論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Automatic sentence segmentation for speech summarization
著者(和文)	ムロジンスキ ヨアンナ, ウィッタッカー エドワード, 古井 貞熙
Authors(English)	Joanna Mrozinski, Pierre Chatain, Edward Whittaker, Sadaoki Furui
出典(和文)	日本音響学会2005年秋季講演論文集, Vol. 〔3-7-9〕, No. , pp. 119-120
Citation(English)	, Vol. 〔3-7-9〕, No. , pp. 119-120
発行日 / Pub. date	2005, 9

Automatic Sentence Segmentation for Speech Summarization*

©Joanna Mrozinski, Pierre Chatain, Edward Whittaker, Sadaoki Furui (Tokyo Institute of Technology)

1 Introduction

One of the major applications of automatic speech recognition is transcribing spontaneous speech. The resulting transcription is often ill-formed and includes redundant information such as disfluencies, word fragments and recognition errors. Automatic speech summarization can be used to produce more understandable and usable content. The acoustic segmentation normally used in speech recognition tasks is typically inadequate when the output of a speech recognizer is used as an input for a speech summarization system. Instead, segmentation to linguistically meaningful units (sentences) is needed. For that goal we present an automatic segmenter of transcribed speech based on N-gram language modelling.

2 Method

For sentence segmentation we used both word-based and class-based statistical language models with different sets of training data to model the distribution of words and sentence boundaries. The probability of a sentence boundary S and a non-sentence boundary NO-S preceding word w_k was estimated with the formulas adapted from [1]:

$$P_S(w_1 \ldots w_k) = P_{NO\text{-}S}(w_1 \ldots w_{k-1})\, p(\langle s\rangle | w_{k-2} w_{k-1})\, p(w_k | \langle s\rangle) + P_S(w_1 \ldots w_{k-1})\, p(\langle s\rangle | \langle s\rangle w_{k-1})\, p(w_k | \langle s\rangle).$$

$$P_{NO\text{-}S}(w_1 \ldots w_k) = P_{NO\text{-}S}(w_1 \ldots w_{k-1})\, p(w_k | w_{k-2} w_{k-1}) + P_S(w_1 \ldots w_{k-1})\, p(w_k | \langle s\rangle w_{k-1}).$$

In computing the probabilities only a history of 2 words preceding w_k was used. The best possible segmentation was found by keeping track of the n^2 best word/sentence boundary paths at each possible word boundary, and finally selecting the most probable complete path.

Automatic summaries with a compaction ratio of 30% were created with different segmentations. The summarization system selected the highest scoring sentences based on a word significance measure, a confidence measure and linguistic likelihood as explained in [2].

2.1 Data

Three different corpora were used: a broadcast news (*BN*) corpus of 160 million words, 16 million words of conference proceedings texts (*PROC*), and 50,000 words of manually transcribed lecture material (*LECT*) [3].

The segmentation evaluation was made on additional sets of lecture material for which both manual transcriptions (TRS) and automatic speech recognizer (ASR) output was available. This lecture material was divided into a development set of 15,000 words (6 lectures), and an evaluation set of 19,500 words (9 lectures). The word error rate of the ASR output was 33.3% for the evaluation set.

In evaluation we used human-made summaries created from manual transcriptions of the evaluation set lectures. For each lecture 8 different manual summaries were used as a target for estimating the summarization accuracy.

2.2 Models

For the automatic segmentation system three word-based trigram language models (*WLM*) were trained on the corpora. The class-based language models (*CLM*) were trained using two data sources; the word classes for the model were generated automatically using the *BN* corpus and the final *CLM* was created using the *LECT*. Finally, all four language models were combined using linear interpolation and the optimal weights for each were determined on the development set.

2.3 Evaluation

The evaluation of sentence segmentation was divided into two stages. First the weights of the different language models were optimized with the development set. To judge the quality of the segmentation we used the following metrics. Precision (*P*) is the ratio of correctly inserted sentence boundaries to the total number of inserted boundaries. Recall (*R*) is the ratio of correctly inserted sentence boundaries to the total number of target sentence boundaries. To combine *P* and *R* we used the F-measure, defined as $F = 2PR/(P+R)$.

Secondly, the system was used to produce sentence segmentation on the evaluation set which was then used as an input to the speech summarization system. Thus, in addition to the sentence boundary detection precision and recall the summarization accuracy results were used as an evaluation method, as proper segmentation was assumed to yield the best summarization results.

3 Results

The two stages of the evaluation process are approached separately: first the accuracy of the segmentation is examined and then the effect it has on the automatic summarization process.

3.1 Sentence segmentation

As a baseline for segmentation a set of 50,000 words of broadcast news transcriptions not included in the language model training data was segmented using only the *BN WLM* and results consistent with the previous experiments on trigram *WLM*s [1] were achieved (Table 1). The same segmentation method was then used to segment the evaluation set lectures. As expected, the results were worse than with *BN* data, lecture data being more spontaneous and ill-formed and containing more disfluencies. With the ASR output the results degraded even further, which is probably due to the high word error rate. As can be seen from Table 1, using different types of training data and language models made the results better. For the best results all four language models were interpolated with different weights. The optimal results were achieved with the following weights: *LECT+BN CLM* 0.4, *PROC WLM* 0.3, *LECT WLM* 0.2 and *BN WLM* 0.1.

*音声要約のための自動文区切りの検討
ヨアンナ ムロジンスキ，ピエール シャタン，エドワード ウィッタッカー，古井貞熙（東工大）

Table 1 Sentence segmentation results.

LM	Test data	Precision %	Recall %	F %
BN word LM	BN	63.4	65.1	64.2
	TRS	29.3	64.0	40.2
	ASR	22.0	45.0	29.6
Interpolated LMs	TRS	46.6	44.1	45.3
	ASR	35.5	28.5	31.6

Table 2 Segmentation evaluation with summarization.

Test data		SumAcc %	SumAcc E-max %	SumAcc E-avg %
TRS RndSeg SL20		55.1	25.5	15.3
TRS RndSeg SL50		58.5	27.1	14.7
TRS	HumSeg	65.4	29.1	16.4
	Autom	61.8	27.1	15.0
ASR RndSeg SL20		40.6	18.2	10.7
ASR	HumSeg	45.5	20.3	11.3
	Autom	45.0	20.7	11.0

Table 3 Summarization baselines.

Test type	SumAcc %	SumAcc E-max %	SumAcc E-avg %
TRS Rnd selection	55.9	21.8	11.9
TRS manual summ	71.1	32.0	19.6

3.2 Summarization

For summary evaluation word networks were built from the human-made summaries. These networks were used to calculate the word accuracy of each automatic summarization result using the most similar word string in the network (*SumAcc*). The automatic summaries were also compared to the individual summaries. When comparing to the individual summaries, both the average accuracy (*SumAcc E-avg*) and accuracy of the best matching summary (*SumAcc E-max*) were considered. The evaluation method is explained more thoroughly in [4]. The summarization scoring was optimized by cross validating on the evaluation set based on the *SumAcc* results.

Automatic summarization was run with different types of sentence segmentation. Lower baselines for the segmentation evaluation by summarization were generated by creating the summaries on randomly segmented data with different sentence lengths (*TRS RndSeg SL20, TRS RndSeg SL50, ASR RndSeg SL20*). On average the results were better when using longer sentences. The original human segmentation served as the upper baseline (*TRS HumSeg, ASR HumSeg*). Finally, the summaries were created with the best segmentations from the automatic sentence segmentation experiments as was determined on the development set experiments (*TRS Autom, ASR Autom*). The results are listed in Table 2.

With the TRS data the human segmentation produced the best summary results by all the evaluation metrics. Also with the ASR output the *HumSeg* outperformed other segmentations in the majority of cases. The randomized segmentations gave the worst results, even though in some cases the results were surprisingly high. The automatic segmentation results lay in between, and especially on the ASR data they were very close to the *HumSeg* results.

To gain a better understanding of the summarization evaluation metrics a lower baseline for summarization was generated by randomly selecting 10% of sentences from the human segmented. An upper baseline was created by comparing the human-made summaries to the word graphs built from the other human-made summaries. The results presented in Table 3 (*TRS Rnd selection, TRS manual summ*) show that the possible variation in *SumAcc, SumAcc E-max* and *SumAcc E-avg*-values is not large, and that the results achieved by replacing the summarization process with random sentence selection are close to those produced by the automatic summarization on randomly segmented data. Also it must be noted that as the parameter optimization for summarization was based on the *SumAcc*-results only, those figures should deemed more important than the *SumAcc E*-values. This indicates that the lead the human-made and automatic segmentations have over the random segmentations is significant.

4 Conclusions

Based on our results proper sentence segmentation is essential for attaining the best possible summarization accuracy. The results achieved by human segmentation give the best results and the automatic segmentation produces better summaries than random segmentations based on *SumAcc*-results. With the evaluation methods used in this study, it is however difficult to determine exactly what is the optimal segmentation or what is the precise effect that the segmentation has on summarization quality. The correctness of the sentences or the readability of the created summaries does not show in the word graph comparisons and in consequence the differences between our final summarization lower and upper baseline is small. Also the comparatively low proficiency of the sentence segmentation system makes the interpretation of the final results difficult. Further research on evaluation methods that would give more weight to the correctness of sentences and on defining the nature of optimal segmentation is needed.

Acknowledgements

We thank Matthias Wölfel, Chiori Hori and the rest of the IWSpS 2004 team for assistance and for preparing the *BN* corpus and the manual summarizations. This work is supported in part by the 21st Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources".

References

[1] A. Stolcke and E. Shriberg. *Automatic linguistic segmentation of conversational speech.* In Proc. ICASSP, vol. 2, pp.1005–1008, Philadelphia, 1996.

[2] C. Hori and S. Furui. *A New Approach to Automatic Speech Summarization.* IEEE Transactions on Multimedia, Vol. 5, NO. 3, SEPTEMBER 2003, pp. 368–378.

[3] Linguistic Data Consortium (LDC). *Translanguage English Database.* www.ldc.upenn.edu/Catalog/LDC2002S04.html

[4] S. Furui, M. Hirohata, Y. Shinnaka and K. Iwano. *Sentence extraction-based automatic speech summarization and evaluation techniques.* Symposium on Large-Scale Knowledge Resources (LKR2005), Tokyo, Japan, pp.33–38 (2005-3).