AUTOMATIC SENTENCE SEGMENTATION OF SPEECH FOR AUTOMATIC SUMMARIZATION

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

Tokyo Institute of Technology
Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

ABSTRACT

This paper presents an automatic sentence segmentation method for an automatic speech summarization system. The segmentation method is based on combining word- and class-based statistical language models to predict sentence and non-sentence boundaries. We study both the performance of the sentence segmentation system itself and the effect of the segmentation on the summarization accuracy. The sentence segmentation is done by modelling the probability of a sentence boundary given a certain word history with language models trained on transcriptions and texts from several sources. The resulting segmented data is used as the input to an existing automatic summarization system to determine the effect it has on the summarization process. We conduct all our experiments with two types of evaluation data: broadcast news and lecture transcriptions. The automatic summarizations are created with different sentence segmentations and different summarization ratios (30% and 40%) and evaluated by comparing them to human-made summaries. We show that a proper sentence segmentation is essential to achieve good performance with an automatic summarization system.

1. INTRODUCTION

Spontaneous speech typically suffers from ungrammatical constructions and contains redundant information such as false starts, word fragments, repetitions and so on. The output of an automatic Speech-To-Text (STT) system has additional problems as the word recognition error rates are still quite high with spontaneous speech. STT-output also has no punctuation or proper segmentation. The readability and usability of such data can be improved by segmenting the text into logical units such as sentences. Furthermore, automatic speech summarization can then be used to remove redundancies and erroneous parts, and to extract the important parts of data.

Both speech segmentation and summarization methods have been widely researched in recent years, but the effect of the segmentation on summarization has not been comprehensively studied before. It has been shown however that the segmentation has a significant effect on the further processing of the speech, such as information extraction and topic detection [1, 2]. The goal of this study is to show the significance of the sentence segmentation on the summarization system and how the sentence segmentation can be improved automatically.

Research has shown that statistical methods developed for segmenting written text are insufficient when processing speech, due to the poor grammatical structure, disfluencies, incorrectly recognized words and other characteristics of speech or STT-output [3]. Even the definition of a sentence in speech is unclear and recent research has instead concentrated on detecting so-called slash or sentence units (SU) [4, 5, 6]. In the light of earlier sentence segmentation research [7, 8] and the aforementioned SU-research the statistical language modeling approach is insufficient for proper automatic sentence segmentation when applied to STT-output, but our study shows that it is enough to see the effect on the automatic summarization process. Our results cannot be directly compared to those in the literature, since our sentence definition is highly dependent on the original segmentation which in turn affects the evaluation target summary creation process.

The paper is organized as follows. We start by explaining the sentence segmentation method (Section 2) and the principles of the automatic summarization (Section 3), then the experimental setup (Section 4) and the results (Section 5), and finally the discussion (Section 6) and the conclusions (Section 7).

2. AUTOMATIC SENTENCE SEGMENTATION

In this paper we perform sentence segmentation using both word-based and class-based statistical language models (LM) trained on different sets of training data to model the probabilities of words and sentence boundaries. To judge the quality of the sentence segmentation we used the following metrics (1) Precision (P) is the ratio of correctly inserted sentence boundaries to the total number of inserted boundaries; (2) Recall (R) is the ratio of correctly inserted sentence boundaries to the total number of target sentence boundaries. To combine P and R we use the F-measure, defined as $F = 2PR/(P+R)$ (harmonic mean of P and R).

The probability of a sentence boundary $P_S(w_1..w_i)$ and a non-sentence boundary $P_{NO-S}(w_1..w_i)$ preceding word w_i was modelled using Eq. (1) and Eq. (2) adapted from [9]:

$$\begin{aligned} P_S(w_1..w_i) &= P_S(w_1..w_{i-1})p(S|Sw_{i-1})p(w_i|S) + \\ &\quad P_{NO-S}(w_1..w_{i-1})p(S|w_{i-2}w_{i-1})p(w_i|S), \quad (1) \\ P_{NO-S}(w_1..w_i) &= P_S(w_1...w_{i-1})p(w_i|Sw_{i-1}) + \\ &\quad P_{NO-S}(w_1..w_{i-1})p(w_i|w_{i-2}w_{i-1}). \quad (2) \end{aligned}$$

where S is a sentence boundary and $NO-S$ a normal word boundary. For each position i a history of $n-1$ words or word boundaries preceding w_i was used to predict the local probabilities. This probability was combined with the matching recursive path probability from w_{i-1}. In the formulae above, a trigram is used, giving a history of 2 words. When the range of the n-gram used grows longer, the complexity of different possible word/word boundary histories increases. As distinct from the method presented in [9] (using Viterbi search to find the best path of words and sentence boundaries) we keep track of the n^2 most probable paths leading from w_{i-n} to w_i

and make a decision between sentence boundary and non-sentence boundary preceding the word w_{i-n}.

Sentence segmentation was done using both word-based and class-based statistical LMs trained on different sets of training data to model the probabilities of words and sentence boundaries. Three LMs were used in sentence segmentation, two word-based LMs and a class-based LM [10]. The LMs were combined by linear interpolation as follows:

$$P(w_i|h) = \sum_m \lambda_m P_m(w_i|h) \quad (3)$$

where h is a word history and P_m is either a word-based n-gram LM

$$P_m(w_i|h) = P(w_i|w_{i-n+1}, ..., w_{i-1}) \quad (4)$$

or a class-based n-gram LM

$$\begin{aligned} P_m(w_i|h) &= P(w_i|C(w_i)) \times \\ & \quad P(C(w_i)|C(w_{i-n+1}), ..., C(w_{i-1})). \end{aligned} \quad (5)$$

We determined the optimal values of λ_m by running sentence segmentation experiments on the development data and finding the highest possible F-measure.

3. AUTOMATIC SUMMARIZATION

Automatic summarization was performed using sentence extraction, which selects the highest scoring sentences based on a combination of word significance score, confidence score and linguistic likelihood [11]. The sentences were extracted as they appeared in the data; they were not compacted.

Summarization evaluation is a difficult task, as even the human-made summaries tend to differ greatly from each other, and it is difficult to determine what is the optimal summary for a given text. Using the average of several subjective human judgments would give a reliable estimate of quality of an automatic summary but this type of evaluation would be too time-consuming and expensive in practice. Therefore the summarization evaluation was done by comparing the automatic summaries to a group of human-made target summaries as follows: several human subjects created summaries through sentence extraction and compaction from manual transcriptions. The subjects were advised to maintain the original sentence segmentation; combining parts of sentences into new ones was not allowed.

The human-made summaries were then used to create word networks, one network for all the summaries from each lecture. These networks were used to calculate the summarization accuracy (SumACCY) of each automatic summarization result using the most similar word string in the network (SumACCY NetW). Because SumACCY NetW tends to be too optimistic when there are many human-made summaries and especially when the summarization ratio is high, the automatic summaries were also compared directly to the individual summaries. When comparing to the individual summaries, both the average accuracy (SumACCY E/avg) and the accuracy of the best matching summary (SumACCY E/max) were considered. In all cases we compared the summaries as two long word strings without sentence boundary markers. The SumACCY for NetW, E/avg and E/max is defined in Eq. (6):

$$SumACCY = (Cmlen - Sub - Ins - Del)/Len \times 100[\%] \quad (6)$$

where $Cmlen$ is the comparison maximum length, Sub is the number of substitution errors, Ins is the number of insertion errors and Del is the number of deletion errors. For SumACCY E/avg and E/max the $Cmlen$ is a maximum length of the two summaries being compared (automatic and target summary). For SumACCY NetW it is the maximum length of the most similar word string in the network and the automatic summary being compared. The evaluation method is explained more thoroughly in [12].

4. EXPERIMENTAL SETUP

Sentence segmentation and summarization were performed on two different data sets: broadcast news stories (CNN) and conference lectures (TED$_e$) from *Translanguage English Database* [13].

4.1. LM training and optimization

Three different corpora were used to train the three LMs used in the sentence segmentation task: LM$_{bn}$ was trained on 160 million words of broadcast news data (BN) and LM$_{proc}$ was trained on 16 million words of conference proceedings texts. The third corpus (TED$_t$) comprised 28 lectures (50K words) from *Translanguage English Database*, disjoint from the evaluation data. TED$_t$ was not large enough to train a robust word-LM and therefore we used the BN corpus to generate word classes and trained the class-based LM (LM$_{ted}$) using these word classes on the TED$_t$.

Finally, all the language models were combined using linear interpolation as in Eq. (3) and the optimal weights (λ_m) for each component were determined experimentally on the TED development set. In the experiments the range of n-grams varied between n=3, 4, 5, depending on the training data and LM combination used.

As development set for the CNN segmentation task we separated 1 million words from the BN training data. A class-model approach with different types of LM interpolation was tested but the best results were obtained using only LM$_{bn}$.

The best combination of word- and class-based LMs to segment the TED$_e$ was determined using a development set of 16 lectures (32K words). The first part of the development set was a fixed set of 8 lectures, into which we added 8 lectures from the evaluation set and used them through a rotating form of cross-validation [14]: the first 8 from the evaluation set and the fixed set of 8 development lectures was used to determine the weights to segment the 9th lecture, and so on.

4.2. Evaluation data

The evaluation of sentence segmentation was divided into two stages. Firstly the weights of the different language models were optimized using the development set. Secondly, the system was used to produce sentence segmentation on the evaluation set which was then used as the input to the speech summarization system. Thus, in addition to sentence boundary detection precision and recall the summarization accuracy results were also used as an evaluation method. Proper sentence segmentation was hypothesized to yield the best summarization results.

The same evaluation data was used in both sentence segmentation and summarization evaluation tasks: the CNN consisted of five news stories and TED$_e$ of nine lectures. Tests were conducted on both manually transcribed data (TRS) and an STT output (STT). The CNN comprised 2K words (125 sentences). The average sentence length was 15.7 words, with standard deviation of 9.6. The WER of the CNN STT was 22%. For each story there were 16 corresponding human-made summaries, which we used as targets in the summarization evaluation. The TED$_e$ comprised 20K words (700 sentences), with average sentence length of 28.7 words (standard de-

LM	Data	Precision	Recall	F-measure
LM_{bn}	CNN	70.4	65.2	67.7
	TED_e	56.7	39.7	46.7
$LM_{bn+proc+ted}$	TED_e	52.9	46.8	49.6

Table 1. Precision [%], recall [%] and F-measure [%] for sentence segmentation on TRS data.

LM	Data	Precision	Recall	F-measure
LM_{bn}	CNN	63.6	63.0	63.3
	TED_e	35.5	28.1	31.3
$LM_{bn+proc+ted}$	TED_e	35.9	38.4	37.1

Table 2. Precision [%], recall [%] and F-measure [%] for sentence segmentation on STT data.

viation 22.7). The sentence lengths between and within the lectures varied significantly, as did the speaking styles. The lectures were selected so that all the speakers were English native speakers, as distinct to the TED_t. The WER of the TED_e STT was 33% on average. For each lecture there were 8 human-made summaries available.

5. RESULTS

The results of the segmentation and summarization experiments are examined separately, first the accuracy of the segmentation (Section 5.1) and then the effect it has on the automatic summarization process (Section 5.2).

5.1. Automatic sentence segmentation

As a lower bound for segmentation experiments the CNN evaluation set was segmented using only the LM_{bn}. The tests were run on both TRS (Table 1) and STT data (Table 2). The results are consistent with previous experiments that used a similar method with a trigram word-based LM [9]. The same method was then used to segment the TED_e. Finally, we used the linear interpolation approach to segment the TED_e data.

The TED_e segmentation results were notably worse than the CNN results. This was to be expected, as the TED_e data is more spontaneous and thus more ill-formed. Using the LM_{bn} only, the TED_e TRS F-measure was 20.0% absolute lower than the CNN TRS, and the TED_e STT 32.0% absolute lower than CNN STT. This large decrease is probably due to the higher word error rate (33.3% on average).

The best TED_e results were obtained using the linear interpolation of LM_{bn}, LM_{proc} and LM_{ted} (Tables 1 and 2: $LM_{bn+proc+ted}$). With TRS data the precision and recall were 52.9% and 46.8%, the precision being lower than with the LM_{bn} approach, but both recall and F-measure being higher. With STT data the precision was only slightly higher than with the LM_{bn} approach, but the increase of recall and F-measure was substantial. With both TED_e TRS and STT data the difference between LM_{bn} and $LM_{bn+proc+ted}$ was statistically significant at the 99% level (McNemar's test).

5.2. Automatic summarization

To determine the range of values the summarization accuracy metrics can have, a lower bound for summarization was generated by randomly selecting sentences from the human segmented original texts (both TRS and STT) according to the desired summarization ratio. An upper bound was created on TRS data by comparing the human-made summaries to the word graphs built from the other human-made summaries. Automatic summarization was then run with three different types of sentence segmentation: for comparison, the summaries were created on randomly segmented data and with the original human-made segmentation. Finally, we created the summaries using the best automatic sentence segmentation as determined on the development set.

SumACCY	RndSel	RndSeg	Autom	HSeg	MSum
CNN, summarization ratio 40%					
E/avg	19.8	21.0	23.9	26.2	22.4
E/max	35.9	43.2	42.2	43.3	62.0
NetW	69.9	72.1	74.3	75.1	82.5
TED, summarization ratio 30%					
E/avg	12.1	14.9	16.0	16.2	20.0
E/max	22.1	24.8	25.9	27.8	37.9
NetW	60.5	59.7	65.2	67.3	69.6

Table 3. SumACCY E/avg [%], E/max[%] and NetW[%] for automatic summarization on TRS data.

SumACCY	RndSel	RndSeg	Autom	HSeg
CNN, summarization ratio 40%				
E/avg	18.6	18.7	27.8	23.5
E/max	34.9	27.0	36.8	40.2
NetW	64.6	63.6	67.6	68.8
TED, summarization ratio 30%				
E/avg	10.3	12.0	12.2	13.6
E/max	18.2	18.5	20.1	24.7
NetW	45.9	49.7	52.3	55.7

Table 4. SumACCY E/avg [%], E/max[%] and NetW[%] for automatic summarization on STT data.

The results of the summarization experiments with different data and different summarization ratios (CNN 40% and TED_e 30%) are presented in Table 3 for TRS and Table 4 for STT. As hypothesized, the summaries generated by random selection (*RndSel*) gave the worst results in most of the cases, with three exceptions. In three cases (TED_e TRS NetW, CNN STT E/max and CNN STT NetW) the summarization on randomly segmented data (*RndSeg*) gave worse results than *RndSel*.

The summarization results on the automatic segmentation (*Autom*) outperformed both the *RndSel* and *RndSeg* in all cases except one: the CNN TRS E/max *RndSeg*-value was lower than the CNN TRS E/max *RndSeg*-value. The exception is not essential: the automatic summarization was conducted without sentence compaction, but the sentences in the target summaries were compacted. This means that the human-made target segmentation cannot produce perfect E/avg or E/max scores. Thus it is possible to find a segmentation that gives better summarization results than the human-made or automatic segmentation. This is especially the case with E/max where one good result makes the difference even when the majority of the results are bad. Both E/avg and NetW include all the summarization accuracies, not only the best one found. The difference between *Autom* and the randomized lower bounds can be seen in both CNN and TED_e results, but it is clearer in CNN results, where also the automatic sentence segmentation results were higher.

The human-made segmentations (*HSeg*) gave the best results in all automatic summarization tests except one: the CNN STT E/avg *Autom* outperformed the *HSeg*. The human-made summaries (*MSum*) were the best as expected and in most cases the difference between

the *HSeg* and the *MSum* was large. The *MSum* results were notably low in one case: the *MSum* E/avg results were lower than the CNN TRS *Autom* and *HSeg* and even the CNN ASR *HSeg*. This emphasizes the difficulties of summarization evaluation: from one text it is possible to create several acceptable summaries that have nothing or only a little in common.

It is not clear how to show statistical significance on summarization results, but our lower and upper bounds showed that possible variation in summarization accuracy values is not large. The trend of the results are as follows: 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. The automatic summarization on *Autom* plainly outperforms the randomized bounds and *HSeg* produces the best automatic summary results.

6. DISCUSSION

Based on our results proper sentence segmentation is essential for attaining the best possible summarization accuracy for both broadcast news and spontaneous lecture speech. Summarizing human-segmented data gave the best results and the automatic segmentation produced better summaries than random segmentations, even with our relatively low sentence segmentation precision and recall. However it is difficult to determine exactly what is the optimal segmentation. Given that the summarization is made through sentence extraction only, even the original human segmentations cannot give very good results especially when the summarization ratio is small because the sentence compaction in the manual summarizations is considerable. Using compaction in the summarization process could compensate bad segmentation but so far our experiments with TRS-data using compaction have always produced worse results than pure sentence extraction.

The grammatical correctness of the sentences or the readability of the created summaries does not show directly in out evaluation metrics and in consequence the differences between our final summarization lower and upper bound is small. 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.

7. CONCLUSIONS

We have presented an automatic segmentation method for automatic speech summarization. Our segmentation method combined word- and class-based statistical language models to model the probability of a sentence boundary given a word history.

We studied both the performance of the segmentation system and the effect of the segmentation on the summarization accuracy. We conducted our segmentation and summarization experiments with CNN broadcast news stories and TED_e conference lectures. The automatic segmentation produced better results than the randomized segmentation, and original human-made segmentation gave the best automatic summarization results. We conclude that proper automatic sentence segmentation is essential to achieve good performance with an automatic summarization system.

8. ACKNOWLEDGEMENTS

We thank Matthias Wolfel, 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".

9. REFERENCES

[1] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1–2, pp. 127–154, September 2000.

[2] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. Ramshaw, D. Stallard, R. Schwartz, and B. Xiang, "The effects of speech regocnition and punctuation on information extraction performance," in *Proc. EUROSPEECH*, Lisbon, Portugal, September 2005, pp. 57–60.

[3] M. Stevenson and R. Gaizauskas, "Experiments on sentence boundary detection," in *Proc. ANLP*, Seattle, April 2000.

[4] D. Hillard, M. Ostendorf, and A. Stolcke, "Improving automatic sentence boundary detection with confusion networks," in *Proc. HLT-NAACL*, 2004, pp. 69–72.

[5] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Comparing and combining generative and posterior probability models," in *Proc. EMNLP*, Barcelona, Spain, July 2004.

[6] Y. Liu, E. Shriberg, A. Stolcke, P. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P.C. Woodland, and M. Harper, "Structural metadata research in the ears program," in *Proc. ICASSP*, Philadelphia, PA., March 2005, vol. V, pp. 957–960.

[7] D. Hakkani-Tur, Z. Tur, A. Stolcke, and E. Shriberg, "Combining words and prosody for information extraction from speech," in *Proc. EUROSPEECH*, Budapest, Hungary., September 1999.

[8] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing*, M. Ostendorf M. Johnson, S. Khudanpur and R. Rosenfeld, Eds., pp. 105–114. Springer, 2002.

[9] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. EUROSPEECH*, Philadelphia, PA, 1996, pp. 1005–1008.

[10] 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.

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

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

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

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