Correlating Newswire Articles with TV News Story using Features of TV News

Yoshimi Suzuki   Yoshihiro Sekiguchi
Dept. of Computer Science and Media Engineering
Yamanashi University
4-3-11 Takeda, Kofu 400 Japan
{ysuzuki,sekiguti}@alps1.esi.yamanashi.ac.jp

Abstract

This paper proposes a method for correlating newswire articles with TV news story. The significant points of our method are:

1. *use of a highly efficient term weighting method*
2. *use of the number of turns of talk in which keywords appear (we defined turns of talk as another speaker begins to talk after someone finished to talk)*
3. *use of overt pronoun resolution*

The method was tested on a corpus of texts from the Reuters newswire articles and the CNN TV news story, and the result can be regarded as a promising the usefulness of the method.

1 Introduction

TV news programs serve a lot of news stories. Many of them are related to other news stories. Correlating newswire articles with TV news story is very useful, as TV watchers can watch TV news program and the related news simultaneously. Figure 1 illustrates a concept of TV news - newswire correlating system. In the bottom of Figure 1, newswire articles which are related to TV news story are displayed.

In this paper, we present a method for correlating the Reuters newswire articles with the CNN news story automatically.

The significant points of our method are:

1. use of a highly efficient term weighting method
2. use of the number of turns of talk in which keywords appear (we defined turns of talk as another

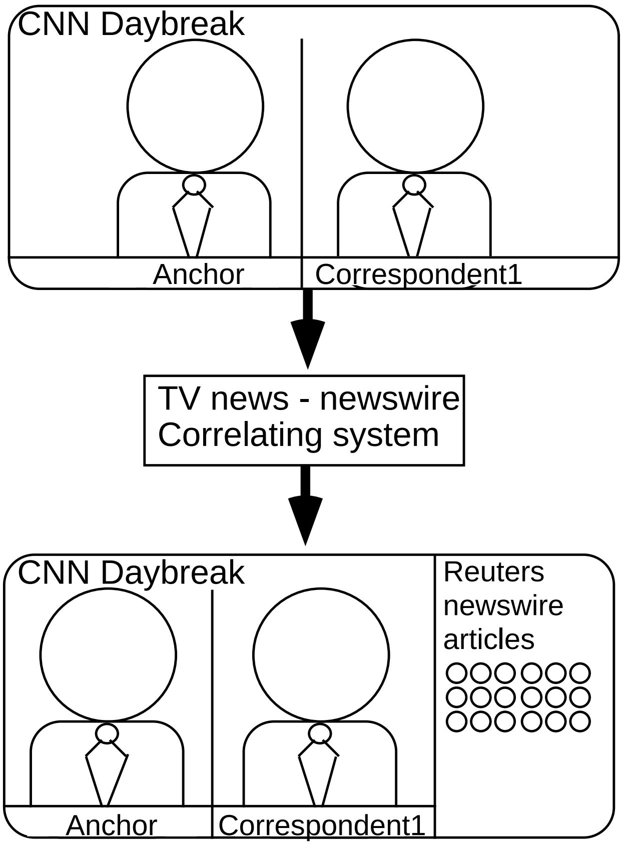

Figure 1: A concept of TV news - newswire correlating system

speaker begins to talk after someone finished to talk)

3. use of overt pronoun resolution

(Henceforth, we call it **feature 1, 2, 3**, respectively.)

Our method consists of two procedures: term weighting and correlating. In the procedure of term weighting, we focused on the Reuters newswire articles which are classified into the 25 target events, and calculate term weighting. We propose a method for term weighting method which is based on standard deviation. In the procedure of correlating, we select related the Reuters newswire for each CNN news story using features of TV news. We propose a correlating method which uses features of TV news. We assume that the words which feature the event appear in most of turns of talk in TV news. However, there are some turns of talk which are not related to the event: greetings, calling out and agreement in actual TV news. Therefore, we removed them and detected the event. There are a large number of pronouns in newswire and TV news, and in many cases, the pronouns are able to be substituted for keywords. Therefore, our system disambiguates it using a very simple overt pronoun resolver (5).

We conducted some correlating experiments using features of TV news. We used the Reuters newswire articles for term weighting and used the CNN news stories for test data. Format of the Reuters newswire articles which we used as training data is like a format of story of newspaper. The CNN news stories which we used as test data consists of conversations between an anchor and correspondents. In each experiment, we compared the result using our term weighting method with four methods: word density, $TF*IDF$, a method based on χ^2 and a method based on entropy. The results demonstrate the viability of the method.

2 Related work

Recently, information retrieval and information extraction are studied by many researchers. Especially, Topic Detection and Tracking (TDT) is studied by the TDT Pilot study [1]. In the TDT Pilot study, there are three major tasks: (1)segmentation, (2)topic correlation and (3)event tracking. Allan and Yang proposed event detection methods [2, 6]. Allan proposed an event detection method using a single pass clustering algorithm and a thresholding model that incorporates the properties of events as a major component. Yang proposed an event detection method using hierarchical cluster and temporal distribution patterns of document clusters. Walls proposed a method for topic detection in broadcast news [5]. He used incremental k-means algorithm and two types of clustering metrics: selection and thresholding. For selection metric, he used probabilistic similarity metric and for thresholding metric, he used combination of cosine distance and mean/sd-normed Tspot. Suzuki [4] proposed event detection using χ^2 value for the TDT corpus. However, most of the methods, described above, are based on word frequency.

In this paper, we propose a method for correlating using features of TV news. We used turns of talk between anchor and correspondents in TV news and overt pronoun resolver which are effective for correlating newswire articles with TV news story.

3 An overview of our system

Our system consists of two procedures, i.e. term weighting and correlation newswire articles with TV news story. Figure 2 illustrates an overview of our system.

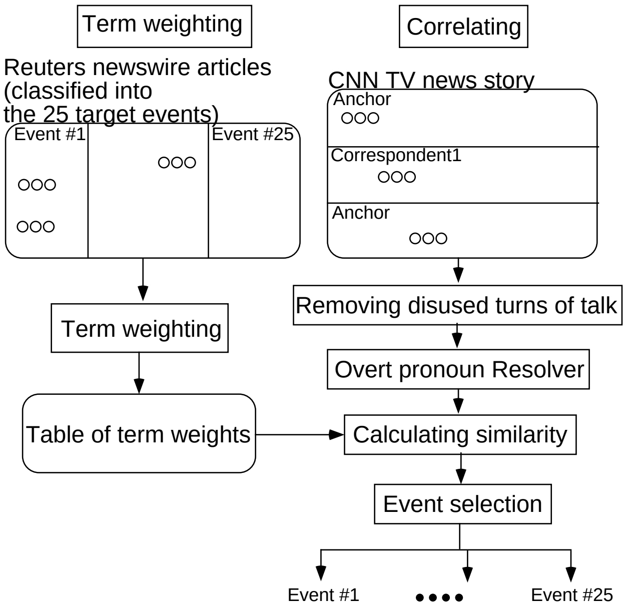

Figure 2: An overview of our system

In the procedure of term weighting, the system calculates weights for each word of the Reuters newswire articles.

In the procedure of correlation, firstly, the system removes some turn of talk which are not involved with the event. Next, overt pronominal anaphora are disambiguated using a very simple resolver. Then, the similarity between each TV news story and newswire articles is calculated, and finally, the most appropriate event is selected using a table of term weights.

4 Term weighting

For term weighting, we calculated weights of words whose parts of speech are noun and verb using Brill's tagger [3]. We assumed that the distribution of density of word$_i$ in an article is right half of normal distribution whose mean is 0. So, we assign weights based on standard deviation. The weights denote how strongly each word relates to each event. We call our weighting method sigma method. Sigma method is calculated by formula (1).

$$sigma(w_i, e_j) \quad = \quad \frac{x_{ij} - m_i}{\sqrt{\frac{\Sigma_{j=1}^{M}(x_{ij}-m_i)^2}{M}}} \qquad (1)$$

where,

$$m_i \quad = \quad \frac{\Sigma_{j=1}^{M} density(w_i, e_j)}{\Sigma_{j=1}^{M}\Sigma_{i=1}^{N} density(w_i, e_j)}$$

$$x_{ij} \quad = \quad \frac{density(w_i, e_j)}{\Sigma_{i=1}^{N} density(w_i, e_j)}$$

M : the number of target events (25 events)
N : the number of words
$density(w_i, e_j)$: density of word$i$ in event$_j$

A part of table of term weights is shown in Table 1.

In Table 1, event number 4 denotes "Cessna on White House", event number 5 denotes "Clinic Murders (Salvi)", event number 15 denotes "Kobe Japan quake" and event number 17 denotes "NYC Subway bombing". The corpus of the Reuters newswire has no articles whose event number is 12. Therefore, all of $sigma(w_i, e_12)$ indicate 0.

5 Keyword extraction from TV news

There are some expressions which are not related to the event in the turns of talk: greetings, calling out and agreement. In order to extract keywords from TV news, firstly, we removed greetings, calling out and agreements. The examples are shown in the following list,

- Good evening.(greetings)
- Bobbie. (calling out)
- Yes., Absolutely. (agreement)
- Charlie Coats, CNN. (Anchor-person's last turn of talk)

Table 1: Table of term weights

Event	Words		
number	police/nn	president/nn	earthquake/nn
1	0.000000	0.000000	0.000000
2	1.719569	0.000000	0.000000
3	0.000000	2.363102	0.000000
4	0.000000	3.152534	0.000000
5	2.290380	0.000000	0.000000
6	0.000000	0.000000	0.000000
7	0.000000	0.000000	0.000000
8	0.381994	1.843523	0.000000
9	1.373089	0.000000	0.000000
10	0.000000	0.000000	0.000000
11	0.000000	0.165320	0.000000
12	0.000000	0.000000	0.000000
13	0.000000	0.000000	0.000000
14	0.000000	0.000000	0.000000
15	0.544898	0.000000	4.999450
16	0.000000	0.000000	0.000000
17	3.538344	0.000000	0.000000
18	0.791813	0.326609	0.000000
19	0.000000	0.000000	0.000000
20	0.000000	2.368956	0.000000
21	0.000000	0.301345	0.000000
22	0.000000	0.283894	0.000000
23	0.000000	0.000000	0.000000
24	0.505478	0.375302	0.000000
25	1.019357	0.000000	0.000000

<SP>JIM CLANCY, International Correspondent</SP><P>
Yasser Arafat got down to work looking relaxed and energetic. Aides and PLO officials said if there had been any anxiety about coming here and facing huge problems, it had long vanished.</P>
<SP>1st AIDE</SP><P>
There's no anxiety in him and I think he is satisfied and he feels that this is a historical moment.</P>
<SP>2nd AIDE</SP><P>
He is very excited. He is very confident and he means business.</P>

Figure 3: A part of article of CNN

In order to remove these words, we used the above expression list. After removing these words, if there are no words in the turn of talk, we remove the turn of talk. For keyword extraction, we extract words whose parts of speech are noun, verb and pronoun. Next, we resolve overt pronouns. There are many overt pronouns in the CNN news stories. On average, there are 15.4 pronouns in a story of the CNN news. Let us take a look at the articles from the CNN (See Figure 3).

In Figure 3, him and he indicate Yasser Arafat. The word "Yasser Arafat" appeared in one turn of talk, but pronoun "he" and "him" which indicate Yasser Arafat appeared in other turns of talk. For overt pronoun resolution, we used noun-pronoun correspondence table which is based on the newswire articles. The examples are shown in following list,

- police : they (their them theirs)
- Simpson : he (his him his) she (her her hers)
- ex-wife : she (her her hers)
- knife : it (its it)

In order to resolve overt pronouns, we trace back the news story and find a noun which is the antecedent of the pronoun.

6 Similarity between TV news story and each event

The system computes similarity between the CNN news and each event using term weighting. The similarity between the TV news and each event is shown by formula (2).

$$sim(c_s, e_j) = \Sigma_{i=1}^{N_j} sigma(w_i, e_j) \times ntt(w_i, c_s) \qquad (2)$$

where, $ntt(w_i, c_s)$: the number of turns of talk in which word$_i$ appears in sth story of the CNN news story.

j : event number ($1 \leq j \leq 25$)
i : news story number
N_j : the number of words in the corpus whose event is j-th target event

For the results of similarity between the CNN news and each event, the system selects a suitable event for the CNN news using formula (3).

$$event_s = \arg\max_j sim(s, j) \qquad (3)$$

7 Experiments

We used the TDT pilot corpus in the experiments. All data are transcribed into texts. We used Brill's tagger [3] for tagging each word with part of speech. For term weighting, we used the 293 Reuters newswire articles whose topics are among the TDT's target events. We used the 1,089 CNN news stories whose topics are among the 25 TDT's target events as test data. The corpus of the Reuters newswire has no articles whose event number is 12. Therefore, we removed CNN news stories whose event number is 12.

7.1 Comparative experiments

This section describes four kinds of metrics, i.e. word density, $TF * IDF$, a method based on χ^2 and a method based on entropy.

(a)Word Density
Word density is calculated by formula (4).

$$density(w_i, e_j) = \begin{cases} \text{\# of } word_i \text{ in the Reuters} \\ \text{newswire articles} \\ \text{whose event number is } j \end{cases} \qquad (4)$$

w_i : word$_i$
e_j : event$_j$

(b)$TF * IDF$
$TF * IDF$ value is calculated by formula (5).

$$TF * IDF(w_i, e_j) \quad = \quad TF(w_i, e_j) \times IDF(w_i) \qquad (5)$$

where,

$$\begin{aligned} TF(w_i, e_j) &= \text{\# of word}_i \text{ in event}_j \\ IDF(w_i) &= \frac{\log(\text{\# of articles})}{\text{\# of articles which includes word}_i} \end{aligned}$$

(c)A method based on χ^2

χ^2 value is calculated by formula (6).

$$\chi^2(w_i, e_j) = \Sigma_{i=1}^{n} \frac{(x_{ij} - m_{ij})|x_{ij} - m_{ij}|}{m_{ij}} \quad (6)$$

where,

$$m_{ij} = \frac{\Sigma_{j=1}^{n} x_{ij}}{\Sigma_{i=1}^{m} \Sigma_{j=1}^{n} x_{ij}}$$

m : the number of words in the corpus
n : the number of stories for training
x_{ij} : the density of word$_i$ in story$_j$
m_{ij} : the expected density of word$_i$ in story$_j$

(d)A method based on entropy

Entropy value is calculated by formula (7).

$$enpy(w_i, e_j) = \frac{p(w_i, e_j)}{entropy(w_i)} \quad (7)$$

where,

$$entropy(w_i) = \frac{-\Sigma_{i=1}^{n} \Sigma_{j=1}^{m} p(i,j) \times \log_2 p(i,j)}{M}$$

7.2 Results and discussion

We conducted three correlating experiments using the Reuters newswire articles and the CNN news stories. In **experiment 2**, we used **feature 2** of our method. In **experiment 3**, we used **feature 2 and 3** of our method. In all experiments, we confirmed that efficiency of **feature 1** by comparing our term weighting method with four methods. Comparison between **experiment 1 and 2** shows efficiency of **feature 2**. Comparison between **experiment 2 and 3** shows efficiency of **feature 3**.

Experiment 1: use of word density in TV news

Table 2 shows the result using word density in TV news instead of $ntt(w_i, c_s)$ in formula (2). The result shows our term weighting method (sigma method) is better than the results using other four methods.

Experiment 2: use of the number of turns of talk in which keywords appear in TV news

Table 3 shows the result using the number of turns of talk in which keywords appear in TV news instead of word density in TV news.

The result shows our term weighting method (sigma method) is better than the results using other four methods.

Table 2: The result of **experiment 1**

Term weighting method	# of news stories (total 1,059)	correlation accuracy (%)
word density	928	87.6%
$TF * IDF$	1,000	94.4%
χ^2	1,001	94.5%
entropy	1,005	94.9%
sigma(our method)	1,008	95.2%

Table 3: The result of **experiment 2**

Term weighting method	# of news stories (total 1,059)	correlation accuracy (%)
word density	932	88.0%
$TF * IDF$	1,001	94.5%
χ^2	998	94.2%
entropy	1,007	95.1%
sigma(our method)	1,012	95.6%

Experiment 3: use of overt pronoun resolution

Table 4 shows the result using overt pronoun resolution.

Table 4: The result of **experiment 3**

Term weighting method	# of news stories (total 1,059)	correlation accuracy (%)
word density	962	90.8%
$TF * IDF$	1,012	95.6%
χ^2	1,015	95.8%
entropy	1,021	96.4%
sigma(our method)	1,028	97.1%

The result shows our term weighting method (sigma method) is better than the results using other four methods. It also shows that each result is better than the result in **experiment 2**.

Figure 4 illustrates contribution of each **feature** of our system to correlation accuracy. In Figure 4, **feature 1** means the result using sigma method in **experiment 1**. **Feature 1+2** means the result using sigma method in **experiment 2**. **Feature 1+2+3** denotes the result using sigma method in **experiment 3**. The accuracy of the result using **feature 1+2** (Table 3) is slightly higher than the result using **feature 1** (Table 2). The result using overt pronoun resolution in test data (Table 4) was better than the result without pronoun resolution (Table 3). However in some stories, the system substituted incorrect nouns for pronouns. For better performance, we have to extend our pronoun resolver to handle these pronouns.

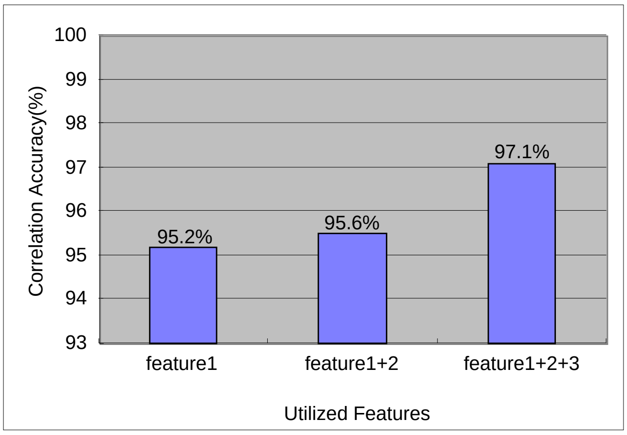

Figure 4: Contribution of each **feature** to correlation accuracy

To summarize the results:

(i) Our term weighting methods (sigma method) is more efficient than other four term weighting methods in these three experiments.

(ii) The accuracy of the method using the number of the turns of talk in which keywords appear is slightly better than the method using word density in test data. For better performance, we have to use other features of TV news, for example, questions from anchor and correspondents' answers.

(iii) The result using overt pronoun resolver is better than the result without it. However our very simple resolver substituted wrong nouns for some pronouns.

8 Conclusion

We have reported on an empirical method for correlating newswire articles with TV news story using features of TV news. The experiments using the Reuters newswire articles and the CNN news articles demonstrate the viability of the method. Future work includes improvement of the overt pronoun resolver. In addition, we plan to classify the training data automatically.

9 Acknowledgements

The authors would like to thank the reviewers for their valuable comments. This work was supported by the Grants from the Government subsidy for aiding scientific researchers (No.11780257) of the Ministry of Education of Japan.

References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Y. Topic detection and tracking pilot study final report. In *the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[3] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the 3nd conference on applied natural language processing*, pages 152–155, 1992.

[4] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi. Segmentation and event detection of news stories using term weighting. In *Proc. of PACLING99*, 1999.

[5] F. Walls, H. Jin, S. Sista, and R. Schwartz. Topic detection in broadcast news. In *DARPA Broadcast News Workshop*, 1999.

[6] Y. Yang, T. Pierce, and J. Carbonell. A study on retrospective and on-line event detection. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.