages of varying length. At the same time we experiment with different sizes of fixed length segments, in combination with a query dependent technique to combine overlapping segments together into segments of different length.

Errors in recognition of the spoken the content may affect the retrieval results. At the same time even a perfect transcript might not contain all the information necessary to describe the semantic content of a spoken passage. For example, the speaker may not explicitly read aloud the text on a set of slides being projected while making their presentation. In this case all those attending the presentation thus have information describing the topic under discussion that does not appear in the spoken soundtrack. Such absence of spoken description can impact on retrieval effectiveness. One way to address this problem is to perform document expansion using additional data to include additional descriptive words relevant to the topic of the content with the objective of increasing its retrievability. Previous work has shown that document expansion for spoken data can reduce the loss in retrieval performance due to recognition errors [8]. An external text document collection well matched topically to the topic of the document being expanded is preferable since it is less likely to distort the topical and detail focus of the expanded document. However, suitable topically corresponding collections may not be available and would be costly and time consuming to create. Therefore researchers have attempted to use general collections for the purpose of expansion as e.g. Wikipedia[1] or its structured shorter

[1]http://www.wikipedia.org

version DBpedia[2]. In order to explore the potential for document expansion using general resources in SpokenDoc2, for our experiments we used the Japanese DBpedia[3] as an expansion source. As the provided dataset does not contain any segment boundaries, we consider each IPU as a document to be expanded. Afterwards the expanded IPUs can be segmented using different segmentation methods which allows us to examine the effect of the use of external knowledge on retrieval effectiveness.

This paper is structured as follows: Section 2 describes the methods we used to prepare, expand, and search the test collection, Section 3 gives details of the results achieved and analysis of the system performance, and finally Section 4 concludes and outlines directions for our future work.

2. RETRIEVAL METHODOLOGY

In this section we give an overview of the tools and methods we applied to perform the NTCIR-10 SpokenDoc passage retrieval task. Task participants were provided with several transcripts of the spoken content. We first extracted words from the transcripts and from the Japanese DBpedia collection. Since there are no topical boundaries marked within the transcripts, we applied alternative segmentation methods to define alternative target retrieval passages.

2.1 SDPWS Transcripts

Task participants were provided with both automatically and manually created transcripts of the oral presentations. Two types of transcripts were created using automatic speech recognition (ASR): n-best word-based and syllable-based [2]. Two forms of each of these transcript types were created using either matched or unmatched language models. For our participation in the task, we used both matched and unmatched 1-best word-based transcripts. For comparison we also used the manual transcript provided by the task organisers.

2.2 Japanese Data Preprocessing

In Japanese the individual morphemes of the sentences need to be recognized for further processing. We used the ChaSen system, version 2.4.0[4], based on the Japanese morphological analyzer JUMAN, version 2.0, with ipadic grammar, version 2.7.0, to extract the words from the sentences in ASR and manual transcripts, and in the external data collection - DBpedia. ChaSen provides both conjugated and base forms of the word, for later processing we used the latter since it avoids the need for stemming of different word forms.

2.3 IPU Expansion

We used the open-source Terrier information retrieval platform[5] to identify the expansion terms for each IPU. We represented each IPU as a query for the DBpedia document collection and carried out standard query expansion implemented in Terrier. We took 5 terms from the 10 top retrieved documents. The use of IPU expansion is marked in the run name by the addition of `_e` to the name of the run, in cases with no expansion `_ne` is added.

[2]http://dbpedia.org/
[3]http://ja.dbpedia.org/
[4]http://chasen-legacy.sourceforge.jp
[5]http://www.terrier.org

Table 1: Average length of relevant passages and passages containing relevant content per run (in IPUs).

Average Length of relevant passages: 7.27.

	manual	asr_matched	asr_unmatched
tt_ne	35.50	46.92	43.17
tt_e	20.78	27.75	23.92
Segm_5_3_ne	–	8.85	8.26
Segm_5_3_e	–	9.58	9.03
Segm_10_5_ne	–	23.86	19.92
Segm_10_5_e	–	23.69	21.48
Segm_15_7_ne	–	56.29	50.29
Segm_15_7_e	–	53.72	47.54

2.4 Text Segmentation

Our previous research on the Japanese data in the SpokenDoc Task at NTCIR-9 [4] showed that TextTiling [6] produces shorter segments than C99 segmentation [3], and that it achieved higher scores in retrieval. Thus we use only TextTiling lexically coherent segmentation for our SpokenDoc2 experiments. TextTiling uses cosine similarities between adjacent blocks of sentences in a text document to predict topical boundary points.

For fixed length segmentation we chose the values of 5, 10, 15 IPUs with a corresponding sliding window of 3, 5 and 7 IPUs. Our runs adopt the following naming convention: `Segm_<Segment_Length>_<Sliding_Step>`.

2.5 Retrieval Setup

For retrieval experiments we used the open-source Terrier information retrieval platform[6] with a standard language modelling method, with *lamda* equal to 0.35.

2.6 Post-editing of the results for fixed length segentation methods

The result list for the fixed length segmentation with sliding window approach contains overlapping segments. Previous research has shown that simple removal of overlapping segments further down in the retrieved ranked list is an effective method to improve retrieval effectiveness [9] [1]. However our assumption is that since our target segments might be longer than fixed length segments, it is reasonable to try to combine the segments instead of removing them. In cases where there are overlapping segments in the retrieved list, we put the overall segment at the highest of the rank positions. We carry out post-editing of the lists individually for each query, thus the new length of the segments varies depending on the search request.

3. RESULTS

In this section we give an overview of the results for the different runs according to the NTCIR-10 metrics. Figures 1, 2 and 3 show the results for the three metrics (uMAP, pwMAP and fMAP) respectively. Table 1 shows the average length of the actual relevant passages and passages containing relevant content per run (in IPUs), while Table 2 contains the average length of all the segments in the retrieved ranked lists.

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

Table 2: Average length of all segments in the retrieved ranked lists (in IPUs).

	manual	asr_matched	asr_unmatched
tt_ne	39.78	52.03	49.80
tt_e	23.07	30.13	27.24
Segm_5_3_ne	–	7.01	7.05
Segm_5_3_e	–	7.26	7.44
Segm_10_5_ne	–	15.26	15.39
Segm_10_5_e	–	15.55	15.71
Segm_15_7_ne	–	27.69	27.99
Segm_15_7_e	–	27.66	28.29

Comparison of Tables 1 and 2 demonstrates an interesting trend: segments containing relevant content are shorter than average segments for lexical coherence based segmentation runs, while the fixed length segmentation runs follow the opposite trend.

Across all metrics the runs that use the language model (LM) that match the collection show better results than the runs which use the unmatched LM. Use of DBpedia for IPU expansion does not help the asr_unmatched runs to achieve the same scores as asr_matched. However the expansion consistently improves the results of longer segments (tt, Segm_10_5, Segm_15_7) for asr_unmatched runs according to the pwMAP score.

The pwMAP metric only counts as relevant segments for which the IPU in the centre of the segment is relevant. Since shorter segments have a greater likelihood of having the relevant content in the centre, runs Segm_5_3 achieve higher pwMAP scores for asr_matched transcripts. Since the fMAP metric is designed to capture the relevancy of segments, the Segm_5_3 runs receive higher fMAP scores as well.

4. CONCLUSION

This paper reports the methods and results for our participation in the NTCIR-10 SpokenDoc2 passage retrieval task. As could have been expected runs using the matched ASR transcript achieve better results than those using the unmatched transcripts. However performance of runs using the unmatched ASR transcript can sometimes be improved with the use of DBPedia as a general external knowledge source for document expansion. This improvement is only captured by one of the benchmark metrics. Further investigation of document expansion will focus on understanding how it modifies retrieval as measured by the other retrieval metrics, and will seek to develop methods to apply it more reliably to improve overall retrieval effectiveness.

5. ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland (Grant 08/RFP/CMS1677) Research Frontiers Programme 2008 and (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at DCU.

6. REFERENCES

[1] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui. Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, Tokyo, Japan, 2011.

[2] T. Akiba, H. Nizhizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, and Y. Yamashita. Overview of the NTCIR-10 SpokenDoc-2 Task. In *Proceedings of NTCIR-10 Workshop Meeting*, Tokyo, Japan, 2013.

[3] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33, Seattle, Washington, USA, 2000.

[4] M. Eskevich and G. J. Jones. DCU at the NTCIR-9 SpokenDoc passage retrieval Task. In *Proceedings of NTCIR-9 Workshop Meeting*, Tokyo, Japan, 2011.

[5] M. Eskevich, G. J. F. Jones, M. Larson, C. Wartena, R. Aly, T. Verschoor, and R. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *Proceedings of the 10th Workshop on Content-Based Multimedia Indexing (CBMI 2012)*, Annecy, France, 2012.

[6] M. Hearst. TextTiling: A quantitative approach to discourse segmentation. Technical Report Sequoia 93/24, Computer Science Department, University of California, Berkeley, 1993.

[7] T. Kaneko, T. Takigami, and T. Akiba. STD based on Hough Transform and SDR using STD results: Experiments at NTCIR-9 SpokenDoc. In *Proceedings of NTCIR-9 Workshop Meeting*, Tokyo, Japan, 2011.

[8] 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*, SIGIR '99, pages 34–41, Berkeley, California, USA, 1999. ACM.

[9] C. Wartena and M. Larson. Rich speech retrieval using query word filter. In *Proceedings of the MediaEval 2011 Workshop*, Pisa, Italy, 2011.

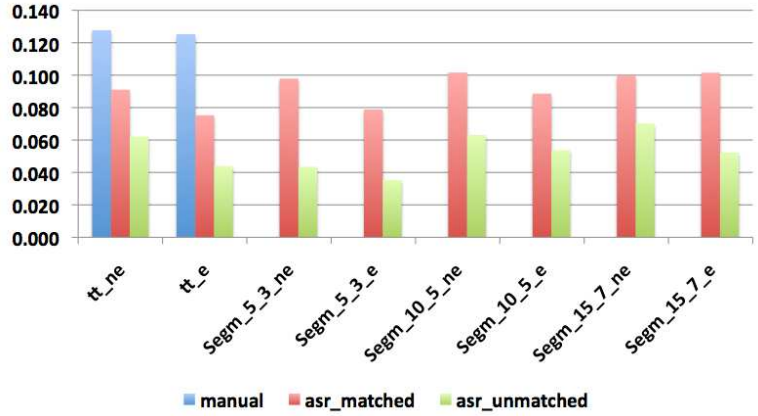

Figure 1: Scores for Utterance-based Measure (uMAP).

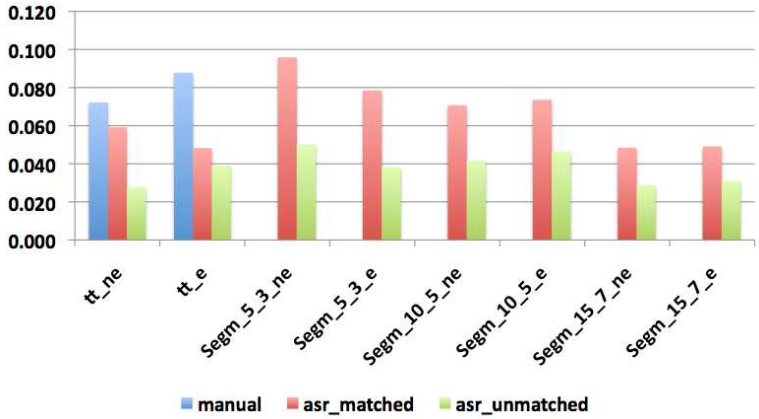

Figure 2: Scores for pointwise MAP (pwMAP).

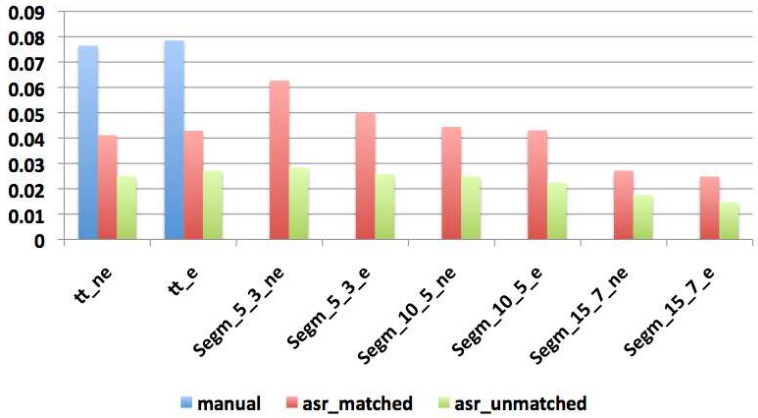

Figure 3: Scores for fraction MAP (fMAP).