PROCEEDINGS OF SCIENCE

Chinese-English Cross-Lingual Text Clustering Algorithm based on Latent Semantic Analysis

Huihong Lan[1]
Guangxi College of Education
Nanning, 530023, China
E-mail: lanlandoll@163.com

Jinde Huang[2]
Guangxi College of Education
Nanning, 530023, China
E-mail: h_jinde@tom.com

Aiming at the problems available in the traditional method of cross-language text clustering, a Chinese-English cross-language text clustering algorithm based on Latent Semantic Analysis is put forward. [Method] With the method of Latent Semantic Analysis, Singular Value Decomposition of characteristic word-text matrix is carried out. The bilingual latent semantic space in Chinese-English is constructed to realize cross-language latent semantic association so as to reduce dimension and noise. The K-means algorithm which chooses the initial cluster center on the basis of the minimum similarity is adopted to avoid the effect of random selection of the initial cluster centers on the clustering effect. [Results] Experiment results show that the number of reserved characteristic words of each text s and the selection of the spatial dimension value k have certain impacts on the clustering result. When each text retains the top 15 characteristic words and k=200, the F-measure can be optimal. Compared to CLTC, 13.96 percentage points can be improved. [Conclusions] This method has greatly reduced the dimension of text space and improved the cross-language text clustering quality effectively. The clustering effect is better than CLTC.

ISCC 2017
16-17 December 2017
Guangzhou, China

PoS(ISCC 2017)007

[1] Speaker
[2] Correspongding Author

This work is supported by National Natural Science Foundation of China (No.: 61262028), and the Scientific Research Project of Education Department of Guangxi Province (No.: KY2015YB337), and the Scientific Research Project of Guangxi College of Education (No.: B2012009), and the Education Science fund of the scientific research projects of Guangxi College of Education (No.: YB2014587).

1. Introduction

With the rapid development of Internet and increasingly more international communication, the network information resources have shown the multi-linguistic characteristics. The limitation of searching with only one language is becoming more and more obvious. Users are no longer satisfied with the retrieval by using the same language, instead, they are eager to search more information about other languages from the internet. Thus, how to cross the language barrier, realize information sharing and communication, and provide multi-linguistic information on the internet for users with diverse language backgrounds effectively has become a hot spot of research on multi-linguistic information mining currently.

The text clustering is considered an effective means of data mining. By far, a lot of researches have been carried out by domestic and overseas scholars on monolingual text clustering. However, researches on cross-language text clustering are still immature. Cross-language text clustering refers to the division of different language types of document collections in the multi-linguistic context according to their similarities. Documents with similar themes and relevant contents are classified to the same cluster, while those with different contents are classified to different clusters. By integrating existing multi-lingual text clustering researches, Zhang have divided the multi-linguistic text clustering method into "clustering first, merging second" and "transformation first, clustering second"[1]. In the first case, multi-lingual texts are clustered respectively before merging and clustering. In the second case, multi-lingual is transformed to mono-language to realize clustering or transformation of the multi-lingual semantic space is conducted to realize clustering. The method of "clustering first, merging second" has been adopted by Chen for clustering of Chinese and English news text collections[2]. However, clustering in the monolingual context needs to be done first with this method, and the relationship of documents and characteristic items cannot be taken into comprehensive consideration, so it is likely to result in the significant defect of weight bias. Lawrence has taken the method of "transformation first, clustering second" to study Russian-English multi-language text clustering by translating the full text via machine translation systems and dictionaries [3]. Nonetheless, the translation highly depends on the performance of the translation system, so favorable translation effects cannot be assured.

Specific to the above problems, a Chinese-English cross-language text clustering algorithm based on Latent Semantic Analysis (CLTC-LSA) is proposed by this paper. Construction of bilingual latent semantic space in Chinese-English by using Latent Semantic Analysis (LSA) to realize cross-lingual latent semantic association so as to reduce dimension and noise. By using K-means algorithm that chooses the initial cluster center on the basis of the minimum similarity, the cross-language text clustering can be realized. The clustering effect is improved.

2. Research Ideas and The Basic Process

The mapping from Chinese-English text corpus to Chinese-English bilingual text space is realized by LSA in this paper. Characteristic words and texts are mapped to a low-dimension vector space via Singular Value Decomposition (SVD), and the noise information in the original semantic space is eliminated by reducing the dimension, thus fulfilling the cross-language semantic association. By using the improved K-means algorithm, the cross-language text clustering is realized as well. The process of CLTC-LSA generally includes the following few steps:

(1) Conduct language recognition of Chinese-English bilingual texts;

(2) According to characteristics of Chinese and English, conduct pre-processing and feature extraction operation to realize the character representation of Chinese-English texts;

(3) Merge all Chinese-English text characteristics and obtain an initial Chinese-English bilingual text representation space;

(4) Carry out the LSA for initial Chinese-English bilingual texts and construct the Chinese-English bilingual latent semantic space;

(5) Use the K-means algorithm which chooses the initial cluster center on the basis of the minimum similarity and conduct cross-lingual text clustering.

The model of CLTC-LSA is shown in Fig. 1.

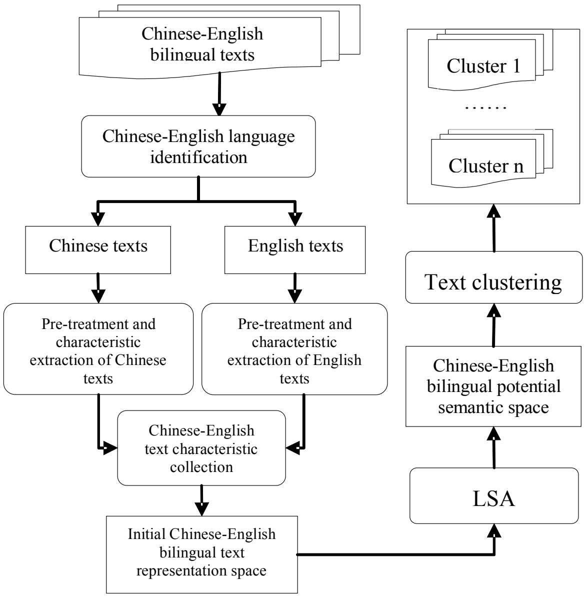
Chinese-English bilingual texts
Chinese-English language identification
Chinese texts
English texts
Pre-treatment and characteristic extraction of Chinese texts
Pre-treatment and characteristic extraction of English texts
Chinese-English text characteristic collection
Initial Chinese-English bilingual text representation space
LSA
Chinese-English bilingual potential semantic space
Text clustering
Cluster 1
……
Cluster n

Figure 1: The Model of CLTC-LSA

3. Establishment of the Bilingual Latent Semantic Space

The basic idea of LSA is: supposing there's implicit correlation between words and words, words and texts, and texts and texts in the text collection, and such potential semantics can be extracted with the method of statistic calculation to realize the purpose of eliminating word association and simplifying text vectors. Via the SVD calculation of the characteristic word–text matrix of the text collection as well as the extraction of k maximum singular values and their corresponding singular vectors, a new matrix can be constructed to represent characteristic word–text matrixes of the original collection approximately.

The preliminary m×n characteristic word–text matrix $X_{m\times n}=[x_{ij}]$ is formed by pre-treatment of texts, where, the weight calculation of x_{ij} is divided into two parts. One is the partial weight L(i,j) that is used to record the frequency of the characteristic word i in the text j; and the other is the overall weight G(i) that is used to record the ability of identifying text semantics in the text collection by different characteristic words.

$$x_{ij} = L(i, j) \times G(i) \qquad (3.1)$$

Where, $L(i, j) = lb(tf_{ij} + 1)$ (3.2)

$$G(i) = 1 - \sum_{j} \frac{p_{ij} lb(p_{ij})}{lbn} \quad (3.3)$$

Where, $p_{ij}=tf_{ij}/gf_i$. It refers to the "frequency of the occurrence of text j in case there is the characteristic word i". tf_{ij} is the frequency of the characteristic word i in the text j, gf_i is the frequency of the characteristic word i in the whole text collection; and n is the total text number of the text collection. As it is the bilingual corpus to be processed, Chinese and English characteristic words–text matrixes are formed respectively:

$$C = (c_{ij})_{mc\times n} \quad (3.4)$$

$$E = (e_{ij})_{me\times n} \quad (3.5)$$

Where, mc is the number of Chinese characteristic words; me is the number of English characteristic words; n is the total number of texts. To form the bilingual latent semantic space, 2 matrixes are merged:

$$M = \begin{bmatrix} C \\ E \end{bmatrix} \quad (3.6)$$

Where, M is the (mc+me)×n matrix. Through SVD of M, it is assumed m=mc+me:

$$M_{m\times n} = U_{m\times m} \sum\nolimits_{m\times n} (V_{n\times n})^T \quad (3.7)$$

Where, U and V are orthogonal matrix, and ∑ is the diagonal matrix.

An appropriate value k is selected, and k interception is conducted for U, ∑ and V matrix, that is, the first k columns of U are selected, the first k maximum singular values of ∑ are, as for the first k columns of V. Finally, the similar matrix M′ of $X_{m\times n}$ is formed:

$$M' = \begin{bmatrix} U_k^c \\ U_k^e \end{bmatrix} \sum\nolimits_k V_k \quad (3.8)$$

Where, U_k^c and U_k^e are mc×k and me×k matrix, representing k-dimension Chinese and English vector matrix; V_k is n×k text matrix.

4. K-means Text Clustering

K-means algorithm is a clustering algorithm based on division. Due to its concise algorithm idea, high clustering speed and positive clustering effect, its application to handling big data is rather extensive. However, because of the randomness of its initial cluster center selection, it is likely to result in local optimum and unstable clustering result. Therefore, this paper takes the K-means algorithm that chooses the initial cluster center on the basis of the minimum similarity for cross-language text clustering experiment.

4.1 Selection of the Initial Cluster Center

When conducting K-means text cluster experiment, it can be found it imposible to calssify the text with the minimum similarity (the farthest) to the same cluster, so two texts with the minimum similarity are taken as the initial cluster centers. Among the rest (N-2) texts, the one having the minimum similarity product to the previous two initial cluster centers is taken as the third initial cluster center. Similarly, among the rest (N-3) texts, the one having the minimum similarity product to the previous three initial cluster centers is taken as the fourth initial cluster center. The rest is deduced by analogy so that K initial cluster centers can be found.

4.2 Description of the K-means

The K-means text clustering algorithm that chooses the initial cluster center on the basis of the minimum similarity is described below:

Input: the text collection N and the cluster number K.

Output: K clusters and K cluster centers.

Step 1: K initial cluster centers are selected according to Section 4.1, the specific process of which is shown below:

Step1.1: calculate $sim(x_i,x_j)$. When $sim(x_1,x_2) \le sim(x_i,x_j)$,(i,j=1,2,…,N), text x_1 and x_2 will be taken as the initial cluster center.

Step1.2: In the rest (N-2) texts, when $sim(x_1,x_3) \times sim(x_2,x_3) \le sim(x_1,x_i) \times sim(x_2,x_i)$, ($x_i$ is any text but x_1, x_2 and x_3), text x_3 will be taken as the third initial cluster center.

Step1.3: In the rest (N-3) texts, when $sim(x_1,x_4) \times sim(x_2,x_4) \times sim(x_3,x_4) \le sim(x_1,x_i) \times$ $sim(x_2,x_i) \times sim(x_3,x_i)$, ($x_i$ is any text but x_1, x_2, x_3 and x_4), text x_4 will be taken as the fourth initial cluster center.

Step1.4: The rest is deduced by analogy, thus, all K initial cluster centers can be found.

Step 2: As for the rest (N-K) texts, the similarity of each text and K cluster centers is calculated, and the text is classified to the cluster with the maximum similarity.

Step 3: K cluster centers are calculated again.

Step 4: Step 2 and 3 are repeated until there's no more change for each cluster.

5. Experiment and Result Analysis

5.1 Source of Bilingual Corpus and Pre-treatment

The open source crawler program Hertrix is used to capture Chinese-English bilingual news on China Daily, 21 Century, VOA bilingual news, FT China, and Financial Network [4]. 727 texts of Chinese-English paralleled news corpus are obtained, as shown in Table 1.

Class	Education	Health	Environment	Sport	Finance	Military
Number of Chinese text	120	90	125	124	120	148
Number of English text	120	90	125	124	120	148
Total size	412KB	285KB	1233KB	395KB	645KB	623KB

Table 1: Chinese-english bilingual corpus

To reduce inconvenience caused by coding of different languages, all Chinese-English bilingual texts are shifted to UTF-8 coded format by the Chinese-English language recognition program. Then, languages are identified by use of the difference in their coding scopes (The Chinese UTF-8 coding scope is between 00000080-000007FF and 00000800-0000FFFF, and the English UTF-8 coding scope is 00000000-0000007F) [5].

After language recognition, the ICTCLAS2014 of Chinese Academy of Sciences is used for pre-treatment of the Chinese text information, that is, word segmentation, stop word removal and characteristic word extraction for Chinese texts. Regarding English texts, PorterStemmer is adopted for deformation, stemming, and characteristic word extraction.

5.2 Experimental Evaluation Indexes and Baseline Algorithm

In this experiment, Precision, Recall and F-measure are taken as evaluation indexes to measure the clustering quality. The "Chinese-English mixed cross-language text clustering (CLTC)" is taken as a contrast to explore the effect of CLTC-LSA.

CLTC: In this experiment, 1454 Chinese and English mixed texts are directly clustered using K-means algorithm that initial cluster center is selected by the minimum similarity.

5.3 Experimental Result Analysis

CLTC-LSA, Firstly, SVD of 1454 Chinese and English bilingual texts is carried out by LSA to construct the bilingual latent semantic space in Chinese-English. Secondly, K-means algorithm that initial cluster center is selected by the minimum similarity is used to cluster. TF*IDF is used to represent the text characteristics in the paper. The number of reserved characteristic words of each text s and the selection of value k in SVD have different impacts on the clustering result. In the experiment, the top 5, 10, 15, and 20 characteristic words are selected as the value s and 100, 150, 200 and 300 are selected as the value k. The Precision, Recall and F-measure of the clustering results are calculated according to different parameters. The result of CLTC and CLTC-LSA as shown in Table 2 and Table 3. The F-measure of different clustering strategies is shown in Fig. 2.

As can be seen from Table 3, When s and k are small, the characteristic word coverage scope will be too small, and the characteristic word–text matrix will be excessively compressed, so the original semantics will be unable to be expressed; when s and k is too large, the noise attenuation effect will be poor, so the clustering quality will be affected. When s=15 and k=200, three indexes, namely, Recall, Precision, and F-measure of the clustering result can be optimal. As can be seen from Fig. 2, when k=150, 200, and 300, the F-measure of CLTC-LSA is better than that of CLTC, especially when s=15 and k=200, the F-measure is optimal. Compared to CLTC, 13.96 percentage points can be improved.

Experimental results show that CLTC-LSA can achieve positive clustering effects. The clustering effect is better than CLTC.

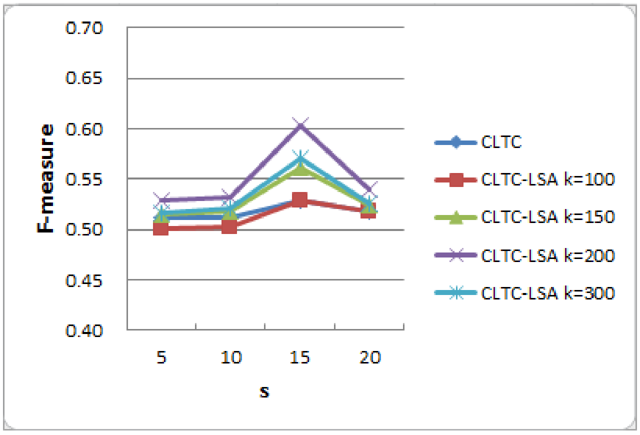

Figure 2: F-measure of Different Clustering Strategies

s	Recall	Precision	F-measure
5	0.5634	0.5282	0.5112
10	0.5741	0.5214	0.5118
15	0.6711	0.5343	0.5294
20	0.5896	0.5208	0.5186

Table 2: Results of cltc

s	k	Recall	Precision	F-measure
5	100	0.5542	0.5123	0.5011
	150	0.5646	0.5292	0.5143
	200	0.5824	0.5401	0.5286
	300	0.5701	0.5307	0.5171
10	100	0.5601	0.5103	0.5018
	150	0.5783	0.5235	0.5184
	200	0.5902	0.5414	0.5321
	300	0.5801	0.5302	0.5204
15	100	0.6702	0.5337	0.5291
	150	0.7013	0.5742	0.5613
	200	0.7432	0.6251	0.6033
	300	0.7051	0.5787	0.5701
20	100	0.5892	0.5203	0.5181
	150	0.5997	0.5314	0.5237
	200	0.6124	0.5437	0.5389
	300	0.6005	0.5328	0.5249

Table 3: Results of cltc-lsa

6. Conclusions

Aiming at the problems available in the traditional method of cross-language text clustering, this paper proposes a algorithm of CLTC-LSA. The main contributions of this paper are: (1) construction of bilingual latent semantic space in Chinese-English by use of LSA to realize cross-linguistic latent semantic association and reduce the overhead of the operation; (2) using minimum similarity method to select K-means initial cluster center for cross-linguistic text clustering and solve the problem that the initial cluster centers are randomly selected which make the clustering results unstable. The clustering effect is improved; (3) each text retains the top 5, 10, 15, and 20 characteristic words, and selection of the different spatial dimension value k, experimental evaluation of CLTC-LSA and CLTC is carried out. When k=150, 200 and 300, CLTC-LSA has better performance than CLTC. Especially, when s=15 and k=200, the F-measure can be optimal. Compared to CLTC, 13.96 percentage points can be improved.

Experiment results show that this method has greatly improved the cross-language text clustering quality effectively. The clustering effect is better than CLTC. In the future, we will continue to expand the scale of the corpus and select more characteristic words for experimental comparison. LSA will be applied to cross-lingual text clustering in more languages.

References

[1] Ch Zh Zhang, H L Wang. *Survey on Multilingual Documents Clustering*[J]. New Technology of Library and Information Service. (6), 31-36(2009)(InChinese)

[2] H H Chen, C J Lin. *A Multilingual News Summarizer*[C]. Proceedings of the 18th International Conference on Computational Linguistics. USA. pp, 159-165(2000)

[3] J L Lawrence. *Newsblaster Russian-English Clustering Performance Analysis*[R]. Columbia Computer Science Technical Reports(2003)

[4] J Shao, Ch Zh Zhang. *Automatic Acquisition of Domain Parallel Corpora from Internet*[J]. New Technology of Library and Information Service. (12), 36-43(2014)(InChinese)

[5] P Han, J X Wan, D B Wang. *Research on English-Chinese Bilingual News Clustering Based on Mixed Strategy*[J]. Information Science. 31(1), 118-122(2013)(InChinese)