Research on the Term Relation Identification Based on Fuzzy Comprehension Evaluation Method*

-- A Case Study on the Natural Disaster Risk Term

LI Hui[1, a], Zhang Jing[1, b], Jin Xiaohong[1, c], Zhang Yunliang[2, d]

[1] Beijing Institute of Science and Technology Information, Beijing 100048, China;

[2] The Institute of Scientific and Technical Information of China, Beijing 100038, China

[a]lisa-lh@126.com, [b]zhangj@bjstinfo.com.cn, [c]jinxh@bjstinfo.com.cn, [d]88642191@qq.com

Keywords: Fuzzy comprehension evaluation; term relation identification; similarity; natural disaster risk term

Abstract: Based on the existing terminology in the Chinese science & technology term system of natural disaster monitoring and defense, in this article, fuzzy identification of term relation is conducted by multi-strategy fuzzy comprehension evaluation. First, calculate similarity through various similarity computing methods; then determine division of relation category and threshold range through continuous attributes discretization method, and determine factor weight through particle swarm algorithm and cross validation method; and finally, integrate and process computed results of all similarity computing methods through fuzzy comprehension evaluation method, so as to achieve fuzzy identification of term relation. Evaluate the results of relation identification via precision, recall and F value, to demonstrate effectiveness of this method. By using the existing terms in natural disaster risk monitoring scientific and technical word system as a test set, it is found that the method can effectively realize the recognition of terms.

Introduction

During the construction of term standardization documents and term classification frameworks or domain knowledge organization systems, it is required to classify term relations and divide terms into the corresponding term relation types. Therefore, term relation identification is of great importance. At present, similarity computing methods focus on simple judgment of the relation among terms, but rarely conduct in-depth subdivision to the similarity relation of terms. The similarity relation of terms collectively refers to specific relations such as synonymy, antonymy, relation between abbreviation and full name, and hyponymy. There are a variety of methods for term similarity relation identification, which can be generally divided into three types: artificial identification, semi-automatic identification and automatic identification. Though the artificial identification method is feasible with high accuracy, it needs to be constructed manually and takes a lot of time and efforts. Semi-automatic identification and automatic identification can be constructed in a short cycle with high efficiency, but they still have the shortcomings below: First, they can only identify a single relation of a small quantity of relations, such as synonymy identification or hyponymy identification [1], with limited number of relations to be identified.

However, in addition to the several relations mentioned above, lexical relations also include

Fund project: This is one of the research results of Beijing municipal finance projects "Information Processing and Analysis Ability Building Against Text Information " (PXM2016-178214-000006) , "Design and Implementation of Specific Entity Relation Extraction and Data Mining Tools"（PXM2016_178214_000007）and "Research on safety risk terms" (PXM2016_178102_000036) in 2016.

composition relation, control relation, temporal relation, spatial relation, etc. Nevertheless, simultaneous identification of multiple relations hasn't been achieved by far. Second, since the relations to be identified are indistinct and non-subdivided, they need to be further processed during the construction of a word system. For example, the similarity relation includes multiple types for similarity relation identification. As long as there is a certain relation between two words, it can be regarded as the similarity relation, but the specific similarity relation type that the relation belongs to cannot be determined. Third, since most of similarity relation identification methods merely use one kind of similarity computing method, the identified result usually has one-sidedness. However, shortcomings of a single algorithm can be avoided by gathering the advantages of several relation identification algorithms, so as to improve both precision and recall effects.

Therefore, based on the information of terms and term definitions in the existing natural disaster risk monitoring scientific and technical word system, this paper proposes to integrate and process computed similarity results by using multiple similarity computing methods through the fuzzy comprehension evaluation method, and then achieves the subdivision of term relations by means of computing characteristics of the fuzzy comprehension evaluation method.

Research Status of Term Relation Identification

Considering the one-sidedness of various identification methods, the current term relation identification usually ensembles multiple methods, which can avoid the shortcomings of a single algorithm and gather the advantages of several algorithms, so as to improve both precision and recall effects. Lu Yong, et al. [0] extract experimental objects from the text corpus of Baidu Baike, and automatically acquire synonyms in the Chinese text corpus of Baidu Baike through the comprehensive use of literal similarity method, characteristic pattern matching method and PageRank link analysis method, then directly collect synonym results of various strategies and delete the repeated ones, and include the words that are recommended as candidate synonyms in the synonym list. It is shown by the experiment that synonym results identified by the three methods are significantly different with low repetition rate, and complementary synonym identification can be effectively achieved in this way. Using the integrated method of direct collection and deletion can avoid the omissions caused by synonym extraction based on a single method as far as possible and increase the recall of synonyms, but may cause deviations in the precision. Ma Haichang, et al. [0] combine the methods of latent semantic analysis and pointwise mutual information, and select the corpus related to news in mid-July of 2008 from the network of Sogou Labs as the research object. First, analyze the term-document matrix of the corpus by the latent semantic analysis (LSA) method; meanwhile, compute the similarity between words by the cosine similarity method in the dimensionality reduction matrix of lexical semantics after singular value decomposition. Second, if a target word "i" appears in the context of another target word "j", compute PMI (i, j) between the target words by the pointwise mutual information (PMI) method. Finally, conduct weighted addition and combination to similarity and PMI according to the result of LSA decomposition of words. If the sum of cosine similarity between words and their PMI (i, j) is more than a certain threshold, the two words are regarded as synonyms of each other. The experiment proves that compared with a single method, the combined method has improved precision, recall and F value.

Later, Ma Haichang, et al. [0] further put forward the syncretic synonym method combining literal similarity and PageRank link for the economic field of Baidu Baike. In the same way, assign the two methods with different weighted values, and then add them up to acquire the similarity value between words after the two methods are integrated. It is shown by the experiment that the syncretic method with weighted values can acquire a large quantity of synonym sets from the large-sized corpus with improved precision, recall and F value in synonym identification compared with a single method. Both the methods integrated can enhance advantages and avoid disadvantages as supplements for each other, and optimize the synonym identification performance. However, with only two synonym identification methods integrated in the paper, the methods are not comprehensive enough and still prone to cause omitted synonyms. Liu Wei, et al. [0] put forward three approaches and implementation technologies used for acquiring synonym term resources from the web, including synonym term extraction based on syntactical pattern, dynamic extraction from retrieval interface of online synonym dictionary, and static crawling for synonym term classification, which realize a web-based synonym term retrieval prototype system used for acquiring Chinese and English synonym term resources in the web. In order to improve the performance of synonym identification algorithms, Curran [0] assembles several synonym identification methods based on the monolingual dictionary. The three assembled methods are respectively based on different context information extracted. The relation instance of context information is defined as a triple (w, r, w’) where "w" represents the target word, "r" is the relation type and "w’" is the adjacent word in the sentence. For example, (dog, direct-obj, walk), (r, w’) is the property of the word "w". The easiest and quickest method is to merely extract the words in a fixed distance from the target word. A complex method is to extract the grammatical relation of words with shallow statistical tools or extensive relation coverage analyzers. The MINIPAR method proposed by Lin [0] is selected for extracting relations in the syntax analysis tree with analyzers in the principle of extensive coverage. Another method is to extract relations by the SEXTANT method proposed by Grefenstette [0]. Finally, the three methods are integrated in the way of voting assembly. It is shown by the experiment that the ensemble method is superior to a single method. Wu Hua, et al. [0] carry out the study on the issue of English synonymy discovery, propose to identify synonymy through the comprehensive use of monolingual dictionaries, bilingual corpora and large-sized monolingual corpora, process the similarity values computed respectively by means of the three kinds of corpora through weighted addition, and finally work out the similarity value between words. This method avoids the defects caused by a single corpus. In order to overcome the shortcoming that a single model only applies to synonym extraction in a specific field, Henriksson, et al. [0] combine different distribution models, then apply the model respectively to monolingual and bilingual corpora, and perform synonym identification by the integration strategy of adding cosine values up. The experiment proves that the combined model is superior to a single model. By means of similarity measure based on WordNet, similarity measure based on window co-occurrence, similarity measure based on syntactic co-occurrence and web-based mutual information measure, Neshati, et al. [0] express similarity computed by various methods for a pair of words and phrases to a 4-dimensional vector, then use the tourism classification system as a training set, compute similarity between a pair of words in the classification system by the path-based similarity algorithm, and use it as a basis for the neural network model to learn and evaluate the weights of various similarity computing measures. Next, they use the classification system of financial field as a test set, and output a compound similarity measure for the similarity vector generated by each pair of words and phrases according to the weights given in the learning stage of the neural network. Bollegala, et al. ^{0} put forward term similarity computation based on web search engines, respectively use four similarity indexes (Jaccard, Overlap, Dice and PMI) to measure the similarity between a pair of words by the search engine hit number method, and use template matching of n-Gram to measure the similarity between a pair of words and phrases in abstracts. 200 templates with the highest frequency of word pair co-occurrence and 4 similarity computing indexes are selected to form 204-dimensional vectors as input of a two-class SVM classifier, so as to work out the final similarity judgment result. However, a set of 204-dimensional vectors formed by similar and dissimilar pairs of words and phrases extracted from WordNet by the two-class SVM classifier is gained during training as the training corpus.

From the above, it can be learned that using multiple methods has more advantages than using a single method for term relation identification. However, most of researches on multi-strategy term relation identification at home and abroad merely focus on synonymy identification, but researches on identification and subdivision of other relations still need to be improved.

Relation Identification Method Based on Fuzzy Comprehension Evaluation

Principle of Fuzzy Comprehension Evaluation

The fuzzy comprehension evaluation method [13, 14] is a kind of comprehension evaluation method based on fuzzy mathematics. This comprehension evaluation method transforms qualitative evaluation into quantitative evaluation according to the membership grade theory of fuzzy mathematics. In other words, it uses fuzzy mathematics to make an overall evaluation on things or objects constrained by multiple factors. First, determine a fuzzy set (factor set U) composed of multiple factors, set review ratings that can be selected for these factors, form a fuzzy set of comments (evaluation set V), respectively work out the membership grade of each single factor to each review rating (fuzzy matrix), and then work out the quantitative solution of evaluation according to the weight distribution of various factors in the evaluation target through computation (fuzzy matrix composition). The process above is exactly fuzzy comprehension evaluation.

Design of Fuzzy Comprehension Evaluation Method

The fuzzy comprehension evaluation method relies on the membership grade theory of fuzzy mathematics. The membership grade is usually determined by synthetic weighted method, comparison ordering method, assignment method, fuzzy statistical method, etc. The process of membership function determination shall be objective in essence, but all of these methods involve certain manual intervention. Therefore, in order to improve the automation degree of membership function determination, this paper introduces the continuous attribute discretization method [15, 16] to divide value ranges of various factors into different intervals, and determines the membership grades corresponding to evaluation sets of various factors in different intervals according to the sample distribution probability. Weights of evaluation factors are usually set by Delphi method, analytic hierarchy process method, etc., but all of these methods have certain subjectivity, and different weight settings may affect the final result of fuzzy comprehension evaluation. Therefore, in order to enhance the objectivity of weighting, this paper trains the training set by particle swarm optimization [17] and cross validation methods, so as to gain optimum weights of various factors.

The computation process of fuzzy comprehension evaluation method is shown in Figure 1:

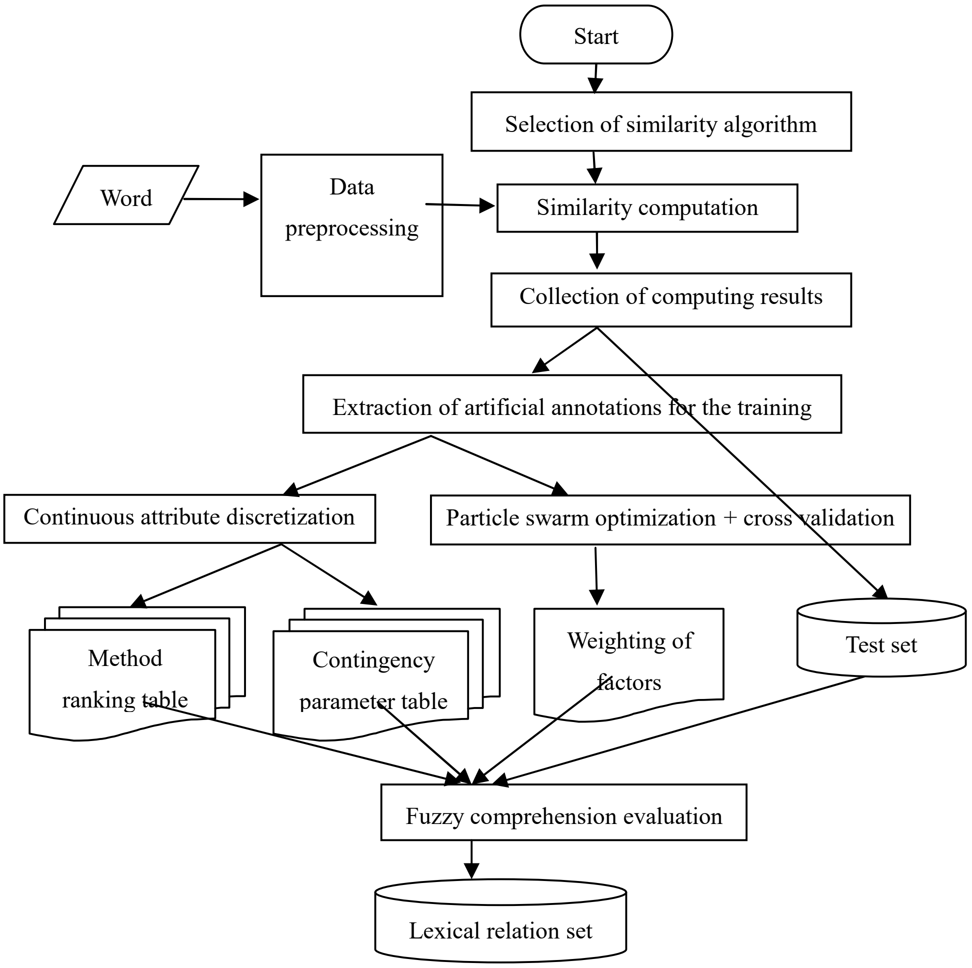
Start
Selection of similarity algorithm
Word
Data preprocessing
Similarity computation
Collection of computing results
Extraction of artificial annotations for the training
Continuous attribute discretization
Particle swarm optimization + cross validation
Method ranking table
Contingency parameter table
Weighting of factors
Test set
Fuzzy comprehension evaluation
Lexical relation set

Figure 1 Computation Process of Fuzzy Comprehension Evaluation Method

Experiment and Result Analysis

Data Source and Preprocessing

Terms and definitions in the "natural disaster risk monitoring scientific and technical word system" [16] are selected as corpora. By January, 2016, there had been 4,943 words and 24,744 term definitions in total (one term may involve two definitions and above). The terms and definitions are processed by denoising and segmentation, and stored in the database for use.

Similarity Algorithm Experiment

Algorithms based on Hownet, literal similarity and Cilin are respectively selected for computing the similarity value of term pairs. The methods based on Hownet and Cilin are referred to in the book of Xia Tian [17], and the literal similarity algorithm adopts the classical weighted similarity computation algorithm proposed by Hou Hanqing [18]. Next, the similarity computing results by the three methods are collected and processed to work out a group of similarity values for each pair of terms, as shown in Table 2:

Table 2 Collection of Results from Three Similarity Computing Methods

term1	term2	sim1	sim2	sim3
Subsidence	Subsidence area	0.000	0.000	0.833
Subsidence	Subsidence-induced earthquake	0.000	0.000	0.750
Frontal fog	Sea fog disaster	0.000	0.000	0.208
Cold wave	Freezing damage	0.114	0.000	0.000
Frost damage	Cold wave	0.369	0.000	0.000
Forecast	Early warning	0.552	0.000	0.500
Maritime Code	Marine losses	0.204	0.000	0.417
Water resources	Fresh water	0.896	0.000	0.250
Tornado	Wind damage	0.109	0.000	0.083
Peach blossom flood	Spring flood	0.167	1.000	0.250
Rainstorm	Meteorological disaster	0.000	0.000	0.000
Social emergency strength	Social emergency potential	0.000	0.000	0.833

Experiment of Fuzzy Comprehension Evaluation

In the fuzzy comprehension evaluation, the factor set U={sim1，sim2，sim3} where sim1 is the gained similarity value based on the definition matching algorithm, sim2 is the gained similarity value based on the literal similarity algorithm, and sim3 is the gained similarity value based on the pattern matching algorithm. Considering it is determined that the lexical relations to be identified are respectively mapped and summarized to four types of synonymy, hierarchical relation, homogeneous relation and no relation, the evaluation set V={v_1, v_2, v_3, v_4}. First, randomly draw 500 samples as the training set and 200 samples as the test set from the collected data, and make artificial annotations (see Table 3) for relations of term pairs according to the relations to be identified. Next, train the training set by the fuzzy comprehension evaluation method, so as to work out factor ranking table and contingency parameter table for a single similarity computing method as well as weights of various factors. Finally, work out the relations corresponding to various terms (see Table 4).

Table 3 Term Relation Annotation Result

relation-id	term1	term2	sim1	sim2	sim3
1	Emergency response	Emergency management	0.000	0.000	0.531
1	Emergency recovery	Emergency management	0.000	0.000	0.500
3	Torrential flood and debris flow	Rainstorm	0.000	0.000	0.000
0	Pestilence	Pandemic disease	1.000	0.000	0.000
1	Volcano	Extinct volcano	0.000	1.000	0.500
2	Geological disaster	Marine disaster	0.000	0.000	0.500
1	Optical radiation	Radiation	0.000	0.786	0.500
3	Periglacial tor	Weathering	0.000	0.000	0.000
0	Emergency fund	Rescue fund	0.000	0.000	0.500
0	Rescue	Emergency rescue	1.000	0.786	0.500

In Table 3, 0 represents hierarchical relation; 1 represents homogeneous relation; 2 represents synonymy; 3 represents no relation. term1 and term2 refer to a pair of terms. sim1, sim2 and sim3 are similarity values computed by the three methods.

Table 4 Relation Identification Result Based on Fuzzy Comprehension Evaluation

relation_id	term1	term2
3	Subsidence	Subsidence area
2	Subsidence	Subsidence-induced earthquake
1	Frontal fog	Sea fog disaster
2	Cold wave	Freezing damage
2	Frost damage	Cold wave
0	Forecast	Early warning
3	Maritime Code	Marine losses
0	Water resources	Fresh water
1	Tornado	Wind damage
0	Peach blossom flood	Spring flood
3	Rainstorm	Meteorological disaster
0	Social emergency strength	Social emergency potential
1	Emergency plan	Emergency operation manual

Result Analysis

The relation identification result based on fuzzy comprehension evaluation is evaluated by precision and recall. Precision refers to the proportion of correctly identified term relations in total identified term relations of a certain type; recall refers to the proportion of correctly identified term relations in total identified term relations of a certain type existing in the test set.

$$\text{Precision}=\frac{\text{correctly found relation} < \text{term1}, \text{term2} >}{\text{total found relation} < \text{term1}, \text{term2} >}$$

$$\text{Recall}=\frac{\text{correctly found relation} < \text{term1}, \text{term2} >}{\text{total relation} < \text{term1}, \text{term2} >}$$

$$\text{F1-measure}=\text{F1}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

Table 5 Term Relation Identification Effect

Relation type	Identified number	Correct number	Number in test set	precision	recall	F1
Hierarchical relation	54	34	63	62.96%	53.97%	0.581
Homogeneous relation	82	51	58	62.20%	87.93%	0.729
Synonymy	31	23	45	74.19%	51.11%	0.605
No relation	33	21	34	63.64%	61.76%	0.627
Total	200	129	200	63.00%	63.00%	0.630

It can be seen from Table 5 that the fuzzy comprehension evaluation has relatively good relation identification effect with relatively high precision and recall for identification of various relations, among which the maximum precision is 74.19% for synonymy identification, the minimum precision is 62.20% for homogeneous relation identification, the maximum recall is 87.93% for homogeneous relation identification, and the minimum recall is 51.11% for synonymy identification. According to the F value, this method has the best effect on synonymy identification. In addition, considering this method is affected by the selected similarity computing method and the number of similarity computing methods, we'll add more similarity computing methods to the experiment or select several good methods for integration in the follow-up research.

Conclusions

Based on multi-strategy fuzzy comprehension evaluation, this paper puts forward the term relation identification method by integrating similarity algorithms, which can make up the shortcomings of a single similarity computing method and gather the advantages of multiple algorithms. In addition, focusing on term relation identification, this method has the advantages such as dynamic adjustment of parameter basis input, objectification of weight distribution and subdivision of identified relations.

References:

[1] Zhang Wei, Yu Yang, You Hongliang. Relation Identification between Conceptual Terms for Automatic Construction of Lexical Knowledge Database [J]. New Technology of Library and Information Service, 2009 (11): 10-16.

[2] Lu Yong, Zhang Chengzhi, Hou Hanqing. Using Multiple Hybrid Strategies to Extract Chinese Synonyms from Encyclopedia Resources [J]. Journal of Library Science in China, 2010, 36 (185): 56-62.

[3] Ma Haichang, Zhang Zhichang, Zhao Xuefeng, et al. Synonym Extraction with Method of Combination of Latent Semantic Analysis and Pointwise Mutual Information [J]. Computer Knowledge and Technology, 2014, 10 (1): 128-132.

[4] Ma Haichang, Zhang Zhichang, Zhao Xuefeng, et al. Research on Syncretic Synonym Acquisition for Economic Field [J]. Science Technology and Engineering, 2014, 14 (15): 207-211.

[5] Liu Wei, Huang Xiaojiang, Wan Xiaojun, et al. Study on Automatic English Synonym Terms Discovery from Web and the System Implementation [J]. Library and Information Service, 2012, 56 (22): 26-31.
[6] Curran J. Ensemble Methods for Automatic Thesaurus Extraction. In Proc. of the Conference on Empirical Methods in Natural Language Processing. 2002: 222-229.
[7] Lin D. Dependency-based evaluation of MINIPAR[C]. In Workshop on the Evaluation of Parsing Systems, Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 1998: 234-241.
[8] Grefenstette G. Explorations in Automatic Thesaurus Discovery [M]. Kluwer Academic Publishers, Boston, USA, 1994.
[9] Wu Hua, Zhou Ming. Optimizing Synonym Extraction Using Monolingual and Bilingual Resource[C]. In: Proceedings of the Second International Workshop on Paraphrasing. Stroudsburg: Association for Computational Linguistics, 2003: 72-79.
[10] Henriksson A, Hans Moen, Skeppstedt M, et al. Synonym extraction and abbreviation expansion with ensembles of semantic spaces [J]. Journal of Biomedical Semantics, 2014: 1-25.
[11] Neshati M, Hassanabadi LS. Taxonomy Construction Using Compound Similarity Measure [J]. Lecture Notes in Computer Science, 2007(4803): 915-932.
[12] Bollegala D, Matsuo Y, Ishizuka M. Measuring Semantic Similarity between Words using Web Search Engines [C]. Banff, Alberta, Canada: International World Wide Web Conference Committee, 2007:757-766.
[13] Li Hongxing, Wang Peizhuang. Fuzzy Mathematics. Beijing: National Defense Industry Press, 1994. 2
[14] Xie Jijian, Liu Chengping. Fuzzy Mathematical Methods and Its Application [M]. Wuhan: Huazhong University of Science & Technology Press, 2005: 31-36.
[15] Chen Bingzheng, Han Chunpeng. Discretization for Inductive Learning from Continuous Sample Data [J]. Systems Engineering-- Theory & Practice, 2001 (4): 1-7.
[16] He Defang, Qiao Xiaodong, Zhu Lijun, et al. Chinese Scientific & Technical Vocabulary System (Major Natural Disasters Monitoring and Defense) [M]. Beijing: Scientific and Technical Documentation Press, 2014.
[17] Xia Tian. Similarity Computing Theories & Methods of Chinese Information [M]. Zhengzhou: Henan Science and Technology Press. 2009.
[18] Hou Hanqing, Wu Zhiqiang. Experiment on Chinese Synonym Identification with Literal Similarity [C]. Informationization of Information Service Industry, 2001: 222-229.