Discovery of Unknown Events From Multi-lingual News

Kin Hui and Wai Lam and Helen Meng

Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong
Shatin, Hong Kong, PRC
Email: khui@se.cuhk.edu.hk, wlam@se.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk

Abstract

We have proposed a new approach to detect topically-related events from multi-lingual news sources. In particular, we are interested in Chinese and English on-line newswire stories. Three categories of named entities terms, namely, people names, geographical location names, and organization names, together with the story content terms constitute the basis for story representation. The named entities are extracted by transformation-based linguistic taggers. One for Chinese and one for English. For Chinese stories, we tackle the unknown word problems by means of a hybrid solution of rule-based and statistical-based methods. To conduct event detection in multi-lingual settings, we conduct gross translation on Chinese story representation into English. One gross translation approach is the basis translation method using only a bilingual dictionary. The second approach makes use of a parallel corpus as an additional resource. Unsupervised learning technique is employed to discover new events.

Keywords: event detection, topic discovery, multi-lingual text processing

1 INTRODUCTION

With the rapid growth of the Internet and wide availability of electronic documents, there are more and more people accessing the newswire stories by electronic means from diverse sources. However, it is impossible to read the entire news collection. It is also infeasible for novice users to generate relevant specific queries if they do not have any knowledge on the event in advance. Therefore, there is a need to develop an intelligent agent to discover related issues from streams of data automatically. This is the objective of the *event detection* problem.

As the Internet is an international media, the content of the stories can be expressed in multi-languages. Therefore, dealing with multi-lingual story content and determining whether the stories in different languages are reporting the same event is a fundamental challenge for many information and knowledge management problems. We are currently interested in handling Chinese and English news. It is one of the major investigations in the Topic Detection and Tracking (TDT) project organized by the Defense Advanced Research Projects Agency (DARPA) [1].

An event is defined as a particular activity or incident happening in certain place at a certain time as well as any follow-up progress. Our task is to explore techniques for discovering new events from a continuous stream of Chinese and English news articles. The detection system aims at identifying whether the incoming stories, which may be in Chinese or in English, belong to a new event or an existing event detected in previous stories chronologically. The detection technique can be applied to various areas, such as the construction of structured guidelines for story navigation of the news collection.

One feature of our event detection approach is to make use of named entity component, automatically extracted from the story, to form the story representation. Named entities are extracted by linguistic taggers and a collection of transformation rules. This kind of representation is able to capture terms conveying useful context for event detection purpose. To handle detection in multi-lingual settings, we conduct gross translation using easily available resources like a bilingual dictionary and a parallel corpus to translate Chinese story terms so that unsupervised learning can be conducted regardless of language limitation. We have developed a modified agglomerative centroid clustering method to conduct the unsupervised learning. It comprises of four main components: Story Preprocessing, Named Entities Extraction, Gross Translation and Unsupervised Learning Module. Story, temporary clusters and final clusters are maintained in the clustering process.

2 NAMED ENTITY EXTRACTION

2.1 Transformation-based Error-driven Learning

There is a named entity extraction for each language. Both taggers are developed based on Transformation-based Error-driven Learning (TEL) approach smiliar to Brill [2]. This approach is a corpus-based natural language processing which combines the merits of rule-based and data-driven techniques in an elegant manner. It has been successfully applied to a number of natural language problems, including part of speech tagging, prepositional phrase attachment disambiguation, and syntactic parsing.

To conduct learning, we prepare a training corpus for each language. The training corpus consists of sample annotations of named entities. We design some tags specifically for representing named entities. During the learning procedure, it produces a compact rule set encoding the contextual and lexical constraints for tagging and it is easily comprehensible by humans. Fig.2.1 illustrates how transformation-based error-driven learning works.

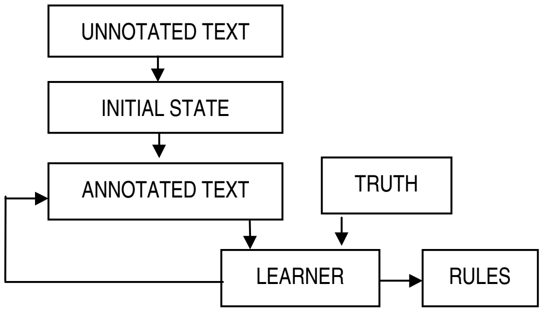

Fig 2.1 Transformation-based Error-Driven Learning

In order to obtain the transformation rules on named entities, we apply the transformation-based error-driven learning to the unannotated text. After the unannotated corpus is passed through the initial-state annotator, the preliminary output is compared with the true tagging as specified in the manually annotated corpus. Transformation rules for maximizing the error correction are proposed. The proposed rules are then applied to the annotated text again and a new set of proposed rules is learned. This iteration continuous until no minimum count of errors can be reduced. Once the rules are learned, they will be adopted by the taggers to tag a story. The TEL tagger consists of two components, start-state tagger and final-state tagger and three data files, namely, lexicon, lexical rules file and contextual rules file.

Initially, the start-state tagger reads each word tokens in the story and assigns the most possible tags from the reference lexicon to them. Next, the state tagger applies the lexical rules to tag the unknown words. The final-state tagger applies the contextual rules to tag the word tokens in the final stage. The terms tagged as specific named entities will be extracted to represent the story.

2.2 Unknown Word Identification

Chinese stories pose an additional problem compared with English counterparts. Specifically, the unknown words may cause incorrect segmentation of named entities which seriously affect the performance of the tagging. To cope with this problem, we employ unknown word identification (UWI) [3] as a post-processing.

Unknown words are defined as the word tokens which are not in the lexicon, called out-of-vocabulary (OOV). The UWI task can be divided into two steps. The first step is to detect the existence of unknown words as most of them are incorrectly segmented into sequences of single character word or short words. The second step is to determine the boundaries of unknown words. Hybrid approach is employed to deal with this problem.

In our hybrid method, rule-based approach and statistical approaches are integrated. Obvious properties, like grouping the cardinal numbers (三, 十, 二) together to form a valid number (三十二), are generalized into the rules. It provides the background knowledge on the tagging. In addition, the statistical information captures the characteristic regularities from the content of the stories.

In the unknown word identification process, n-gram grouper and entity formulator are used for extracting the unknown named entities. First, a tagged sentence is passed to the n-gram grouper to detect the boundaries of the unknown words with the assistance of the stopword list. After the boundaries have been recognized, the n-gram grouper will utilize the word list of person name to determine candidates for person names. Then, the n-gram candidates and tagged sentence will be passed to the entity formulator to extract organization names, location names and time by the formulation rules. Together with the word candidates produced from the unknown words identifier and the named entities word tokens tagged by TEL tagger, we will assign weights to candidates.

2.3 Story Representation

Traditionally, a story is represented by a set of weighted terms. In our approach, we augment this by a two-dimensional semantic representation, comprising of named entity feature representation and content term representation. Named entity feature representation is composed of three components: people name component, geographical location name component and organization name component. Each component is represented by a set of terms with corresponding weights. Specifically, the people name component is expressed as:

$$\{w(S, p_1), w(S, p_2), \cdots\}$$

where
$w(S, p_i)$ represents the weight of the corresponding people name p_i in story S.

Similarly, the geographical location name and organization name components are expressed as:

$$\{w(S, l_1), w(S, l_2), \cdots\}$$

$$\{w(S, o_1), w(S, o_2), \cdots\}$$

Relative location of the terms in the story, term frequencies and the incremental document frequencies are the factors contributing to the weight of the terms. For news articles, the content in the first few lines are usually more important. Thus, an Importance Factor has developed to represent the degree of importance of the story terms with in a story. Higher scores will be assigned to the terms located in earlier sentences in the story. The Importance Factor, *I(S ,t)*, of term *t* of story *S* as :

$$I(S,t) = \sum_{i=1} \frac{(E-1) + (L-1)(s_i - 1)}{(E-1)}$$

where
E is the total number of sentences in the story;

s_i is the sentence number of the *i-th* occurrence of the term t; and,

L is a predefined bound associated with the last sentence and it should be set in [0,1].

Furthermore, traditional term frequency and incremental document frequency also contribute to the weight determination. Let $D(S,t)$ denote the incremental document frequency of the term t in story S. The final weight of the story terms representing the story will be:

$$w(S,t) = I(S,t)D(S,t)$$

Top ranked terms are selected to form the basis for the named entity and content term feature representation.

3. GROSS TRANSLATION

3.1 Basic Translation

One requirement of the multi-lingual event detection task is to determine whether multi-lingual stories are reporting the same event. To tackle this problem, we conduct translation on the story representation of Chinese stories into English so that we can perform unsupervised learning on a uniform data representation. Gross translation approach is adopted instead of full-fledged machine translation since our objective is to translate Chinese terms adequately for event detection purpose. As a result, Chinese stories can be direct compared with English stories in a mono-lingual manner.

A bilingual dictionary and a pin-yin file provided by Linguistic Data Consortium (LDC) are used to obtain the English translation of Chinese terms. We first look up the Chinese term in the bilingual dictionary and obtain the corresponding set of English translation terms. If the corresponding dictionary entry does not exist for the people name and geographical location name terms, we obtain the translation by means of the pin-yin of the Chinese terms.

In general, a set of English terms is obtained when looking up the bilingual dictionary for a Chinese term. The translated English terms will then replace the original Chinese terms to represent the story. As a result, the whole story collection will be represented by English story terms after the appropriate translation is applied.

3.2 Enhanced Translation

In addition to the use of the bilingual dictionary, we propose an enhanced approach for improving the translation using a parallel corpus as an additional resource. Specifically, the weights of translated English terms are adjusted by using the information contained in the parallel corpus. A typical parallel corpus is usually aligned by stories. We currently make use of the parallel corpus collection on Chinese and English news items form the Information Services Department of Hong Kong Special Administration Region (HKSAR). The current collection contains news items, which are story-aligned, released by HKSAR from July1, 1997 to June 30, 1998.

We transform a typical story-aligned parallel corpus into a passage-aligned corpus via an automated procedure. A statistical length-based approach is employed to identify matching passages between English and Chinese stories. A passage is basically a sentence or a sequence of related sentences. We follow length-based text alignment technique proposed by Wu [4] to perform such translation.

The passage-aligned parallel corpus provides a means to enhance the translation quality in addition to solely using the bilingual dictionary and the pin-yin-list. For each Chinese story term, we use a similarity-based retrieval engine to retrieve the top ranked Chinese passages in the parallel corpus that are relevant to the Chinese term by posing it as a query. Then we retrieve the corresponding English passages, aligned from the above mapping method, of those Chinese passages. For each English translated term, we count the occurrence frequency of the translation term in the retrieved English passages. Finally, we adjust the weight of the translated terms by assigning a higher weight for terms with better translation quality.

4. UNSUPERVISED LEARNING MODULE

The event discovery module conducts incremental unsupervised learning [5]. It aims to automatically detect the new event or existing events for stories in chronological order. At the beginning of the detection task, the system will read a number of batches of stories up to a window period. The stories in the first batch of the story list are examined and their events are identified according to the clusters to which they belong. After the related events of those stories in the first batch are obtained as output, these stories are removed and new stories are loaded form the next incoming batch.

Our unsupervised learning approach is based on agglomerative clustering approach to group stories in the window period and the current clusters. We adopt the centroid clustering method to calculate similarities between clusters since the number of stories in a cluster could be very large. Three categories of elements, namely, stories, temporary clusters, and final clusters, will be maintained during the clustering process. The representation of temporary clusters and final clusters are similar to story representation.

5. EXPERIMENTS AND RESULTS

In order to examine the performance of our event generation system, we evaluate our method by the corpus used in the latest Topic Detection and Tracking 2000 (TDT2000) project organized by DARPA [6]. The corpus consists of text and transcribed speech news data, distributed by the National Institute of Standards and Technology, in Chinese and English spanning from January 1, 1998 to June 30, 1998. We followed the topic detection evaluation methodology designed in the TDT2000 project. The detection performance is measured by the metric $(C_{DET})_{Norm}$. Smaller values correspond to better performance.

We have conducted two sets of experiments to investigate the detection performance of our event detection approach. In the first set of experiments, we have processed the Chinese stories. The stories expressed by the original

Chinese terms (CHI), both the simple (BR) and enhanced (ER) translation with and without the unknown word identifier (UWI), are conducted under various parameter settings. We run our event detection system to compare the effects of translation and the unknown word identifier in terms of the detection performance. The number of terms for each kind of named entity component was set to 15. The number of content terms was set to 20 for Chinese stories and 30 for multi-lingual stories. Table 5.1 reveals the detection performance under different similarity thresholds, *q*, by varying from 0.05 to 0.15 for Chinese stories.

Table 5.1 Performance on Translated Chinese Story Detection Under Different Similarity Threshold

	Similarity Threshold				
	0.05	**0.08**	**0.1**	**0.13**	**0.15**
ER - no UWI	0.5906	0.6307	0.5224	0.5198	0.5184
BR - no UWI	0.5975	0.539	0.5568	0.5459	0.5137
ER - UWI	0.6362	0.5961	0.5036	0.5214	0.4705
BR - UWI	0.6307	0.5224	0.5198	0.4952	0.5452
CHI - no UWI	0.495	0.482	0.4725	0.4876	0.5116

From the result, we discover that the best performance for the enhanced translation integrating with the unknown word identification when the threshold is 0.15. Moreover, the results obtained from the enhanced translation combining with unknown word identification is more encouraging than that of the basic translation approach and without using unknown word identification when the similarity tends larger. This combination is preferred when the detection is more sensitive to the novelty topics. We also observe that this translation approach achieve the best performance with the detection cost 0.4705. And if the number of content terms increased to 25, it can even achieve 0.4470 in the detection cost. Therefore, the enhanced translation integrating with the unknown word identification gives the most preferable performance.

Table 5.2 Performance on Multi-lingual Story Detection Under Different Similarity Threshold

	Similarity Threshold				
	0.05	**0.08**	**0.1**	**0.11**	**0.12**
ER - no UWI	0.403	0.3795	0.3902	0.4016	0.4034
BR - no UWI	0.3769	0.3945	0.3993	0.41	0.4101
ER - UWI	0.4082	0.3777	0.3924	0.3766	0.3772

In the second set of experiments, we have conducted a series of runs on multi-lingual news stories. Table 5.2 summarized the results on detecting multi-lingual stories by varying the similarity threshold *q*, from 0.05 to 0.12. The table indicates the trend that the enhanced translation with unknown word identification is preferred when the similarity threshold tends larger. The best detection cost obtained from the enhanced translation with unknown word handling element is 0.3766 when the similarity threshold equals to 0.11 compared with the value of 0.4016 and 0.4100 of the enhanced translation and basic translation without unknown word identification respectively. Similar to the observation from the Chinese story detection, the enhanced translation approach together with the unknown word identification give us better performance in general.

6. CONCLUSIONS

We describe the construction of an intelligent system to discover new events from incoming sources of news stories automatically. We develop an innovative characteristic on representing a story by deploying the named entities extracted by transformation-based error-driven linguistic taggers. Furthermore, we integrate the unknown word identification approach in Chinese stories as the post-processing in extracting the missed entity terms due to the incorrect segmentation. To deal with the multi-lingual news, we propose two gross translation approaches. One is called the basic translation approach which replaces the Chinese word tokens from the corresponding bilingual dictionary entries directly. Another one is called enhanced translation approach which employs a parallel corpus for providing the additional information. In the detection module, we modify the traditional agglomerative centroid clustering approach by introducing new elements like temporary clusters and final clusters. The experiment results reveal that the performance of the detection from the Chinese stories translated by the enhanced translation approach integrated with unknown word identification is better than those translated by the basic translation approach and without applying unknown word identification. When the similarity threshold tends larger, the enhanced translation approach with the unknown word identification gives us a more encouraging result. Event detection is a challenging and interesting investigation area which still have many rooms to improve in terms of efficiency and error reduction. We anticipate the event detection technology can be widely utilized in various applications in the real world.

REFERENCES

[1] http://www.itl.nist.gov/iani/894.01/tests/tdt/tdt2000/index.htm TDT2000 Home

[2] Brill, E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. In *1995 Association for Computational Linguistics*, Volume 21, Number 4, 1995, p. 543 - 565,.

[3] Ip, C.W. Transformational Tagging for Topic Tracking in Natural Language. Dept. of Systems Engg. & Engg Mngt of CUHK Master Thesis, June 2000.

[4] Wu, D. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In *Proceedings of the 32nd Annual Conference of the Association for computational Linguistics, Las Cruces, New Mexico, U.S.A.* pages 80 - 87, 1994.

[5] Wong, K.L. Automatic Topic Detection of Multi-lingual News Stories. Dept. of Systems Engg. & Engg. Mngt. of CUHK Master Thesis, June 2000.

[6] The Year 2000 Topic Detection and Tracking (TDT2000) Task Definition and Evaluation Plan.