Story Link Detection Based on Event Words

Letian Wang and Fang Li

Department of Computer Science & Engineering
Shanghai Jiao Tong University
Shanghai, China
{koh,fli}@sjtu.edu.cn

Abstract. In this paper, we propose an event words based method for story link detection. Different from previous studies, we use time and places to label nouns and named entities, the featured nouns/named entities are called event words. In our approach, a document is represented by five dimensions including nouns/named entities, time featured nouns/named entities, place featured nouns/named entities, time&place featured nouns/named entities and publication date. Experimental results show that, our method gain a significant improvement over baseline and event words plays a vital role in this improvement. Especially when using publication date, we can reach the highest 92% on precision.

Keywords: story link detection, event words, multidimensional model, nouns/named entities, featured nouns/named entities.

1 Introduction

Story link detection, which was first defined in the Topic Detection and Tracking (TDT) [1,2,12,14,16] competition program, is the task of determining whether two stories, such as news articles and/or radio broadcasts, are about the same event, or linked. Story link detection is important for many applications. For example, there are three reports whose titles are:

- Midterm election polls open in United States
- US presidential vote is underway
- Voting in parliamentary election starts in Japan

The content of three news stories above are very similar, because they are all about the election, they have many common words in the text such as "election", "vote", "candidate" and so on. But actually they are different because they are not the same event. The first one is related to the election in U.S.A. in 2006 while the second one is about the election in U.S.A in 2008 and the last one refers to the election in Japan in 2007. The task of story link detection is to find out if the two stories are about the same event even though they may have the same content.

According to TDT, two stories are linked if the events in the stories happened at some specific time and place. In this paper, we give a more explicit definition:

A. Gelbukh (Ed.): CICLing 2011, Part II, LNCS 6609, pp. 202–211, 2011.

Definition 1. *Two stories are linked if they contain the same event words.*

Where event words (EW) is defined as:

Definition 2. *Nonus/named entities with time or place labels.*

There are three types of labels for event words including time, place and both (time&place). We use five dimensions to represent a story, including nouns/named entities, time featured nouns/named entities, place featired nouns/named entities, time&place featured nouns/named entities and the publication date, where all the featured nouns/named entities are event words. Cosine similarity, resemblance function and date similarity function are used to calculate similarity of each dimension. A combined story similarity function is used to calculate the similarity of two stories. Experimental results show that our approach gain a significant improvement over pure text similarity method and each dimension has its own contribution.

The following section contains the previous studies on story link detection. Section 3 shows our multidimensional model for representing stories and describes how to calculate similarities between two stories. Experimental results and discussions are described in Section 4. We give our conclusions in the last section.

2 Related Work

There were two kinds of methods for story link detection. One is based on vector space model and the other is based on language model.

Based on vector space model, Chen and Farahat et al. used incremental tf-idf instead of traditional static tf-idf in vector space model, and used several similarity method including Cosine, Hellinger, Tanimoto and Clarity to find out linked stories [4,5,7]. They also proposed a source-pair specific method for story link detection to avoid linking two stories from the same media because of the customary words. Shah et al. used named entities and traditional vector space model to represent a document [14]. They also used a graph based method to extend named entities for each document. Zhang et al. used an event model, which is actually a multi-vector model, to represent a story including time, number, person, location, organization, abstract and content description [16]. Brown et al. proposed a method to ignore common event to discriminate among similar events [3]. Chen et al. considered several important issues for monolingual and multilingual link detection [6]. They used nouns, verbs, adjectives and compound nouns to represent news stories, and used story expansion, topic segmentation and a translation model to help the detection process. Ferret et al. proposed a method to combines word repetition and the lexical cohesion stated by a collocation network to compensate for the respective weaknesses of segmentation and link detection [8].

For the second method, Nallapati et al. used a semantic language model for story link detection, they used named entities and part of speech tag to classify features into different categories, defined a semantic class for each document

and used likelihood as features with perceptron learning algorithm to distinguish stories [13]. Yu et al. defined a semantic domain as a collection of semantic related terms for Chinese corpus [9]. With the semantic domain language model, two stories will be linked if they have similar semantic domains, the distance calculation is based on Kullback-Leibler divergency. Lavrenko et al. proposed a relevance model for story link detection and also used Kullback-Leibler divergency to calculate distance between two stories [11].

Our method is similar with traditional vector space model but different from it. We use event words to improve the performance of story link detection. Besides using Cosine similarity like previous studies, we use a resemblance function and define a date similarity function to help link detection.

3 Multidimensional Model for Story Representation

3.1 Event Words (EW)

Time and place information is important for story link detection. Our work is to maximize the using of the time and place information. In our study, time and place information is used to establish event words in order to distinguish same words in different documents, such as the "earthquake" at Sichuan and the "earthquake" at Yunnan. All the labels are divided into three types:

1. **time:** Nouns/named entities with only time label.
2. **place:** Nouns/named entities with only place label.
3. **time&place:** Nouns/named entities with both time and place labels.

For example, "earthquake@2008" is featured with time, "earthquake@Sichuan" is featured with place and "earthquake@Sichuan@2008" is featured with time-&place.

Nouns/Named Entities Featured with Time ($\mathbf{NN_{time}}$). We use only date format as time information, because words like "today" or "yesterday" is hard to distinguish between different stories. Any time in documents can be represented as a triple in our model:

$$< year, month, day >$$

Regular expressions are used to extract time from documents. Only four types of combinations of *year*, *month* and *day* are used to label nouns/named entities, they are *year*, *year.month*, *year.month.day* and *month.day*. For example, "earthquake@2008.05.12" is a time featured event words.

Nouns/Named Entities Featured with Place ($\mathbf{NN_{place}}$). A place is a structure rather than a word in our model. Like time information, it can also be presented as a triple:

$$< city, region, country >$$

We have a place database which contains places information with triple format above. With a word presenting a place, our approach extends it to a triple. The *city* and *region* may be null if the word originally represent a *region* or *country*. After extension, a noun/named entity will be labeled with at most three places. For example, if a sentence contains "earthquake" and "Wenchuan", we will have three featured nouns, "earthquake@Wenchuan", "earthquake@Sichuan" and "earthquake@China", where Wenchuan is located in Sichuan, China.

Nouns/Named Entities Featured with Time&Place ($\mathbf{NN_{time\&place}}$). We also use time&place labels in our method. "earthquake@Sichuan@2008" is more representable than "earthquake@Sichuan" and "earthquake@2008".

How to Produce Event Words. Two words are more related if they are close to each other in a document. So, in our research, nouns/named entities are labeled with time or places from the same sentence. For the label of time&place, all the possible combination of time and places within the same sentence will be used to label the nouns/named entities.

3.2 Modeling

Five dimensions are used to represent each document, which are shown in Table 1. Publication date is used in our method because it is an important feature to distinguish two news stories. For a publication date, we just care about *year*, *month* and *day*.

Table 1. Five Dimensions in Multidimensional Model for Story Link Detection

Abbreviation	Description
NN	Nouns/Named Entities
$\mathbf{NN_{time}}$(event words)	Nouns/Named Entities Featured with Time
$\mathbf{NN_{place}}$(event words)	Nouns/Named Entities Featured with Place
$\mathbf{NN_{time\&place}}$(event words)	Nouns/Named Entities Featured with Time&Place
PD	Publication date

Different modeling method and similarity calculation approaches are used for different dimensions. Vector space model and Cosine similarity based on tf-idf is used for NN dimension. A resemblance function is used for event words and a date similarity function is used for publication date.

tf-idf. We use tf-idf in the NN dimension. The tf-idf is often used in information retrieval and text mining. This weight is a statistical measure used to evaluate

how important a word is to a document in a collection or corpus. Equation 1 shows the calculation of tf-idf,

$$(tfidf)_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \times \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, the denominator is the sum of number of occurrences of all terms in document d_j, $|D|$ is total number of documents in the corpus, and $|\{d : t_i \in d\}|$ is number of documents where the term t_i appears (that is $n_{i,j} \neq 0$).

3.3 Similarity Calculation

In our approach, we use similarity function to compare two news stories. We first calculate similarities for each dimension respectively. Since five dimensions are modeled differently, three similarity methods are used in our approach. Cosine similarity is used for the NN dimension, resemblance function is used for the event words dimensions and date similarity function is used for PD dimension.

Cosine Similarity. Cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them. Given two document vectors d_1 and d_2, the similarity can be represented as:

$$sim_{cosine}(d_1, d_2) = \frac{\sum_{i=1}^{n} w_{i,1} \times w_{i,2}}{\sqrt{\sum_{i=1}^{n} w_{i,1}^2}\sqrt{\sum_{i=1}^{n} w_{i,2}^2}} \tag{2}$$

where $w_{i,1}$ and $w_{i,2}$ are weights (tf-idf values here) of term t_i in document d_1 and d_2, and n is the total number of terms in the corpus. The similarity for the NN dimension between two news stories is:

$$sim_{NN}(i, j) = sim_{cosine}(d_{i_{NN}}, d_{j_{NN}}) \tag{3}$$

where $d_{i_{NN}}$ is the vector of the NN dimension of document d_i.

Resemblance Function. We choose the resemblance as our similarity metric for the dimensions of event words. The reason we use resemblance here is that featured nouns/named entities need more accurate comparison. The resemblance r of two documents d_1 and d_2 is defined as follows:

$$r(d_1, d_2) = \frac{|d_1 \cap d_2|}{|d_1 \cup d_2|} \tag{4}$$

where $|d_1 \cap d_2|$ is the number of terms both occur in d_1 and d_2, and $|d_1 \cup d_2|$ is the number of all the distinct terms in d_1 and d_2. We use resemblance for the

dimensions with featured nouns and named entities, so the similarities of these three dimensions can be represented as:

$$sim_{NN_{time}}(d_i, d_j) = r(d_{1_{NN_{time}}}, d_{2_{NN_{time}}}) \tag{5}$$

$$sim_{NN_{place}}(d_i, d_j) = r(d_{1_{NN_{place}}}, d_{2_{NN_{place}}}) \tag{6}$$

$$sim_{NN_{time\&place}}(d_i, d_j) = r(d_{1_{NN_{time\&place}}}, d_{2_{NN_{time\&place}}}) \tag{7}$$

Date Similarity Function. We first calculate time difference $time_{diff}$ between two date and represent in days. Then, the date similarity function sim_{date} can be represented as:

$$sim_{date}(t_1, t_2) = \frac{1}{time_{diff}(t_1, t_2) + 1} \tag{8}$$

Where t_1 and t_2 is two different dates. We use date similarity for the PD dimension, so the similarity of the PD dimensions can be represented as:

$$sim_{PD}(d_1, d_2) = sim_{date}(d_{1_{PD}}, d_{2_{PD}}) \tag{9}$$

Similarity Function for News Stories. In order to calculate the similarity between two news stories, similarities for each dimension are combined together with the following equations:

$$sim_{i,j} = \sqrt[\lambda]{\frac{\alpha \cdot sim_{NN} + sim_{EW}}{\alpha + \beta + \gamma + \delta} \times sim_{PD}^{\theta}} \tag{10}$$

where

$$sim_{EW} = \beta \cdot sim_{NN_{time}} + \gamma \cdot sim_{NN_{place}} + \delta \cdot sim_{NN_{time\&place}} \tag{11}$$

We have radical sign here because we need to avoid the similarity falling into a too small interval. In our experiment, the parameters of the Equation (11) are respectively set to: $\alpha = 1$, $\beta = 2$, $\gamma = 4$, $\delta = 4$, $\theta = 2$ and $\lambda = 8$.

For some reason, some news stories may not have publication date. So we need a method to calculate similarities if there is no publication date in the corpus. The similarity function without sim_{PD} is:

$$sim_{i,j_{withoutPD}} = \sqrt[\lambda]{\frac{\alpha \cdot sim_{NN} + sim_{EW}}{\alpha + \beta + \gamma + \delta}} \tag{12}$$

In our experiment, the parameters of the Equation (14) are respectively set to: $\alpha = 1$, $\beta = 2$, $\gamma = 4$, $\delta = 4$ and $\lambda = 4$.

We do some experiments over several group of parameters, and the above ones achieve the best result.

4 Experiment and Discussion

4.1 Data Set and Experimental Procedures

We use a Chinese corpus from SINA[1] which contains 1591 news stories on 148 topics. There are about 10 news stories for each topic. We assume that news stories from the same topic are linked with each other, because the topics are collected by people manually and each topic is refer to an event.

To get the result, we first do the word segmentation work and extract named entities from all the documents including recognizing time and places information. Then, we extract event words from all the processed news stories and establish multidimensional model for each story. We use similarity value to verify if the two stories are linked.

4.2 Evaluation Methods

F-score. The traditional F-measure or balanced F-score is the harmonic mean of precision and recall:

$$F = \frac{2pr}{p+r} \tag{13}$$

where p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned.

Detection Cost. Detection cost is a evaluation method in TDT project. It can be represented as:

$$C_{det} = C_{miss} \cdot P_{miss} \cdot P_{target} + C_{fa} \cdot P_{fa} \cdot P_{non-target} \tag{14}$$

where $P_{miss} = \frac{number\ of\ missed\ detection}{number\ of\ targets}$, $P_{fa} = \frac{number\ of\ false\ alarms}{number\ of\ non\text{-}targets}$, C_{miss} and C_{fa} are the costs of a missed detection and a false alarm respectively, and are pre-specified, P_{target} is the a priori probability of finding a target and $P_{non-target} = 1 - P_{target}$.

4.3 Experimental Results

Table 2 shows the results of our experiment. The first two rows are baseline systems using traditional vector space model with Cosine similarity while the first one is based on terms and the other is based on nouns/named entities. The last two methods are event words based methods (EWM) while the first one is the multidimensional model without PD dimension and the last one makes the final result.

Vector space model can get higher recall but lower precision, because it can not distinguish stories with similar contents but different events. With nouns/named

[1] http://www.sina.com.cn

Table 2. Experiment Results

	$p(\%)$	$r(\%)$	$F(\%)$	$(C_{Det})_{Norm}$
VSM	50.66	67.35	58.54	0.5282
$\mathbf{VSM_{NN}}$	58.60	60.10	59.35	0.4202
$\mathbf{EMW_{withoutPD}}$	72.71	63.23	67.64	0.3795
$\mathbf{EMW_{withPD}}$	**92.37**	**67.61**	**78.08**	**0.3266**

entities, we can get higher precision but lower recall. With event words, we can achieve a higher precision and an acceptable recall, because event words describe stories more accurate and detailed. With publication date, a significant high precision can be achieved because it reduce the candidates to be linked.

4.4 Discussion

In our method, we choose five features each as a dimension to represent a document. Instead of these five features, there are lots of other information can be used to represent a document. We discuss here to show why we choose these five features instead of others, such as title, time words and place words. We use nouns/named entities as baseline here, and each time add one additional feature to represent documents. Table 3 shows the results. where double line arrow in

Table 3. Results with different dimensions

	$p(\%)$	$r(\%)$	$F(\%)$	$(C_{Det})_{Norm}$
NN	58.60	60.10	59.35	0.4202
NN + Title($\Downarrow$)	57.22	50.93	53.89	0.5097
NN + Time($\Downarrow$)	43.36	57.48	49.43	0.4628
NN + Place	76.09($\uparrow$)	47.56($\downarrow$)	58.53	0.5318($\downarrow$)
$\mathbf{NN + NN_{time}}$	65.02($\uparrow$)	57.95	61.28($\uparrow$)	0.4361
$\mathbf{NN + NN_{place}}$($\Uparrow$)	67.60	64.37	65.94	0.3717
$\mathbf{NN + NN_{time\&place}}$	62.33($\uparrow$)	59.38	60.82($\uparrow$)	0.4241

the table means the feature has a significant impact on the result while single line arrow means it has a limited impact.

Generally, we may think that title is a good feature to distinguish news stories, since a title is an accurate summary of a story. But from the result above, we find out that all the performance decreases with titles taken into account. It is because different stories may have similar titles and titles are too short to distinguish when they have the same words.

We do not use time and places alone in our method. The time dimension make no improvements just implicate the performance. The place dimension gain a good performance in precision, but the recall and detection cost are unacceptable

for story link detection. Actually, time and place alone may bring noises as well as title. Many events happen in the same place or at the same time.

From the result above, we can see that all the event words help to improve the performance especially for place labels. The reason why event words make such improvement is that they can describe a story more accurate and contains more information.

5 Conclusion

We propose a event words based method for story link detection in this paper. The main contribution of our work is:

1. Event words are used to distinguish stories with similar contents.
2. A multidimensional model based on event words is used in our approach.
3. A combined similarity method is used in our model.

Three similarity methods are used in our approach, Cosine similarity for nouns/named entities, resemblance for event words and date similarity function for publication date. Our method gain a significant improvements over baseline systems and the results prove that:

1. Nouns/named entities are more helpful to story link detection than other content words.
2. Event words can improve the performance of story link detection.
3. Publication date is also a useful information for story link detection.

Acknowledgement

The research is supported by the National Science Foundation of China under Grant No.60873134, Threads and topics detection for news events.

References

1. Allan, J.: Topic detection and tracking: event-based information organization. Kluwer Academic Publishers, Norwell (2002)
2. Allan, J., Lavrenko, V., Swan, R.: Explorations within topic tracking and detection, pp. 197–224 (2002)
3. Brown, R.D.: Dynamic stopwording for story link detection. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 190–193. Morgan Kaufmann Publishers Inc., San Francisco (2002)
4. Chen, F., Farahat, A., Brants, T.: Story link detection and new event detection are asymmetric. In: NAACL 2003: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 13–15. Association for Computational Linguistics, Morristown (2003), doi:10.3115/1073483.1073488
5. Chen, F., Farahat, A., Brants, T.: Multiple similarity measures and source-pair information in story link detection. In: In HLT-NAACL 2004, pp. 2–7 (2004)

6. Chen, Y.-J., Chen, H.-H.: Nlp and ir approaches to monolingual and multilingual link detection. In: Proceedings of the 19th International Conference on Computational Linguistics, pp. 1–7. Association for Computational Linguistics, Morristown (2002)
7. Farahat, A., Chen, F., Brants, T.: Optimizing story link detection is not equivalent to optimizing new event detection. In: ACL 2003: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pp. 232–239. Association for Computational Linguistics, Morristown (2003)
8. Ferret, O.: Using collocations for topic segmentation and link detection. In: Proceedings of the 19th International Conference on Computational Linguistics, pp. 1–7. Association for Computational Linguistics, Morristown (2002)
9. Hong, Y., Zhang, Y., Fan, J., Liu, T., Li, S.: Chinese topic link detection based on semantic domain language model. Journal of Software, 2265–2275 (2008)
10. Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., Thomas, S.: Relevance models for topic detection and tracking. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 115–121. Morgan Kaufmann Publishers Inc., San Francisco (2002)
11. Luo, W., Liu, Q., Chen, X.: Development and analysis of technology of topic detection and tracking. In: Sun, M.S. (ed.) Proc. of the JSCL 2003, Beijing, China, pp. 560–566 (2003)
12. Nallapati, R.: Semantic language models for topic detection and tracking. In: NAACL 2003: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, pp. 1–6. Association for Computational Linguistics, Morristown (2003)
13. Schultz, J.M., Liberman, M.Y.: Towards a "universal dictionary" for multilanguage information retrieval applications, pp. 225–241 (2002)
14. Shah, C., Croft, W.B., Jensen, D.: Representing documents with named entities for story link detection (sld). In: CIKM 2006: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 868–869. ACM, New York (2006)
15. Wayen, C.L.: Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Athens, Greece, pp. 1487–1494 (2000)
16. Zhang, X., Wang, T., Chen, H.: Story link detection based on event model with uneven svm. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 436–441. Springer, Heidelberg (2008)