TOPIC AND STYLISTIC ADAPTATION FOR SPEECH SUMMARISATION

Pierre Chatain, Edward W.D. Whittaker, Joanna A. Mrozinski and Sadaoki Furui.

Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan

ABSTRACT

Contemporary approaches to automatic speech summarisation comprise several components, among them a linguistic model (LiM) component, which is unrelated to the language model used during the recognition process. This LiM component assigns a probability to word sequences from the source text according to their likelihood of appearing in the summarised text. In this paper we investigate LiM topic and stylistic adaptation using combinations of LiMs each trained on different adaptation data. Experiments are performed on 9 talks from the TED corpus of Eurospeech conference presentations, as well as 5 news stories from CNN broadcast news data, for all of which human (TRS) and speech recogniser (ASR) transcriptions along with human summaries were used. In all ASR cases, summarisation accuracy (SumACCY) of automatically generated summaries was significantly improved by automatic LiM adaptation, with relative improvements of at least 2.5% in all experiments.

1. INTRODUCTION

One of the major applications of automatic speech recognition is to transcribe spontaneous speech such as found in conversations, lectures and presentations. Although speech is the most natural and effective method of communication between human beings, it is not easy to review speech documents quickly if they are simply recorded as an audio signal. Transcribing and condensing speech from presentations and lectures by removing irrelevant or inaccurately recognised words and phrases and extracting the important parts is therefore an important issue. Automatic speech summarisation is an approach towards accomplishing this goal.

Techniques for automatically summarising written text have been actively investigated in the field of natural language processing [1], and more recently new techniques have been developed for speech summarisation [2]. However it is still very hard to obtain good quality summaries. Moreover, recognition accuracy is still around 30% on spontaneous speech tasks, in contrast to speech read from text such as broadcast news. Spontaneous speech is characterised by disfluencies, repetitions, repairs, and fillers, all of which make recognition and consequently speech summarisation more difficult [3]. In this paper we extend the work done on the two-stage summarisation method described in [2] by focusing on adapting the linguistic component, which is not related at all to the language model used during the recognition process, to make it more suited for the summarisation task. If appropriate LiMs can be used, sentences related to the topic, as well as grammatically correct sentences are more likely to be extracted from the speech. In particular we examine methods for adapting the LiMs automatically to improve performance.

Experiments were performed both on spontaneous speech, using 9 talks taken from the Translanguage English Database (TED) corpus [4], and speech read from text, using 5 talks from CNN broadcast news from 1998.

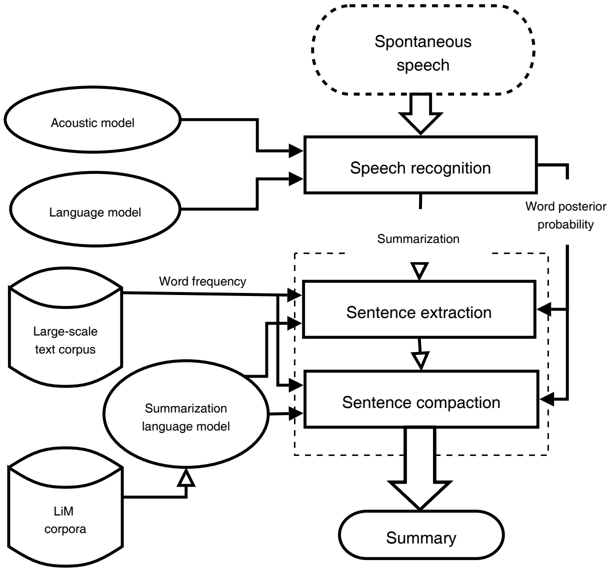

Fig. 1. Automatic speech summarisation system.

2. SUMMARISATION METHOD

The summarisation system used in this paper is basically the same as the one described in [2]. It involves a two step summarisation process, consisting of sentence extraction and sentence compaction, as shown in Figure 1.

Important sentences are first extracted according to the following score for each sentence $W = w_1, w_2, ..., w_n$, obtained from the automatic speech recognition output (ASR):

$$S(W) = \frac{1}{N}\sum_{i=1}^{n}\{\alpha_C C(w_i) + \alpha_I I(w_i) + \alpha_L L(w_i)\}, \quad (1)$$

where N is the number of words in the sentence W, and $C(w_i)$, $I(w_i)$ and $L(w_i)$ are the confidence score, the significance score and the linguistic score of word w_i, respectively. α_C, α_I and α_L are the respective weighting factors of those scores, determined experimentally.

The compaction is then done by selecting important words among the extracted sentences using a similar score (composed of the same components) computed for each word in the talk to be summarised.

For each word from the ASR, a logarithmic value of its posterior probability, the ratio of a word hypothesis probability to that of all other hypotheses, is calculated using a word graph obtained from the speech recogniser and used as a confidence score.

For the significance score, the frequencies of occurrence of 115k words were found using the WSJ and the Brown corpora. Important

1-4244-0469-X/06/$20.00 ©2006 IEEE

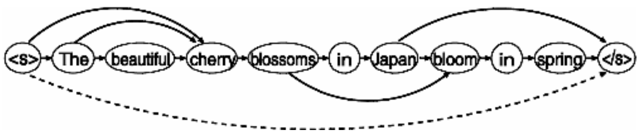

Fig. 2. Word network made by merging manual summarisation results.

keywords get a higher weight and common words unrelated to the gist of the talk are effectively de-weighted by this score.

In the experiments in this paper we modified the linguistic component, which indicates the linguistic likelihood of word strings in the sentence, so as to be able to use combinations of different linguistic models. Starting with a baseline LiM (LiM$_B$) we perform LiM adaptation by linearly interpolating the baseline model with other component models trained on different data. The probability of a given n-gram sequence then becomes:

$$P(w_i|w_{i-n+1}..w_{i-1}) = \lambda_1 P_1(w_i|w_{i-n+1}..w_{i-1}) + \\ ... + \lambda_n P_n(w_i|w_{i-n+1}..w_{i-1}), \qquad (2)$$

where $\sum_k \lambda_k = 1$ and λ_k and P_k are the weight and the probability assigned by model k. Different types of component LiM are built, coming from different sources of data, and using either unigram, bigram or trigram information. The LiM$_B$ and component LiMs are then combined for adaptation using linear interpolation as in Equation (2). The linguistic score is then computed using this modified probability as in Equation (3):

$$L(w_i) = \log P(w_i|w_{i-n+1}..w_{i-1}). \qquad (3)$$

3. EVALUATION CRITERIA

The measure of summary quality used in this paper is summarisation accuracy (SumACCY). To automatically evaluate the summarised speeches, correctly transcribed talks were manually summarised, and used as the correct targets for evaluation. Variations of manual summarisation results are merged into a word network as shown in Figure 2, which is considered to approximately express all possible correct summarisations covering subjective variations. The word accuracy of automatic summarisation is calculated as the summarisation accuracy using the word network [5]:

$$Accuracy = (Len - Sub - Ins - Del)/Len * 100[\%], \qquad (4)$$

where Sub is the number of substitution errors, Ins is the number of insertion errors, Del is the number of deletion errors, and Len is the number of words in the most similar word string in the network.

4. EXPERIMENTAL SETUP

Due to lack of data we had to use the talks both for development and evaluation with a rotating form of cross-validation [6]: all talks but one are used for development, the remaining talk being used for testing. This process is repeated for all combinations of development and evaluation sets. During the development phase, summaries from the development talks are generated automatically by the system using different sets of parameters. These summaries are evaluated using SumACCY and the set of parameters which maximises the weighted average for the LiM$_B$ is chosen for evaluation on the remaining talk. The LiM$_B$ is a trigram model built on a corpus consisting of around ten years of conference proceedings (17.8M words) on the subject of speech and signal processing. The purpose of the development phase is to choose the most effective combination of weights α_C, α_I and α_L for the confidence (in the case of speech recognition), significance and linguistic scores, as well as the optimal sentence extraction/compaction ratio. The summary generated for each talk using its set of optimised parameters is then evaluated using SumACCY, which gives us our baseline for this talk.

Using the same parameters as those that were selected for the baseline, we generate summaries for the lectures in the development set for different LiM interpolation weights λ_k. Values between 0 and 1 in steps of 0.1, were investigated for the latter, and an optimal set of λ_k is selected. Using these interpolation weights, as well as the set of parameters determined for the baseline, we generate a summary of the test talk, which is also evaluated using SumACCY, giving us our final adapted result for this talk. Averaging those results over the test set (i.e. all talks) gives us our final adapted result.

Lower bound results are given by random summarisation (Rnd sum) i.e. randomly extracting sentences and words, without use of the scores present in Equation (1) for appropriate summarisation ratios for both TRS and ASR. Upper bound results are determined by evaluating the human made summaries (Human sum) against word-graphs (as described in Section 3) built using the remaining human-made summaries, again in a rotating validation process, with one summary held out each time.

4.1. The TED data

Nine talks from the TED corpus were used in this paper. Speech recognition transcriptions were obtained for each talk. The latter were produced using the Janus Recognition Toolkit (JRTk) with an acoustic model trained on 300 hours of Broadcast News (BN) data merged with the close talking channel of meeting corpora [7]. The acoustic model used 42 features and consisted of 300k gaussians with diagonal covariances organised in 24k distributions over 6k codebooks [8]. The language model (LM) used for the speech recogniser was generated by interpolating a word 3-gram and a class-based 5-gram LM each trained on BN data (160M words) and the proceedings corpus described above, and a 3-gram LM based on talks (60k words) by the TED adaptation speakers. The overall OOV rate is 0.3% with a vocabulary size of 25000 words including multi-words and pronunciation variants. The average word error rate of the TED talks is 33.3%. Nine talks, each transcribed and manually summarised by nine different humans for both 10% and 30% summarization ratios were used for both development and evaluation using the rotating form of cross-validation described above.

Different types of component LiM are built, coming from two different sources of data, using either unigram, bigram or trigram information. The LiM$_B$ and component LiMs are then combined for adaptation using linear interpolation as in Equation (2).

The first type of component linguistic models (LiM$_{S1}$, LiM$_{S2}$ and LiM$_{S3}$ for unigrams, bigrams and trigrams, respectively, where S denotes Summary) are built on the small corpus of hand-made summaries described above, made for the same summarisation ratio as the one we are generating. For each talk the hand-made summaries of the other eight talks (i.e. 72 summaries) were used as the LiM training corpus. This type of LiM is expected to help generate automatic summaries in the same style as those made manually.

The second type of component linguistic model (LiM$_{T1}$, LiM$_{T2}$ and LiM$_{T3}$ for unigrams, bigrams and trigrams, respectively, where T denotes Talk) are built from the papers in the conference proceed-

10%	TRS	Rnd sum	34.4
		Human sum	59.6
		Baseline	63.1
	ASR	Rnd sum	33.9
		Baseline	48.6
30%	TRS	Rnd sum	71.2
		Human sum	77.7
		Baseline	81.6
	ASR	Rnd sum	56.1
		Baseline	66.7

Table 1. Reference results for TED.

ings for the talk we want to summarise. This type of LiM, used for topic adaptation, is investigated because key words and important sentences that appear in the associated paper are expected to have a high information value and should be selected during the summarisation process. The LiM weights are optimised during the development phase along with the other system parameters.

Experiments were first made on the human transcriptions, investigating interpolation of the baseline language model with each component separately. We also investigated a three way interpolation, combining LiM$_{B}$, LiM$_{T1}$ and LiM$_{S3}$ for both TRS and ASR.

4.2. The CNN data

The CNN data consists of five talks from broadcast news of 1998, transcribed by the JANUS speech recognition system. The acoustic model was trained on 66 hours of BN (different from the data used in Section 4.1), and consisted of 105k gaussians organized in 6000 distributions and sharing 2000 codebooks. The LM used by the speech recognizer was an interpolation of bigram and trigram models based on a BN corpus with a vocabulary of 40k words. The average word error rate for the ASR of these news stories read from written text was 22%. The five talks were transcribed and manually summarised by sixteen different humans for a 40% summarization ratio. Again, they were used for both development and evaluation using the previously described cross-validation process.

For the CNN news stories we only built one type of component linguistic model, using hand made summaries in the same manner as for the TED data: for each talk the hand-made summaries of the other four talks (i.e. 64 summaries in total) were used as the training corpus. LiM$_{C1}$, LiM$_{C2}$ and LiM$_{C3}$ (for unigram, bigram, and trigram, respectively, C standing for CNN) were interpolated separately with LiM$_{B}$ as above.

5. RESULTS

5.1. Reference Results

Reference results, consisting of the results for random summarisation, the human summaries and the baseline are given for appropriate summarisation ratios in Tables 1 and 2 for the TED and CNN data, respectively.

5.2. TED Results

Initial experiments focused on summarising the speech transcribed by humans rather than the ASR. In previous studies, the confidence score proved to be very important [2] and often outweighed the other factors so it would have been more difficult to see the effect of LiM adaptation. Using human transcribed speech is equivalent to having 100% recognition accuracy, although the characteristics associated with spontaneous speech are largely preserved.

40%	TRS	Rnd sum	80.7
		Human sum	88.2
		Baseline	81.1
	ASR	Rnd sum	68.2
		Baseline	71.3

Table 2. Reference results for CNN.

		N-gram		
Ratio	LiM	1	2	3
10%	LiM$_{TN}$	67.6	61.6	61.4
	LiM$_{SN}$	62.0	62.0	65.9
30%	LiM$_{TN}$	82.6	83.7	82.5
	LiM$_{SN}$	83.2	82.1	83.7

Table 3. TED adaptation results on TRS.

Table 1 shows that experiments conducted using LiM$_{B}$ achieve better results than both random summarisation and more surprisingly human summarisation. This indicates that the summarisation system performs above expectations and that our baseline linguistic model is appropriate and well trained, which makes the increase witnessed using adapted models all the more significant.

Results for the adapted LiMs are given in Table 3 for the interpolations combining the different component LiMs with the baseline one by one. Results using LiM$_{T1}$ seem to indicate that in the case of talk-based adaptation, the unigram adaptation model has the greatest effect. This is because the unigram information coming from such a small amount of data is more robust than bigram or trigram information. This unigram information helps selecting important key and/or technical words present in the lecture which are uttered during the talk. This complements the role of the significance score, which is already supposed to select important words, emphasising topic and key words.

However, with summary-based adaptation, trigram models yield the best results. This confirms our idea of realising stylistic adaptation: human-made summaries used in the model actually help us select utterances likely to be used by human subjects when making summaries. The unigram information in this case is not as important as the trigram information, which helps to select frequently used expressions and phrases likely to appear in multiple summaries.

We also investigated a three way interpolation combining LiM$_{B}$, LiM$_{S3}$ and LiM$_{T1}$, the results of which are shown in Table 4, for both TRS and ASR. For TRS, relative improvements of 7.4% and 2.1% were obtained for the 10% and 30% summarisation ratios respectively. ASR experiments are similar to the TRS ones, except that the confidence score was also optimised during development. Results are much lower than in the transcription case, as the word error rate averaged over the nine talks was 33.3%. With the ASR, relative improvements of 2.5% and 3% are observed for the 10% and 30% summarisation ratios, respectively.

10%	TRS	67.8
	ASR	49.8
30%	TRS	83.3
	ASR	68.7

Table 4. Three way interpolation for TED.

	N-gram		
	1	2	3
LiM$_{CN}$ TRS	80.9	80.8	80.8
LiM$_{CN}$ ASR	72.5	73.4	73.5

Table 5. CNN adaptation results.

5.3. CNN Results

Results for LiM adaptation on the CNN data are given in Table 5 for both TRS and ASR.

LiM adaptation did not yield any improvement on the CNN TRS. The baseline results were already high, and adaptation had almost no effect. Moreover, random summarisation performance was also very high.

With the ASR though, a 3.1% relative improvement was obtained using LiM adaptation, the greatest increase coming from the trigram adaptation model as was also observed on the TED data.

6. DISCUSSION

The fact that human summaries scored less than our automatically generated summaries in the case of the TED data can be explained by the fact that they are evaluated against a network built upon eight summaries, as opposed to nine for our automated results, thus allowing fewer possibilities of correctly evaluated summaries. This is a problem with the SumACCY evaluation metric, meaning that if too many summaries are used to build the wordgraph used for evaluation, inaccurate summaries can eventually be considered correct, which is why the random summarisation score is so high in the CNN data case, where 16 summaries are being used for evaluation.

In all ASR cases, relative improvements of at least 2.5% could be observed. These improvements are significant, since the metric used does not leave much margin for improvement, as the upper bound results given by the human made summaries indicate.

It is to be noted that even though we tried optimising the compaction ratio along with the other parameters, in almost all cases only pure sentence extraction was picked during the development phase. This underlines a problem with the development phase, which does not select the best possible set of parameters for a given talk. Tables 6 and 7 show the results we would have had with perfect development for the baseline and adaptation experiments respectively. Ideally, similar results could be obtained by using more data in the development set.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have investigated combinations of models used to compute the linguistic score in an automatic speech summarisation system. It was found that summarisation performance was improved by at least 2.5% relative increase over the baseline on the ASR output of both spontaneous speech data coming from the TED corpus and speech read from text from CNN broadcast news.

TED	Baseline 10%	70.6
	Baseline 30%	85.8
CNN	Baseline 40%	84.8

Table 6. Baseline results assuming perfect development.

			N-gram		
Data	Ratio	LiM	1	2	3
TED	10%	LiM$_{TN}$	75.2	75.5	75.5
		LiM$_{SN}$	74.4	77	75.6
	30%	LiM$_{TN}$	88.7	88.9	89.2
		LiM$_{SN}$	88.2	87.8	87.8
CNN	40%	LiM$_{CN}$	86.6	86.4	86.3

Table 7. Adaptation results assuming perfect development.

Topic adaptation was performed using unigram information from a small source of data related to the talk we were trying to summarise, and stylistic adaptation was realised using the trigram information coming from a small corpora of hand-made summaries. Both adaptations proved beneficial, and there are many other types of linguistic model and sources of data that can be investigated in the context of improving summarisation performance in future work.

8. ACKNOWLEDGEMENTS

The authors would like to thank Matthias Woelfel for the recogniser transcriptions and Chiori Hori for her previous work on two stage summarisation and for gathering data from the TED corpus. This work is supported in part by the 21st Century COE Program.

9. REFERENCES

[1] I. Mani, *Automatic summarization*, John Benjamins publishing company, Amsterdam, Netherlands, 2001.

[2] S. Furui T. Kikuchi and C. Hori, “Automatic speech summarization based on sentence extraction and compaction,” *Proc. ICASSP, Hong Kong, China*, vol. 1, pp. 236–239, 2003.

[3] K. Zechner, “Summarization of spoken language-challenges, methods, and prospects,” *Speech Technology Expert eZine, Issue.6*, 2002.

[4] A. Fourcin J. Mariani L. Lamel, F. Schiel and H. Tillmann, “The translanguage english database (ted),” *Proc. ICSLP, Yokohama, Japan*, vol. 4, pp. 1795–1798, 1994.

[5] T. Hori C. Hori and S. Furui, “Evaluation method for automatic speech summarization,” *Proc. Eurospeech, Geneva, Switzerland*, vol. 4, pp. 2825–2828, 2003.

[6] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[7] V. Maclaren S. Burger and H. Yu, “The ISL meeting corpus: the impact of meeting type on speech style,” *Proc. ICSLP, Denver, USA*, vol. 1, pp. 301–304, 2002.

[8] M. Wolfel and S. Burger, “The ISL baseline lecture transcription system for the TED corpus,” Tech. Rep., Karlsruhe University, 2005.