DCU at the NTCIR-9 SpokenDoc Passage Retrieval Task

Maria Eskevich
Centre for Digital Video Processing
School of Computing
Dublin City University
Dublin 9, Ireland
meskevich@computing.dcu.ie

Gareth J. F. Jones
Centre for Digital Video Processing
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 "IR for Spoken Documents (SpokenDoc) Task" at NTCIR-9. The focus of our participation in this task was the investigation of the use of segmentation methods to divide the manual and ASR transcripts into topically coherent segments. The underlying assumption of this approach is that these segments will capture passages in the transcript relevant to the query. Our experiments investigate the use of two lexical coherence based segmentation algorithms (TextTiling, C99). These are run on the provided manual and ASR transcripts, and the ASR transcript with stop words removed. Evaluation of the results shows that TextTiling consistently performs better than C99 both in segmenting the data into retrieval units as evaluated using the centre located relevant information metric and in having higher content precision in each automatically created segment.

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

Team Name: DCU

Subtask: SpokenDoc Passage Retrieval

External Resources Used: ChaSen, SMART with language modeling, TextTiling, C99

1. INTRODUCTION

The rapid increase in the availability of digital audio data collections is creating growing interest in the development of effective spoken content retrieval technologies. Spoken datasets can differ in style and the form of the contents, leading to differing challenges to effective search. Earlier work on spoken document retrieval focused mainly on well structured spoken content recorded in controlled recording environments such as broadcast news [1]. Current interest focuses on less formally structured speech such as lectures, conversational interviews and socially contributed recordings. Speech search tasks can vary from seeking to locate individual spoken terms to the retrieval of passages, whole documents or even playback jumpin points in these items. Since processing of speech data itself requires a lot of computational power, the speech retrieval process is usually divided into two parts: automatic speech recognition (ASR) (or manual transcription of the content which is time consuming and therefore used mostly in creating datasets for research development and not real applications); and the retrieval process that is performed over the transcripts.

The NTCIR-9 "IR for Spoken Documents (SpokenDoc)" has two tracks for search of spoken content in 2011 - Spoken Term Detection (STR) and Spoken Document Retrieval (SDR) which in turn has two sub-tasks - lecture retrieval and passage retrieval [2]. DCU participated in the SDR passage retrieval sub-task.The target was to find relevant passages in 2702 lectures from the Corpus of Spontaneous Japanese (CSJ) [7]. Three official evaluation metrics were used: utterance-based measure (uMAP), passage-based measures: pointwise MAP (pwMAP) and fraction MAP (fMAP).

This paper is structured as follows: Section 2 describes the methods we used to prepare 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

Speech retrieval involves several data processing steps. In this section we give an overview of the tools and methods we applied to perform speech retrieval for the NTCIR-9 SpokenDoc passage retrieval task.

2.1 Lecture Transcripts

Task participants were provided with n-best word-based and syllable-based automatic recognition transcriptions of the lectures [2]. For our participation in the task, we used only the 1-best word-based transcripts. For comparison we also used the manual transcript of the lectures taken from the Corpus of Spontaneous Japanese [7].

2.2 Transcript 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[1], based on the Japanese morphological analyzer JUMAN, version 2.0, with ipadic grammar, version 2.7.0, to extract the words from the sentences

[1] http://chasen-legacy.sourceforge.jp

in ASR and manual transcripts. 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 words forms.

2.3 Text Segmentation

Our investigation focused on the segmentation of the transcripts into topically coherent passages to be used as retrieval units. Our objective was to explore the use of segment units to retrieve relevant content on the assumption that these units will capture relevant passages. We explored the application of two segmentation algorithms originally developed for segmentation of written text documents - C99 [4] and TextTiling [5].

The C99 algorithm computes similarity between sentences using a cosine similarity measure to form a similarity matrix. Cosine scores are then replaced by the rank of the score in the local region and segmentation points assigned using a clustering procedure. TextTiling looks at the cosine similarities as well, but only between adjacent blocks of sentences.

Both algorithms work with the fundamental unit of the sentence placing segment boundaries between the end of one sentence and the start of the next one. Since the ASR transcripts did not contain punctuation, we considered each Inter-Pausal Unit (IPU) to be a sentence on its own. We ran the segmentation algorithms on both ASR and manual transcripts, and on the ASR transcript when stop words had been removed from the text[2] (asr_nsw).

2.4 Retrieval Setup

The segments obtained using each segmentation technique from the manual transcripts were indexed for search using a version of SMART information retrieval system [3] extended to use language modelling (a multinomial model with Jelinek-Mercer smoothing) with a uniform document prior probability [6]. Equation 1 shows how a query q is scored against a document d within the SMART framework.

$$P(q|d) = \prod_{i=1}^{n} (\lambda_i P(q_i|d) + (1 - \lambda_i) P(q_i)) \qquad (1)$$

where $q = (q_1, \dots q_n)$ is a query comprising of n query terms, $P(q_i|d)$ is the probability of generating the i^{th} query term from a given document d being estimated by the maximum likelihood, and $P(q_i)$ is the probability of generating it from the collection and is estimated by document frequency. The retrieval model used $\lambda_i = 0.3$ for all q_i, this value being optimized on the TREC-8 ad hoc dataset.

Separate retrieval runs were carried out for each topic for each segmentation scheme for segments created from the manual and ASR transcripts.

3. RESULTS

The official evaluation metrics for this task are variations of the standard Mean Average Precision (MAP). These are applied to the list of the retrieved items after expanding the retrieved passages into IPUs. In case of the uMAP metric, relevance is assigned to individual IPUs in a relevant region of the lecture, and uMAP is calculated for relevant segments at the level of IPUs. For pwMAP, relevance is assigned to the whole passage retrieved at a certain rank if its centre IPU is part of the relevant content, the score is then calculated for retrieved passages classified as relevant according to this criteria. Recall of a passage and precision up to its rank at IPU level are taken into consideration for the fMAP calculation.

Table 1: Scores for official metrics

Transcript type	Segmentation type	uMAP	pwMAP	fMAP
BASELINE		0.0670	0.0520	0.0536
manual	tt	0.0859	0.0429	0.0500
manual	C99	0.0713	0.0209	0.0168
ASR	tt	0.0490	0.0329	0.0308
ASR	C99	0.0469	0.0166	0.0123
ASR_nsw	tt	0.0312	0.0141	0.0174
ASR_nsw	C99	0.0316	0.0138	0.0120

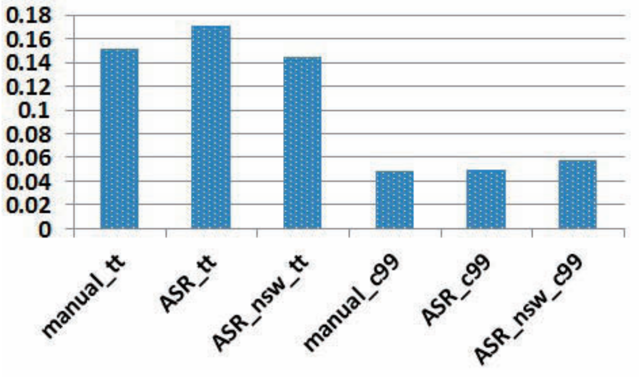

Figure 1: Average of Precision for all passages with relevant content.

Table 1 shows our experimental results for these metrics along with the baseline scores provided by the task organisers. It can be seen that, as would be expected, runs using manual transcripts show better results than those based on ASR trascripts. However manual runs outperform the baseline only for one metric (uMAP): 0.0859 and 0.0713 for TextTiling and C99 respectively versus 0.0670. It can be seen that transcript segmentation using TextTiling consistently achieves higher scores than segmentation using the C99 algorithm for all types of transcript.

The remainder of this section provides a more detailed analysis of our results for each of the evaluation metrics.

3.1 uMAP Results

The uMAP metric calculates MAP on the level of IPUs after each retrieved passage has been expanded into its constituent IPUs and they have been rearranged so that the relevant IPUs are at the beginning of the sequence.

In order to better understand the relationship between our retrieved segments and the amount of relevant content that we had actually retrieved, we calculated the precision of the content for each retrieved segment which contained at least one relevant IPU. We then calculated the average of these precision values for each topic and then the average of these values across the completed topic set. Although we process the transcripts and return as output the numbers of start

[2] Stop words were taken from SpeedBlog Japanese Stop-words: `dnnspeedblog.com/SpeedBlog/PostID/3187/Japanese-Stop-words`

[3] ftp://ftp.cs.cornell.edu/pub/smart/

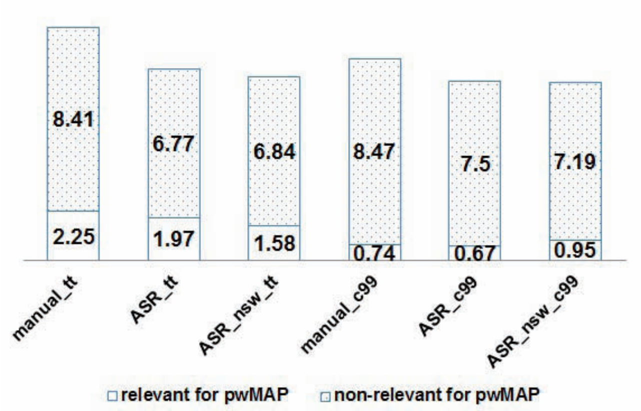

Figure 2: Number of ranks with relevant content that are taken or not taken into account for calculation pwMAP

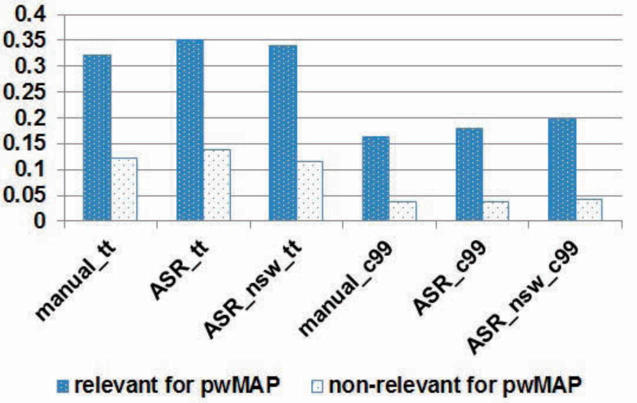

Figure 3: Average of Precision for the passages with relevant content that are taken or not taken into account for calculation pwMAP

and end IPUs (passages), our ultimate goal is to provide the user with segments to listen to. Therefore the actual timing information of the beginning and end points of relevant data are important for the analysis of results. This is especially true since IPUs may differ considerably in time length and this is not included by any of the metrics. Thus the precision value of each segment was calculated using the length in time for each IPU unit provided with the ASR transcript. Figure 1 shows these averaged values for both TextTiling and C99 for manual, ASR and ASR with stop words removed transcripts. From these results it can be seen that, similar to the official results in Table 1, TextTiling outperforms C99 in all cases. Comparing the results for the three different transcripts in each case, no clear trend emerges in terms of precision of the contents of the individual segments, which is perhaps a little surprising, since the results in Table 1 show a clear trend that manual transcripts outperform ASR with respect to uMAP which outperforms ASR without stop words.

3.2 pwMAP Results

pwMAP metric counts as relevant only segments for which the IPU in the centre of the segment is relevant. The results in Table 1 show that none of our methods was competitive with the provided baseline result with respect to pwMAP. This contrasts with the uMAP results, and indicates that although we are able to retrieve similar amounts of relevant content at similar ranks, the content segmentation methods that we are applying do not reliably place relevant content at the centre of the retrieved segments.

In order to analyze the scores further, we calculated the number of the segments in each run that were counted by the metric as relevant and the ones that had relevant content, but it was not located in the centre of the retrieved segment and was therefore overlooked by the pwMAP metric. Figure 2 shows the average numbers of these relevant captured and relevant non-captured retrieved segments. From the figure, it can be seen that the runs on the manual transcript (manual_tt and manual_c99) contain more segments with relevant content. All of the runs using TextTiling segmentation (manual_tt, ASR_tt, ASR_tt_nsw) have more retrieved segments with relevant content that are included in the pwMAP score than C99 segmentation runs. This means that in general TextTiling segmentation is more likely to have the relevant content in the centre of the retrieved segment than C99 segmentation, and that thus the boundaries formed using TextTiling are not just more effective for retrieval of relevant content, but are more likely to place the relevant content towards the centre of the segment. However, it should be noted that in all cases the proportion of segments containing some relevant content, but where it is not in the centre of the segment is very high.

Since the pwMAP metric is based on standard MAP, it gives higher scores to techniques that place relevant documents higher in the ranked list. Therefore a larger number of retrieved segments containing relevant content does not automatically imply that the run will be scored better. The pwMAP scores of the runs using TextTiling segmentation on manual and ASR transcripts have better rankings than all the other methods, including C99 segmentation of the manual transcript. The same trend exists between the C99 runs: the average number of retrieved segments considered relevant for each topic using C99 segmentation is the highest for ASR_nsw, but apparently the rank of the relevant passages is better for both manual and standard ASR transcripts, since their pwMAP values are higher, suggesting that ASR_nsw is the worst one in terms of content ranking.

Comparing the numbers of retrieved segments containing relevant information and the breakdown by content included and not included in pwMAP calculations in Figure 2, it can be seen that while TextTiling and C99 segmentation retrieve similar numbers of segments containing relevant content, that the number of included segments is much lower in the case of C99. This indicates that the balance of many of these segments is poor, i.e. that they are not centred on relevant material. Looking at this finding in the context of Figure 1, we can see that poor segmentation in this way correlates with the rankings of relevant segments even where all available segments containing relevant content are taken into account when calculating uMAP.

Figure 3 shows the average of the precision of segment content for segments averaged across the topic set, counted as relevant for the pwMAP calculation and those not included by pwMAP. It can be seen that on average the precision is much higher in all cases for segments which are included

Table 2: Average relevant and total length of segments with relevant central IPU and segments with non-centered relevant content (in seconds)

Run	Rel Length		Total Length	
	centre	non centre	centre	non centre
manual_tt	83.65	142.06	260.73	1153.38
ASR_tt	73.88	110.53	210.01	805.26
ASR_nsw_tt	71.99	117.59	212.19	1026.69
manual_c99	60.45	171.32	372.34	4710.93
ASR_c99	59.57	154.46	332.99	4203.04
ASR_nsw_c99	65.36	144.87	332.48	3496.67

in the pwMAP calculation than those which are not. This could be expected since segments for which the central IPU is not relevant are likely to have lower precision on average than those for which the central IPU is relevant. It can further be noted that all results for segmentation using TextTiling are superior to the corresponding results for C99 segmentation. Precision for the passages that have the relevant segment in the middle is always more than twice as high as that for passages that do not. Again these results indicate that these segments are associated with segments which are topically consistent, as measured against their relevance to the topics. Looking again at Figure 1, this further emphasizes the role of good segmentation in superior ranking in retrieval as measured by uMAP.

3.3 fMAP Results

The fMAP metric is designed to capture the relevancy of the segments. In this evaluation none of our segmentation methods outperformed the baseline, shown in Table 1. This result is probably caused by the low precision of the segments containing relevant content, as observed in Figure 1 (low average of precision) and in Figures 2 and 3 (where number of the segments having lower precision due to the fact that the relevant content is not located in the centre of the segment, is considerably higher than the number of the segments with centered relevant content). It is interesting to note that for this metric TextTiling segmentation not only shows better results for each of the same transcript types as C99, but even its ASR transcript outperforms C99 scores for the manual transcript.

To calculate fMAP score precision and relevance is counted in IPU units. Following the same reasoning as in Section 3.1 when calculating the average of precision from the user perspective, the actual length of the segments which must be auditioned is important, we decided to look at the precision in terms of the length in time (in seconds). Table 2 shows the average lengths of relevant content retrieved per topic in each run and the average of the total length of the passages containing the relevant content per topic. We keep the distinction between the segments with a relevant central IPU and with non centred relevant content. The average lengths of the relevant content for segments with relevant central IPU are figures of the same order for both segmentation schemes, with TextTiling segmentation runs being slightly higher. In the case of non-centred relevant IPU, C99 segmentation runs have longer relevant content than TextTiling ones. The total average lengths of the relevant content retrieved in the list is higher for all C99 runs. Unfortunately due to less accurate segmentation, retrieving more relevant content is correlated with having much longer segments: the total lengths of C99 segmentation runs are considerably higher than the TextTiling ones and therefore a metric focused on precision gets lower scores for the C99 segmentation runs. Also they contain more non-relevant content and are thus likely to be ranked more unreliably as observed in the uMAP results in Figure 1.

4. CONCLUSION AND FUTURE WORK

This paper has reported and analysed results for our participation in the NTCIR-9 SpokenDoc passage retrieval subtask track. Our experiments show that for the task of retrieving passages from the Japanese lecture archive, TextTiling segmentation is a more suitable algorithm than C99 for preprocessing the data collection in order to obtain retrieval units better coinciding with actual relevant content.

The removal of the stop words from the transcript before segmentation did not have any positive effect on the results. The reason for this finding is not clear.

For our future work, we plan to explore the application of other segmentation methods to the provided transcripts and the combination of multiple segmentation methods. In this study the influence of the ASR errors was not investigated. We think that this is an important area of further investigation since it may help explain the behaviour of both segmentation and retrieval systems.

5. ACKNOWLEDGMENTS

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008 Grant No: 08/RFP/CMS1677.

6. REFERENCES

[1] Garofolo, J. S., Auzanne, C. G. P., and Voorhees, E. M.:The TREC spoken document retrieval track: A success story. In Proceedings of RIAO 2000. Paris, France. pp1-20. (2000)

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

[3] Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y., and Itoi, K.: Test Collections for Spoken Document Retrieval from Lecture Audio Data. Proceedings of LREC 2008, Marrakech, Morocco. (2008)

[4] Choi F.Y.Y.: Advances in domain independent linear text segmentation. Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (ACL 2000), Seattle, Washington, USA. pp26-33. (2000)

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

[6] Hiemstra, D.: *Using Language Models for Information Retrieval.* Ph.D. thesis, Center of Telematics and Information Technology, AE Enschede, The Netherlands (2000)

[7] Maekawa K., Koiso H., Furui S., Isahara H.: Spontaneous speech corpus of Japanese. In Proceedings of LREC 2000, Athens, Greece. pp. 947-952 (2000)