DCU at MediaEval 2011: Rich Speech Retrieval (RSR)

Maria Eskevich
CDVP, School of Computing
Dublin City University
Dublin 9, Ireland
meskevich@computing.dcu.ie

Gareth J. F. Jones
CDVP & CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

We describe our runs and results for the Rich Speech Retrieval (RSR) Task at MediaEval 2011. Our runs examine the use of alternative segmentation methods on the provided ASR transcripts to locate the beginning of the topic, assuming that this will capture or get close to the starting point of the relevant segment; combination of various types of queries and weighting of metadata to move the relevant segment higher in the ranked list; and different ASR transcripts to compare the influence of the ASR transcripts quality. Our results show that newer versions of the transcripts and use of metadata produce better results on average. So far we have not used information about the illocutionary act type corresponding to each query, but analysis of the retrieval results shows difference in behaviour for queries associated with certatin classes of act.

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, information retrieval, automatic speech recognition

1. INTRODUCTION

The Rich Speech Retrieval (RSR) Task at MediaEval 2011 seeks to open discussion of a new task in the search of spoken content. The information to be found has special features - a certain speaker's intention (illocutionary act[1]). This new way of setting the problem of speech search raises the question of uniformity of the structures of naturally produced queries for different speech acts and how belonging to certain type of acts affects retrieval behaviour. This dataset contains 5 basic speech acts: 'apology', 'definition', 'opinion', 'promise' and 'warning'. Two of these ('definition' and 'opinion') are more neutral and appear more as simple textual requests for information, while the otherw are more emotional and subjective, and therefore less similar to the usual textual query style. A full description of the task can be found in [3]. The official metric of the RSR task was used to evaluate our results - mGAP which reflects how close the predicted jump in point of the run result is to the manual ground truth within a certain window. The following sections summarise our methods and results.

[1]http://en.wikipedia.org/wiki/Speech_acts

MediaEval 2010 Workshop, September 1-2, 2011, Pisa, Italy

2. APPROACH DESCRIPTION

The videos in the data setare diverse in their structure, style of language and length. Both ASR transcripts and confusion networks are provided for all videos. This information can be used as input for the retrieval process. We treated both the 2010 transcripts and the 2011 confusion networks in the same way: creating clean text out of the words and punctuation from the transcripts. The next step was to preprocess the data for retrieval. We first automatically segmented the data into topically coherent segments. For this we examined the use of two existing text segmentation algorithms: C99 [1] and TextTiling [2].

Most videos in the collection are accompanied by metadata relating to the whole video regardless of its length or the number of topics discussed. This metadata tag information was added once ('m1') or 5 times ('m5', to give it more weight) to all of the segments in the file. Segment indexing and retrieval were carried out using the lemur[2] Indri toolkit.

As queries we used only the naturally formulated full query ('title') and the short query similar to the query for an internet search engine ('google') and the combination of both ('title + google'). For these experiments, the starting time of the segment was selected as the jump-in point the results.

3. RESULTS

Table 1 shows the results of our runs. As could have been anticipated, larger window size shows better scores, since more of the results have non zero GAP; more complicated queries ('title + google') make the request for information more detailed and consequently relevant segments are found better; addition of metadata, and especially allocation of more weight to the metadata can overcome the problem of some keywords being misrecognized or not uttered at all in the segment and therefore improves the overall results. The confusion networks provided for 2011 dataset have a restriction that the second variant is reported only if its confidence

[2]`http://www.lemurproject.org/`

Table 1: mGap results on the test set

Transcript type	Segmentation type	Metadata used	Query type	Window size	Granularity	mGAP
2011	tt	+ (5)	title + google	60	10	0.2043
2011	c99	+ (5)	title + google	60	10	0.1622
2011	c99	+ (1)	title + google	60	10	0.1603
2011	tt	+ (5)	title + google	30	10	0.1394
2010	c99	–	title	60	10	0.1344
2011	c99	+ (5)	title + google	30	10	0.1193
2011	c99	+ (1)	title + google	30	10	0.1192
2010	c99	–	title	30	10	0.1078
2011	c99	–	google	60	10	0.1068
2011	c99	–	google	30	10	0.0686
2011	tt	+ (5)	title + google	10	10	0.0646
2011	c99	–	google	30	10	0.0686
2011	c99	+ (5)	title + google	10	10	0.0554
2011	c99	+ (1)	title + google	10	10	0.0554
2010	c99	–	title	10	10	0.0542
2011	c99	–	google	10	10	0.0061

measure is higher than 50%, in most cases this second variant is either the same word written with a capital letter or is another grammatical form of the same word. Since we were taking all the words from the confusion networks to prepare our text, these variants do not bring new terms into the document, but increase the weight of the term that has multiple entries.

4. ILLOCUTIONARY ACT BREAKDOWN

mGAP over all the queries shows average performance for a specific combination of different system parameters, but it is also interesting to look into the results of the same combinations separated into illocutionary act type. When simple queries ('title') are used on the 2010 transcript not enriched with metadata information, the results fall into two classes: 'definition' and 'opinion' have scores of the same level for window size 60, while the three other act types have significantly lower scores. In the case of the other simple query type ('google'), the difference in speech acts types is not so distinct, however with the small number of queries for certain types (only 1 for apology), it is hard to argue that the query type is the reason for the results achieved or the dataset itself.

Our runs enable us to compare the affect of using metadata with different weight (2011_c99_m5_title_and_google and 2011_c99_m1_title_and_google). In general the 'm1' run has lower scores than the 'm5', but in reality the scores are the same for all window sizes for 'apology', 'definition' and 'promise' and higher for 'm5' for 'opinion' and 'warning'.

5. CONCLUSIONS

This investigation has shown that queries that have several dimensions - not only requesting specific data in the transcript, but also certain emotion or illocution related to it, that have to be treated in a different way depending on the type of the speech act. When the illocution is less neutral more data needs to be combined in order to find the relevant segments. While the distribution of the illocutionary acts in the query set models real life, perhaps we need to create more queries of specific less popular types in order to develop better ways of processing the different query types.

Preliminary experiments suggested C99 to be the better algorithm for segmenting the data, hence more runs were submitted with C99. However, the results from the full runs show that TextTiling can outperform C99, more runs with different combinations of transcripts and queries will be carried out in further work.

6. FUTURE WORK

In future work we plan to compare all the possible combinations of query types, use of metadata and transcript segmentation to be able to demonstrate our results more solidly. Segmentation algorithms that have been developed for other types of spoken content (i.e. meetings, broadcast news) can be applied to the data in order to examine alternative ways of splitting the transcripts into search units. Since so far we were calling the beginning of the segment the jump-in point, another potential research direction may be to postprocess the retrieved segment locate the assigned jump-in point closer to the manually assigned position.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] 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, 2000.

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

[3] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmeideke, and G. J. F. Jones. Overview of mediaeval 2011 rich speech retrieval task and genre tagging task. In *Proceedings of the MediaEval Workshop 2011*, 2011.