Pacific Association for Computational Linguistics

SEGMENTATION AND EVENT DETECTION OF NEWS STORIES USING TERM WEIGHTING

YOSHIMI SUZUKI AND FUMIYO FUKUMOTO AND YOSHIHIRO SEKIGUCHI

Department of Computer Science and Media Engineering,
Yamanashi University
4-3-11 Takeda, Kofu 400-8511 Japan
{ysuzuki@alps1.esi, fukumoto@skye.esb, sekiguti@alps1.esi†}.yamanashi.ac.jp

In this paper, we propose a segmentation method and event detection method for news stories. In our segmentation method, we calculate term weighting automatically and use a suitable window function to locate the boundaries between adjacent news stories correctly. Using the window function, we can locate the boundaries between short stories in news source correctly. In the experiments, we examined the results using four kinds of window functions and compared the results. We also propose an event detection method. In our method, we used χ^2 value as term weighting which is the degree of bias for specific events. We used many words (nouns, adjectives, adverbs and verbs) in news stories for segmentation and event detection. The results of segmentation and event detection experiments show that our method is effective.

Key words: segmentation, event detection, term weighting.

INTRODUCTION

Recently, information retrieval and information extraction are studied by many researchers. Especially, Topic Detection and Tracking (TDT) is studied by the TDT Pilot study (Allan TDT 1998, Yang 1998, Allan SIGIR 1998). In TDT Pilot study, there are three major tasks: (1)segmentation (2)topic identification (3)event tracking.

In TV news and radio news, there are some different news stories. In speech data, it is difficult to locate the boundaries correctly between adjacent news stories. To segment each story is important for information retrieval and information extraction. For event detection and event tracking of radio news and TV news, highly efficient story segmenter is indispensable.

Dragon Systems studied segmentation by using TDT corpus. They used HMM for segmentation of a news source into stories. But using HMM, it is difficult to detect which words contribute segmentation and to improve the segmentation system. Hearst (Hearst 1997) segmented text into subtopic passages. But her target was multi-paragraph segmentation of expository texts, and she did not mentioned window function. Nomoto (Nomoto 1994) segmented editorial articles of Japanese newspaper using $tf \cdot idf$. We tried to locate the boundaries between adjacent stories of news source using term weighting with the suitable window function.

We proposed a term weighting method for topic identification and keyword extraction of Japanese radio news (Suzuki 1998). Our basic idea is : in a same stories of news, some keywords (they identify the topic of a story) are frequently appear, but in different stories, they are not.

In this paper, we propose the segmentation method and report the experiments with transcribed speech data (The Topic Detection and Tracking: TDT corpus). For training, the system calculates term weighting using a part of TDT corpus. Training text which we used are analyzed morpheme by Brill's part-of-speech tagger (Brill 1998). Then the system extracts words which appear frequently, and counts frequency of each word. For term weighting, we calculate χ^2 values of each word. In order to find a suitable window function, we examined the results which are obtained by using 4 kinds of window functions. For stories segmentation, the system shifts the window word by word, and calculate similarity between a word sequence and its adjacent word sequence. We also propose event detection method and conducted the experiments with transcribed broadcast news and newswire. On training phase, the system calculates term weighting using a part of TDT corpus. Training text which we used are analyzed morpheme by Brill's part-of-speech tagger (1992 Brill). Then the system extracts words which appear in the training data frequently, and counts frequency of each word. For term weighting, we calculate χ^2 values of each word and the 25 events which were selected by TDT Pilot Study. On event detection phase, we detect which event is suitable for each news story by calculation of similarity between each news story and each of the 25 events.

1. TRAINING PHASE

We used χ^2 values for term weighting for a segmenter of news stories. χ^2 value of the words which appeared in specific stories are large (e.g. Simpson). On the other hand, χ^2 value of words which appeared in almost news stories frequently are small (e.g. today). However, some words (said, we, do and so on) appeared in many news stories frequently, while other words (Simpson, trial, currency and so on) appeared in specific events. Therefore, we used χ^2 values in order to weight for specific words.

2. SEGMENTATION

On segmentation phase, we used the window function which are shown in Figure 1. Using two adjacent windows: the first half of a window (Window A) and the second half of the window (Window B), we calculated similarity between word sequences in the two adjacent windows for segmentation. In order to locate correct boundaries, we shift the windows word by word and search local minima of similarity between the two adjacent windows.

We used Hanning window, Hamming window and Blackman window in order to segment news sources into news stories. Formula (1) shows similarity between Window A and Window B.

$$Sim(i,j) = \sum_{i=1}^{\frac{N}{2}} \sum_{j=\frac{N}{2}+1}^{N} win[i]\chi^2_{wi} \times win[j]\chi^2_{wj} \times eq(i,j) \qquad (1)$$

where

$$eq(i,j) = \begin{cases} 1 & \text{if } wi = wj \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

In formula (1), N indicates the number of words in a window, wi and wj indicates words of i-th and j-th of a window, respectively. win[] indicates window function (rectangle window, Hanning window, Hamming window and Blackman window).

When the boundary of Window A and Window B agrees with the correct boundary of a story, the similarity between Window A and Window B is remarkably small. When the correct boundary of story is off to the slide of the boundary of Window A and Window B, the similarity between Window A and Window B is large.

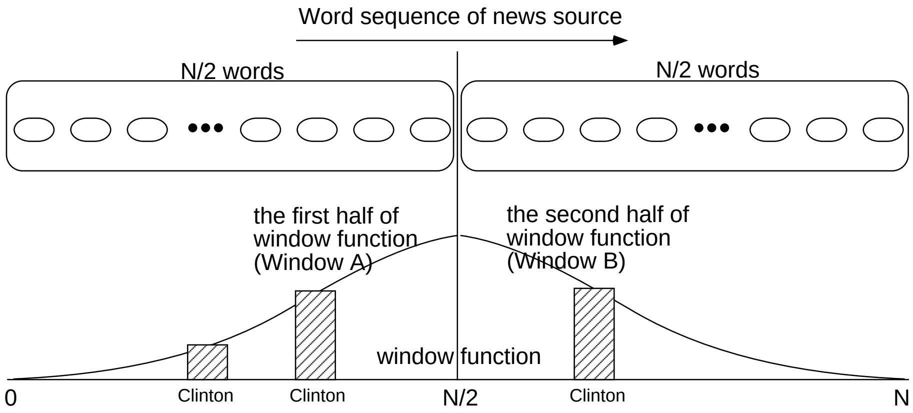
Word sequence of news source
N/2 words
N/2 words
the first half of window function (Window A)
the second half of window function (Window B)
window function
0
Clinton
Clinton
N/2
Clinton
N

Figure 1. Windows for segmentation

The other advantage of the proposed window is that the window is not affected by length of each story. If we used rectangle window whose width is 300 words and length of an story is 200 words, we use 100 words in next story with same weight as other 200 words. Using the proposed window, weight of the edge of the window is smaller than the weight of words which are neighborhoods of center of the window.

For segmentation, compound words have very effective information. We used adjacent two and three words for segmentation and event detection. For example, 'New' is not very effective for event detection, but 'New York' is effective for story segmentation and event detection. We give 4 times weight to the same adjacent two words in Window A and Window B. We give 9 times weight to the same adjacent three words in Window A and Window B.

For decision of boundaries of news stories, the system shifts a window word by word and the system extracts a boundary where $Sim(i,j)$ indicates local minimum.

3. EVENT DETECTION

On event detection phase, we use χ^2 value of each word and each event. Then, we calculate similarity between feature vector of each event and each news story.

$$Sim(j,i) \quad = \quad \sum_{k}^{N_i} \chi^2_{kj} \qquad (3)$$

where j shows an event number: $1 \leq j \leq 25$, and i is news story number. N_i indicates the number of words in the i-th news story. Detected event of each news story is shown in Formula (4),

$$Event_i = \begin{cases} \arg_j \max Sim(j,i) & \text{if } \max_j Sim(j,i) > thre \\ 26(other\ events) & \text{otherwise} \end{cases} \qquad (4)$$

where *thre* is the threshold which was examined by a preliminary experiment.

4. EXPERIMENTS

This section describes the results of the segmentation and event detection experiments. The result of each experiment is shown in 4.2 and 4.3, respectively.

4.1. Training Data and Data for Evaluation

We used 1001 stories of TDT corpus (TDT001000-TDT002000) for evaluation and used the rest stories of TDT corpus (TDT000001-TDT00999 and TDT002001-TDT015863) for training data. In the data for evaluation, there are 426 REUTERS newswire and 575 CNN broadcast news. In the training data, there are 7539 REUTERS newswire and 7323 CNN broadcast news. In evaluation data, there are 429,712 words. In each story in evaluation data has about 429 words on the average.

4.2. Segmentation

For segmentation experiment, We used 4 kinds of windows: rectangle window, Blackman window, Hanning window and Hamming window. We used 3 kinds of window widths: 300 words, 400 words and 500 words. Cut-off frequency of word is 100.

Table 1. Segmentation result

Method		P_{miss}	$P_{FalseAlarm}$	error rate
Hamming window	(N=300)	14.3%	12.7%	27.0%
	(N=400)	16.1%	8.2%	24.3%
	(N=500)	20.6%	7.9%	28.5%
Hanning window	(N=300)	17.1%	15.7%	32.8%
	(N=400)	21.7%	10.8%	32.5%
	(N=500)	27.9%	9.0%	36.9%
Blackman window	(N=300)	18.3%	17.0%	35.3%
	(N=400)	21.2%	13.6%	34.8%
	(N=500)	27.9%	14.0%	41.9%
rectangle window	(N=300)	15.8%	12.1%	27.9%
	(N=400)	18.7%	9.7%	28.4%
	(N=500)	27.4%	5.4%	32.8%
Dragon Systems (Allan TDT 1998)				12.9%

In Table 1, P_{miss} and $P_{FalseAlarm}$ are defined by formula 5 and formula 6, respectively.

$$P_{miss} = \frac{\sum_{i=1}^{N-k} \delta_{hyp}(i, i+k)(1 - \delta_{ref}(i, i+k))}{\sum_{i=1}^{N-k}(1 - \delta_{ref}(i, i+k))} \qquad (5)$$

$$P_{FalseAlarm} = \frac{\sum_{i=1}^{N-k}(1-\delta_{hyp}(i,i+k))\delta_{ref}(i,i+k)}{\sum_{i=1}^{N-k}\delta_{ref}(i,i+k)} \quad (6)$$

where the summations are over all the words in the corpus and where

$$\delta(i,j) = \begin{cases} 1 & \text{when words } i \text{ and } j \text{ are from the same story} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

We defined k as 250 (It depends on TDT segmentation task).

In Table 1, the result using Hamming window whose window width was 400 was the best of all our results. However, each result is not much different with each other. As compared with the result of Dragon Systems, our result was slightly worse than the system.

4.3. Event Detection

In the event detection experiments, the system sorted 1001 news of the corpus to 25 target events and other events. Table 2 shows the result of event detection.

Table 2. result of event detection

	χ^2	$tf \cdot idf$	word frequency
A (M=30)	30	28	28
B (N=971)	944	864	851
$\frac{\frac{A}{M}+\frac{B}{N}}{2}$	98.6%	91.2%	90.5%

In Table 2, A indicates the number of the stories whose events are judged to be one of the 25 events correctly. B indicates the number of the stories whose events are not one of the events and recognized correctly. M indicates the number of the stories whose events are the 25 target events (30) and N indicates the number of the stories whose events are not the 25 target events (971). We used the average of $\frac{A}{M}$ and $\frac{B}{N}$ for estimation of event detection results. According to the results, the method using χ^2 was better than that of using $tf \cdot idf$ and that of using word frequency. When rearranging the experimental results, the result using χ^2 values (our method) is better than the result using $tf \cdot idf$ and the result using word frequencies. Using χ^2 method, all news stories which are about the selected 25 events are detected correctly. However some news stories which are not about the 25 events are detected as the selected 25 events.

5. DISCUSSION

In segmentation experiment, the result is slightly lower than the result of Dragon systems, since our method used only weight which indicate the degree of the feature of each event. If we combine our method with other methods, i.e. with rhetorical information etc., the result will be better. For further improvement in this system, we will use the combination of two window functions, such as Blackman window and rectangle window.

In the event detection experiment, some news stories which are not about the 25 events used for training are detected as the selected 25 events incorrectly. If we use larger training data and tune term weighting accurately, we will obtain better results. As a result of experiment, the method using χ^2 values was better than the result using $tf \cdot idf$ and the result using word frequency. The reason is χ^2 values have more feature than $tf \cdot idf$ and word frequency.

6. CONCLUSIONS

In this paper, we proposed a method for segmentation and event detection by using term weighting for transcribed speech data. Our method is based on term weighting which indicate degree of the feature of each word and each event. It is difficult to segment and detect suitable event for an unique news story with our method. For better result, we have to use rhetorical information and local context analysis with our method. In future, we will conduct the segmentation and event detection experiment using speech data.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

ALLAN, J., CARBONELL, J., DODDINGTON, G., YAMRON, J., and YANG, Y. 1998. Topic Detection and Tracking Pilot Study Final Report. the DARPA Broadcast News Transcription and Understanding Workshop.

ALLAN, J., PAPKA, R., and LAVRENKO, V. 1998. On-line New Event Detection and Tracking. In 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 37–45.

BRILL, E. 1992. A simple rule-based part of speech tagger. Proceedings of the 3nd conference on applied natural language processing, 152-155.

HEARST, A. M. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics **23**:33–64.

NOMOTO, T., and NITTA, Y. 1994. A Grammatico-Statistical Approach to Discourse Partitioning. Fifteenth International Conference on Computational Linguistics(COLING), 1145-1150.

SUZUKI, Y., FUKUMOTO, F., and SEKIGUCHI, Y. 1998. Keyword Extraction using Term-Domain Interdependence for Dictation of Radio News. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, **2**:1272–1276.

YANG, Y., PIERCE, T., and CARBONELL, J. 1998. A Study on Retrospective and On-Line Event Detection. In 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 28–36.