ATLANTIS PRESS

Advances in Engineering Research, volume 140

2nd International Conference on Electrical, Control and Automation Engineering (ECAE 2017)

Synonym Relationship Discovery for Knowledge Organization System Based on Paired Translation Information

Yunliang Zhang[1,2,*]

[1]Institute of Scientific and Technical Information of China

[2]Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, SAPPRFT, China

[*]Corresponding author

***Abstract*—Synonym relationship is one of the most important relationships for knowledge organization systems. In this paper, we propose a series of methods on synonym relationship discovery based on paired translation information already in knowledge organization systems. The translations amount and the translation concurrence strength are inputs of similarity, which then calculated by Boolean comparison, conditional mutual information and cosine similarity of vector space model. With the output of these methods, synonym relationships are discovered and constructed. Precision and richness are used as the indexes to evaluate the methods. With an experiment in new energy vehicles domain, it is found that the effects of the result are acceptable and have the potential to be improved by more translation pairs.**

Keywords—knowledge organization system; Chinese scientific and technical vocabulary system; synonym relationship discovery; paired translation information

I. INTRODUCTION

It is very important for the construction and maintenance of knowledge organization systems about the application range and generalization performance in specific intelligence application, but there are at least two problems in knowledge organization system development. Firstly, knowledge organization systems are knowledge intensive and strongly relied on the experts of specific domain, and it should take a lot of time and expense to deal with the increasing need in intelligence analysis, which is severe for the conditions of new emerging or unexpected demands. Secondly, the update speed of knowledge organization system is very slow, and the most fast update speed we know up to now is a time per week, which belongs to the Library Congress Classification [1]. Most knowledge organization systems update in several years even never update in the whole life cycle, which leads to bigger and bigger gap from the knowledge organization systems to the advances of the disciplines. Thus, to cope with the problems, we should develop some methods to construct and maintain the knowledge organization systems with low cost, especially on automation method.

Different knowledge organization systems have different knowledge infrastructures, but for most of them, synonyms are compulsory. For that in natural language, there are too many synonyms because of the complexity of language itself. By synonym relationship discovery, the synonyms are clustered and which make language understanding easier. So synonym relationship discovery is the key problem of the construction and maintenance of the knowledge organization systems and which can give the whole construction work a leg up.

In addition, as an international language, English is very popular in academic interchange, and Chinese knowledge organization systems , especially in science and technology domains, usually have English translations of Chinese terminologies. Then to discover the synonym relationships between Chinese terminologies with the paired translation information is well reasoned. In this paper, we use the paired translation information from Chinese scientific and technical vocabulary systems that we developed to study and analyze the methods to know the feasibility on technologies, and to provide the theory and sample basis for development of the semi-automatic or automatic discovery tools.

Synonym relationship discovery is a task of some disciplines such as linguistics, intelligence sciences and natural language processing and so on. In different discipline, the definition and constraints of synonym are different. In intelligence sciences and related knowledge organization systems, the definition is relatively broad [2]. The basis of synonym relationship discovery is corpus, usually including normalized dictionaries or definitions from real texts [3-6]. Of course, the explanatory knowledge segments of wiki or general language resources [7][8]. There are also some researchers use aliened bilingual corpus for statistics and further analysis [5], and obtain the Chinese terminology probability vectors of English terminologies, then get the similarity of English terminologies by the calculation of the vectors. On algorithms, different similarity or semantic distance calculation methods are adopted. The most common algorithm is to calculate the cosine similarity, of course, the Jaccard distance, Mahalanobis distance [9] and Dice distance [5] is also effective. And there is also similarity calculation based on graph and supervised machine learning and so on [2][10][11]. These previous researches give us the basic method and algorithm hints of synonym relationship discovery, though the inputs are different.

II. METHOD ON SYNONYM DISCOVERY AND ITS EVALUATION

A. Method on Synonym Discovery

The basic inputs are the Chinese and English translation pairs. Assume that there are two Chinese terminologies c1 and c2, each of them have one or more same English translations, then c1 and c2 may have synonym relationship. In fact in linguistics, c1 and c2 are called translation equivalent [12], and because only a part of translation equivalents are synonyms, we also use the amount of English translations of Chinese terminologies and the frequency of the concurrence of the same translation of different terminologies, which we use a new terminology of translation concurrence strength to name and use a variable ρ to represent it. In an ordinary knowledge organization system, ρ is usually very small. But the background of this paper is a knowledge organization system that called Chinese scientific and technical vocabulary systems (CSTVS) developed by Institute of Scient8ific and Technical Information of China. In the construction process, we consider that English is not the native language of the users and the knowledge workers and experts, and the mastery and use of English is very difficult, so we use 2 or more knowledge workers and experts to give and check the translations and it is permitted that one Chinese terminology can have more than one English translations.

In detail, suppose that Chinese terminology c1 has m pieces of English translations as e11; e12 …e1m, and c2 has n pieces of English translations as e21; e22 …e2n. The similarity function f of c1 and c2 can be calculated by formula (1), in which function f' is the similarity of two English terminologies and there are a lot of algorithms can be used to calculate it. A basic method called Method A is Boolean algorithm that shown with formula (2), that is, if two English translation e1i and e2j are equal, the value of function f' is 1, and the value is 0 otherwise. Of course there are more complicated algorithms, one of which is indicated in formula (4). That is, in an English thesaurus or dictionary, if e1i and e2j are synonyms or sometimes may be expanded the scope to quasi-synonyms, the value of function g is 1, and the value is 0 otherwise.

$$f(c_1, c_2) = \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} f'(e_{1i}, e_{2j}) \tag{1}$$

$$f'(e_{1i}, e_{2j}) = \begin{cases} 1, e_{1i} = e_{2j} \\ 0, e_{1i} \neq e_{2j} \end{cases} \tag{2}$$

$$f'(e_{1i}, e_{2j}) = g(e_{1i}, e_{2j}) \tag{3}$$

$$g(e_{1i}, e_{2j}) = \begin{cases} 1, e_{1i} \cong e_{2j} \\ 0, else \end{cases} \tag{4}$$

In consideration of the respective translation amount of c1 and c2 and refer to the conditional mutual information formula [8], replace the variables with attributes of c1 and c2, we can have a new formula as indicated in formula (5). Formula (6) is the equivalent transformation of formula (5), and in which N is the amount of samples. In fact, N is a fixed value in some specific condition and not affect the comparison of function I, so in this paper, we use function I' as indicated in formula (7) to replace function I and the method is called Method B.

$$I(c_1 \mid c_2) = I(c_2 \mid c_1) = \log\left(\frac{Nf(c_1, c_2)}{f(c_1, c_1) f(c_2, c_2)}\right) \tag{5}$$

$$\log\left(\frac{Nf(c_1, c_2)}{f(c_1, c_1) f(c_2, c_2)}\right) = \log N + \log\left(\frac{f(c_1, c_2)}{f(c_1, c_1) f(c_2, c_2)}\right) \tag{6}$$

$$I' = \log\left(\frac{f(c_1, c_2)}{f(c_1, c_1) f(c_2, c_2)}\right) \tag{7}$$

In addition, VSM (Vector Space Model) can also be used for the representation of Chinese terminologies [5]. That is, construct a m-dimension vector space (e1, e2, e3 …em) with all English terminologies used to translate the Chinese terminologies. On the basis, every Chinese terminology can be represented in the vector space with the value of every dimension is 1 or 0. The similarity then can be calculated with cosine algorithm as indicated in formula (8) and the method is called Method C.

$$\cos(c_1, c_2) = \frac{\sum_{1 \le k \le m} e_{1k} e_{2k}}{\sqrt{\sum_{1 \le k \le m} e_{1k}^2} \sqrt{\sum_{1 \le k \le m} e_{2k}^2}} = \frac{f(c_1, c_2)}{\sqrt{f(c_1, c_1) f(c_2, c_2)}} \tag{8}$$

With the similarity calculation algorithms, we can get the positive examples and negative examples of different ρ based on training set. With the average value of the examples, the threshold values δ can be conducted and series two-class classifiers will be built. With the classifier the synonym relationship can be recommended.

B. Method on Result Evaluation

In some text processing applications such as information retrieval, text classification and information extraction, precision and recall are commonly used indexes to evaluate the results. But in this condition, no one knows the exact synonyms of a terminology, so there are no meaningful recall evaluation values. So precision is the selected index to evaluate the results. In consideration of that the key problem of this paper is a two-class classification problem, the calculation of the precision as indicated in formula (9), in which Nrr is the amount of the synonym pairs judged by algorithms and in fact correct, and Nee is the account of the non-synonym pairs judged by algorithms and in fact correct, and Nall is the account of all pairs that need to be judged with at least a common translation.

$$p = \frac{N_{rr} + N_{ee}}{N_{all}} \tag{9}$$

In addition, a new index named richness should be included in to evaluate the relative ration of the final correct synonyms discovered to the amount of Chinese terminologies in these candidate pairs. In condition of different using scenarios such as initial construction and revision, we use two different methods to calculate as indicated in formula (10) and (11), in which Nc is the number of Chinese terminologies in these candidate pairs and Nrn is the new discovered synonyms not included by the original knowledge organization system.

$$q = \frac{N_{rr}}{N_c} \tag{10}$$

$$q' = \frac{N_{rn}}{N_c} \tag{11}$$

III. EXPERIMENTAL STUDY

A. Data Sources and Scale

The experimental data is divided into two parts, that is, training set and test set. The test set is the Chinese –English terminology pairs collected from the new energy vehicles domain CSTVS [13]. Until Jan. 2013, the core terminology number of this CSTVS is 6126 and the amount of all the English translation is 13190. Use the query processing in a MySQL database, there is 599 distinct Chinese terminologies have equivalent English translation and the amount of Chinese-English terminology pairs is 1000. The amount of translations of these Chinese terminologies ranged from 1 to 8. After a merge wok of the 1000 pairs, we get 826 distinct synonym candidates and the translation concurrence strength ranged from 1 to 5. The training set is selected from another CSTVS that in cleaning energy domain with positive and negative examples half to half, but there is no example that has the translation concurrence strength of 4 or 5. The detailed amount of training set and test set is listed in table Ⅰ.

TABLE Ⅰ. AMOUNT OF EXAMPLES IN TRAINING SET AND TEST SET WITH DIFFERENT TRANSLATION CONCURRENCE STRENGTH

Translation concurrence strengthρ	Example numbers	
	Training set	Test set
Total	150	826

The thresholds of similarity under different ρis different, so we use a piecewise function to describe the δ. In consideration of the distribution of the training set and test set, we use a 3-piece function as indicated in formula (12).

$$\delta = \begin{cases} \delta_1, \rho = 1 \\ \delta_2, \rho = 2 \\ \delta_3, \rho \geq 3 \end{cases} \tag{12}$$

It is predictable and verified by experiment that the average similarity of positive examples is larger than that of negative examples on condition of the same ρ. But for a specific example, the opposite may appear. In this paper, we use the average similarity value of negative examples as the thresholds. Thresholds of different similarity calculation functions are shown in table II.

TABLE II. δ VALUE OF DIFFERENT SIMILARITY FUNCTION WITH THE PVARIES

Similarity function	δvalue		
	δ_1	δ_2	δ_3
f	1	2	3
I’	-0.944	-0.795	-0.634
cos	0.36	0.586	0.836

B. The Experiment Results

After a manual judge, there are 500 Chinese-Chinese pairs are synonyms, and the other 326 pairs are non-synonyms as listed in table III.

TABLE III. δ THE REAL DISTRIBUTION OF TEST SET

ρ	synonyms	Non-synonyms	All in test set
1	389	293	682
2	90	30	120
3	17	3	20
4	2	0	2
5	2	0	2
Total	500	326	826

With the δ values in table 2, judge the test set and get the precision and richness of different methods as shown in table IV, table V.

TABLE IV. PRECISION VARIES WITHPBASED ON METHOD A (F) ON TEST SET

ρ	1	2	3	4	5
Precision	57%	75%	85%	100%	100%

TABLE Ⅴ. PRECISION & RICHNESS OF DIFFERENT METHODSAND RELATED DATA

method	Precision	Richness q	Richness q’
A(f)	60.53%	**83.47%**	**53.59%**
B(I’)	**64.53%**	56.42%	34.22%
C(cos)	61.99%	48.91%	30.22%

C. Results Analysis

Through the comparison of the results from table4, table5 and table6, it is found that precision increase with the translation concurrence strength, and with method A, high richness and acceptable precision are got, and with method B

and C, the precision will increase comparably and the richness decrease. Method B gets the highest precision in all three methods.

Of course, the richness will increase with the increase of the number of translations of a specific Chinese terminology. But can the precision get a great increase? We do an experiment to answer the question.

It is known that the main inputs of the three methods are numbers of common English translations, translations of Chinese terminology c1 and c2. We use variables x1, x2 and x3 to represent these numbers and use the perceptron algorithm to find a best classifier. The main formula of perceptron algorithm is indicated in formula (13), with an initial vector of w= (w0, w1, w2, w3) and a lot of times of iteration, we can get the final vector as the classifier. And then we can judge with the output y value with the input variables. In the succeeding experiment, it is found that after iteration more than 10,000 times, still there is no convergence. So about the test set, maybe we need more time to get a better convergence result, but perhaps it is more likely to be linear inseparable. But with the intermediate results of the vector, the judge results are very similar in number. Randomly Select one of the vectors from the calculation process as w=(-7.0653,6.4693,1.1463,-7.0653), and we can get that Nrr is 390, Nee is 92, the precision is 58.35%, which is less than any of the three methods. So we can know that the three methods we adopted are relatively high in precision.

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \tag{13}$$

$$sim(c_1, c_2) = \begin{cases} 1, y \geqslant 0 \\ 0, y < 0 \end{cases} \tag{14}$$

IV. CONCLUSION

With the three methods of similarity calculation, we can discover synonym relationship to construct or update a knowledge organization system to some extent. The precision of the synonym discovery increase with the translation concurrence strength and the richness increase with the number of the English translations of candidate Chinese terminologies. Through an experiment on new energy vehicles CSTVS, the results of the three methods of Boolean algorithm, conditional mutual information and cosine similarity on vector space model are acceptable, which can fit different scenarios of Chinese terminology pairs whether they are similar in form or not. But from an overview, the amount of synonyms discovered by these methods is still not very large. So other methods such as concurrence in corpora, definition analysis and so on should be used together. Of course, the precision and richness have the space to shift, which is our future work.

ACKNOWLEDGMENT

This work is partially supported by National Science and Technology Support Program (Grant No. 2015BAH25F01) and CKCEST Project Program (Grant No. CKCEST-2017-1-12). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Ma Zhanghua, Hou Xiaoqiong, Development Overview of Management and Maintenance of the Major Foreign Classifications under Network Environment (In Chinese), In Research Report on Classification Revision , pp.1-10, 2007

[2] Zhong Weijin, Automatic Recognition of Synonymous about Keyword-Descriptor Co-occurrence Based on Comparative Analysis Between Mutual Information Method and Probability Method (In Chinese), Library and Information Service, Vol. 56(18), pp.122-126, 2012

[3] Falk I, Gardent C, Jacquey E, et al, Grouping Synonyms by Definitions. In Proceedings of International Conference RANLP, pp.76-81, 2009

[4] Muller P, Hathout N, Gaume B. Synonym Extraction Using a Semantic Distance on a Dictionary, In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing pp. 65-72, 2006

[5] Wu Hua, Zhou Ming. Optimizing Synonym Extraction Using Monolingual and Bilingual Resources, IN Proceedings of the Second International Workshop on Paraphrasing, pp. 72-79, 2003

[6] Zhang Yunliang, Liang jian, Zhu Lijun, et al. Key Techniques Study on Automatic Enrichment of KOS from Scientific and Technical Definitions (In Chinese), Library and Information Service, Vol. 54(7/8), pp. 66-71, 2010

[7] Liu Jiangming, Xu Jinan, Wu peihao, et al., Automatic Acquisition of Lexical Semantic Relationship based on Web Resource (In Chinese) In http:// tcci.ccf.org.cn/conference/2012/dldoc/NLPCC2012papers/workshoppapers/sem/001.pdf , pp .1-7, 2012

[8] Oleksandr Grushetsky, Steven D. Baker, Document-based Synonym Generation: United States, US 008161041, pp.1-12,2012

[9] Shimizu N, Hagiwara M, Ogawa Y, et al, Metric Learning for Synonym Acquisition, In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop, pp. 793-800,2008

[10] Minkov E, Cohen W W, Graph Based Similarity Measures for Synonym Extraction from Parsed Text, In Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing, pp. 20-24, 2012

[11] Sun xia, Dong lehong, Automatic Extraction of Synonymy Relation Using Supervised Learning (In Chinese) , Journal of Northwest University (Natural Science Edition Vol.38(1), pp. 35-39, 2008

[12] Su yingxia, The several discrimination methods on translation equivalent (In Chinese), Chinese language Study vol 2, pp. 60-62, 2000.

[13] He defang, Qiao Xiaodong, Zhu Lijun, et al. Chinese Scientific and Technical vocabulary System–New Energy Vehicles (In Chinese), Scientific and Technical Documentation Press , China, 2012