論文 / 著書情報
Article / Book Information

Title	Class Model Adaptation for Speech Summarisation
Author	Pierre Chatain, Edward W. D. Whittaker, Joanna A. Mrozinski, Sadaoki Furui
Journal/Book name	Human Language Technology Conference of the NAACL, Companion Volume, Vol. , No. , pp. 21-24
発行日 / Issue date	2006, 6

Class Model Adaptation for Speech Summarisation

Pierre Chatain, Edward W.D. Whittaker, Joanna Mrozinski and Sadaoki Furui
Dept. of Computer Science
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan
{pierre, edw, mrozinsk, furui}@furui.cs.titech.ac.jp

Abstract

The performance of automatic speech summarisation has been improved in previous experiments by using linguistic model adaptation. We extend such adaptation to the use of class models, whose robustness further improves summarisation performance on a wider variety of objective evaluation metrics such as ROUGE-2 and ROUGE-SU4 used in the text summarisation literature. Summaries made from automatic speech recogniser transcriptions benefit from relative improvements ranging from 6.0% to 22.2% on all investigated metrics.

1 Introduction

Techniques for automatically summarising written text have been actively investigated in the field of natural language processing, and more recently new techniques have been developed for speech summarisation (Kikuchi et al., 2003). 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 (Zechner, 2002). In a previous study (Chatain et al., 2006), linguistic model (LiM) adaptation using different types of word models has proved useful in order to improve summary quality. However sparsity of the data available for adaptation makes it difficult to obtain reliable estimates of word n-gram probabilities. In speech recognition, class models are often used in such cases to improve model robustness. In this paper we extend the work previously done on adapting the linguistic model of the speech summariser by investigating class models. We also use a wider variety of objective evaluation metrics to corroborate results.

2 Summarisation Method

The summarisation system used in this paper is essentially the same as the one described in (Kikuchi et al., 2003), which involves a two step summarisation process, consisting of sentence extraction and sentence compaction. Practically, only the sentence extraction part was used in this paper, as preliminary experiments showed that compaction had little impact on results for the data used in this study.

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:

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

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.

For each word from the automatic speech recogni-

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 21–24,
New York City, June 2006.

tion transcription, 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.

In the experiments in this paper we modified the linguistic component to use combinations of different linguistic models. The linguistic component gives the linguistic likelihood of word strings in the sentence. 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:

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

where $\sum_k \lambda_k = 1$ and λ_k and P_k are the weight and the probability assigned by model k.

In the case of a two-sided class-based model,

$$P_k(w_i|w_{i-n+1}..w_{i-1}) = P_k(w_i|C(w_i)) \cdot P_k(C(w_i)|C(w_{i-n+1})..C(w_{i-1})), \quad (3)$$

where $P_k(w_i|C(w_i))$ is the probability of the word w_i belonging to a given class C, and $P_k(C(w_i)|C(w_{i-n+1})..C(w_{i-1}))$ the probability of a certain word class $C(w_i)$ to appear after a history of word classes, $C(w_{i-n+1}), ..., C(w_{i-1})$.

Different types of component LiM are built, coming from different sources of data, either as word or class models. 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 (4):

$$L(w_i) = \log P(w_i|w_{i-n+1}..w_{i-1}). \quad (4)$$

3 Evaluation Criteria

3.1 Summarisation Accuracy

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, 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 (SumACCY) using the word network (Hori et al., 2003):

$$Accuracy = (Len - Sub - Ins - Del)/Len * 100[\%], \quad (5)$$

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.

3.2 ROUGE

Version 1.5.5 of the ROUGE scoring algorithm (Lin, 2004) is also used for evaluating results. ROUGE F-measure scores are given for ROUGE-2 (bigram), ROUGE-3 (trigram), and ROUGE-SU4 (skip-bigram), using the model average (average score across all references) metric.

4 Experimental Setup

Experiments were performed on spontaneous speech, using 9 talks taken from the Translanguage English Database (TED) corpus (Lamel et al., 1994; Wolfel and Burger, 2005), each transcribed and manually summarised by nine different humans for both 10% and 30% summarization ratios. Speech recognition transcriptions (ASR) were obtained for each talk, with an average word error rate of 33.3%.

A corpus consisting of around ten years of conference proceedings (17.8M words) on the subject of speech and signal processing is used to generate the LiM_B and word classes using the clustering algorithm in (Ney et al., 1994).

Different types of component LiM are built and combined for adaptation as described in Section 2.

The first type of component linguistic models 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.

		Baseline				Adapted			
		SumACCY	R-2	R-3	R-SU4	SumACCY	R-2	R-3	R-SU4
10%	Random	34.4	0.104	0.055	0.142	-	-	-	-
	Word	63.1	0.186	0.130	0.227	67.8	0.193	0.140	0.228
	Class	65.1	0.195	0.131	0.226	72.6	0.210	0.143	0.234
	Mixed	63.6	0.186	0.128	0.218	71.8	0.211	0.139	0.231
30%	Random	71.2	0.294	0.198	0.331	-	-	-	-
	Word	81.6	0.365	0.271	0.395	83.3	0.365	0.270	0.392
	Class	83.1	0.374	0.279	0.407	92.9	0.415	0.325	0.442
	Mixed	83.1	0.374	0.279	0.407	92.9	0.415	0.325	0.442

Table 1: TRS baseline and adapted results.

The second type of component linguistic models are built from the papers in the conference proceedings 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.

Three sets of experiments were made: in the first experiment (referred to as Word), LiM_B and both component models are word models, as introduced in (Chatain et al., 2006). For the second one (Class), both LiM_B and the component models are class models built using exactly the same data as the word models. For the third experiment (Mixed), the LiM_B is an interpolation of class and word models, while the component LiMs are class models.

To optimise use of the available data, a rotating form of cross-validation (Duda and Hart, 1973) is used: all talks but one are used for development, the remaining talk being used for testing. Summaries from the development talks are generated automatically by the system using different sets of parameters and the LiM_B. These summaries are evaluated and the set of parameters which maximises the development score for the LiM_B is selected for the remaining talk. The purpose of the development phase is to choose the most effective combination of weights α_C, α_I and α_L. The summary generated for each talk using its set of optimised parameters is then evaluated using the same metric, 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 evaluated using the same evaluation metric, 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.

This process is repeated for all evaluation metrics, and all three experiments (Word, Class, and Mixed).

Lower bound results are given by random summarisation (Random) i.e. randomly extracting sentences and words, without use of the scores present in Equation (1) for appropriate summarisation ratios.

5 Results

5.1 TRS Results

Initial experiments were made on the human transcriptions (TRS), and results are given in Table 1. Experiments on word models (Word) show relative improvements in terms of SumACCY of 7.5% and 2.1% for the 10% and 30% summarisation ratios, respectively. ROUGE metrics, however, do not show any significant improvement.

Using class models (Class and Mixed), for all ROUGE metrics, relative improvements range from 3.5% to 13.4% for the 10% summarisation ratio, and from 8.6% to 16.5% on the 30% summarisation ratio. For SumACCY, relative improvements between 11.5% to 12.9% are observed.

5.2 ASR Results

ASR results for each experiment are given in Table 2 for appropriate summarisation ratios. As for

		Baseline				Adapted			
		SumACCY	R-2	R-3	R-SU4	SumACCY	R-2	R-3	R-SU4
10%	Random	33.9	0.095	0.042	0.140	-	-	-	-
	Word	48.6	0.143	0.064	0.182	49.8	0.129	0.060	0.173
	Class	50.0	0.133	0.063	0.170	55.1	0.156	0.077	0.193
	Mixed	48.5	0.134	0.068	0.176	56.2	0.142	0.077	0.191
30%	Random	56.1	0.230	0.124	0.283	-	-	-	-
	Word	66.7	0.265	0.157	0.314	68.7	0.271	0.161	0.328
	Class	66.1	0.277	0.165	0.324	71.1	0.300	0.180	0.348
	Mixed	64.9	0.268	0.160	0.312	70.5	0.304	0.192	0.351

Table 2: ASR baseline and adapted results.

the TRS, LiM adaptation showed improvements in terms of SumACCY, but ROUGE metrics do not corroborate those results for the 10% summarisation ratio. Using class models, for all ROUGE metrics, relative improvements range from 6.0% to 22.2% and from 7.4% to 20.0% for the 10% and 30% summarisation ratios, respectively. SumACCY relative improvements range from 7.6% to 15.9%.

6 Discussion

Compared to previous experiments using only word models, improvements obtained using class models are larger and more significant for both ROUGE and SumACCY metrics. This can be explained by the fact that the data we are performing adaptation on is very sparse, and that the nine talks used in these experiments are quite different from each other, especially since the speakers also vary in style. Class models are more robust to this spontaneous speech aspect than word models, since they generalise better to unseen word sequences.

There is little difference between the Class and Mixed results, since the development phase assigned most weight to the class model component in the Mixed experiment, making the results quite similar to those of the Class experiment.

7 Conclusion

In this paper we have investigated linguistic model adaptation using different sources of data for an automatic speech summarisation system. Class models have proved to be much more robust than word models for this process, and relative improvements ranging from 6.0% to 22.2% were obtained on a variety of evaluation metrics on summaries generated from automatic speech recogniser transcriptions.

Acknowledgements: The authors would like to thank M. Wölfel for the recogniser transcriptions and C. Hori for her work on two stage summarisation and gathering the TED corpus data. This work is supported by the 21st Century COE Programme.

References

P. Chatain, E.W.D. Whittaker, J. Mrozinski, and S. Furui. 2006. Topic and Stylistic Adaptation for Speech Summarization. *Proc. ICASSP, Toulouse, France.*

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

C. Hori, T. Hori, and S. Furui. 2003. Evaluation Method for Automatic Speech Summarization. *Proc. Eurospeech, Geneva, Switzerland*, 4:2825–2828.

T. Kikuchi, S. Furui, and C. Hori. 2003. Automatic Speech Summarization based on Sentence Extraction and Compaction. *Proc. ICASSP, Hong Kong, China*, 1:236–239.

L. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillmann. 1994. The Translanguage English Database (TED). *Proc. ICSLP, Yokohama, Japan*, 4:1795–1798.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. *Proc. WAS, Barcelona, Spain.*

H. Ney, U. Essen, and R. Kneser. 1994. On Structuring Probabilistic Dependences in Stochastic Language Modelling. *Computer Speech and Language*, (8):1–38.

M. Wolfel and S. Burger. 2005. The ISL Baseline Lecture Transcription System for the TED Corpus. Technical report, Karlsruhe University.

K. Zechner. 2002. Summarization of Spoken Language-Challenges, Methods, and Prospects. *Speech Technology Expert eZine, Issue.6.*