ICIC Express Letters
Volume **10**, Number **8**, August **2016**

ICIC International ©2016 ISSN 1881-803X
pp. 1837–1842

ONTOLOGY CONCEPT POPULATION AND NON-TAXONOMIC RELATION LEARNING BASED ON DEPENDENCY SYNTACTIC INFORMATION

Jing Qiu[1], Ke Jia[1], Jingfang Su[1] and Junkang Hao[2]

[1]Department of Information Science and Engineering
[2]School of Animation
Hebei University of Science and Technology
No. 26, Yuxiang Street, Shijiazhuang 050018, P. R. China
qiujing.ch@gmail.com; jiake_2004@sina.com; suejingfang@foxmail.com
haojunkang@hebust.edu.cn

Received January 2016; accepted April 2016

Abstract. *Most of the ontology learning research works focus on extraction of concepts and taxonomic relations. This paper presents a method to use dependency syntactic information together with statistic information to automatically identify non-taxonomic relations and suggest relation labels for Chinese domain ontology based on Web news corpus. Concept instances are extracted out based on two patterns as the first step. Then system takes the domain corpus, concepts and concept instances as input. The input instances can help to find more objects sentences which may contain relation label verbs. Four different types of verbs are defined to help extract labels of non-taxonomic relations. A confidence score function is defined to compute verb scores. Verbs that contain the same sense are combined into a label set, and the sum of each verb score in the set acts as the label score. All label sets of each non-taxonomic relationship are ranked according to the label scores. Experimental results show the effectiveness of our approach.*

Keywords: Ontology learning, Concept population, Non-taxonomic relation learning, Dependency structure

1. Introduction.

Ontology plays an important role in fulfilling semantic interoperability, and it is the core of Semantic Web. Therefore, in recent years many research works focus on ontology building, learning and population.

Ontology learning (OL) and population aims at automatic or semi-automatic ontology construction, which can save more time and resources than manual ontology building. Natural language processing, machine learning, information extraction and text mining methods are often combinedly used in three main tasks: concept extraction, taxonomic relation extraction, and non-taxonomic relation extraction. The data source for ontology learning can be structured, semi-structured, and non-structured data. Non-structured data can be easily obtained from WWW, and lots of useful information and semantic information is hidden in the natural language texts. Information extraction (IE) techniques can automatically extract information from text, such as identifying named entities, and finding various predefined semantic relations between pairs of entities, which are the very similar tasks with the OL. Therefore, IE techniques can naturally be used for non-structured data OL.

There are already many ontology learning tools and systems available now [1-6]. Methods in the fields of information extraction, machine learning, and natural language processing are often used or combinedly used to solve ontology learning problem. All the methods can be classified into three types, statistics-based method, linguistic method, and hybrid method. Statistics-based methods, such as clustering and Latent Semantic

Analysis, are often used for the tasks of concept extraction and taxonomic relation extraction [7]. Linguistic-based methods include part of speech (POS) tagging, syntactic structure analysis, and language model, and can be used in all sub-tasks of ontology learning [8]. In recent years, hybrid-based methods are widely used in ontology learning, especially in the complex tasks such as non-taxonomic relation learning and axioms learning [9-11].

Non-taxonomic relation learning is considered to be one of the most challenging tasks in ontology learning, and is often neglected. Wong et al. [9] present an ontology learning framework based on a multi-phase correlation search strategy to learning non-taxonomic relations. Concept pairs are allowed to be located in different sentences. Association rule mining which is one of the most well-known data mining techniques is used to identify concepts. Pattern-based method is used to extract relation labels. Serra and Girardi [12] use NLP and statistic methods to extract non-taxonomic relations semiautomatically. Several constraints are defined to extract candidate relationships. Two statistic solutions are used in refinement process. Villaverde et al. [13] propose a semiautomatic method to discover and label non-taxonomic relations. Syntactic structure information and dependency information among concepts are analyzed to find candidate relation pairs and labels. A constraint that two concepts are separated by no more than N terms is made to ensure concept pairs have high probability to have semantic relationship. Association rules are used to suggest the candidate concept relationships. Sanchez and Moreno [14] present a new approach for learning non-taxonomic relations from Web. Domain relevant verbs are extracted out firstly, and then combined with domain key words to construct extraction patterns. The verbs act as the labels of non-taxonomic relationshiops naturally. They also propose an evaluation method that evaluates the results against WordNet.

This paper proposes an automatic method for concept population and non-taxonomic relation learning from Chinese Domain Web news corpus. A set of domain-relative concepts and domain-specific texts corpus are collected as input of the system. Firstly, concept instances are extracted in concept population process. Then, concept pairs which may have non-taxonomic relationships are identified using statistic information. Finally, verbs are extracted out and act as the labels of the non-taxonomic relations based on dependency syntactic information.

The remainder of this paper is organized as follows. Section 2 describes the methods that are used in our system. Section 3 discusses the evaluation results. Concluding remarks are made in Section 4.

2. **Our Method.**

2.1. **Dependency tree structure.**

In this paper, dependency parser is used to parse the sentences, and to find the dependency relationship between two words. Dependency grammar can describe the relationship of two words directly. And the dependency relationship has direction: a word depends on another, except the root word of the sentence. Dependency grammar emphasizes the relationship between words. And dependency relation types can be mapped into semantic expression naturally.

2.2. **Concept population.**

Domain concepts and concept instances are two different things. Some research works which focus on concept relations extraction often collect domain relative concepts and concept instances as input, since they can help to capture more candidate concept pairs by using both concepts and instances. However, constructing the domain concepts set is already not a trivial task, and it will be a very tedious process to construct concept instances set manually. Therefore, the first step of this project is to extract concept instances automatically.

We use a pattern-based linguistic method to extract concept instances. Two patterns are built based on word POS and dependency relationship.

The first pattern is called ATT-based pattern, where ATT is the dependency relation type name of "attribute". A string I will be extracted out if $I = i_1, i_2, \ldots, i_n$ is a substring of sentence with the beginning or ending word to be a certain concept, and all the other words have ATT dependency relationship with concept word. The second pattern is called SBV_VOB-based pattern, where SBV and VOB are the dependency relation type names of "subject-verb" and "verb-object" respectively. In this pattern, concept instance contains just one word; instance and concept are connected by SBV and VOB relations, and act as subject and object in the string I respectively. At the same time, all the extracted words should be nouns, but not labeled by POS tag "n", since we believe the concept instances are different from the general nouns words. Therefore, we only focus on the following POS tags: "ni", organization name; "nl", location noun; "ns", geographical name; "nt", temporal noun; "nz", other proper noun; "nd", direction noun; "nh", person name. For each concept, many candidate instances can be extracted out and divided into several groups according to the different POS tags described above. The group that has the biggest amount is identified as the real concept instances set.

Figure 1 and Figure 2 show the examples of these two patterns. Concepts are marked by dotted box. Chinese sentence is translated into English word by word without thinking about grammar.

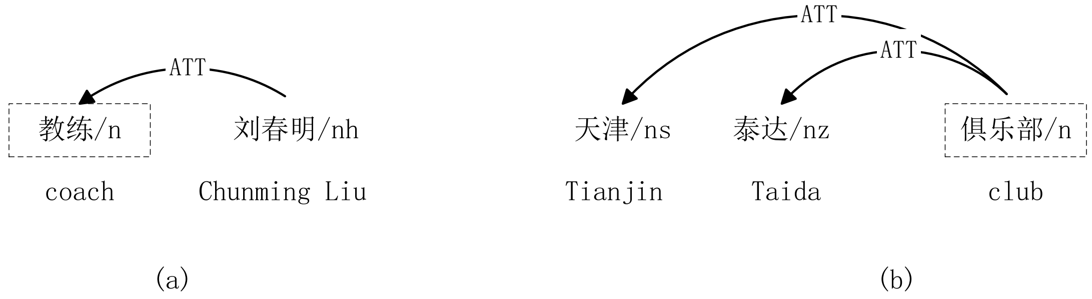

Figure 1. Example of ATT-based pattern

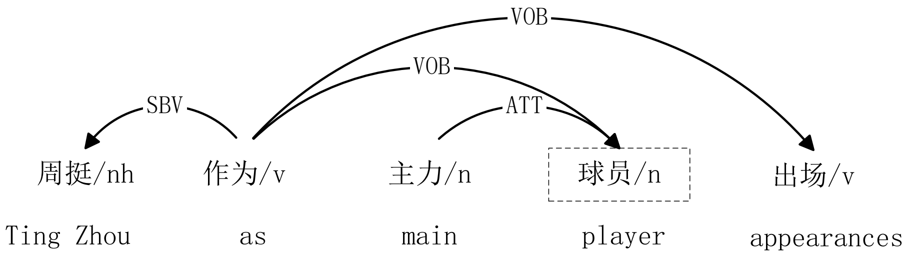

Figure 2. Example of SBV_VOB-based pattern

2.3. **Candidate relationship identification.** A simple statistics-based method is used to identify the candidate concept pairs. Sentences are considered to be analysis units. Two concepts are identified as a candidate relationship when the frequency of their co-occurrence in the same sentence is above a given threshold. Synonym sets are used instead of concepts to increase the matching rates.

2.4. **Non-taxonomic relation label extration.** Verbs in the sentence often act as the relation labels of non-taxonomic relations. In this paper, dependency structure information is used to help find out the most appropriate verb labels. Because we believe verbs that have dependency relationship with concepts have more important role than the verbs just located between two concepts. In order to capture more verbs, we collect all the concepts and the instances extracted in previous step as the input.

Four different verb types are defined as follows, and four corresponding verb sets are extracted out for each candidate relationship.

VB: Verbs located between the two concepts of a candidate relationship.

OVB: Only verb between the two concepts.

CFV: If the nearest common ancestor is verb in the sub-dependency tree of the concept pair.

CFVB: If the nearest common ancestor is verb and located between the two concepts in the sentence.

Therefore, for each candidate relationship we can obtain a verb set V.

$$V = \{v_1, v_2, \ldots, v_n\} = VB \bigcup OVB \bigcup CFV \bigcup CFVB$$

A score function is defined to compute the confidence score of each v_i.

$$Score(v_i) = \sum_{set \in \{VB \bigcup OVB \bigcup CFV \bigcup CFVB\}} feq_{set}(v_i)$$

where $feq_{set}(v_i)$ is the frequency of v_i in set.

After we get the score for each verb, then the verbs which contain the same semantic sense are combined as a verb label set $L_j \in L = \{L_1, L_2, \ldots, L_m\}$. The confidence score of each label set L_j is computed as the sum of verb scores in the label set.

$$V = \{v_1, v_2, \ldots, v_n\} = \{L_1, L_2, \ldots, L_m\}$$

$$L\_Score(L_j) = \sum_{v_i \in L_j} Score(v_i)$$

All the label sets of each concept pair are ranked according to L_Score. The top 3 label sets are returned to users as the non-taxonomic relation labels.

3. Evaluation.

We focus on the domain of football game. Total 2600 Web documents about the competition news of China Football Association Super League (CSL) are collected as the domain corpus. 42 domain concepts are used as input of non-taxonomic relation learning system. 19 candidate relationships are identified since the co-occurrence frequencies are above 100 which is the threshold value we set. We use HIT-SCIR Language Technology Platform (LTP) [15] as dependency parser. The synonym sets building and semantic sense searching are based on Hownet [16,17] which is a lexical database and semantic repository for Chinese language.

3.1. Concept population results.

We use our method to extract instances for concepts “教练” (“coach”), “球队” (“team”), “球员” (“player”) and “裁判” (“referee”). The extracted results are shown in Table 1. The number of extracted instances contains the redundant items. We only care about the precision of the extraction, since the purpose of instance extraction is to help find out non-taxonomic relation labels. Because of the Chinese grammar habits, most of the time, ATT-based pattern receives better results than SBV_VOB-based pattern.

3.2. Non-taxonomic relation label extraction results.

Table 2 shows the part of non-taxonomic relation learning results. Domain expert was asked to rate the candidate label sets as “good” or “bad”. We can observe that there are often more than one “good” label sets for each concept pair. Using label set can improve the performance of the system, because the results are ranked by label scores instead of single verb scores. For example, the noisy verb “加上” (“add”) has a high verb score, the second high score; however, it is ranked as the third label set according to label scores.

Table 1. Concept population results

Concept	Pattern	#Extracted	#Correct	Precision
教练 (“Coach”)	ATT-based	960	946	98.5%
	SBV_VOB-based	79	72	91.1%
球队 (“Team”)	ATT-based	763	722	94.6%
	SBV_VOB-based	68	66	97.0%
球员 (“Player”)	ATT-based	321	314	97.8%
	SBV_VOB-based	276	255	92.4%
裁判 (“Referee”)	ATT-based	140	138	98.6%
	SBV_VOB-based	22	12	54.5%

Table 2. Non-taxonomic relation learning results

Concept Pair	Label Set	Common Sense	Score	Label Rating
教练, 球队	1. 担任 (“served as”)	担任 (“served as”)	43	Good
“Coach”, “Team”	担任 (“served as”)		28	
	兼任 (“concurrently served as”)		12	
	兼 (“concurrently served as”)		3	
	2. 执教 (“coaching”)	从事 (“engaged in”)	25	Good
	执教 (“coaching”)		15	
	参加 (“participate”)		7	
	参与 (“participate”)		3	
	3. 加上 (“add”)	增加 (“increase”)	20	Bad
	加上 (“add”)		20	
教练, 球员	1. 领 (“lead”)	引导 (“guide”)	31	Good
“Coach”, “Player”	领 (“lead”)		12	
	带领 (“lead”)		10	
	带 (“lead”)		9	
	2. 当 (“as”)	担任 (“served as”)	19	Good
	当 (“as”)		10	
	担任 (“served as”)		7	
	兼 (“concurrently served as”)		2	
	3. 训练 (“training”)	训练 (“training”)	16	Good
	训练 (“training”)		16	
球队, 球员	1. 比赛 (“match”)	比赛 (“match”)	85	Bad
“Team”, “Player”	比赛 (“match”)		85	
	2. 成为 (“become”)	成为 (“become”)	61	Bad
	成为 (“become”)		61	
	3. 引进 (“introduce”)	引进 (“introduce”)	47	Good
	引进 (“introduce”)		47	

4. **Conclusions.** This paper proposes a new method of non-taxonomic relation learning and concept population. Concept instances are extracted out and together with concepts as the input of system to help find more object sentences. Dependency information combined with statistic information is used to find appropriate verbs to act as relation labels. Although the precision of concept population is not very high, it helps to obtain more analysis objects. Therefore, more verbs which may become labels are extracted out to help improve the performance of the system.

In the future, more evaluations and comparisons have to be conducted. And more complex patterns are needed to improve the precision of concepts instance extraction.

Acknowledgment. This paper is supported by the National Natural Science Foundation of China (61300120), Hebei Natural Science Foundation (F2012208016), and Colleges of Science and Technology Research Foundation in Hebei Province (YQ2013032, YQ2014036). We thank HIT-SCIR for providing us with LTP modules.

REFERENCES

[1] A. Maedche and S. Staab, The text-to-onto ontology learning environment, *Proc. of Software Demonstration at the 8th International Conference on Conceptual Structures*, pp.14-18, 2000.

[2] P. Cimiano and J. Volker, Text2Onto: A framework for ontology learning and data-driven change discovery, *Proc. of the 10th International Conference on Applications of Natural Language to Information Systems*, pp.227-238, 2005.

[3] C. Nedellec, Corpus-based learning of semantic relations by the ILP system, Asium, *Proc. of Learning Language in Logic*, pp.259-278, 2000.

[4] M. Shamsfard and A. Barforoush, Learning ontologies from natural language texts, *International Journal of Human-Computer Studies*, vol.60, no.1, pp.17-63, 2004.

[5] P. Velardi, R. Navigli, A. Cucchiarelli et al., Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies, *Ontology Learning from Text: Methods, Applications and Evaluation*, pp.92-106, 2005.

[6] A. Schutz and P. Buitelaar, RelExt: A tool for relation extraction from text in ontology extension, *Proc. of the 4th International Semantic Web Conference*, pp.593-606, 2005.

[7] C. Brewster, S. Jupp, J. Luciano, D. Shotton, R. Stevens and Z. Zhang, Issues in learning an ontology from text, *BMC Bioinformatics*, vol.10, no.5, 2009.

[8] F. Colace, M. De Santo, L. Greco et al., Terminological ontology learning and population using latent Dirichlet allocation, *Journal of Visual Languages and Computing*, vol.25, pp.818-826, 2014.

[9] M. K. Wong, S. S. R. Abidi and I. D. Jonsen, A multi-phase correlation search framework for mining non-taxonomic relations from unstructured text, *Journal of Knowledge and Information System*, vol.38, no.3, pp.641-667, 2014.

[10] A. Weichselbraun, G. Wohlgenannt and A. Scharl, Refining non-taxonomic relation labels with external structured data to support ontology learning, *Journal of Data & Knowledge Engineering*, vol.69, no.8, pp.763-778, 2010.

[11] V. H. Ferreira, L. Lopes, R. Vieira and M. J. Finatto, Automatic extraction of domain specific non-taxonomic relations from Portuguese Corpora, *Proc. of the 12th IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, pp.135-138, 2013.

[12] I. Serra and R. Girardi, PARNT: A statistic based approach to extract non-taxonomic relationships of ontologies from text, *Proc. of the 10th International Conference on Information Technology – New Generations*, pp.561-566, 2013.

[13] J. Villaverde, A. Persson, D. Godoy and A. Amandi, Supporting the discovery and labeling of non-taxonomic relationships in ontology learning, *Journal of Expert Systems with Applications*, vol.36, no.7, pp.10288-10294, 2009.

[14] D. Sanchez and A. Moreno, Learning non-taxonomic relationships from web documents for domain ontology construction, *Journal of Data & Knowledge Engineering*, vol.64, no.3, pp.600-623, 2008.

[15] W. Che, Z. Li and T. Liu, LTP: A Chinese language technology platform, *Proc. of Coling*, pp.13-16, 2010.

[16] Z. Dong and Q. Dong, *Hownet Knowledge Database*, http://www.keenage.com/.

[17] Q. Liu and S. Li, Word similarity computing based on How-net, *Computational Linguistics and Chinese Language Processing*, vol.7, no.2, pp.59-76, 2002.