Correlating TV news stories with a newswire article

Yoshimi Suzuki, 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@esb, sekiguti@alps1.esi}.yamanashi.ac.jp

Abstract

In this paper, we propose a method for correlating TV news stories with a target newswire article. The significant points of our method are using the two techniques; overt pronoun resolver and two step correlating method. We compared our method with other methods. The results of experiments show that our method is effective for correlating TV news stories with a target newswire article.

Introduction

Today, lots of news are broadcasted on TV and radio. There are many news articles on some web site. When an user checks the web site and is interested in a news articles, he will want to watch news related to it. We are trying to the system which correlates TV news stories with a newswire article. However, it is difficult to correlate TV or radio news stories with a newswire article. The difficulties are showed as follows.

1. Pronouns are often used instead of nouns which are keywords for the newswire articles or TV news stories.
2. Each newswire article is short and there are few keywords.
3. The event of some newswire articles are not suitable for any TV news.
4. An anchor or announcer does not mention the background of the news in some news stories.
5. There are some words which are not related to the story in conversation between anchor and correspondents.

In this paper, we propose a method which correlates TV news stories with a target newswire article.

Figure 1 shows a concept of our newswire-TV news correlating system. When an user selects one of the article of newswire, the system picks up some TV news related to the selected article.

For better results, we used an overt pronoun resolver and two step correlation. We compared our method with other methods by some experiments. The result using dictated CNN TV news demonstrates the effectiveness of our method.

Related work

Our study is based on topic detection. There are lots of papers which mentioned topic detection. Topic detection and tracking is studied by the TDT project (Allan et al., 1998a). Allan proposed an event detection method using a single pass clustering algorithm and a thresholding model (Allan

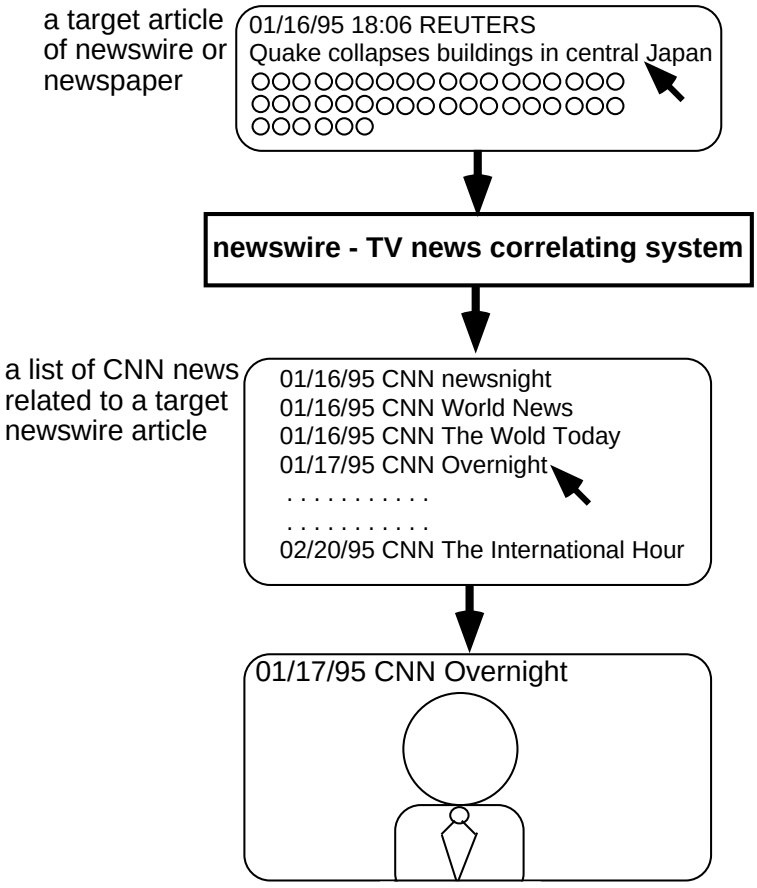

Figure 1: A concept of our newswire-TV news correlating system

et al., 1998b). Yang proposed an event detection method using hierarchical cluster and temporal distribution patterns of document clusters (Yang et al., 1998). Walls proposed a topic detection method for broadcast news (Walls et al., 1999). 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 proposed a topic detection and segmentation using χ^2 method (Suzuki et al., 1999). However there are few application-oriented studies of topic detection. We selected the newswire - TV news correlating system as the goal and used overt pronoun resolver and selecting newswire articles related to a target article.

An overview of our newswire - TV news correlating system

Figure 2 shows an overview of our method.

Our TV news correlating system consists of the following steps:

1. Calculating term weighting using classified TV news stories for the 25 events and other events (Section)
2. Calculating similarity between the selected newswire article and TV news stories (Section)

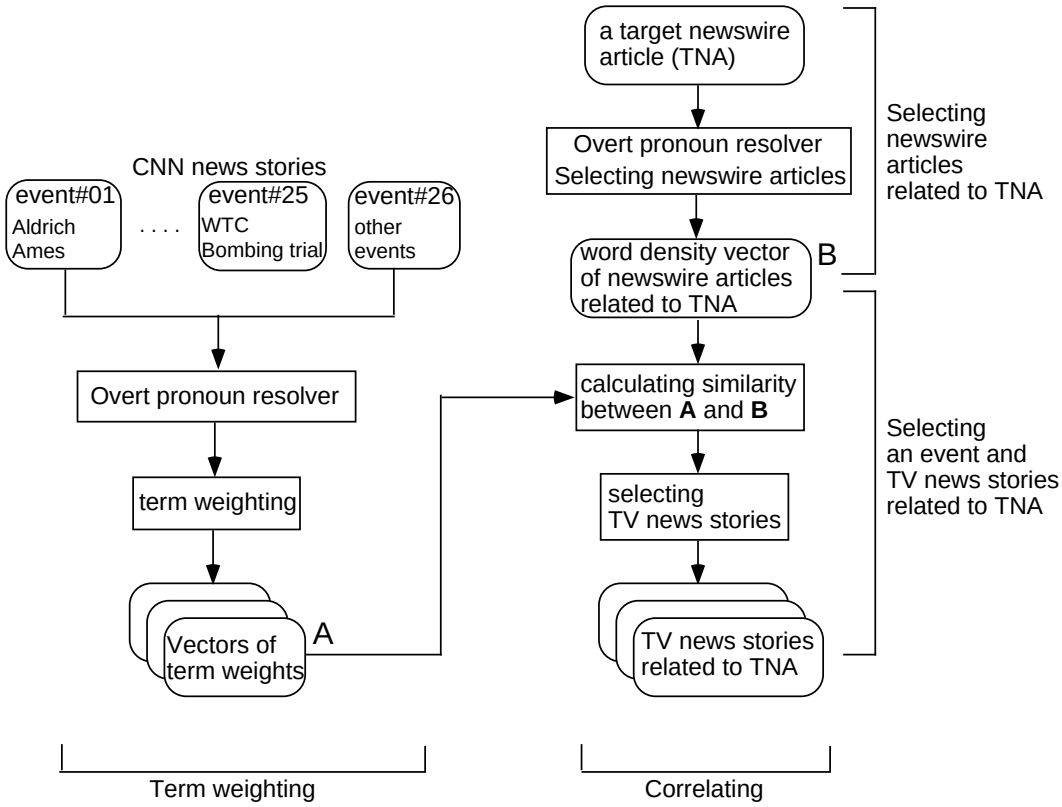

Figure 2: The proposed method

<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 story of the CNN (underlines illustrate overt pronouns)

3. Selecting TV news related to selected newswire article (Section)

Term weighting

We used CNN news stories whose events are the 25 events and other events for term weighting. Firstly we select nouns, verbs and pronouns using Brill's tagger (Brill, 1992) from CNN news stories.

Then, we resolve overt pronouns. There are many overt pronouns in the CNN news stories. On average, there are 24.9 pronouns in a story of the CNN news. Most of pronouns are used instead of keywords. For term weighting based on word frequency, overt pronoun resolver is necessary. Let us take a look at the stories from the CNN (See Figure 3).

In Figure 3, "Yasser Arafat" is one of the keywords, but it appeared at once. However him and he

appear six times and all of them indicate "Yasser Arafat". If all of him and he are able to resolve correctly, The word "Yasser Arafat" appeared seven times. We used a term weighting method based on word frequency. Therefore, overt pronoun resolver must be an effective pre-processing. For overt pronoun resolution, we used noun-pronoun correspondence table. The examples are shown in Table 1.

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

Table 1: A part of noun-pronoun correspondence table

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

Then, we calculate term weighting of nouns and verbs. Our term weighting method is based on standard deviation. We assume that the distribution of density of word$_i$ in event#j is right half of normal distribution whose mean is 0. Our term weighting method is calculated by formula (1).

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

where,

$x_{ij} = \frac{density(w_i, e_j)}{\sum_{i=1}^{N} density(w_i, e_j)}$
$density(w_i, e_j)$: the density of word$_i$ in event#j
M: the number of events (26=the 25 events + other events)
N: the total number of different words

Selecting TV news stories with newswire articles related to the target newswire article (TNA)

Selecting newswire articles related to TNA

Firstly, we resolve overt pronouns of newswire articles. On average, there are 8.6 pronouns in each article of the Reuters newswire. In order to resolve overt pronouns, we trace back the article and find a noun which is the antecedent of the pronoun. After resolving overt pronouns, we extract nouns and verbs. Each newswire article is short and there are few keywords. Therefore, we select newswire articles related to TNA by using word density. And the word density vectors are made for selecting an event of TNA.

Selecting an event and TV news stories related to TNA

Then the system calculates similarity between CNN TV news stories and a group of newswire articles (B in Figure 2) using Formula (2).

$$sim(a_s, e_j) = \sum_{i=1}^{N} sigma(w_i, e_j) \times \frac{density(w_i, a_s)}{\sum_{i=1}^{N} density(w_i, a_s)} \quad (2)$$

where, $density(w_i, a_s)$: the density of word$_i$ in $s-th$ article of the Reuters newswire.

After calculating similarity between the vectors of term weighting by CNN news stories (A in Figure 2) and the word density vector of newswire article related to TNA (B in Figure 2), the system selects a suitable event for the target newswire article, and selects CNN news stories whose event is the selected event using formula (3). The threshold for selection is selected by the prior experiments.

$$event_s = \begin{cases} \arg\max_j \frac{sim(a_s,e_j)}{\sqrt{\frac{\sum_{j=1}^{M} sim(a_s,e_j)}{M}}} & \text{if } \max_j \frac{sim(a_s,e_j)}{\sqrt{\frac{\sum_{j=1}^{M} sim(a_s,e_j)}{M}}} > \text{threshold} \\ \text{other events} & \text{otherwise} \end{cases} \quad (3)$$

where, $sim(a_s, e_j)$: similarity between the $s-th$ Reuters newswire article and event #j.

Finally, the system selects TV news whose event is event$_s$.

Experiments

Data

We used the TDT Pilot corpus in the experiments. All data were transcribed into texts. We used Brill's tagger (Brill, 1992) for tagging each word with part of speech. We used 7898 CNN news stories (the 25 event:1089, other events: 6809) and used 390 articles of the Reuters newswire (the 25 event:290, other events:100).

Comparative experiments

For comparative experiments, we used 3 kinds of methods for correlating TV news with a target newswire article (method A, B and C), and our method (method D). Figure 4 shows 3 comparative methods and our method.

Method A: TNA → keywords → TV news stories

Several keywords are selected from the title of each article of the Reuters newswire. The system decides a suitable event and selects TV news stories which include the keywords.

Method B: TNA → keywords → newswire articles related to TNA→ TV news stories related to TNA

Firstly the system selects some newswire articles which include selected keywords, and decides suitable events and selects TV news which are similar to the selected newswire articles.

Method C: TNA → TV news stories related to TNA

The system decides the suitable event using word density of the target article as term weighting method.

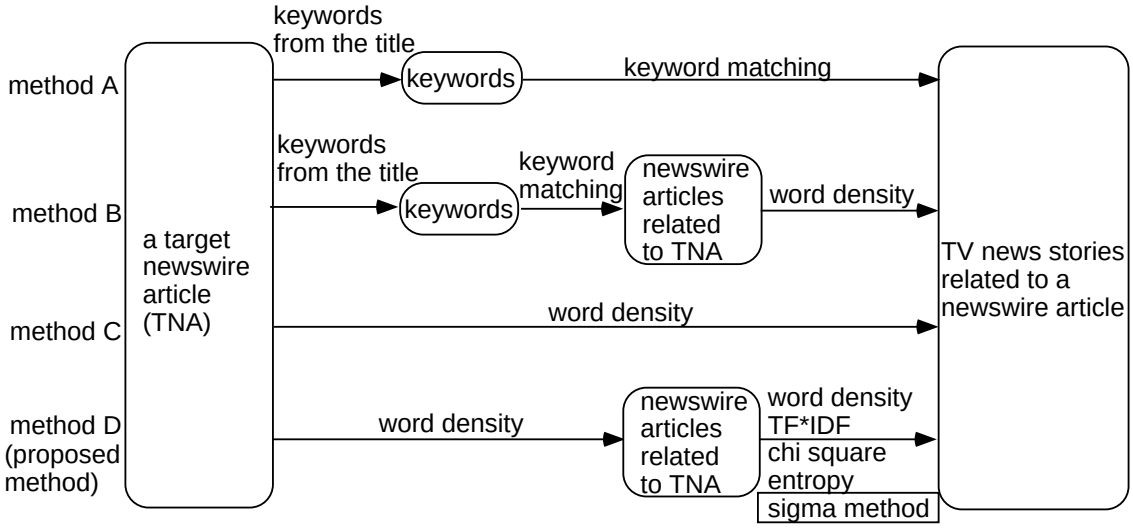

Figure 4: Our method and 3 comparative methods

Method D: TNA → newswire articles related to TNA→ TV news stories related to TNA(proposed method)

Firstly the system selects some newswire articles which are similar to a target newswire article, and selects TV news which are similar to the selected newswire articles.

In method D, we employed 5 kinds of methods for term weighting.

1. word density
2. $TF * IDF$
3. a method based on χ^2 method (Suzuki et al., 1999)
4. a method based on entropy
5. sigma method (proposed term weighting method)

(1)Word Density

Word density is calculated by formula (4).

$$density(w_i, e_j) = \frac{\#\text{of word}_i \text{ in the CNN news stories whose event number is } j}{\Sigma_{i=1}^{N} \#\text{of word}_i \text{ in the CNN news stories whose event number is } j} \quad (4)$$

where, N : the total number of different words in the corpus

(2)$TF * IDF$

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

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

where,

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

(3)A method based on χ^2

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

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

where,

$$m_{ij} = \frac{\Sigma_{j=1}^{M} x_{ij}}{\Sigma_{i=1}^{N} \Sigma_{j=1}^{M} x_{ij}} \times \Sigma_{i=1}^{N} x_{ij}$$

M : the number of events
x_{ij} : the density of word$_i$ in event$_j$
m_{ij} : the expected density of word$_i$ in event$_j$

(4)A method based on entropy

Entropy value is calculated by formula (7).

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

where,

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

Table 2 illustrates the accuracy of correlating experiments. Method D using sigma method is our method.

The result using method D and sigma method is better than the result using other methods.

Discussion

We examined which procedure contributes accuracy rate and which procedure leave room for improvement.

Term weighting

We examined 5 term weighting methods in method D (Table 2). The results using entropy and sigma method are better than the results using word density, $TF * IDF$ and χ^2 method. Because

Method	Method of term weighting	accuracy rate
A	keywords	78.5% (306/390)
B	word density	75.1% (293/390)
C	word density	77.2% (301/390)
D	word density	77.7% (303/390)
	$TF * IDF$	80.8% (315/390)
	χ^2 method	80.0% (312/390)
	entropy	82.8% (323/390)
	sigma method	83.3% (325/390)

Table 2: Result of the correlating experiment

distributions of keywords are used by entropy and sigma method effectively. Table 4 illustrates the results using overt pronoun resolver and without overt pronoun resolver. Using overt pronoun resolver, the result is better than that without it. Overt pronoun resolver is effective, however, our overt pronoun resolver is very simple.

Selecting newswire articles related to TNA

Table 3 illustrates the result of selecting newswire articles related to TNA.

F-measure is calculated by the formula (8).

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

without overt pronoun resolver			with overt pronoun resolver		
Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
58.8% (5637/9590)	61.2% (5637/9213)	0.60	66.5% (6381/9590)	69.9% (6381/9128)	0.68

Table 3: The result of selecting newswire articles related to TNA

We obtained better precision and recall by using overt pronoun resolver. Therefore overt pronoun resolver is effective for selecting newswire articles related to TNA. However the part of selecting newswire articles related to TNA has room for improvement.

Selecting an event and TV news stories related to TNA

We examined correlating experiment using correct group of newswire articles. Table 4 illustrates the result. The accuracy when the system used overt pronoun resolver are 100%. For better results, we have to select newswire articles related to the target article correctly.

Conclusion

In this paper, we proposed a method for correlating TV news stories with a newswire article. Our method employed overt pronoun resolver and two step selection method for using term weighting method effectively. Future work includes clustering TV news automatically. In addition, we plan to apply our method to actual spoken TV news.

	without overt pronoun resolver	with overt pronoun resolver
Term weighting	accuracy rate	accuracy rate
word density	96.6% (280/290)	100% (290/290)
$TF * IDF$	97.2% (282/290)	100% (290/290)
χ^2 method	97.2% (282/290)	100% (290/290)
entropy	97.6% (283/290)	100% (290/290)
sigma	97.6% (283/290)	100% (290/290)

Table 4: Result of correlating experiment using correct group of newswire articles related to the target article

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

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Yang Y. (1998)a. Topic detection and tracking pilot study final report. In *the DARPA Broadcast News Transcription and Understanding Workshop*, page http://www.itl.nist.gov/iaui/894.01/proc/darpa98/index.htm.

J. Allan, R. Papka, and V. Lavrenko. (1998)b. On-line new event detection and tracking. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45.

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

Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi. (1999). Segmentation and event detection of news stories using term weighting. In *Proc. of PACLING99*, pages 149–154.

Frederick Walls, Hubert Jin, Sreenivasa Sista, and Richard Schwartz. (1999). Topic detection in broadcast news. In *DARPA Broadcast News Workshop*, page http://www.itl.nist.gov/div894/894.01/proc/darpa99/.

Y. Yang, T. Pierce, and J. Carbonell. (1998). A study on retrospective and on-line event detection. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36.