EXTRACTING SEMANTICALLY-COHERENT KEYPHRASES FROM SPEECH

Diana Inkpen[1], and Alain Désilets[2]
[1]School of Information Technology and Eng., University of Ottawa, 800 King Edward St., Ottawa, ON, Canada, K1N6H5
diana@site.uottawa.ca
[2]Institute for Information Technology, National Research Council, 1200 Montreal Rd., Ottawa, ON, Canada, K1AOR6
Alain.Desilets@nrc-cnrc.gc.ca

1. INTRODUCTION

Browsing through large volumes of spoken audio is known to be a challenging task for end users. One way to facilitate this task is to provide keyphrases extracted from the audio, thus allowing users to quickly get the gist of the audio document or sections of it.

Previous methods for extracting keyphrases from spoken audio have used text-based summarisation techniques on automatic speech transcription. The method of Désilets et.al. (2000) was found to produce accurate keyphrases for transcriptions with Word Error Rates (WER) of the order of 25%, but performance was less than ideal for transcripts with WERs of the order of 60%. With such transcripts, a large proportion of the extracted keyphrases included serious transcription errors.

In this paper, we extend those previous methods by taking advantage of the fact that the mistranscribed keyphrases tend to have a low semantic coherence with the correctly transcribed ones. For each pair of extracted keywords, we determine their semantic coherence by computing a Pointwise Mutual Information (PMI) score based on a very large web corpus. We then use those semantic coherence scores to identify semantic outliers and filter them from the set of extracted keyphrases. The effect of the method on the accuracy of the extracted keyphrases is evaluated. We also use the same approach to filter semantic outliers in the speech on transcripts, before extracting keyphrases from it.

1.1 Data

We used a subset of the ABC and PRI stories of the TDT2 English Audio data that had correct transcripts generated by humans. We conducted experiments with two types of automatically-generated speech transcripts. The first ones were generated by the NIST/BBN time-adaptive speech recogniser and have a moderate WER (27.6%), which is representative of what can be obtained with a state of the art SR system tuned for the Broadcast News domain. See an example of a transcribed paragraph in Fig.1. The second set of transcripts was obtained using the Dragon NaturallySpeaking speaker dependant recogniser. Their WER (62.3%) was much higher because the voice model was not trained for speaker independent broadcast quality audio, in order to approximate the type of high WER seen in more casual less-than-broadcast quality audio.

1.2 Extracting keyphrases

Our approach to extracting keyphrases from spoken audio is based on the Extractor system developed for text by Turney (2000). Extractor uses a supervised learning approach to maximise overlap between machine extracted and human extracted keyphrases and it was estimated to be approximately 80% accurate. A keyphrase consist of one, two, or three keywords. See Fig.1 for some examples of keyphrases, extracted from the manual transcripts and from the BBN transcripts.

2. METHOD

Our algorithm detects the semantic outliers to be filtered out from keyphrases. It declares as outliers all the keywords with low similarity to the other keywords.

For a set of keyphrases containing the keywords (w_1, w_2, ... w_n), the algorithm has the following steps:
1. Compute semantic similarity scores $S(w_i, w_j)$ between all the pairs w_i, w_j, for all $1 <= i, j <= n$, $i \neq j$, using PMI.
2. For each keyword w_i, compute its semantic coherence score $SC(w_i)$ by summing up all $S(w_i, w_j)$, $1 <= j <= n$, $i \neq j$.
3. Compute the average score of all keywords.
4. Declare as outliers the keywords with score $SC(w_i) < K\%$ of the average score. The value of K in the threshold is chosen empirically, as shown in Section 3.

The **semantic similarity score between two words** w_1 and w_2 is their pointwise mutual information score, defined as the probability of seeing the two words together over the probability of each word separately. $PMI(w_1, w_2) = \log P(w_1, w_2) / (P(w_1) \cdot P(w_2)) = \log C(w_1, w_2) \cdot N / (C(w_1) \cdot C(w_2))$, where $C(w_1, w_2)$, $C(w_1)$, $C(w_2)$ are frequency counts, and N is the total number of words in the corpus. The scores were computed using the Waterloo Multitext system with a very large corpus of Web data (Clarke and Terra 2003).

A variant of this algorithm detects each keyword with low similarity to its closest semantic neighbour, by using the maximum score in Step2, instead of the sum. The threshold is chosen differently, as a function of the minimum $SC(w_i)$.

Manual transcript: Time now for our geography quiz today. We're traveling down the Volga river to a city that, like many Russian cities, has had several names. But this one stands out as the scene of an epic battle in world war two in which the Nazis were annihilated.
Keyphrases:
- Russian cities --> (22.752942)
- city --> (22.752942)
- Volga river --> (22.752942)
- Nazis --> (11.376471)
- war --> (11.376471)
- epic battle --> (11.376471)
- scene --> (11.376471)

NIST/BBN transcript: time now for a geography was they were traveling down river to a city that like many russian cities has had several names but this one stanza is the scene of ethnic and national and world war two in which the nazis were nine elated
Keyphrases:
- russian cities --> (22.752942)
- city --> (22.752942)
- river --> (22.752942)
- elated --> (11.376471)
- nazis --> (11.376471)
- war --> (11.376471)
- scene --> (11.376471)
- stanza --> (11.376471)

Detected outlier keywords: stanza, elated
Lost keywords: --none--

Fig.1. Fragment of a manual transcript and the extracted keyphrases; the BBN transcript, the extracted keyphrases, and the detected outliers.

3. RESULTS

Table 1 shows the results of our outlier detection algorithm on the keyphrases extracted from the BBN transcripts and from the Dragon Naturally Speaking transcripts. The second column shows the WER in the speech transcripts, measured with the standard NIST tool as a function of the number of insertions, deletions, and substitutions. The third column shows the word error rate in the keyphrases extracted by Extractor, and the last column shows the error rate after the outliers were eliminated. The word error rate in the keyphrases (kWER) is measured as the number of words that are in the keyphrases but not in the manual speech transcript. Table 1 shows that the number of wrong keywords caused by recognition errors reduces by almost half when the outliers are eliminated (for K=80%).

The variation of the error rate in keyphrases with K is shown in Fig.2. The higher the threshold, the more outliers are eliminated, but some good keywords can also be lost. Fig 2 shows the percent of lost keywords, computed as the percent of the keywords that are were wrongly declared as outliers (they are considered good keywords because they are in the keyphrases extracted from the manual transcripts). The variant of the algorithm that uses maximum scores in Step 2 produces better reduction in kWER, but higher loss of good keywords. Its results are not shown here because of space limitations.

Table 1. Word error rate in the transcripts, in the initial keyphrases, and in the filtered keyphrases (plus % lost keywords, for K=80%).

Tran-scripts	WER transcripts	kWER initially	Filtered keyphrases % Lost k.	kWER
BBN	27.6%	10.6%	14.5%	**5.4%**
Dragon	62.3%	43.3%	4.3%	**27.6%**

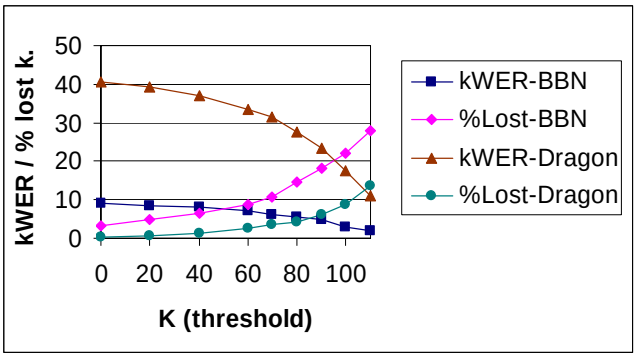

Fig.2. The variation of the kWER and percent of lost keywords for the two sets of data, in function of the threshold K% of the average similarity of a set of keyphrases.

4. CONCLUSION

We presented a method for filtering semantic outliers from keyphrases that summarize speech. Future work includes experimenting with other methods for computing semantic outliers (Jarmasz and Barrière 2004), and building small domain models from reliable keywords in order to detect the outliers relative to them (Turney 2003). We also plan to run the outlier detection algorithm directly on the speech transcripts. In this case the input to the algorithm is all the content words in the transcript.

REFERENCES

Clarke, C. and Terra, E. (2003). Passage retrieval vs. document retrieval for factoid question answering. ACM SIGIR'03, 327-328.
Désilets, A. and de Brujin, B. and Martin. J. (2001). Extracting keyphrases from spoken audio documents. SIGIR Workshop on Information Retrieval Techniques for Speech Applications, 36-50.
Jarmasz, M. and Barrière, C. (2004). Keyphrase extraction: enhancing lists. Proceedings of CLINE'04.
Turney, P. D. (2003). Coherent keyphrase extraction via Web mining. Proceedings of IJCAI'03, 434-439.
Turney, P.D. (2000). Learning algorithms for keyphrase extraction, *Information Retrieval*, 2 (4), 303-336.

ACKNOWLEDGEMENTS

We wish to thank Peter Turney and Gerald Penn for their useful feedback and discussions. We thank Egidio Terra, and Charlie Clarke for allowing us to use the Multitext System, the NRC copy.