Automatic Event Generation From Multi-lingual News Stories

Kin Hui
The Chinese University of Hong Kong
Shatin, Hong Kong, PRC
khui@se.cuhk.edu.hk

Wai Lam
The Chinese University of Hong Kong
Shatin, Hong Kong, PRC
wlam@se.cuhk.edu.hk

Helen M. Meng
The Chinese University of Hong Kong
Shatin, Hong Kong, PRC
hmmeng@se.cuhk.edu.hk

ABSTRACT
We propose a novel approach for automatic generation of topically-related events from multi-lingual news sources. Named entity terms are extracted automatically from the news content. Together with the content terms, they constitute the basis of representing the story. We employ transformation-based linguistic tagging approach for named entity extraction. Two methods of gross translation on Chinese story representation into English have been implemented. The first approach uses only a bilingual dictionary. The second method makes use of a parallel corpus as an additional resource. Unsupervised learning is employed to discover the events.

Keywords
Event Discovery, Event Detection, Multi-lingual Text Processing

1. INTRODUCTION
Interests in the event generation have grown rapidly in recent years. While most existing studies have concentrated on handling relevant specific queries, there is a need to explore techniques for classifying relevant issues from a continuous stream of data. Moreover, the rapid growth of the Internet and wide availability of electronic media allows the use of electronic means to access the newswire stories from diverse sources. Thus, the capability of identifying whether multi-lingual news are discussing the same event is one of the challenges of the event discovery task. In this paper, we particularly focus on dealing with English and Chinese news stories.

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. In the event generation problem, news stories are arriving from different sources in chronological order. Some stories are in Chinese and some are in English. Event generation aims at identifying whether the incoming stories belong to a new event or an existing event detected in previous stories. The detection can contribute to the construction of structured guidelines for story navigation of the whole news collection.

Traditionally, stories are represented solely on raw terms appeared in their content. In our approach, we augment the representation by named entity components, namely person names, geographical location names and organization names. This story representation is able to capture terms conveying useful context for event generation purpose. We look into two named entity extraction approaches and investigate their impacts on event generation. The first approach utilizes the named entity extraction module provided in a commercial text mining product. The second method makes use of the transformation-based linguistic tagger and a collection of transformation rules to extract named entity terms. The transformation rules are learned from a training corpus.

Another feature of our approach lies in the online gross translation. We have developed two methods to translate Chinese terms into English. The first one is the basic translation, which involves looking up each term in a bilingual dictionary and replaces it with possible term translation. Another approach utilizes the easily available resources like a parallel corpus to perform translation. Discovery of new events is tackled by unsupervised learning. Our unsupervised learning is based on a modified agglomerative clustering algorithm.

2. STORY REPRESENTATION
As raw stories contain a sequence of words, we need to identify sentence boundaries for subsequent processing. Lexical clues, such as punctuation marks, are used to locate the boundaries. To address the absence of word boundaries for Chinese stories, we employ the dynamic programming technique based on the tool provided by Linguistic Data Consortium [1] to locate it.

We augment the story representation by a two-dimensional semantic expression, comprising of named entity feature representation and content term representation. Named entity feature representation is constituted by people name component, geographical location name component and organization name component. Each component is represented by a set of terms with corresponding weights expressed as a vector.

3. NAMED ENTITY EXTRACTION
We look into two approaches for extracting named entity terms. The first method is taken from the Feature Extraction Tool in the Intelligent Miner for Text Analysis package from IBM. Presently, this tool can only be applied for processing English newswire stories. The second method is based on transformation-based linguistic tagging (TEL) approach derived from Brill [2]. This approach can deal with English and Chinese as well as newswire and broadcast stories.

For Chinese stories, unknown words may cause incorrect segmentations of named entities which degrade the results of tagging. Therefore, we apply unknown word identification [3] to cope with this problem. Together with the word candidates produced from the unknown word identifier and the named entity word tokens tagged by tagger [4] , we associate a weight to each term to represent the story for the event discovery task.

4. GROSS TRANSLATION

In order to determine whether multi-lingual news stories are topically-related, we conduct translation on the story representation of Chinese stories to English so that we can perform unsupervised learning on a uniform data representation. Two approaches of translation have been developed, namely the basic translation and the enhanced translation.

4.1 Basic Translation

We make use of a bilingual dictionary and a pin-yin file provided by Linguistic Data Consortium (LDC) in finding the English translation terms. First, we look up the Chinese term in the bilingual dictionary and retrieve the corresponding set of English terms. If there is no dictionary entry for a term in the people name component or the geographical location name component, we obtain the translation by means of the pin-yin of the Chinese term. The English translation terms will replace the original Chinese term to represent the story with the assigned weights based on the original Chinese term weight.

4.2 Enhanced Translation

In addition to the bilingual dictionary, we utilize a parallel corpus collection on Chinese and English news items as an additional resource to improve the accuracy of translation. An automated procedure is developed to transform a parallel corpus to passage-aligned setting. 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 the Chinese term as a query. Then we retrieve the corresponding English passages of those Chinese passages. Finally, we adjust the weight of the translated terms by assigning a higher weight for terms with better translation quality based on the statistics obtained from the retrieval English passages.

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 [5] . 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 performance of our event generation approach. In the first set of experiments, we have processed the Chinese stories by the original Chinese terms, both the basic and enhanced translation are conducted on the stories under various settings to compare the effects of translation. Table 5.1 reveals the detection performance, in terms of $(C_{DET})_{Norm,}$ under various similarity thresholds, q, from 0.05 to 0.15.

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**
Original Chinese Terms	0.4950	0.4820	0.4725	0.4876	0.5116
Enhanced Translation	0.6362	0.5961	0.5036	0.5214	0.4705
Basic Translation	0.6307	0.5224	0.5198	0.5184	0.4952

The result shows that, the detection performance of the enhanced translation method is more encouraging than that of the basic translation method. We also observe that the best result achieves a detection cost of 0.4705. If the number of content terms is increased to 25, it can even achieve 0.4470.

Table 5.2 summarizes the detection performance under different similarity threshold, q, by varying from 0.05 to 0.13 for event generation of all English stories and the multi-lingual stories.

Table 5.2 Performance on All English Stories and Multi-lingual Stories Detection Under Different Similarity

	Similarity Threshold								
	0.05	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**	**0.11**	**0.12**	**0.13**
ENG IMT	0.4308	0.4279	0.4379	0.4365	0.4346	0.4276	0.4349	0.4181	0.4376
ENG TAG	0.4132	0.3801	0.3864	0.3986	0.3854	0.3809	0.3802	0.3897	0.3881
MUL IMT	0.4545	0.4387	0.4223	0.5057	0.4127	0.4288	0.4291	0.4449	0.4252
MUL TAG	0.4082	0.3861	0.3882	0.3777	0.3913	0.3924	0.3766	0.3777	0.3744

From the result, it depicts that the detection cost of the stories with the named entity term extracted by the transformation-based linguistic approach (TAG) is better than those extracted by the commercial product (IMT) for both multi-lingual (MUL) and English (ENG) stories. The best results obtained from the TAG approach from both data stream are 0.3744 and 0.3801 respectively compared with 0.4127 and 0.4181, the best costs produced by IMT method.

6. FUTURE WORKS

We will conduct more experiments to optimize the parameter settings. Moreover, we will focus on investigations on unsupervised learning module and explore other techniques on clustering to further improve the performance.

7. REFERENCES

[1] http://morph.ldc.upenn.edu/TDT Linguistic Data Consortium

[2] E. Brill. 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, pages 543-565, 1995.

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

[4] H. Meng and C.W. Ip. An Analytical Study of Transformational Tagging for Chinese Text. In *Proceedings of Research on Computational Linguistics Conference (ROCLING XII),* Taipei, Taiwan, ROC,pages 101-122, 1999.

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