MULTILINGUAL TOPIC DETECTION USING A PARALLEL CORPUS

Wai Lam and Helen M. Meng and Kin Hui

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin
Hong Kong
{wlam,hmmeng,khui}@se.cuhk.edu.hk

ABSTRACT

We have developed an approach for topic detection from multi-lingual news, in particular Chinese and English. We extract named entities such as people names, geographical location names, and organization names automatically from the news content by transformation-based linguistic taggers. These sets of named entities together with the remaining content terms form the basis of news representation. Gross translation of Chinese story representation into English is conducted using easily available resources. We have investigated two approaches for gross translation. One is a basic method using only a bilingual dictionary. The second approach makes use of a parallel corpus as an additional resource. The topic discovery task uses a modified agglomerative clustering algorithm to group stories. One difference between our clustering approach and the standard agglomerative one is that we maintain three kinds of elements in the clustering process, namely, story, temporary clusters, and final clusters.

1. DESIGN OVERVIEW

We have developed an approach for topic detection from multi-lingual news, in particular Chinese and English. Our approach employs a story representation scheme based on named entities and content terms. Named entities including people names, geographical location names, and organization names are extracted by transformation-based linguistic taggers, one for English and one for Chinese. The corresponding set of transformation rules is learned from a training corpus. These sets of named entities together with the remaining content terms form the basis of news representation. This representation is able to capture important terms conveying useful context for topic detection purpose.

Another feature of our approach lies in the online gross translation. The news stream contains English stories as well as Chinese stories. It is necessary to determine whether the stories, regardless of the language, are topically related. To tackle this problem, gross translation is performed on the Chinese story representation so that unsupervised learning can be conducted under a uniform representation scheme.

We have investigated two approaches for gross translation using easily available resources. One is a basic method using only a bilingual dictionary. The second approach makes use of a parallel corpus as an additional resource.

Unsupervised learning is adopted for the topic discovery task. We have developed a modified agglomerative centroid clustering method for grouping related stories. We maintain three kinds of elements in the clustering process, namely, story, temporary clusters and final clusters. Stories and clusters share a similar representation. To control the influence of old stories, a time adjustment scheme is introduced in the clustering process.

2. STORY REPRESENTATION

Since each raw story only comes in a sequence of tokens, we need to identify sentence boundaries for subsequent processing. Lexical clues, such as punctuation marks, are used to locate the sentence boundaries for newswire stories. The length of silence period is used to locate the boundaries for broadcast news. For Chinese stories, word segmentation is performed using the tool provided by Linguistic Data Consortium (LDC) to locate word boundaries.

The next step is to automatically extract named entities from the news. In particular, we extract three kinds of entities, namely, people names, geographical location names and organization names. We utilize a transformation-based linguistic tagger which tags a sentence according to a set of transformation rules, one set for English and one set for Chinese. Each set of rules is learned from a training corpus, also one for English [1] and one for Chinese [2]. Each training corpus contains sample annotations of named entities. To start the training phase, we apply the transformation-based error-driven learning to the text. After the text has been passed through the initial-state annotator, it is compared with the sample tagging as specified in the annotated training corpus. This process is repeated until an acceptable set of transformation rules is obtained.

Traditionally, a story is represented by a set of weighted terms. In our approach, we adopt a two-dimensional semantic representation, which comprises 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.

There are several factors determining the weight of the terms.

The first factor is the relative location of the terms in the story. Those terms appearing in the first few sentences will get a higher weight. The other factors are traditional term frequency and incremental document frequency. The top ranked terms are selected as the basis for the named entity feature representation and content term representation.

3. GROSS TRANSLATION

3.1 Basic Translation

One requirement of the multilingual topic detection problem is to determine whether stories, regardless of the language, are topically related. To tackle this problem, we conduct translation on the story representation of Chinese stories to English so that we can perform unsupervised learning on a uniform data representation. Full-fledged machine translation is not necessary since our objective is to translate Chinese terms adequately for topic detection purpose. Therefore, gross translation will be performed on the Chinese story terms to allow for direct comparison between stories.

In order to obtain the English translation of the Chinese terms, a bilingual dictionary and a pin-yin file provided by Linguistic Data Consortium (LDC) are used. We first look up the Chinese term in the bilingual dictionary and retrieve the corresponding set of English translation 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 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. The weight of those translated English terms will be computed as:

$$w(S,t_t) = \frac{w(S,t)}{N}$$

where N is the total number of translated English terms for the Chinese term t; and $w(S, t_t)$ is the term weight of the translated English term t_t for the Chinese story term t in story S.

For example, if there is a Chinese term ' •• ' with term weight 1.000. Suppose there are three translation entries, namely, 'well-known', 'famous' and 'celebrated' in the bilingual dictionary. Then, the system will assign the value 1.000 / 3 to the weight of each translated English term. After the appropriate translation is applied on the Chinese terms, the whole story will be represented by English story terms.

4.2 Enhanced Translation

In addition to the use of the bilingual dictionary, we propose an enhanced approach for improving the translation by using a parallel corpus as an additional resource. Specifically, the weights of translated English terms are adjusted 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 from the Information Services Department of Hong Kong Special Administration Region (HKSAR) of the People's Republic of China. The current collection contains news items, which are story-aligned, released by HKSAR from July 1, 1997 to April 30, 1999.

We propose to 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 the text alignment technique proposed by Wu [3]. The length-based alignment approach is described below.

Our aim is to maximize the probability over all possible alignments, given a pair of parallel stories as:

$$Max(\Pr(A|S_e,S_c)) \tag{1}$$

where A is a particular passage alignment for the English story S_e with the Chinese story S_c. Let P_e and P_c denote an English passage and a Chinese passage respectively in S_e and S_c. By assuming that the probabilities of the individual aligned pairs of passage within a story are independent, the maximum approximation in Eq.1 can be converted into a minimum-sum problem as:

$$Max\left(\Pr\left(A|S_e,S_c\right)\right) \approx Max\prod\left(\Pr\left(A|P_e,P_c\right)\right)$$
$$= Min\sum -\log\left(\Pr\left(A|P_e,P_c\right)\right)$$

The minimization problem can be solved by dynamic programming approach.

The first step is to find the value of $\Pr(A|P_e,P_c)$ by analyzing the length of the passages. The distributions can be easily estimated by applying the Bayes' Rule:

$$\Pr\left(A|P_e,P_c\right) = \frac{\Pr\left(P_e,P_c|A\right)\Pr\left(P_e,P_c\right)}{\Pr\left(A\right)}$$

where $\Pr(A)$ is a normalizing constant that can be ignored during minimization. The value of $\Pr(P_e, P_c|A)$ can be estimated according to a distribution of the lengths of an English passage and its corresponding Chinese passage. The GB hybrid English-Chinese encoding system is used for Chinese stories. Thus, we count Chinese characters including occasional English proper names and abbreviations, as well as Chinese punctuation marks as having a length of two and each English or punctuation character as having a length of one. A Gaussian distibution is formulated based on these observations. For estimating $\Pr(P_e, P_c)$, the priors over six categories of alignment on passages are established. They are 0-for-1, 1-for-0, 1-for-1, 2-for-1, 1-for-2 and 2-for-2 where i-for-j denotes i English passages match to j Chinese passages. These prior probabilities are taken from [3]. By applying the above statistical length-based alignment approach, we can refine a story-aligned parallel corpus into passage-aligned level.

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 the Chinese term 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. An example is shown below.

Recall from the above example, a Chinese term ' ••' can be translated into 'well-known', 'famous' and 'celebrated' in English from the bilingual dictionary. We first retrieve the relevant Chinese passages containing the Chinese term ' ••'. Suppose two Chinese passages are returned as follows:

"•• • •• • •• •• • • • • • • • •••• • • • •• ••• • • • • ••• • •• •• •••"

"• • • •••• • • • • • • • • • • •• • •• •• •• • • • •• • •••• •• • •• • • • •• • • • • • • •••"

Then, the corresponding English passages in the parallel corpus are retrieved as follows:

"The inspiration of this play comes from a book written by Carl A. Hammerschlag. The theatre piece has been created by the famous mime artist Wong Kwok Chung with members of the Theatre of the Deaf."

"Mr. Leung noted that the new vehicular bridge and the approach viaduct would be a well-known and much admired landmark in Hong Kong, as well as a more than worthy neighbour to the nearby Lantau Link."

We count the occurrence frequencies of all translated English terms within these two retrieved English passages. 'famous' appears once in the first passage and 'well-known' appears once in the second passage. But the term 'celebrated' does not appear in the above passages.

By accumulating the number of occurrences of the translation terms in all retrieved English passages, we find that 'well-known' appears once and 'famous' appears once as well. We compute the Usage Factor for the translated terms as follows:

$$U\left(t_{\ },t_t\right)=\frac{f_t+1}{\sum_{i=1}^{N}\left(f_i+1\right)}$$

where N represents the number of translation terms of the Chinese term; t_c represents the Chinese term; t_t represents the translated English term; and f_i represents the accumulated term frequency of the *i-th* English translated term within the retrieved English passages. Thus,

U(' •• ','famous') = 0.4;

U(' •• ','well-known') = 0.4; and

U(' •• ','celebrated') = 0.2

This factor is then combined with the original Chinese term weight and the final weight of each translated term can be obtained as:

$$w(S,t_t)=U(t_c,t_t)w(S,t_c)$$

where $w(S, t_t)$ represents the final term weight of the translated story term t_t in story S.

5. TOPIC DISCOVERY MODULE

5.1 Detection Overview

The topic discovery module conducts incremental unsupervised learning. It aims to automatically detect the new topic or existing topics for stories in chronological order. At the beginning of the detection task, the system will read a number of batches of stories up to the deferral period. The stories are stored in a story list. The next step is to conduct clustering on stories in the story list. The stories in the first batch in the story list are examined and their topics are identified according to the clusters to which they belong. After the topics of those stories in the first batch are obtained as output, these stories are removed and new stories are loaded from the next incoming batch.

Our unsupervised learning approach is based on a modified agglomerative clustering approach to group stories in the deferral window 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 kinds of elements, namely, stories, temporary clusters, and final clusters, are maintained during the clustering process. The nature of these three kinds of elements will be described in the following subsections. The representations of temporary clusters and final clusters are similar to the story representation.

5.2 Unsupervised Learning

Let us define S as the story list containing current batches of stories in the deferral window; F as the first batch of stories in S. Two kinds of similarity matrices are maintained during the detection process. One is the story-story similarity matrix which stores the pair-wise similarities between stories in S. The second one is the story-cluster similarity matrix which stores the similarities between stories and clusters. However, not all the data in S are included in these matrices. We only need to compute the pair-wise similarities of the stories from F to all the remaining stories in S. Also the similarities between the stories in F and the existing clusters will be computed.

Our similarity calculation is based on the cosine similarity measure. We consider two factors in our similarity formulation. The first factor is the named entity relevance score, s_n, computed by processing the weights in the named entity feature representation. The second factor is the content term relevance score, s_s, computed by processing the weights in the content term representation. The final similarity score, s_f, is a weighted sum of these two scores controlled by a parameters, w_n, as shown below:

$$s_f = s_n w_n + s_s(1-w_n)$$

where w_n is a user-defined parameter, in the range of [0,1], controlling the relative degree of contribution for the named entity component and the content term component.

After similarities are calculated, the system will select the pair of elements attaining the highest similarity. If this similarity score is larger than a user-defined threshold, q, the system will proceed. By changing this threshold, we can adjust the sensitivity of our topic discovery module. If the score comes from the story-cluster comparison, it shows that the story is very similar to the final cluster. Therefore, we group the story into the final cluster and update the content of the final cluster by taking the average of the weights of terms in the cluster and the story representations. If the similarity comes from the story-story comparison, the two stories will be grouped into a temporary cluster. To determine the representation of the new temporary cluster, we take the average of the weights of terms of the story and the cluster representations. Then we will further compare this temporary cluster with the remaining stories in the story list, existing temporary clusters, and final clusters, by finding the similarities between them and proceed as follows:

1. If the similarity between the temporary cluster and the existing final cluster is the highest among those similarity scores and larger than q, we will merge the temporary cluster into the final cluster. The new cluster representation will be updated by taking the average of the weights of terms of the temporary cluster representation and the final cluster representation.

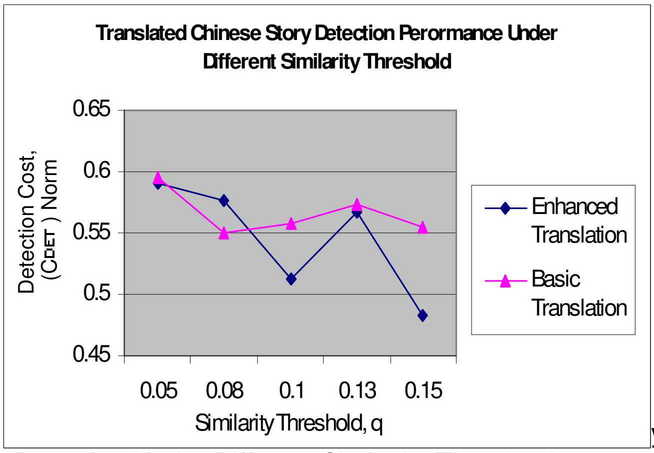
2. If the highest score comes from a temporary cluster and a story, the content of the temporary cluster will be updated by taking the average of the weights of the terms in the cluster and the story representations.

This process repeats until all the stories in F have been assigned to either a final cluster or a temporary cluster. After the clustering task, each temporary cluster is converted into a new final cluster. We check each story in a temporary cluster and only stories, which belong to F, are transferred to a final cluster. Then we compute the average of the weights of the stories as the representation of the final cluster. Finally, the temporary clusters are removed and only final clusters are retained.

6. EXPERIMENTAL RESULTS

6.1 Results for TDT2 Corpus

Two sets of experiments have been conducted on the TDT2 corpus (Jan–June, 1998) to investigate the detection performance of our topic detection approach. Manual story boundaries and native languages were used in both sets of experiments. In the first set of experiments, we only processed the Chinese stories. Both the simple and enhanced translation were conducted on the stories under various parameter settings. Then, we ran our topic detection system on the translated stories to compare the effects of translation on 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. Fig.6.1 depicts the detection performance on different similarity threshold, q, by varying from 0.03 to 0.15.



y
Detection Under Different Similarity Threshold

We find that the best performance for the enhanced translation when the threshold is set to 0.15. Moreover, it shows that the result obtained from the enhanced translation approach is more encouraging than that of the basic translation approach when the similarity threshold tends larger. It seems that the enhanced translation approach is preferred when the detection is more sensitive to the novelty topics. We also observe that the enhanced translation can achieve the best performance with the detection cost of 0.4837. For the basic translation approach, its best detection cost can only achieve 0.5508.

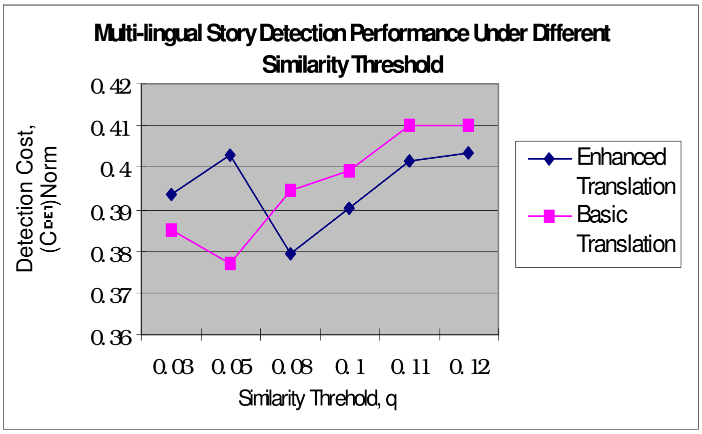


Fig. 6.2 Performance on Multi-lingual Story Detection Under Different Similarity Threshold

In the second set of experiments, we have conducted a series of runs on multi-lingual news stories. In Fig.6.2, we summarize the results on detecting multi-lingual stories by varying the similarity threshold, q, from 0.03 to 0.12. From the graph, it shows that the basic translation approach has a lower detection cost of 0.3769 when the similarity threshold is very small. However, there is a trend that the enhanced translated approach is preferred when the similarity threshold becomes larger. The lowest detection cost obtained from the enhanced translation method is 0.3795 when the similarity threshold is 0.08. By comparing the performance obtained for the translated Chinese stories and multi-lingual stories by the enhanced translation approach versus the basic translation approach, we can observe that the enhanced translation approach can give us better performance in general.

6.2 Results for TDT3 Evaluation Corpus

We have also conducted detection experiments on the TDT3 corpus (Oct-Dec, 1998). To follow the instructions suggested by NIST, we submitted experiment runs under the conditions of automatic story boundaries and native languages. The story boundaries were obtained from NIST. The detection cost for translated Chinese stories only is 0.4079. The detection cost for multilingual stories is 0.6588.

REFERENCES

1. 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.
2. 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.
3. D. Wu. 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.