Perplexity Based Linguistic Model Adaptation for Speech Summarisation

Pierre Chatain, Edward Whittaker, Joanna Mrozinski, Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology, Tokyo, 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. One of the problems encountered was the high computational cost and low efficiency of the development phase. In this paper we compare our original development approach of evaluating summaries produced by an exhaustive search over all parameters with a much faster development method using an expectation maximization algorithm that minimizes perplexity in order to find the optimal combination of linguistic models for the speech summarisation task. Perplexity proves to be sufficiently correlated to the objective evaluation metrics used in the summarisation literature that it can be used in this fashion. For a much reduced computational cost (approximately 500 times faster), final relative improvements are very similar to those previously obtained, ranging from 1.5% to 21.3% on all investigated metrics for summaries made from automatic speech recogniser transcriptions.

Index Terms: Speech summarization, language modeling, perplexity, class models, adaptation, ROUGE, SumACCY.

1. Introduction

Text summarisation continues to receive increasing attention from the language processing community [1], and more recently this interest has been extended to speech summarisation [2]. However it is still very difficult to obtain good quality summaries, especially in the case of spontaneous speech, which is characterised by disfluencies, repetitions, repairs, and fillers. All of this makes speech recognition and consequently speech summarisation even more difficult than summarisation of speech read from text [3].

In a previous study [4], linguistic model (LiM) adaptation using different types of word models was shown to be useful in order to improve summary quality. This work has since been extended by investigating class models [5], which further improved performance. However the development process that was used previously was an exhaustive-search approach, that required substantial computation time and restricted the range of parameters that could be investigated. Moreover, this process was far from effective, as analysis of the results showed that the set of parameters determined by the development set for the test set was far from being optimal. In this paper we investigate another way of determining optimal LiM interpolation weights to perform LiM adaptation without having to build large numbers of summaries for different parameter combinations during the development phase.

A common metric used to evaluate the quality of language models used in speech recognition is perplexity, and in this study we use an expectation maximization (EM) algorithm to determine LiM interpolation weights that minimize the perplexity on the text of summaries in the development set. If the perplexity of LiMs used to generate automatic summaries is highly correlated with evaluation metrics used in the summarisation literature, this would provide a faster and smoother development, permitting more fine-grained parameter optimisation and making it easier to combine a greater number of different LiMs.

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. In practice, only the sentence extraction step 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 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.

For each word from the ASR 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. Important keywords receive 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 to use combinations of different linguistic models.The linguistic component gives the linguistic likelihood of word strings in the sentence. Starting with a baseline trigram 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.

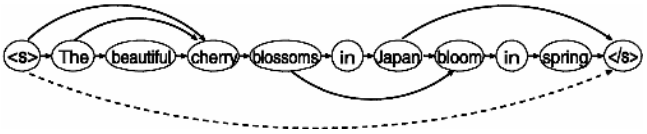

Figure 1: Word network made by merging manual summarisation results.

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

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

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 either as word or class models, coming from different sources of data, and either as unigram, bi-gram or trigram 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

Two objective measures of summary quality are used in this paper, summarisation accuracy (SumACCY) and ROUGE.

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 as shown in Figure 1, 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 [6]:

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

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

The version 1.5.5 of the ROUGE scoring algorithm [7] is also used to corroborate results. ROUGE F-measure scores are given for ROUGE-2 (bigram) and ROUGE-SU4 (skip-bigram), using the model average (average score across all target summaries) metric.

4. Experimental Setup

4.1. The TED Data

Experiments were performed on spontaneous speech, using 9 talks taken from the Translanguage English Database (TED) corpus [8, 9], each transcribed and manually summarised by nine different humans for both 10% and 30% summarization ratios. ASR transcriptions were obtained for each talk, with an average word error rate of 33.3%. 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 [10]. The acoustic model used 42 features and consisted of 300k gaussians with diagonal covariances organised in 24k distributions over 6k codebooks. 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.

4.2. The Linguistic Models

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 a thousand word classes using the clustering algorithm in [11]. In these experiments the class models were built using a fixed number of a thousand classes so as not to add an extra variable to the problem, but ideally this number should also be optimised.

Different types of component LiM are built either as word or class models, coming from two different sources of data. 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 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 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. 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$_B$ and the component models are interpolations of class and word models built on the same data as above.

4.3. Parameter Selection

To optimise use of the available data, a rotating form of cross-validation [12] 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 maximizes the development score for the LiM$_B$ is selected for the remaining talk. The purpose of this 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 base-

		Baseline			Exhaustive Adaptation			PP Adaptation		
		SumACCY	R-2	R-SU4	SumACCY	R-2	R-SU4	SumACCY	R-2	R-SU4
10%	Random	34.4	0.104	0.142	-	-	-	-	-	-
	Word	63.1	0.186	0.227	67.8	0.193	0.228	67.4	0.196	0.232
	Class	65.1	0.195	0.226	72.6	0.210	0.234	72.9	0.217	0.242
	Mixed	63.6	0.186	0.218	71.8	0.211	0.231	70.3	0.214	0.240
30%	Random	71.2	0.294	0.331	-	-	-	-	-	-
	Word	81.6	0.365	0.395	83.3	0.365	0.392	82.5	0.369	0.399
	Class	83.1	0.374	0.407	92.9	0.415	0.442	93.5	0.422	0.449
	Mixed	83.1	0.374	0.407	92.9	0.415	0.442	91.3	0.409	0.448

Table 1: TRS baseline and adapted results.

line for this talk.

In order to determine the optimal set of LiM interpolation weights λ_k to be used for LiM adaptation, we compare the method previously used in [4, 5] (Exhaustive Adaptation) to a new method that uses an EM algorithm to minimize perplexity (PP Adaptation) on the human summaries in the development set. The previous method involved generating 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 (for a total of 66 combinations per talk in the case where the LiM$_B$ is combined with two component models). The set of λ_k that maximized the average score (obtained using SumACCY or ROUGE) over the eight development talks was then selected.

In this study, for each talk of the development set we evaluate seperately the perplexity of each LiM (LiM$_B$ and the component LiMs) with respect to the concatenated human made summaries of that talk for the appropriate summarisation ratio (9 summaries). Using an EM algorithm we determine the set of λ_k for those LiMs that minimizes the perplexity of the adapted model with respect to the human summaries made for that talk. Combining the results from the eight development talks we select an optimal set of λ_k for the remaining test talk. This process is approximately 500 times faster (assuming 1 CPU) than the previous one since it does not require the creation and evaluation of summaries (66 summaries created per development talk, for a total of 528 summaries).

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 metric as the one that is used during the development phase (i.e. SumACCY, ROUGE-2, ROUGE-SU4), 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. Human Transcription Results

Initial experiments were made on the human transcriptions (TRS), and results are given in Table 1. Exhaustive Adaptation results, as stated in [5], show that summarisation performance is improved by performing LiM adaptation, the best improvements obtained when using class models. PP Adaptation yields very similar results, in some cases a little better, others a little lower, with relative improvements compared to the Exhaustive Adaptation approach ranging from -2.1% to +4.9% over all metrics. In all cases, summarisation performance is improved by LiM adaptation, with relative improvements over the baseline ranging from 1.0% to 15.1%.

5.2. Automatic Speech Recognition Results

ASR results for each experiment are given in Table 2 for appropriate summarisation ratios. Results do not show significant improvements over the baseline by performing linguistic adaptation using word models, especially for the 10% summarisation ratio. Linguistic adaptation using class models, however, improves the performance for both Exhaustive and PP Adaptation in a similar way, with relative differences between the two development methods ranging from -4.6% to 7.8%. In all cases, improvements in terms of SumACCY and ROUGE metrics using PP Adaptation with class models are observed, ranging from 1.5% to 21.3% relative increase over the baseline.

5.3. Correlation

For each talk, summarisation ratio and evaluation metric, we also computed the correlation factor between the perplexities of the 66 linguistic models used during the Exhaustive Adaption development phase with the evaluation results of the automatically generated summaries made using those models. Average correlation factors over the nine talks are given in tables 3 and 4, for the Word and Class experiments, respectively. The negative correlation factors come from the nature of the correlation: the lower the perplexity of a given LiM, the higher the score a summary made using this LiM recieves. Results show that the perplexity of class models have a much higher correlation factor (in absolute value) with the evaluation metrics used in this study than word models. Correlation between perplexity and evaluation metrics is also stronger for the 30% summarisation ratio than for the 10% summarisation ratio.

		SumACCY	ROUGE-2	ROUGE-SU4
TRS	10%	-0.102	-0.108	-0.210
	30%	-0.170	-0.291	-0.268
ASR	10%	-0.341	-0.212	-0.213
	30%	-0.409	-0.416	-0.377

Table 3: Correlation between word model perplexity and evaluation metrics.

		Baseline			Exhaustive Adaptation			PP Adaptation		
		SumACCY	R-2	R-SU4	SumACCY	R-2	R-SU4	SumACCY	R-2	R-SU4
10%	Random	33.9	0.095	0.140	-	-	-	-	-	-
	Word	48.6	0.143	0.182	49.8	0.129	0.173	47.5	0.139	0.177
	Class	50.0	0.133	0.170	55.1	0.156	0.193	57.3	0.159	0.194
	Mixed	48.5	0.134	0.176	56.2	0.142	0.191	53.6	0.145	0.185
30%	Random	56.1	0.230	0.283	-	-	-	-	-	-
	Word	66.7	0.265	0.314	68.7	0.271	0.328	65.7	0.275	0.325
	Class	66.1	0.277	0.324	71.1	0.300	0.348	72.2	0.297	0.347
	Mixed	64.9	0.268	0.312	70.5	0.304	0.351	72.2	0.307	0.354

Table 2: ASR baseline and adapted results.

		SumACCY	ROUGE-2	ROUGE-SU4
TRS	10%	-0.579	-0.577	-0.582
	30%	-0.600	-0.591	-0.597
ASR	10%	-0.358	-0.255	-0.226
	30%	-0.617	-0.654	-0.593

Table 4: Correlation between class model perplexity and evaluation metrics.

6. Discussion

Improvements obtained by performing LiM adaptation using the perplexity-based development method are very similar to improvements obtained using the exhaustive-search approach, especially in the case of adaptation using class models, which perform much better than word models. The use of perplexity as a linguistic model evaluation metric is adapted to this task, yielding similar results while saving a lot of computation, and making the combination of a much larger number of component LiMs possible. The fact that the perplexity of class models is more highly correlated to the evaluation metrics used in this paper than word models explains why PP adaptation performs even better than Exhaustive Adaptation on experiments involving class models. A possible explanation for the higher correlation factors of the class models is that the data we are using to perform adaptation is very sparse in addition to being transcribed from spontaneous speech. This makes it difficult to obtain reliable estimates of word n-gram probabilities, whereas class models are more robust in such cases. This probably also explains why higher correlation factors are observed for the 30% summarisation ratio, since there is more data than for the 10% case. However we still expected combinations of word and class models to perform significantly better than class models alone. Even though the reductions in perplexity for each talk are greater for the Mixed case than for the Class one, the summary improvements are not much better, and sometimes worse, which shows that even though PP Adaptation is efficient for this task, it is still not perfect.

7. Conclusions

In this paper we investigated a perplexity-based development method to improve linguistic model adaptation using different sources of data for a speech summarisation system. Perplexity was found to be highly correlated to objective evaluation metrics used in the summarisation literature in the case of class models, which yield the best results in this task, with relative improvements ranging form 1.5% to 21.3%. Results are only very slightly improved, but the computation time saved is a significant advantage of the approach.

8. Acknowledgments

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] T. Kikuchi, S. Furui, 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] P. Chatain, E.W.D. Whittaker, J. Mrozinski, and S. Furui, "Topic and Stylistic Adaptation for Speech Summarization," *To appear in Proc. ICASSP, Toulouse, France*, 2006.

[5] P. Chatain, E.W.D. Whittaker, J. Mrozinski, and S. Furui, "Class Model Adaptation for Speech Summarization," *To appear in Proc. HLT-NAACL, New York, USA*, 2006.

[6] C. Hori, T. Hori, and S. Furui, "Evaluation Method for Automatic Speech Summarization," *Proc. Eurospeech, Geneva, Switzerland*, vol. 4, pp. 2825–2828, 2003.

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

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

[9] M. Wolfel and S. Burger, "The ISL Baseline Lecture Transcription System for the TED Corpus," Tech. Rep., Karlsruhe University, 2005.

[10] S. Burger, V. Maclaren, 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.

[11] H. Ney, U. Essen, and R. Kneser, "On Structuring Probabilistic Dependences in Stochastic Language Modelling," *Computer Speech and Language*, , no. 8, pp. 1–38, 1994.

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