The Fourth Symposium on Natural Language Processing 2000

CORRELATING TV NEWS STORIES WITH A NEWSWIRE ARTICLE USING OVERT PRONOUN RESOLVER

Yoshimi Suzuki, Fumiyo Fukumoto and Yoshihiro Sekiguchi

Department of Computer Science and Media Engineering Yamanashi University 4-3-11 Takeda, Kofu 400-8511 Japan

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 correlating steps. We compared our method with other methods by some experiments. The result using dictated the CNN TV news demonstrates the effectiveness of our method.

Key words: overt pronoun resolver, correlating TV news

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. Overt 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.

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

The goal of this work is that to make a system which displays the CNN VT news related to a target newswire article (TNA). 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 the CNN TV news demonstrates the effectiveness of our method.

1. 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 [1]. Allan proposed an event detection method using a single pass clustering algorithm and a thresholding model [2]. Yang proposed an event detection method using hierarchical cluster and temporal distribution patterns of document clusters [3]. Walls proposed a topic detection method for broadcast news [4]. 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 [5]. However there are few topic detection studies using overt pronoun resolver even though lots of overt pronouns appear in TV news. 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.

2. AN OVERVIEW OF OUR NEWSWIRE - TV NEWS CORRELATING SYSTEM

Figure 1 shows an overview of our method.

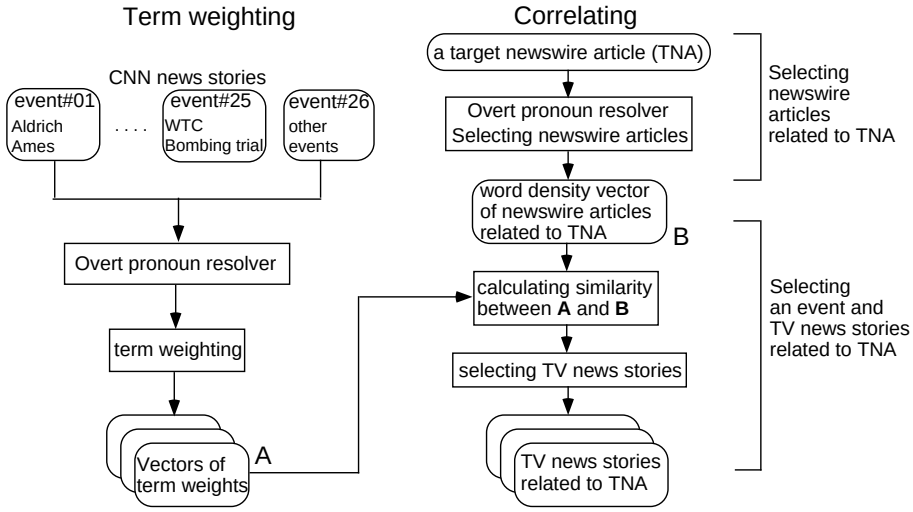

Figure 1. The proposed method

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 3.2)
2. Calculating similarity between the selected newswire article and TV news stories (Section 4.1)
3. Selecting TV news related to selected newswire article (Section 4.2)

3. TERM WEIGHTING

3.1. Overt pronoun resolver

We used the 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 [6] from the 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 2).

In Figure 2, "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.

<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 2. A part of story of the CNN (underlines illustrate overt pronouns)

Table 1. A part of noun-pronoun correspondence table

nouns	pronouns
Arafat	He (his him himself)
Aides	they (their them theirs)
PLO	they (their them theirs)
officials	they (their them theirs)
anxiety	it (its it)
problems	they (their them theirs)
anxiety	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.

3.2. Calculating term weights

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

4. SELECTING TV NEWS STORIES WITH NEWSWIRE ARTICLES RELATED TO THE TARGET NEWSWIRE ARTICLE (TNA)

4.1. 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.

4.2. Selecting an event and TV news stories related to TNA

Then the system calculates similarity between the CNN TV news stories and a group of newswire articles (B in Figure 1) 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 the CNN news stories (A in Figure 1) and the word density vector of newswire article related to TNA (B in Figure 1), the system selects a suitable event for the target newswire article, and selects the CNN news stories whose event is the selected event using formula (3). The threshold for selection is decided 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$_s$.

5. EXPERIMENTS

5.1. Data

We used the TDT Pilot corpus in the experiments. All data were transcribed into texts. We used Brill's tagger [6] for tagging each word with part of speech. We used 7,898 CNN news stories (the 25 event:1,089, other events: 6,809) and used 390 articles of the Reuters newswire (the 25 event:290, other events:100).

5.2. 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 3 shows 3 comparative methods and our method.

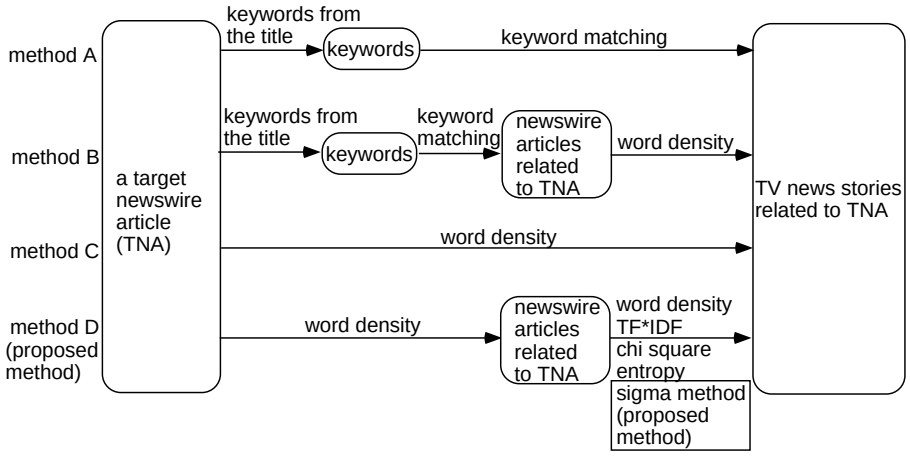

Figure 3. Our method and 3 comparative methods

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 [5]
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).

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

where,

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

(3)A method based on χ^2

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

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

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).

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

where,

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

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

Table 2. Result of the correlating experiment

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)

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

6. DISCUSSION

6.1. 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 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, although our overt pronoun resolver is very simple.

6.2. 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).

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

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

without overt pronoun resolver			with overt pronoun resolver		
Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
58.8% (5,637/9,590)	61.2% (5,637/9,213)	0.60	66.5% (6,381/9,590)	69.9% (6,381/9,128)	0.68

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.

6.3. 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.

Table 4. Correlating accuracy using correct group of newswire articles related to TNA

Term weighting	not using overt pronoun resolver	using overt pronoun resolver
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)

6.4. Overt pronoun resolver

We used overt pronoun resolver. Table 5 shows the result of our overt pronoun resolution. The success rate of our system was 69.1%. And the errors were classified into 3. Error(1) is

that the system selects a noun incorrectly which is situated between the overt pronoun and the antecedent. Error(2) is the error for example "It's hot". Error(3) is the error for example "It was interesting to see ....".

Table 5. The result of our overt pronoun resolution

	overt pronoun resolution result	
Correct	69.1%	(168/243)
Error(1)	24.3%	(59/243)
Error(2)	2.1%	(5/243)
Error(3)	4.5%	(11/243)

7. 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.

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

Allan, J., J. Carbonell, G. Doddington, J.P. Yamron, and Y. Yang. 1998. Topic Detection and Tracking Pilot Study Final Report. the DARPA Broadcast News Transcription and Understanding Workshop. http://www.itl.nist.gov/iaui/894.01/proc/darpa98/index.htm.

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

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

Walls, F., H. Jin, S. Sista, and R. Schwartz. 1999. Topic Detection in Broadcast News. DARPA Broadcast News Workshop. http://www.itl.nist.gov/div894/894.01/proc/darpa99/.

Suzuki, Y., F. Fukumoto, and Y. Sekiguchi. 1999. Segmentation and Event Detection of News Stories using Term Weighting. Proc. of PACLING99. 149–154.

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