CLTC: A Chinese-English Cross-lingual Topic Corpus

Yunqing Xia[1], Guoyu Tang[1], Peng Jin[2], Xia Yang[2]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, Beijing
E-mail: yqxia@tsinghua.edu.cn, sweetyuer@gmail.com
[2]Lab of Intelligent Information Processing and Application, Leshan Normal University, Leshan 614004, China
E-mail: jandp@pku.edu.cn; yangxia113@gmail.com

Abstract

Cross-lingual topic detection within text is a feasible solution to resolving the language barrier in accessing the information. This paper presents a Chinese-English cross-lingual topic corpus (CLTC), in which 90,000 Chinese articles and 90,000 English articles are organized within 150 topics. Compared with TDT corpora, CLTC has three advantages. First, CLTC is bigger in size. This makes it possible to evaluate the large-scale cross-lingual text clustering methods. Second, articles are evenly distributed within the topics. Thus it can be used to produce test datasets for different purposes. Third, CLTC can be used as a cross-lingual comparable corpus to develop methods for cross-lingual information access. A preliminary evaluation with CLTC corpus indicates that the corpus is effective in evaluating cross-lingual topic detection methods.

Keywords: cross-lingual topic detection, document clustering, corpus annotation

1. Motivation

Internet brings people convenience due largely to the multimedia content that covers almost everything. Statistics show that text remains as a dominating media on the Internet. A great many of articles are found on the Internet and the number is increasing every day. The article collection has nowadays become so huge that knowledge discovered from the content become stable and reliable. For example, people go through online news every day to track hot topics and breaking events. A traditional way to achieve this goal is that we follow the newspaper agencies. In the new Internet era, news articles are released in Web portals such as Yahoo[1]. Organizing topics and events becomes a laborious and challenging issue. Meanwhile, news articles are usually presented in different languages. For example, Yahoo! operates Web portals in different languages. Serious language barrier occurs when people want to browse news in the languages other than their mother languages. The huge demand of cross-lingual information access thus makes the research on cross-lingual topic detection very hot.

Topic detection and tracking (TDT) started to attract research interests in late 1990's (Allan et al., 1998) based mainly on military consideration. Today, TDT applications upgrades to a household demand. Very recently, research on cross-lingual topic detection (CLTD) appears in workshops and conference tracks on cross-lingual information access (Pattabhi et al., 2010; Ding, 2011; Jones, 2008; Khaitan et al., 2007). The published work indicates that cross-lingual topic detection is attracting more research interests.

Evaluation on CLTD relies on benchmark dataset. Currently, the only dataset for CLTD is TDT datasets, in which the most widely used ones are TDT1999, TDT2000, TDT2002 and TDT2003 (Graff et al., 1999, Strassel, 2005). Statistics on TDT datasets are given in Table 1.

Dataset	TDT 1999	TDT 2000	TDT 2002	TDT 2003
# of CN articles / topics	2663/ 60	572/ 50	690/ 38	570/ 32
# of EN articles / topics	6023/ 60	1835/ 56	1284/ 37	622/ 34
# of CN-EN cross-lingual articles / topics	8686/ 60	2011/ 46	1947/ 35	1152/ 28

Table 1: Statistics on TDT datasets.

We summarize drawbacks of the TDT datasets as follows. First, a small number of topics are covered. For example, only 28 Chinese-English cross-lingual topics are contained in TDT2003 dataset. Second, a small number of articles are included in the topics. For example, TDT2003 dataset contains only 41 articles on average in a topic. At last, Chinese articles and English ones are not balanced. For example, 572 Chinese articles and 1835 ones are included in TDT2000 dataset.

To address the above problems, a Chinese-English cross-lingual topic corpus, referred to as CLTD, was compiled semi-automatically in this work. Some open-source natural language processing tools were deployed to achieve automation. Finally, 58,657 Chinese articles and 56,003 English ones were organized in 150 Chinese-English cross-lingual topics.

Contributions of this work are summarized as follows. First, CLTD corpus is suitable for evaluation of large-scale cross-lingual topic detection approaches since the topics cover more domains such as finance, entertainment, politics, and so on. Articles are evenly distributed in topics so that topic detection approaches will not suffer from imbalanced data problem. Second, some cross-lingual topic detection baseline approaches are evaluated in this work, which show that CLTC corpus is potential to promote the research on cross-lingual topic detection.

[1] www.yahoo.com

Some work has already been published to achieve the goal of cross-lingual text clustering task using CLTC corpus. For example, Tang et al. (2011) evaluate cross-lingual document clustering in our CLTC corpus. (Tang, 2010).

The rest of this paper is organized as follows. In Section 2, annotation procedure is described. In Section 3, corpus analysis is given. In Section 4, evaluation of cross-lingual topic detection on the CLTD corpus is presented. This paper concludes in Section 5.

2. Annotation Procedure

2.1 Annotation Scheme

Topics and articles are stored in files. Every topic is given a unique id (topic_id) and so is every article (article_id). Two files are created to store topics and topic-article relations separately. The format of the two files is as follows:

```
<topic>
  <topic_id>topic_id</topic_id>
  <topic_path>topic_path</topic_path>
</topic>
```

In the topic file, topic_path gives where the topic is stored. Format of the topic-document relation file is given below.

```
<topic>
    <topic_id>topic_id</topic_id>
    <article>
      <article_id>article_id</article_id>
      <article_path>article_path</article_path>
      <article_lang>article_lang</article_lang>
      <article_label>article_label<article_label>
    </article>
    ……
    <article>
      ……
    </article>
</topic>
```

In this work, one of the following four labels is assigned to each article by the annotators:

– Y: indicates that the article is related to the topic;
– N: indicates that the article is un-related to the topic;
– U: indicates that the annotator is uncertain whether the article is related to the topic;
– I: indicates that the article is ignored by the annotator.

2.2 Annotation Approach

We designed a semi-automatic annotation approach to improve efficiency. The annotation work is accomplished in eight steps.

Step 1: Select articles from Chinese Gigaword Second Edition (LDC2009T27) and English Second Edition (LDC2009T13) with timestamp ranging from 1994 to 2005.

Step 2: Run text clustering algorithm (e.g., K-means) on collections with either language to find mono-lingual text clusters. Then as to have enough clusters in two languages to align, we select 500 clusters for Chinese articles and 500 clusters for English articles, which contain bigger

Step 3: Run keyword extraction tool (e.g., statistical tool) on each of the 1,000 clusters to find 5 words/phrases that can represent the cluster.

Step 4: Run cross-lingual word similarity tool (e.g., HowNet) to align the clusters of Chinese articles and English articles based on the keywords. Merging the aligned clusters, we obtained a few Chinese-English cross-lingual clusters.

Step 5: Two human annotators are assigned to compile each Chinese-English cross-lingual cluster. Around 200 Chinese articles and 200 English articles were carefully selected from each Chinese-English cross-lingual cluster. Finally, we selected 150 clusters.

Step 6: The human annotators were also assigned to label the clusters and define keywords. We finally obtained 150 Chinese-English cross-lingual topics.

Note that the corpus being produced in this way may overfit the clustering algorithm. To make the corpus fair to other techniques, we need to incorporate some articles that are not collected by the clustering tool.

Step 7: Full-text search was conducted with each of the 150 queries that list the keywords in the clusters within the remaining articles to expand the clusters. To save manpower, each cluster was assigned 1,000 candidates. Note one article may be selected to expand a few clusters.

Step 8: Assign human annotators to select 200 Chinese articles and 200 English ones to expand each of the 150 topics. Finally, we obtained 120,000 articles in total.

2.3 Annotation Tools

In order to help human annotators compile cross-lingual cluster efficiently, we developed an annotation tool (see Figure 1). It displays topic list as well as article list. The English and Chinese articles are presented in two lists. So annotators can compare English articles and Chinese articles easily.

Each article is double-blind annotated by human annotators and a verification tool is developed for checking annotation consistency (Figure 2). It compares annotation results from the two annotators and automatically finds inconsistent results.

2.4 Annotation Results

Statistics on CLDT corpus is given in Table 2.

Item	Value
# of Chinese articles/topics	58,657/150
# of English articles/topics	56,003/150
# of Chinese-English cross-lingual topics	114,660/150

Table 2: Statistics on CLDT corpus.

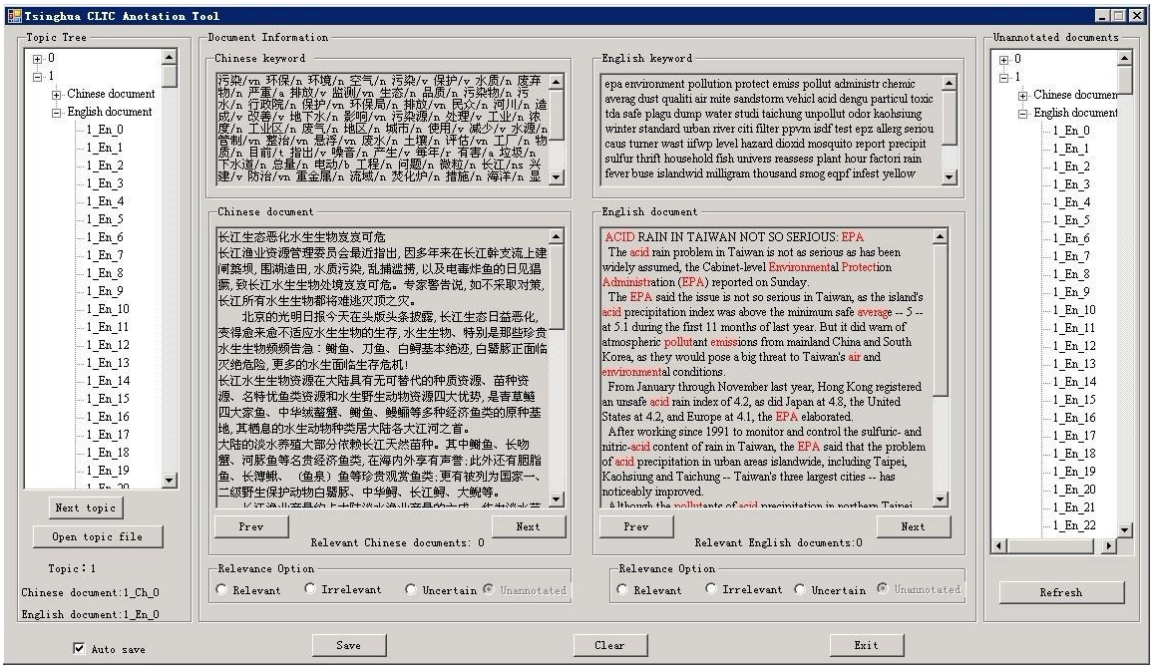

Figure 1：Interface of annotation tool

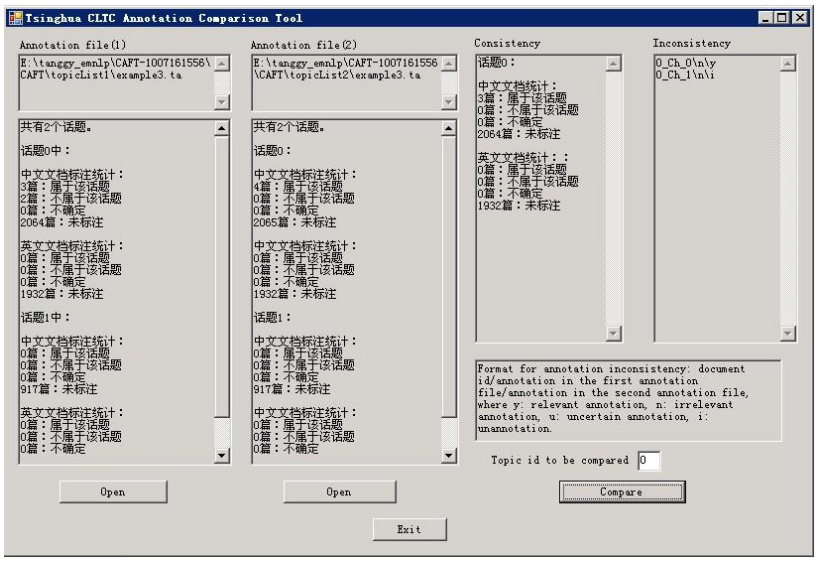

Figure 2：Interface of consistency tool

Figure 3 shows how articles distribute in the 150 topics and the two languages.

We can see from Figure 3 that number of articles within topic ranges from 0 to 600, and most topics contain about 400~500 Chinese articles and 300~400 English articles.

Figure 4 shows an example of a topic in our corpus. It has a topic id and keywords both in Chinese and English. Examples of Chinese and English articles are also displayed in Figure 4.

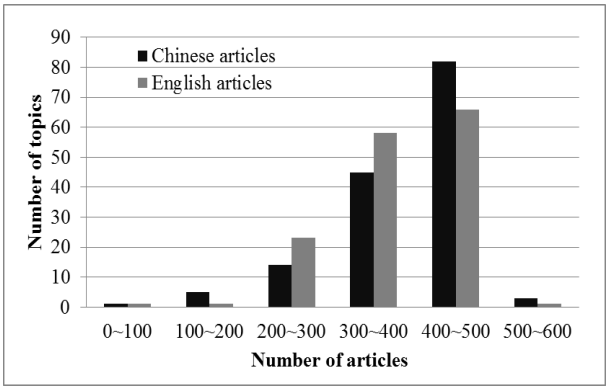

Figure 3: Article distribution within the 150 topics and the two languages.

Topic_id: 0

English Keyword: *EPA environment pollution protect pollute*

Chinese Keyword: *污染 环保 环境 空气 水质*

Example of English article:
article_id:0
article_lang : English

ACID RAIN IN TAIWAN NOT SO SERIOUS: EPA

The acid rain problem in Taiwan is not as serious as has been widely assumed, the Cabinet-level Environmental Protection Administration (EPA) reported on Sunday.
The EPA said the issue is not so serious in Taiwan, as the island's acid precipitation index was above the minimum safe average -- 5 -- at 5.1 during the first 11 months of last year. But it did warn of atmospheric pollutant emissions from mainland China and South Korea, as they would pose a big threat to Taiwan's air and environmental conditions.

Example of Chinese article:
article_id:1
article_lang : Chinese

长江生态恶化水生生物岌岌可危
长江渔业资源管理委员会最近指出,因多年来在长江幹支流上建闸築坝,围湖造田,水质污染,乱捕滥捞,以及电毒炸鱼的日见猖獗,致长江水生生物处境岌岌可危。专家警告说,如不采取对策,长江所有水生生物都将难逃灭顶之灾。
北京的光明日报今天在头版头条披露,长江生态日益恶化,变得愈来愈不适应水生生物的生存,水生生物、特别是那些珍贵水生生物频频告急:鲥鱼、刀鱼、白鲟基本绝迹,白鳖豚正面临灭绝危险,更多的水生面临生存危机!

Figure 4: Example of a topic in CLTC corpus.

3. Corpus Analysis

3.1 Annotation Agreement

The annotation agreement is computed based on the four annotation labels (see Section 2.1). We first brief the agreement measures as follows.

3.1.1. Agreement Measures

The most intuitive way for the agreement measure is the percentage of agreement between two annotators. However, the shortage of this measure is also apparent: the chance agreement of two coders is ignored. So, many researchers proposed methods to *estimate* the chance agreement. Agreement measures can be grouped into two categories: the un-weighted methods, which consider all categories are equally likely such as π (pi) (Scott 1955) and κ (kappa) (Cohen 1960); and the weighted methods, which consider the disagreement on all category is not equal-weighted such as α (alpha) (Krippendorff 1980, 2004).

Given two annotators c_1 and c_2, they classify the items into *K* categories.

$$A_\pi^e = A_\kappa^e = \sum_{k \in K} P(k|c_1) \times P(k|c_2) \qquad (1)$$

The difference between π and κ is the estimation of $P(k|c_i)$. For π, it "*uses the actual behavior of the coders to estimate the prior distribution of the categories*" and "*on the assumption that random assignment of categories to items, by any coder, is governed by the distribution of items among categories in the actual world*" (Artstein and Poesio 2008). So,

$$P(k|c_1) = P(k|c_2) = \hat{P}(k) = \frac{n_k}{2 \times i} \qquad (2)$$

where n_k denotes the total number of assignments to category k by two coders, and i the number of the items.

For κ, it "*assumes that random assignment of categories to items is governed by prior distributions that are unique to each coder, and which reflect individual annotator bias*"(Artstein and Poesio 2008). So,

$$P(k|c_i) = \hat{P}(k|c_i) = \frac{n_{c_i k}}{i} \qquad (3)$$

here, $n_{c_i k}$ denotes the number of assignments to category k by coder c_i. Finally, the two statistics are computed as follows:

$$\pi, \kappa = \frac{A_o - A_e}{1 - A_o} \qquad (4)$$

where A_o denotes the observed agreement.

The disadvantage of both π and κ is that all disagreements are treated equally. However, for our case, disagreement between Y (related to the topic) and N (un-related to the topic) of a document is obviously more serious than a disagreement between U (uncertainty to the topic) and I (ignored by the coder). To overcome this disadvantage, Krippendorff's α (alpha) (Krippendorff 1980, 2004) is used which is a weighted and more versatile method for the agreement measure.

3.1.2. Agreement Results

On the CLTD corpus, the observed agreement A_o is 0.896. Then, π and κ for the CLTD corpus are computed, considering all the four labels equally. The κ value is 0.680 and the π value is 0.670. According to Carletta (1996), the annotation is "*allowing tentative conclusions to be drawn*". Furthermore, we compute π and κ on Chinese and English topics separately. We obtain κ=0.696 and π=0.689 on Chinese, and κ=0.664 and π=0.651 on English.

All the disagreements are shown in Table 3. The values were assigned because we believe the items in Y and N categories are clearly distinct classification, while U and I are deemed vaguely distinct compared to Y and N. Obviously, Y and N have the same weight, we set 1and so U should be 1/2 because the annotator cannot tell it from Y and N. It is difficult to assign the weight to I, for simplify, we assigned 1/3 to it.

	Y	N	U	I
Y	-	1	1/2	1/3
N	1	-	1/2	1/3
U	1/2	1/2	-	1/3
I	1/3	1/3	1/3	-

Table 3: The Weights for all Disagreements

With the weights in Table 3, we finally obtain α=0.685. We also compute the α values for Chinese and English articles separately. We obtain α =0.688 for Chinese and α=0.682 for English. The slight difference occurs because the annotators are Chinese native speakers.

4. Cross-lingual Text Clustering Evaluation

In this work, we evaluated some popular text clustering algorithms with the CLTC corpus. The intention is to investigate how algorithms perform on the CLTC corpus.

4.1 Cross-lingual Text Clustering Methods

We use two clustering methods in CLUTO (Karypis, 2002).

- ***Bisecting K-Means***: An extension of K-means, which is proved better than standard K-Means and hierarchical agglomerative clustering (Steinbach et al.,2000). It begins with a large cluster consisting of every element to be clustered and iteratively picks the largest cluster in the set, split it into two.
- ***Graph based Clustering***: A method based on graph-partition. It first models the objects using a nearest-neighbor graph and then splits the graph into k-clusters using a min-cut graph partitioning algorithm.

Note that we do not evaluate the Hierarchical Agglomerative Clustering method because of its quadratic time and space complexity.

4.2 Cross-lingual Text Similarity

Similarity between two documents is a prerequisite for text clustering. Previous work mostly handles documents in the same language with the vector space model and compute text similarity using cosine distance. In this work, we use HowNet (Dong and Dong, 2006) to match words in different languages.

4.3 Evaluation Metric

Two evaluation metrics are adopted in this experiment:

- ***Entropy***

The entropy (Zhao and Karypis, 2001) measures how the various classes of documents are distributed within each cluster. Given a particular cluster S_r of size n_r, the entropy of this cluster is defined to be

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \quad (5)$$

where q is the number of class in the dataset, and n_r^i is the number of documents of the ith class that were assigned to the rth cluster. The entropy of the entire clustering solution is then defined to be the sum of the individual cluster entropies weighted according to the cluster size. That is,

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r) \quad (6)$$

where n is the size of dataset, k is the number of clusters.

- ***Purity*:**

The purity (Zhao and Karypis, 2001) measures the extent to which each cluster contained data points from primarily one class. Given a particular cluster S_r of size n_r, the purity of this cluster is defined to be

$$P(S_r) = \frac{1}{n_r}\max_i(n_r^i) \quad (7)$$

where n_r^i is the number of documents of the ith class that were assigned to the rth cluster. The purity of a clustering is calculated as a weighted sum of individual cluster purities.

$$Purity = \sum_{r=1}^{k}\frac{n_r}{n}P(S_r) \quad (8)$$

4.4 Results and Discussions

4.4.1. Influence of Number of Articles.

We randomly selected articles within topics and altered maximum number of articles within one topic from 100 to 1,000 and run the two clustering methods. Experiment results are presented in Figure 5 and Figure 6.

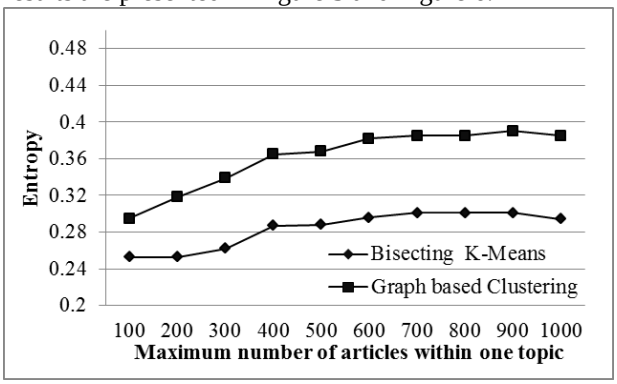

Figure5: Entropy of the two clustering methods with different maximum number of articles within one topic

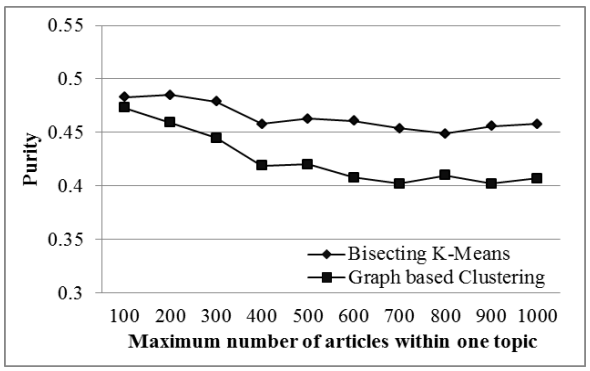

Figure 6: Purity of the two clustering methods with different maximum number of articles within one topic

We can see from Figure 5 that entropy with both methods converges to 0.39 when the maximum numbers

of articles within one topic is bigger than 800. That is, entropy remains when number of articles increase from 800 to 900. The similar observation happens with purity in Figure 6. It can thus be concluded that at least 800 articles should be contained in every topic so as to make the evaluation be convincing.

4.4.2. Influence of Number of Topics

We randomly selected topics and altered maximum number of topics from 10 to 150 and run the two clustering methods. Experimental results are presented in Figure 7 and Figure 8.

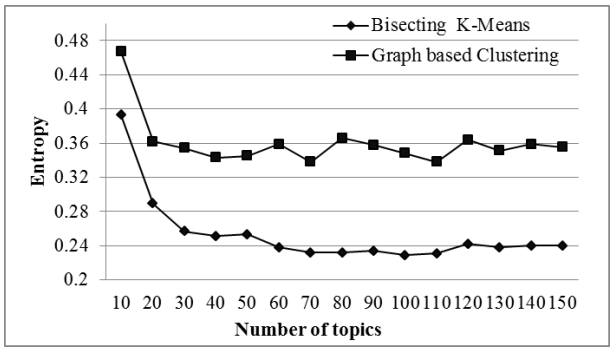

Figure 7: Entropy of the two clustering methods with different maximum number of topics

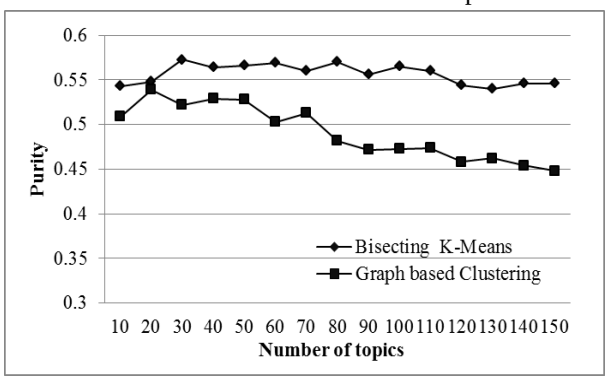

Figure 8: Purity of the two clustering methods with different maximum number of topics

We can see from Figure 7, entropy with both methods stops dropping when number of topic is bigger than 70. Similarly, Figure 8 indicates that purity stop dropping with both two clustering method when number of topic is bigger than 100. That means that, when number of topic is bigger than 100, two methods remains stable in predicting topics. This shows that a topic corpus with more than 100 topics is suitable for topic detection evaluation.

5. Conclusion

Research on cross-lingual topic detection relies on large-scale language resource, which is rather limited nowadays. CLTC corpus is a newly developed Chinese-English cross-lingual topic corpus, covering 58,657 Chinese articles and 56,003 English ones in 150 cross-lingual topics. Experiments on cross-lingual topic detection show that CLTC corpus fits into the evaluation task well. The corpus will be released to the research community shortly. It is also worth noting that annotation labor is decreased by a semi-automatic annotation approach that incorporates natural language processing tools such as text clustering, keyword extraction and information retrieval.

6. Acknowledgements

This work is partially supported by NSFC (60703051, 61003206) and MOST of China (2009DFA12970). We thank the reviewers for the valuable comments and advices.

7. References

Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194--218.

Dong, Z. and Dong, Q. (2006). *HowNet* and the Computation of Meaning. World Scientific Publishing Co. Inc., River Edge, NJ, USA.

Duo Ding. (2011). Integrate Multilingual Web Search Results using Cross-Lingual Topic Models. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 20–24, Chiang Mai, Thailand, November 8-12, 2011.

Graff, D., Cieri, C., Strassel, S. and Martey N. (1999). The Tdt-3 Text And Speech Corpus. *Proceedings of DARPA Broadcast News Workshop*

Jones, G.J.F., et al. (2008). Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia. *Proceedings CLIA2008. 2008*. Hydrabad, India.

Karypis, G. (2002). CLUTO - A Clustering Toolkit. Dept. of Computer Science, University of Minnesota, May 2002. http://www-users.cs.umn.edu/~karypis/cluto/

Pattabhi, T., Rao, R. K. and Devi, S. L. (2010). How to Get the Same News from Different Language News Papers, *Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010*, pp. 11–15

Sanjeet Khaitan, Kamaljeet Verma, Rajat Mohanty, and Pushpak Bhattacharyya. (2007). Exploiting semantic proximity for information retrieval. In *IJCAI '07: Workshop on Cross Lingual Information Access*, 2007.

Strassel, S. (2005). TDT4 Multlingual Broadcast News. *Linguistic Data Consortium.*

Steinbach, M.; Karypis, G.; and Kumar, V. (2000). A comparison of document clustering techniques. *KDD Workshop on Text Mining*.

G. Tang, Y. Xia, M. Zhang, H. Li, F. Zheng. 2011 CLGVSM: Adapting Generalized Vector Space Model to Cross-lingual Document Clustering. P*roc. of IJCNLP'2010*: 580-588.

Zhao, Y., Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis (Technical Report). Department of Computer Science, University of Minnesota.