NLP and IR Approaches to Monolingual and Multilingual Link Detection

Ying-Ju Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, 106
yjchen@nlg2.csie.ntu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, 106
hh_chen@csie.ntu.edu.tw

Abstract

This paper considers several important issues for monolingual and multilingual link detection. The experimental results show that nouns, verbs, adjectives and compound nouns are useful to represent news stories; story expansion is helpful; topic segmentation has a little effect; and a translation model is needed to capture the differences between languages.

Introduction

In the digital era, how to assist users to deal with data explosion problem becomes emergent. News stories on the Internet contain a large amount of real-time and new information. Several attempts were made to extract information from news stories, e.g., multi-lingual multi-document summarization (Chen and Huang, 1999; Chen and Lin, 2000), topic detection and tracking (abbreviated as TDT hereafter, http://www.nist.gov/TDT), and so on. Of these, TDT, which is a long-term project, proposed many diverse applications, e.g., story segmentation (Greiff *et al.*, 2000), topic tracking (Levow *et al.*, 2000; Leek *et al.*, 2002), topic detection (Chen and Ku, 2002) and link detection (Allan *et al.*, 2000).

This paper will focus on the link detection application. The TDT link detection aims to determine whether two stories discuss the same topic. Each story could discuss one or more than one topic, and the sizes of two stories compared may not be so comparable. For example, one story may contain 100 sentences and the other one may contain only 5 sentences. In addition, the stories may be represented in different languages. These are the main challenges of this task. In this paper, we will discuss and contribute on several issues:

1. How to represent a news story?
2. How to measure the similarity of news stories?
3. How to expand a story vector using historic information?
4. How to identify the subtopics embedded in a news story?
5. How to deal with news stories in different languages?

The multilingual issue was first introduced in 1999 (TDT-3), and the source languages are mainly English and Mandarin. Dictionary-based translation strategy is applied broadly. In addition, some strategies were proposed to improve the translation accuracy. Leek *et al.*, (2002) proposed probabilistic term translation and co-occurrence statistics strategies. The algorithm of co-occurrence statistics tended to favour those translations consistent with the rest of the document. Hui *et al.*, (2001) proposed an enhanced translation approach for improving the translation by using a parallel corpus as an additional resource. Levow *et al.*, (2000) proposed a corpus-based translation preference. English translation candidates were sorted in an order that reflected the dominant usage in the collection. Most of these methods need extra resources, e.g., a parallel corpus. In this paper, we will try to resolve multilingual issues with the lack of extra information.

Topic segmentation is a technique extensively utilized in information retrieval and automatic document summarization (Hearst *et al.*, 1993; Nakao, 2001). The effects were shown to be valid. This paper will introduce topic

Table 1. Performance of Link Detection under Different Feature Selection Strategies (I)

	Similarity Threshold								
	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.11	0.12
All	1.6234	1.274	1.0275	0.8440	0.7245	0.6463	0.5911	0.5528	0.5268
N	0.7088	0.5547	0.4553	0.4012	0.3815	0.3743	0.3775	0.3834	0.3883
N&V	0.8152	0.6028	0.4899	0.4254	0.3922	0.3803	0.3780	0.3870	0.4002
N&J	0.6126	0.4671	0.3918	0.3624	0.3485	**0.3437**	0.3481	0.3628	0.3780
N&V&J	0.6955	0.5121	0.4200	0.3720	0.3498	0.3474	0.3480	0.3617	0.3795

segmentation in link detection. Several experiments will be conducted to investigate its effects.

1 Environment

LDC provides corpora to support the different applications of TDT (Fiscus *et al*., 2002). The corpora used in this paper are the TDT2 corpus and the augmented version of TDT3 corpus. We used the TDT2 corpus as training data, and evaluated the performance with the augmented version of TDT3 corpus. Both corpora are text and transcribed speech news from a number of sources in English and in Mandarin. The TDT2 corpus spans January 1, 1998 to June 30, 1998. There are 200 topics for English, and 20 topics for Mandarin. The TDT3 corpus spans October 1, 1998 to December 31, 1998. There are 120 topics for both English and Mandarin. In the augmented version of TDT3 corpus, additional news data is added. These data spans from July 1, 1998 to December 31, 1998.

There are 34,908 story pairs (Fiscus *et al.*, 2002) for link detection in both monolingual and multilingual tasks. Of these, the numbers of target and non-target pairs are 4,908 and 30,000, respectively. In the monolingual task, Mandarin news stories are translated into English ones through a machine translation system. In the multilingual task, Mandarin news stories are represented in the original Mandarin characters. In both tasks, all the audio news stories are transcribed through an automatic speech recognition (ASR) system.

We adopt the evaluation methodology defined in TDT to evaluate our system performance. The cost function for the task defined by TDT is shown as follows. The better the link detection is, the lower the normalized detection cost is. In the next sections, all experimental results are evaluated by this metric.

$$C_{Det}=C_{Miss}\times P_{Miss}\times P_{target}+C_{FA}\times P_{FA}\times P_{non\text{-}target},$$

where C_{Miss} and C_{FA} are the costs of Miss and False Alarm errors, and P_{Miss} and P_{FA} are the probabilities of a Miss and a False Alarm, and P_{target} and $P_{non\text{-}target}$ are *a priori* probabilities of a story pair chosen at random discuss the same topic and discuss different topics. The cost of detection is normalized as follows:

$$(C_{Det})_{Norm}=C_{Det}/min(C_{Miss}\times P_{target},C_{FA}\times P_{non\text{-}target})$$

2 Basic Link Detection System

2.1 Basic Architecture

The basic algorithm is shown as follows. Each story in a given pair is represented as a vector with *tf***idf* weights, where *tf* and *idf* denote term frequency and inverse document frequency as traditional IR defines. Then, the cosine function is used to measure the similarity of two stories. Finally, a predefined threshold, $TH_{decision}$, is employed to decide whether two stories discuss the same topic or not. That is, two stories are on the same topic if their similarity is larger than the predefined threshold. The *idf* values and the thresholds are trained from TDT2 corpus. Each English story is tagged using "Apple Pie Parser" (version 5.9). In addition, English words are stemmed by Porter's algorithm, and function words are removed directly.

2.2 Story Representation

The noun terms denote interesting entities such as people names, location names, and organization names, and so on. The verb terms denote the specific events. In general, noun and verb terms are important features to identify the topic the story discusses. We conducted several experiments to investigate the performance of different story representations. Table 1 shows the performance of different story representation schemes under different similarity thresholds. The row denotes which lexical items are used. "All" means any kind of lexical items is

Table 2. Performance of Link Detection under Different Feature Selection Schemes (II)

	Similarity Threshold						
	0.04	0.05	0.06	0.07	0.08	0.09	0.1
N&CNs	0.3825	0.3564	0.3612	0.3754	0.4026	0.4377	0.4700
N&V&CNs	0.4090	0.3572	0.3520	0.3658	0.3917	0.4279	0.4617
N&J&CNs	0.3372	0.3361	0.3353	0.3568	0.3845	0.4163	0.4471
N&V&J&CNs	0.3451	0.3398	**0.3283**	0.3446	0.3751	0.4055	0.4360

considered. N, V and J denote nouns, verbs, and adjectives, respectively.

The experimental results show that the best performance is 0.3437 when only noun and adjective terms are used to represent stories, and the similarity threshold is 0.09. Examining why nouns and adjectives terms carry more information than verbs, we found that there are important adjectives like "Asian", "financial", *etc*., and some important people names are mis-tagged as adjectives. And the matched verb terms, such as "keep", "lower", *etc*., carry less information and the similarity would be overestimated.

In the next experiments, we investigate the effects of compound nouns (abbreviated as CNs) in the story representation. The results are shown in Table 2. All performances are improved when using CNs. The best one is 0.3283 when nouns, verbs, adjectives and CNs are adopted and the similarity threshold is 0.06. The performance is better than the result (i.e., 0.3437) in Table 1. We found that the threshold for the best performance decreased in the CNs experiments. This is because matching CNs in two different news stories is more difficult than matching single terms, but the effect is very strong when matching is successful, such as "Red Cross", "Security Council", *etc*.

2.3 Story Expansion

The length of stories may be diverse. With the method proposed in Section 2.1, there may be very few features remaining for short stories. And different reporters would use different words to describe the same event. In such situations, the similarity of two stories may be too small to tell if they belong to the same topic. To deal with the problems, we try to introduce a story expansion technique in the basic algorithm. The method we employed is quite different from that proposed by Allan (2000), which regarded local context analysis (LCA) as a smoothing technique. Each story is treated as a "query" and is expanded using LCA.

Our method is described below. When the similarity of two stories is higher than a predefined threshold $TH_{expansion}$, which is always larger than or equal to $TH_{decision}$, the two stories are related to some topic in more confidence. Thus, their relationship is kept in a database and will be used for story expansion later. For example, if the similarity of a story pair (A, B) is very high, we will expand the vector of A with B when a new pair (A, C) is considered. Table 3 shows our experiments on TDT2 data. We conducted different lexical combinations and different weighting schemes for the expanded terms.

Story expansion with the non-relevant terms would reduce the performance of a link detection system. That is, it may introduce some noise into the story and make the detection more difficult. We assigned the expanded terms two different weights. One is using the original weights, and the other one is using half of the original weights, which is denoted as "half" in Table 3.

The results show that story expansion

Table 3. Performance of Link Detection with Story Expansion Strategy

$TH_{decision}$	0.06					
$TH_{expansion}$	0.06	0.07	0.08	0.1	0.11	0.13
N&J&CNs	0.3713	0.3580	0.3392	0.3260	0.3230	0.3278
N&V&J&CNs	0.3342	0.3363	0.3155	0.3061	0.3057	0.3073
N&J&CNs (half)	0.2691	**0.2638**	0.2654	0.2785	None	None
N&V&J&CNs (half)	0.2797	0.2751	0.2826	0.3259	None	None

Table 4. Performances of Topic Segmentation in Link Detection

	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Strategy (I)	None	None	None	0.4338	0.3891	0.3766	0.3857	0.4063
Strategy (II)	0.3581	0.3490	0.3983	0.4629	0.5226	None	None	None
Strategy (III)	None	0.3309	0.3280	0.3282	0.3288	None	None	None

outperforms the basic method, and assigning expanded terms half weights would be better. The best performance when applying story expansion achieves 0.2638. The total miss rate was decreased to third fourths of the original amount. Sum up, story expansion is a good strategy to improve the link detection task.

3 Topic Segmentation

There is no presumption that each story discusses only one topic. Thus, we try to segment stories into small passages according to the discussing topics and compute passage similarity instead of document similarity. The basic idea is: the significance of some useful terms may be reduced in a long story because similarity measure on a large number of terms will decrease the effects of those important terms. Computing similarities between small passages could let some terms be more significant.

The first method we adopted is text tiling approach (Hearst, 1993). TextTiling subdivides text into multi-paragraph units that represent passages or subtopics. The approach uses quantitative lexical analyses to segment the documents. After through TextTiling algorithm, a file will be broken into tiles. Suppose one story is broken into three tiles and the other one is broken into four tiles. There are twelve (i.e., 3*4) similarities of these two stories. We conducted three different strategies to investigate the effect of topic segmentation. Strategy (I) is computing the similarity using the most similar passage pair. Strategy (II) is computing the similarity using passage-averaged similarity. Strategy (III) is computing the similarity using a two-state decision (Chen, 2002). But the result is not so good as we expected. Up to now, the best performance is almost the same as the original method without text tiling.

Next, we applied another topic segmentation algorithm developed by Utiyama *et al*. (2001). The results show that this segmentation algorithm is better than TextTiling. But the improvement is still not obvious. Table 4 shows the experimental results for topic segmentation. For strategy (III), the first threshold is 0.06, which is also the best threshold for the basic method, and the second threshold varies from 0.04 to 0.07 for segmentation. After applying topic segmentation, topic words would be centred on small passages. The amount of news stories discussing more than one topic is few in the test data and the overall performance depends on the segmentation algorithm. We make an index file similar to the original TDT index file. In this file, at least one story of each pair discusses multi-topics. We conducted different strategies to investigate the effect of topic segmentation. The experimental results demonstrate that topic segmentation is useful in this task (Chen, 2002).

4 Multilingual Link Detection Algorithm

The multilingual link detection should tell if two stories in different languages are discussing the same topic. In this paper, the stories are in English and in Chinese. Comparing to English stories, there is no apparent word boundary in Chinese stories. We have to segment the Chinese sentences into meaningful lexical units. We employed our own Chinese segmentation and tagging system to pre-process Chinese sentences. Similar to monolingual link detection, each story in a pair is represented as a vector and the cosine similarity is used to decide if two stories discuss the same topic.

In multilingual link detection, we have to deal with terms used in different languages. Consider the following three cases. E and C denote an English story and a Chinese story, respectively. (E, E) denotes an English pair; (C, C) denotes a Chinese pair; and (C, E) or (E, C) denotes a multilingual pair.

(a) (E, E): no translation is required.

(b) (C, E) or (E, C): C is translated to E'. The new E' could be an English vector or the vector is mixed in two languages if the original

Chinese terms are included in the new English vector.

(c) (C, C): No translation is required; or both stories are translated into English and use English vectors; or these new English terms are added into the original Chinese vectors.

The reason that we included the original Chinese terms in the new English vector is that we could not find the corresponding English translation candidates for some Chinese words. Including the Chinese terms could not lose information.

We employed a simple approach to translate a Chinese story into an English one. A Chinese-English dictionary is consulted. There are 374,595 Chinese-English pairs in the dictionary. For each English term, there are 2.49 Chinese translations. For each Chinese term, there are 1.87 English translations. In this dictionary, English translations are less ambiguous. Therefore, we translated Chinese stories into English ones. If a Chinese word corresponds to more than one English word, these English words are all selected. That is, we did not disambiguate the meaning of a Chinese word. To avoid the noise introduced by many English translations, each translation term is assigned a lower weight. The weight is determined as follows. We divided the weight of a Chinese term by the total number translation equivalents.

$$w(d, t_e) = w(d, t_c) / N,$$

where $w(d, t_c)$ is the weight of a Chinese term in story d, $w(d, t_e)$ is the weight of its English translation in story d, and N is the number of English translation candidates for the Chinese term.

Table 5 shows the performances of multilingual link detection. We conducted three experiments using different story representation schemes for Chinese stories. “E” denotes Chinese stories are translated into English ones. “C” denotes Chinese stories are compared directly without translation, but Chinese stories are translated into English ones in multilingual pairs. “EC” denotes Chinese stories are represented in Chinese terms and their corresponding English translation candidates. The threshold for English story pairs is set to 0.12. The threshold for the other pairs varies from 0.1 to 0.5. The results reveal that “E” is better than “C” and “EC”.

Table 5. Performance of Multilingual Link Detection with Different Translation Schemes

	Similarity Threshold				
	0.1	0.2	0.3	0.4	0.5
E	0.9925	0.6760	**0.6359**	0.6558	0.6864
C	1.0971	0.7204	0.6546	0.6701	0.6969
EC	1.1525	0.7712	0.7146	0.7410	0.7694

Comparing stories in translated English terms could bring some advantages. Some Chinese terms which denote the same concept but in different forms could be matched through their English translations, for example, "屠殺" and "殺害" (kill), as well as "行為" and "行徑" (behaviour).

The effect of English translations for Chinese stories is similar to the effect of thesaurus. We employed the CILIN (Mei *et al.*, 1982) in multilingual link detection. We use the small category information and synonyms to expand the features we selected to represent a news story. The experimental results are shown in Table 6.

Table 6. Performance of Multilingual Link Detection with Different Thesaurus Expansion Schemes

	Similarity Threshold				
	0.1	0.2	0.3	0.4	0.5
Small Category	1.6576	0.9196	0.6656	0.6500	0.6832
Synonyms	0.9486	**0.6260**	0.6342	0.6734	0.7059

We found that the performances of “E” translation and synonyms expansion schemes are very close. In our consideration, a good bilingual dictionary can be regarded as a thesaurus.

The results of multilingual link detection are apparently worse than those of monolingual link detection. When the threshold is 0.2, the best performance is 0.6260 and the miss rate is 0.4547. The value of miss rate is very high. To improve the performance, we have to reduce the miss rate. We found the similarity of two stories in different languages is very low in comparison with the similarity of two stories in the same language. It is unfair to set the same threshold for different languages, thus we introduced a two-threshold method to resolve this problem. The performance of the two-threshold method for synonyms expansion (denotes as "Syn") is shown in Table 7. "Chinese" means the

threshold for Chinese pairs and "Multi" means the threshold for multilingual pairs.

Table 7. Performance of Multilingual Link Detection with a Two-threshold Method

	Similarity Threshold					
Chinese	0.2					
Multi	0.01	0.02	0.03	0.04	0.05	0.06
Syn	1.2929	0.7804	0.5818	0.5166	**0.5033**	0.5124

The result reveals that there is a great improvement when applying the two-threshold method. The threshold for Chinese story pairs is 0.2, the threshold for English story pairs is 0.12, and threshold for multilingual story pairs is 0.05. The similarity distributions for story pairs in different languages vary. As monolingual link detection, we did experiments about the combinations of different lexical terms. The results of these different combinations are shown in Table 8. It shows that the representation of the best performance in the multilingual task is different from that in the monolingual task. CNs bring positive influence. But using nouns, verbs and adjectives to represent a story is better than using nouns and adjectives only in multilingual link detection. Words in Chinese are seldom tagged as adjective. They are tagged as verbs in Chinese, but are tagged as adjectives in English ("安全" vs. “safe”).

We also adopted story expansion mentioned in Section 2.3 before computing the similarity. Note that only stories in the same language are used to expand each other. In Table 9, “One” denotes the weights of expanded terms are the same as the original ones, and “Half” denotes the weights of the expanded terms are only half of the original ones. The results reveal that expanded terms with half weights are better than with original ones. Giving expanded terms half weights could reduce the effect of noise. Nouns, verbs, adjectives and compound nouns are used to represent stories in Table 9, and the thresholds are set as the best ones in the previous experiments. The expansion threshold for Chinese pairs varies from 0.2 to 0.3.

Table 9. Performances of Multilingual Link Detection with All the Best Strategies

$TH_{expansion}$	0.2	0.25	0.3
One	0.3852	0.3873	0.3916
Half	0.3721	**0.3718**	0.3734

5 Results of the Evaluation on TDT3 corpus

We applied the best strategies and the trained thresholds in above experiments for both monolingual and multilingual link detection tasks to TDT3 corpus. The results of our methods and of the other sites participating the TDT 2001 evaluation are shown in Table 10. In this evaluation, both published and unpublished topics are considered.

For monolingual task, nouns, adjectives and CNs are used to represent story vectors. And the thresholds for decision and expansion are 0.06 and 0.07, respectively. For multilingual task, nouns, verbs, adjectives and CNs are used to represent story vectors. The thresholds for English pairs are set the same as those in the monolingual task, and for Chinese pairs, they are 0.2 and 0.25, respectively. The decision threshold for multilingual pairs is 0.05.

Table 10. Link Detection Evaluation Results

	CMU	CUHK	NTU	UIowa
Monolingual	0.2734	None	0.2963	0.3375
Multilingual	None	0.4143	0.3269	None

Table 8. Performances of Multilingual Link Detection under Different Feature Selection Scheme

	Similarity Threshold			
Chinese	0.2			
Multi	0.03	0.04	0.05	0.06
N	0.4707	0.4421	0.4319	0.4389
N&J	0.4600	0.4162	0.4082	0.4126
N&V	0.5162	0.4459	0.4233	0.4299
N&V&J	0.5116	0.4248	0.4042	0.4093
N&CNs	0.4685	0.4399	0.4297	0.4366
N&J&CNs	0.4570	0.4193	0.4106	0.4199
N&V&CNs	0.5010	0.4386	0.4162	0.4219
N&V&J&CNs	0.4886	0.4152	**0.3931**	0.3978

In the multilingual task, our result (NTU) is better than The Chinese University of Hong Kong (CUHK). And the multilingual result is close to the monolingual result. This is a significant improvement.

Conclusion and Future Work

Several issues for link detection are considered in this paper. For both monolingual and multilingual tasks, the best features to represent stories are nouns, verbs, adjectives, and compound nouns. The story expansion using historic information is helpful. Story pairs in different languages have different similarity distributions. Using thresholds to model the differences is shown to be usable.

Topic segmentation is an interesting issue. We expected it would bring some benefits, but the experiments for TDT testing environment showed that this factor did not gain as much as we expected. Few multi-topic story pairs and segmentation accuracy induced this result. We made an index file containing multi-topic story pairs and did experiments to investigate. The experimental results support our thought.

We examined the similarities of story pairs and tried to figure out why the miss rate was not reduced. There are 919 pairs of 4,908 ones are mistaken. The mean similarity of miss pairs is much smaller than the decision threshold. That means there are no similar words between two stories even they are discussing the same topic. None or few match words result that the similarity does not exceed the threshold. That is the problem that we have to overcome.

We also find that the people names may be spelled in different ways in different news agencies. For example, the name of a balloonist is spelled as “Faucett” in VOA news stories, but is spelled as “Fossett” in the other news sources. And for machine translated news stories, the people names would not be translated into their corresponding English names. Therefore, we could not find the same people name in two stories. In substance, people names are important features to discriminate from topics. This is another challenge issue to overcome.

References

Allan J., Lavrenko V., Frey D., and Khandelwal V. (2000) *UMass at TDT 2000*. In Proceedings of Topic Detection and Tracking Workshop.

Chen H.H. and Huang S.J. (1999). *A Summarization System for Chinese News from Multiple Sources.* In Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages, Taiwan, pp. 1-7.

Chen H.H. and Lin C.J. (2000) *A Multilingual News Summarizer*. In Proceedings of 18th International Conference on Computational Linguistics, University of Saarlandes, pp. 159-165.

Chen H.H. and Ku L.W (2002) *An NLP & IR Approach to Topic Detection*. In "Topic Detection and Tracking: Event-based Information Organization", Kluwer Academic Publishers, pp. 243-261.

Chen Y.J (2002) *Monolingual and Multilingual Link Detection.* Master Thesis. Department of Computer Science and Information Engineering, National Taiwan University, 2002.

Fiscus J.G., Doddington G.R. (2002) *Topic Detection and Tracking Evaluation Overview.* In "Topic Detection and Tracking: Event-based Information Organization", Kluwer Academic Publishers, pp. 17-32.

Greiff W., Morgan A., Fish R., Richards M., Kundu A. (2000) *MITRE TDT-2000 Segmentation System.* In Proceedings of TDT2000 Workshop.

Hearst M.A. and Plaunt C. (1993) *Subtopic Structuring for Full-Length Document Access.* In Proceedings of the 16th Annual International ACM SIGIR Conference.

Hui K., Lam W., and Meng H.M. (2001) *Discovery of Unknown Events From Multi-lingual News.* In Proceedings of the International Conference on Computer Processing of Oriental Languages.

Leek T., Schuartz R., Sista S. (2002) *Probabilistic Approaches To Topic Detection and Tracking.* In "Topic Detection and Tracking: Event-based Information Organization", Kluwer Academic Publishers, pp. 67-84.

Levow G.A. and Oard D.W. (2000) *Translingual Topic Detection: Applying Lessons from the MEI Project.* In the Proceedings of Topic Detection and Tracking Workshop (TDT-2000).

Mei, J. *et al*. (1982) tong2yi4ci2ci2lin2 (CILIN), Shanghai Dictionary Press.

Nakao Y. (2000) An Algorithm for One-page Summarization of a Long Text Based on Thematic Hierarchy Detection. In Proceeding of ACL 2000, pp. 302-309.

Utiyama M. and Isahara H. (2001) *A statistical Model for Domain-Independent Text Segmentation.* ACL/EACL-2001, pp. 491-498.