Research on Topic Link Detection Method Based on Semantic Domain

Pei-Yu Liu[1,2], Yu-Zhen Yang[1,2,*], Shao-Dong Fei[1,3], and Zhen Zhang[1,2]

[1] School of Information Science and Engineering
Shandong Normal University
Shandong, Jinan, China
[2] Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology
Shandong Province, Jinan, China
No. 88 East Wenhua Road, Jinan 250014, P.R. China
zscyyz@126.com
[3] Shandong University of Finance and Economics
Shandong, Jinan, China
No.7366, East Second Ring Road, Jinan, P.R. China

Abstract. Topic Link Detection aims to detect whether a pair of random stories discuss the same topic, which is an important subtask of Topic Detection and Tracking. In previous works, statistical method and machine-learning approach are used more often than not, however, the semantic distribution of a story and the structure relationship of contents are ignored. A new method based on the semantic domain is proposed for the purpose of improved the precision. In this method, every story is divided some semantic domain through analyzing internal semantic distribution and structure relationships of contexts. The results of experiment proved that the proposed method can improve performance of system.

Keywords: topic link detection, semantic domain, topic model.

1 Introduction

TDT aims to detect the new event and tracking its development dynamically from the timing report flow[1]. Detecting and tracking the new event is need to determine the relationship of old events and new events. If the new report directly is related to the prior, this report is used to track the prior; otherwise, it is defined as new event. Obviously, every step of TDT is a process of topic link detection. Thus, Link Detection Task (LDT) is an important subtask of Topic Detection and Tracking (TDT), which task is to detect whether a pair of random stories discuss the same topic.

Different from text classification, LDT not only calculate the similarity of two random events, but take the semantic relationship of two events into account. By far, the existing methods of LDT are based on statistical or machine-learning, just as literatures [2-6]. However, the methods based on statistics usually reflect the space similarity, but not describe the semantics similarity of the reports. So literatures

* Corresponding author.

Q. Zu et al. (Eds.): ICPCA/SWS 2013, LNCS 8351, pp. 325–334, 2014.

[7-10] introduced semantic information to represent the topics. Such methods well solve the problem of link between the report pair; however, the semantic information introduced by these methods increases the complexity of the system and is very likely to cause unnecessary noise by the Chinese parsing which is incomplete currently.

Thus, the researchers integrated the evolution features of topic[11] and the structural features[12] to the construction of topic model. Lakshmi[13] built an aggregation model which based on subtopics. It is a pity that this model did not build the link relation among the subtopics. Zhang[14] used the hierarchical clustering to partition the topics and describe the hierarchical relationship among the topics. Nomoto[15] built a two-layer similar model for LDT, and compared with the hierarchical model built by Zhang[16], the experiments of results demonstrate they have the similar effect. Such models can describe the relationship between the internal structures of topics statically, but it is hard for them to update the model according to the dynamic evolution of topics.

Inspired by these, we aim to detect the relationship of random a pair of stories via diving semantic domain. In this work, we analyze the internal structure of stories, based on this, every story is divided some semantic domains. We identify the relevance of two random stories via using these semantic domains. In order to catch the relationship among stories, the relationship from stories and semantic domains, and the relationship from topic and stories, the words are divided into topic terms and semantic domain terms. In addition, we construct a loss function to measure the loss during semantic domain divided.

2 The Proposed Method

2.1 Definition of Semantic Domain

This paper believes that each semantic domain is a collection of the semantic slices which have the same semantic feature, each subtopic consists of several semantic domains and each topic is a collection of several subtopics. Based on this theory, one story d_1 can be represented as $d_1 = \{t_1, t_2, ..., t_n\}$; Here, $t_i \in d_1$ is a semantic domain in d_1; there is $t_i \neq t_j$ in any pair of semantic domains $< t_i, t_j >$, and each semantic domain is described by the semantic distribution $\{w_1, w_2, ..., w_n\}$ of the features in the report.

According to the above-mentioned definition, the relationship of the report pair $< d_1, d_2 >$ can be translated to the relationship of $< t_i, t_j >$; where, $t_i \in d_i, t_j \in d_j$. As there are only few features in the semantic domain and there is insufficient information, it is very likely to deem the uncorrelated topic as correlated topic. For example, it is likely to consider "defend the territorial sovereignty of Huangyan Island" and "defend the territorial sovereignty of Diaoyu Island" as the same topic.

Therefore, this paper partitions the semantic domain feature to two dimensionalities of topic feature $E = \{e_1, e_2, ..., e_n\}$ and semantic domain feature $W = \{w_1, w_2, ..., w_3\}$; among which, the topic feature is used to describe the relation

between the semantic domain and the topic and the unit feature is used to distinguish each semantic domain in the report. Finally, the semantic domain is defined as $T_i =< E_i, W_i >$.

The key of partitioning semantic domain is calculating the relevance. This paper regards the calculation of relevance between the semantic domains as a process of information compression. The higher the relevance between two semantic domains is, the lower the information loss is after they are condensed to one semantic domain. Therefore, this paper introduces the loss function to solve the problem of partitioning the semantic domain.

2.2 Constructing the Loss Function

Topic Feature Extraction

The purpose of defining the topic features is describing the relationship between the semantic domain and the topic, and taking the topic features as the first layer judging the correlation among the semantic domains. Therefore, the topic feature shall have relatively more semantic descriptiveness and relatively higher weight both in the whole report and the semantic domain. This paper uses the tf method to calculate the weight of the topic feature for reference. Different from the traditional tf calculation method, this paper not only calculates the tf value of the feature in the semantic domain, but also takes the tf value of the feature in the report into consideration. The calculation formula is as followed:

$$tf_E(t) = \frac{tf_L}{l} \times \frac{tf_D}{N} \tag{1}$$

In the above formula, tf_L means the time of the feature item t appearing in the semantic domain; l is the length of the semantic domain; tf_D is the time of the feature item t appearing in the report; N is the length of the report. As noun and named entity have relatively higher semantic descriptiveness, the named entity and noun are two times weighted in the experiments of this paper.

Unit Feature Extraction

Unit feature is the symbol of semantic domain in the report. The features which not only embody the semantic domain, but also can distinguish the other semantic domains shall be selected. The $tf - idf$ method calculates the weight of the feature in semantic domain and selects the feature which has a higher weight as the unit feature according to the sequence of weight size.

$$w_{ij} = \frac{tf_{ij} \times \log(N / n_i)}{\sqrt{\sum_{t_k \in S_j} [tf_{ij} \times \log(N / n_i)]^2}} \tag{2}$$

In the above formula, tf_{ij} is the frequency of the feature item t_i appearing in the semantic domain S_j; n_i is the number of all semantic domains containing the feature item t_i.

The Method of Constructing Loss Function

It is assumed that two subtopics can be merged into one semantic domain. In accordance with IB theory, the minimum feature loss will be realized before and after mergence. Division of semantic domains of subtopics can be converted into text clustering based on IB theory through this assumption.

Since feature extraction is a subject worth studying, this paper only indicates methods to extract theme features and semantic domain features in the laboratory, which is not demonstrated more than necessary. The following tasks are performed on the premise that theme features and semantic domain features are already known.

To make illustration easier T_i represents certain semantic domain to be merged; W_i represents feature words space; E_i represents news entity space of semantic domains; T_i^* represents semantic domains during clustering process; W_i^* represents feature words space of semantic domains during clustering process; and E_i^* represents news entity space of semantic domains during clustering process.

$I(T_i;W_i)$ is defined as the relationship between news feature words and topic feature words; $I(T_i;E_i)$ is defined as the relationship between news theme and topic feature words; $I(T_i^*;W_i^*)$ indicates the relationship between news feature words and topic feature words during clustering process of semantic domains. $I(T_i^*;E_i^*)$ is defined as the relationship between news theme and topic feature words during clustering process of semantic domains.

Definition 1. $f(T_i,W_i)$ is defined as joint probability distribution of T_i and W_i, then:

$$f(t_i,w_i)=p(t_i,w_i) \tag{3}$$

Definition 2. $f^*(T_i,W_i)$ is defined as joint probability distribution of T_i and W_i in (T_i^*,W_i^*) during clustering process of semantic domains, then:

$$\begin{aligned} f^*(T_i,W_i) &= p(t_i^*,w^*)p(t_i \mid t_i^*)p(w_i \mid w_i^*) \\ &= p(t_i^*,w_i^*)\frac{P(t_i)}{p(t_i^*)}\frac{p(w_i)}{p(w_i^*)} \end{aligned} \tag{4}$$

$k(T_i,E_i)$ is defined as joint probability distribution of T_i and E_i in the same method; $k^*(T_i,E_i)$ represents joint probability distribution of T_i and E_i during clustering process (T_i^*,E_i^*) of semantic domains, then:

$$\begin{cases} k(t_i, e_i) = p(t_i, e_i) \\ k^*(t_i, e_i) = p(t_i^*, e_i^*) \dfrac{p(t_i)}{p(t_i^*)} \dfrac{p(e_i)}{p(e_i^*)} \end{cases} \tag{5}$$

Through the above definition, knowledge on feature space and news theme is introduced into the information bottleneck method and final loss function is formed as follows:

$$\alpha[I(T_i;E_i) - I(T_i^*;E_i^*)] + (1-\alpha)[I(T_i;W_i) - I(T_i^*;W_i^*)] \tag{6}$$

In the above formula, $\alpha > 0$, which indicates the guidance intensity of news theme and feature words on semantic domains.

For easy calculation, this paper converts Formula (11) into Kullback-Leibler distance, i.e.,

$$I(T_i;W_i) - I(T_i^*;W_i^*) = D_{KL}(f(T_i,W_i) \| f^*(T_i,W_i)) \tag{7}$$

2.3 The Algorithm of Dividing Semantic Domain

Algorithm for division of topic semantic domains is listed as follows:

Input: T, story grouping of the given topic, λ, a parameter
Output: T_i, a story semantic domain of the given topic
Begin:

1. Dividing each story in T into paragraph grouping $P = \{p_1, p_2, \ldots, p_n\}$ and building theme feature semantic space and semantic domain feature space for each paragraph;
2. For paragraph pair p_i, p_j in each story in T, conducting following calculation for each paragraph:

$d_{ij} = \alpha[I(T_i;E_i) - I(T_i^*;E_i^*)] + (1-\alpha)[I(T_i;W_i) - I(T_i^*;W_i^*)]$
While Ture
Finding i, j that induces the minimum d_{ij};
Merging p_i, p_j into p^*;
Deleting p_i, p_j from P and adding p^* into P;
Updating d_{ij} related to p^*;
If $\forall p \in P$ and, $d_{ij} \geq \lambda$
Break.

3 Topic Link Detection Method

According to division of semantic domains, link detection of the given topic can be converted into correlation detection of semantic domains. Assuming D_1 and D_2, two pieces of news to be determined, among which D_1 is divided into three semantic domains $T_1 = \{t_1, t_2, t_3\}$; D_2 is divided into two semantic domains $T_2 = \{t_4, t_5\}$; each semantic domain t_i includes feature words w_i and news theme e_i. We still adopt Kullback-Leibler distance to assess correlation between all semantic domains.

For semantic domains t_1 and t_4 in the above D_1 and D_2, the feature space Kullback-Leibler distance is calculated as follows:

$$D_{KL}(t_1 \| t_4) = \sum_{w} p(w_i \mid t_1) \log \frac{p(w_i)}{p(w_i \mid t_2)} \tag{8}$$

In the above formula, $p(w_i \mid t_1)$ and $p(w_i \mid t_2)$ represents probability distribution of feature w_i in semantic domains t_1 and t_4 respectively.

The news theme Kullback-Leibler distance is calculated as follows:

$$D_{KL}(t_1 \| t_4) = \sum_{e} p(e_i \mid t_1) \log \frac{p(e_i)}{p(e_i \mid t_2)} \tag{9}$$

In the above formula, $p(e_i \mid t_1)$ and $p(e_i \mid t_2)$ represents probability distribution of feature e_i in semantic domains t_1 and t_4 respectively.

Since KL distance indicates degree of difference of two random distributions, when the difference increases, KL distance increases correspondingly. Therefore, it is usually used to measure degree of approximation of random distribution.

4 Experiments and Results

4.1 Experimental Corpus

The experiments of this paper adopt the TDT4 Chinese text corpus provided by the Language Standards Institute as experimental corpus. In this corpus, totally 26066 report pairs are included; among which, 3075 pairs are the correlated and the others are uncorrelated. In this paper, 1400 pairs of correlated reports and 8000 pairs of uncorrelated reports are extracted from the above corpus to compose a training set; the rest pairs compose a test set.

4.2 Experimental Process

This paper sets three groups of experiment separately over two corpus sets in order to verify the following questions respectively.

1. Experiment 1: It is used to detect the feasibility of the hierarchical model of the semantic domain proposed in this paper. Four systems are set in this paper.
   Here, it mainly uses the vector space model to represent the report and uses the cosine distance to calculate the correlation among the topics in System-1; In System-2, the report is represented by unigram language model proposed by literature[3]; In Sytem-3, topics is used semantic domains to represent, and semantic domains are used unigram language model to represent; In Sytem-4, the method of present topic is proposed in this paper.
2. Experiment 2: It is used to verify whether the proposed method is better than the other in LDT. We compared with the SDLM model proposed in the literature [9], the EMW model proposed in the literature [10] and the TTSM model proposed in the literature [15] respectively. Here, the proposed method is still represented by System-4.
3. Experiment 3: It is used to verify whether the semantic domain condensed by loss function construed in this paper is feasible. Thus, this paper adopts the LDA partitioning strategy, cosine distance (COS), KL distance, JS distance and Euclidean distance (EUC) respectively for comparison; the method in this paper is still represented by System-4.

4.3 Experimental Result and Analysis

Parameter Setting

The experimental results is affected by parameterαandθ, so we estimated the value of the parameters firstly. The Fig. 1 describes distribution of numbers of semantic domains and relevance of sentence pairs with the variation of α. Obviously, when αis 0.5, the performance of system is best.

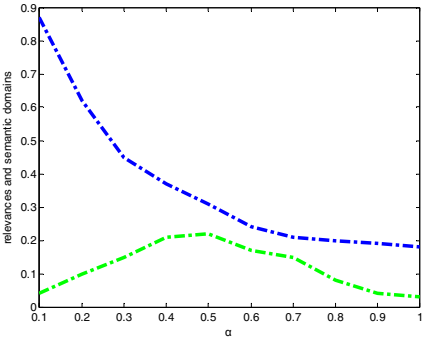

Fig. 1. Performance of system affected by α

In addition, we estimated the value of θ. Fig.2 depicts with θincreasing, Visibly, the performance of the system achieves optimal, when θis 0.07.

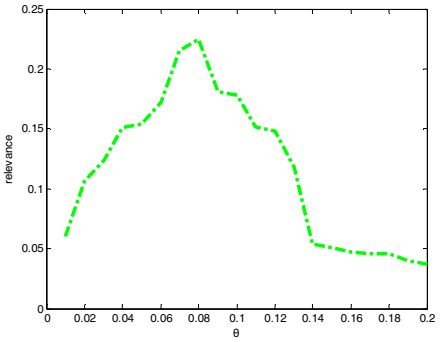

Fig. 2.

Experiment 1

Table 1 describes the evaluation results of each system LDT in Chinese TDT4 corpus and the authentic corpus. It is seen from the evaluation results that the evaluation results of System-4 and System-3 are obviously better than those of the other two systems. This result shows that it is feasible to represent a report as a hierarchical model. Compared System-4 with System-3, System-4 is superior to System-3 because System-4 faces to each semantic unit of topic to establish the topic model and represents the topic model as two layers of topic feature and unit feature.

Table 1. Comparison of LDT evaluation results

system	θ	News subjects	features	Min Norm(Cost)
System-1	-	-	55	0.58021
System-2	-	-	68	0.5983
System-3	0.38	-	63	0.2354
System-4	0.07	11	48	0.18406

Experiment 2

Table 2 describes the results comparing the link detection model with the other kinds of model. In TDT4 corpus set, the method in this paper and the SDLM model are obviously superior to TTSM model and EMW model. Compared with the SDLM method, the method in this paper does not improve greatly in TDT4 but improves by nearly 4% higher than SDLM in the authentic corpus. In addition, the method in this paper avoids the complicated syntactic analysis and calculation and lowers the system cost. Viewing from this point of view, the method in this paper is superior to the method of Baseline.

Table 2. LDT Result Comparison of Different Model

system	θ	News subjects	features	Min Norm(Cost)
SDLM	0.09	-	35	0.18453
TTSM	-	-	-	0.2033
EMW	-	-	-	0.2554
System-4	0.07	11	48	0.18406

Experiment 3

Table 3 describes the LDT evaluation result of each semantic condensation strategy. In the process of condensing the semantic domain, it is seen that the IIB method is optimal by comparing EUC, COS, KL, JS and IIB. The reason is that the condensation strategies based on distance such as EUC, COS, KL and JS are relatively applicable to centrally distributed data, but the distribution of features of topic, especially the topic of semantic domain is a kind of discrete state. Therefore, the effect of the condensation strategies based on distance is worse. The IIB method carries out the overall optimal search from the local optimal search successively and adopts the joint distribution of topic and feature to represent the semantic domain which is good for describing the distribution status of features.

Table 3. LDT evaluation result of each semantic condensation strategy

CLUSTING	θ	News subjects	features	Min Norm(Cost)
COS	0.37	9	41	0.2301
EUC	0.45	11	51	0.5109
KL	0.09	8	49	0.2487
JS	0.07	9	55	0.2108
LDA	-	-	45	0.1857
System-4	0.07	11	48	0.18406

5 Conclusion

This paper divides news into several semantic domains through constructing a cost function, and proposed a method of topic link detection based on semantic domain. The experiment proves that the method avoids complex syntax analysis during semantic modeling and effectively improves precision of relevant topic detection.

During research, this paper also finds that complex syntax analysis can improve precision of the system, but not to a large extent. Comparatively speaking, introducing syntax knowledge to topic modeling increases cost greatly. The statistics-based method is relatively simple, but hard to elaborate relations between all internal features. Therefore, we shall seek a statistics-based approach that introduces shallow semantic knowledge for topic modeling.

Acknowledgement. This work was mainly supported by Taishan Scholar Program of Computer Application, National Natural Science Foundation of China (61373148), Natural Foundation of Shan Dong Province (ZR2012FM038), and Science and Technology Development Plan of Shan Dong Province (2012GGB01194), National Natural Science Foundation of China (61373148).

References

1. The 2004 Topic Detection and Tracking (TDT 2004) Task Definition and Evaluation Plan (EB/OL) (2004), http://www.nist.gov
2. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceeding of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 297–304. ACM, New York (2004)
3. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yiming, Y.: Topic detection and tracking pilot study final report. In: Proceedings of the Broadcast News Transcription and Understanding Workshop, vol. 2, pp. 1–25 (1998)
4. Naptali, W., Tsuchiya, M., Nakagawa, S.: Topic-dependent language model with voting on noun history. ACM Transactions on Asian Language Information Processing(TALIP) 9(7) (2010)
5. Ponte, J.M., Bruce Croft, W.: A language modeling approach to information retrieval. In: Proc. SIGIR, pp. 275–281 (1998)
6. Ha-Thuc, V., Srinivasan, P.: Topic models and a revisit of text-related applications. In: Proceedings of the 2nd PHD Workshop on Information and Knowledge Management, pp. 25–32 (2008)
7. Nallapati, R., Feng, A., Peng, F., Allan, J.: Event threading within news topics. In: Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management (CIKM), pp. 446–453 (2004)
8. Chaitanya, C., Smyth, P., Steyvers, M.: Combining Concept Hierarchies and Statistical Topic Models. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 1469–1470 (2008)
9. Hong, Y., Zhang, Y., Fan, J.L., Liu, T., Sheng, L.: Chinese topic link detection based on semantic domain language model. Journal of Software 19(9), 2265–2275 (2008)
10. Wang, L., Li, F.: Story Link Detection Based on Event Words. In: Gelbukh, A. (ed.) CICLing 2011, Part II. LNCS, vol. 6609, pp. 202–211. Springer, Heidelberg (2011)
11. Shah, C., Eguchi, K.: Improving document representation for story link detection by modeling term topicality. Information and Media Technologies 4(2), 433–441 (2009)
12. Hong, Y., Zhang, Y., Fan, J.L., Liu, T., Li, S.: New Event Detection Based on Division Comparison of Subtopic. Chinese Journal of Computers 31(4), 687–695 (2008)
13. Lakshmi, K., Mukherjee, S.: Using Cohesion-model for story link detection system. Intenational Journal of Computer Science and Network Security 7(3), 59–66 (2007)
14. Zhang, K., Zi, J., Wu, L.G.: New event detection based on indexing-tree and named entity. In: Proceedings of the 30th Annual Int'1 ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 215–222 (2007)
15. Nomoto, T.: Two-tier Similarity Model for Story Link Detection. In: Proceedings of the 19th ACM International Conference Information and Knowledge Management, pp. 789–798 (2010)
16. Zhang, K., Li, J.Z., Wu, G., Wang, K.H.: Term-Committee-Based Event Identification Within Topics. Journal of Computer Research and Development 46(2), 245–252 (2009)