Wuhan University
Journal of Natural Sciences

2015, Vol.20 No.2, 129-133

Article ID 1007-1202(2015)02-0129-05
DOI 10.1007/s11859-015-1070-4

Automatic Ontology Construction Based on Clustering Nucleus

☐ **ZHAO Ling[1, 2], REN Han[1, 2†], WAN Jing[2]**

1. School of Foreign Languages and Literature, Wuhan University, Wuhan 430072, Hubei, China;

2. Hubei Research Establishment of Language and Intelligent Information Processing, Wuhan University, Wuhan 430072, Hubei, China

Abstract: Ontology construction is the core task of ontology-based knowledge representation. This paper explores a semantic description approach based on primitive structure, which benefits ontological relation description in a more precise and concrete way. In view of primitive structure, this paper introduces an approach to extract primitive structures of words based on a multi-label learning model, correlated label propagation. Also, this paper proposes an approach to recognize clustering nucleuses in word clusters heuristically. By this approach, more precise ontological relations are able to be discovered automatically.

Key words: clustering nucleus; ontology construction; primitive structure; multi-label learning

CLC number: TP 391.1

Received date: 2014-10-03

Foundation item: Supported by the National Natural Science Foundation of China (61402341, 61173095, 61173062), the Major Projects of National Social Science Foundation of China (11&ZD189) and the China Postdoctoral Science Foundation Funded Project (2014M552073)

Biography: ZHAO Ling, female, Associate professor, Ph.D., research direction: applied linguistics and Chinese information processing. E-mail: lingzhao2006 @126.com

† To whom correspondence should be addressed. E-mail: hanren@whu.edu.cn

0 Introduction

Ontology-based knowledge representation refers to taxonomy of knowledge, representing relations among entities, concepts, events and their attributes. As a widely used model in semantic technologies, ontology-based knowledge representation has been much concerned in the research areas of artificial intelligence and computational linguistics [1, 2].

The core task of ontology-based knowledge representation is ontology construction. Currently, research on ontology construction focuses on automatic approaches, most of which employ thesauruses and dictionaries such as WordNet and HowNet to access word hyponymy as ontological relations [3-5]. However, since ontological relations describe common characteristics rather than inclusion relations of word clusters, which can be represented by word hyponymy, such ontological relations are hardly described using only word hyponymy in many cases. Therefore, how to find out common characteristics of word clusters becomes a key problem of ontological relation recognition in automatic ontology construction.

This paper explores common characteristics of word clusters based on primitive, a semantic description approach proposed by Xiao *et al* [6] and Hu [7]. By this approach, ontological relations are observed through a more elaborated view, i.e., primitive structure, and clustering nucleuses of concepts, which represent ontological relations of word clusters, can be identified. This paper also introduces an approach to annotate clustering nucleus of concept, and to recognize clustering nucleus of concept. Using this approach, more precise ontological relations of word clusters can be discovered automati-

cally.

The rest of this paper is organized as follows. Section 1 discusses related work of ontology construction. Section 2 shows the system architecture of the approach. Section 3 describes an analysis as well as identification method of primitive and clustering nucleus. Section 4 discusses the automatic construction approach of ontology based on recognizing clustering nucleus automatically. Finally, a conclusion is drawn in Section 5.

1 Related Work

Generally, there are two strategies for ontology construction: manual construction and automatic construction [8]. Since pure manual construction is a time consuming process and highly depends on expert experience, research on automatic or semi-automatic construction of ontology is more focused. Noy *et al* [9] proposed a top-down approach, which first constructed an upper structure of an ontology based on expert experience, then added concepts extracted from a knowledge base of such domain to every layer under the top one. Pazzaglia *et al* [10] gave an opposite way, i.e., a bottom-up approach for ontology construction by combing small ontologies to a large scale one using conceptual similarity. Wache *et al* [11] proposed a hybrid approach, that is, expanding concepts and relations to the top and the bottom respectively. Such approaches need less manual work though they acquire concepts and relations from thesauruses and lexicons such as WordNet for ontology expansion [12].

However, in many cases, concepts and relations acquired from thesauruses and lexicons hardly describe ontological relations of different semantic clusters. Take the concept *transport* as an example, there are five hyponyms: *public transport*, *private transport*, *air transport*, *water transport* and *land transport*. Apparently, the first two hyponyms and the last three ones are not in a same semantic cluster. In fact, there are two semantic clusters: one is ownership and the other is spatial range of use. As to the former, the concept *transport* has two hyponym concepts, i.e., *public transport* and *private transport*; as to the latter, such concept contains three hyponym concepts, i.e., *air transport*, *water transport* and *land transport*. Figure 1 shows the ontological structures of such two semantic clusters.

To build a unified ontological framework containing such two clusters, two methods can be adopted: 1) Rename the node *transport* according to the semantic clusters separately; 2) Add nodes between the node *transport* and its children according to the semantic cluster for each tree so that the node *transport* keeps unchanged and such two trees are easy to be combined. But both these methods need to access semantic clusters, which are essentially derived from common characteristics of word clusters.

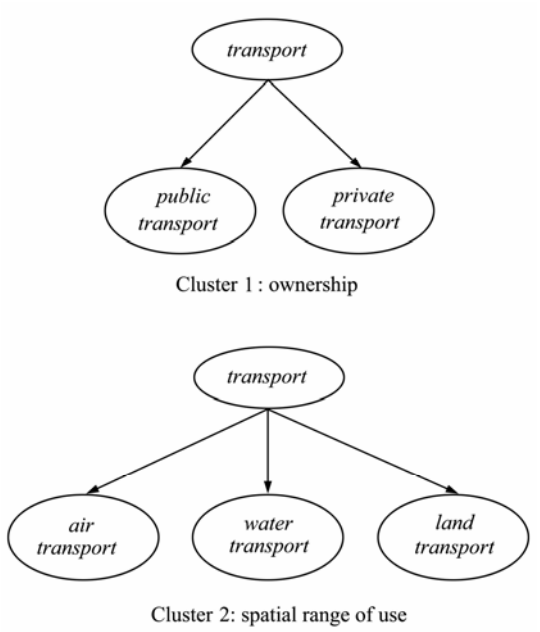

Fig. 1 Ontological structures of concept *transport*

To this end, some researches extracted concepts in words as characteristics of words using Wikipedia. Cui *et al* [13] accessed concepts derived from words using infobox, a conceptual description for each term. For a term, there are probably more than one infobox, each of which represents a concept that the term holds. However, since only a small set of terms has been described using infobox, this approach is hard to be practically used in ontology construction.

For describing characteristics of words more clearly, a primitive-based view is explored. Primitive refers to basic concept or structure in linguistics. The basic constituent of a concept is primitive and it is feasible to represent a concept by all primitives that the concept holds [6, 7, 14]. If two concepts have a same primitive, it is suggested that these two concepts probably have a same characteristics, in other words, they are in a same word cluster. More specifically, once a common primitive in a word cluster is found, it is probably the most precise and concrete description to the cluster. In comparison with the approaches that extract concepts from dictionaries or online knowledge bases, this approach helps find common concepts of word clusters, a.k.a, clustering nucleuses, as more precise ontological relations.

2 Proposed Approach

This section gives an overall description of the proposed approach in this paper. Figure 2 shows the system architecture of the proposed approach.

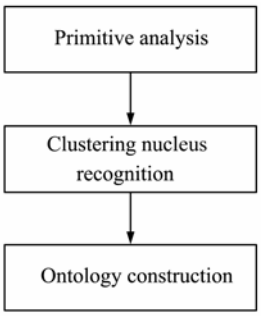

Fig. 2 System architecture of the proposed approach

As shown in Fig. 2, the approach presented in this paper includes three steps. First, primitives, which are predefined, are extracted for each word. Then, clustering nucleuses are identified from words with same primitives. More specifically, a word holding all or most of the primitives with other members in one word cluster is treated as the clustering nucleus for such word cluster. Finally, an ontological structure is built using clustering nucleuses with a bottom-up way.

3 Clustering Nucleus Analysis

3.1 Primitive Analysis

Analyzing primitive structure is a prior task for identifying clustering nucleus. Essentially, primitives are concepts profiling words with conceptual and pragmatical views. Xiao *et al* [6] classified primitives as conceptual ones, which were indispensable, and pragmatical ones, which were optional. Conceptual primitives contained core concept primitives and class-attributive ones, while pragmatical primitives contained general pragmatical primitives and discriminative pragmatical ones. Clustering nucleuses in different word set indicated different conceptual primitives, while clustering nucleuses in different hyponym set indicated different pragmatical primitives.

The annotation method of primitive follows the above taxonomy. More specifically, core concept primitives and class-attributive ones are first picked out for the construction of primitive structures. For example, the primitive structure of the word set {*bus*, *subway*, *taxi*, ···} is described as [*a device* + *public* + *city area* + *carrying from one place to another* + *somebody*], namely *city transport*. The clustering nucleus of this structure is [*public*+ *city area*], which is the key factor against its hypernym concept *transport* and its sibling concept *distance transport*.

3.2 Clustering Nucleus Identification

The approach of constructing ontologies based on clustering nucleus defines a set of representation terms called concept. This approach suggests that clustering nucleus for each node in the structural tree should be firstly found. Here, clustering nucleus refers to a primitive (attributes) or a group of primitives (attributes) shared by a conceptual synset. A concept contains many semantic attributes representing different primitives. Concepts based on the same primitives are grouped in one category and those primitives form a concept's clustering nucleus. In other words, all members in a minimal subset have the same clustering nucleus.

For example, *transport* is interpreted as *all kinds of man-made devices for carrying somebody or something*. The concept can be described like [*a device* + *carrying from one place to another* + *somebody/something*]. But in different domains, *transport* may be either Cluster 1 like *public transport* and *private transport* or Cluster 2 like *air transport*, *water transport* and *land transport*. In Cluster 2, the primitive of *range of use* is the inseparable attribute, which makes Cluster 2 a unique category. Additionally, every concept in Cluster 2 comprises a shared primitive, *range of use*. Moreover, the shared primitive can be inherited by its sub-subset, such as *city transport*.

It can be seen from the example that, a clustering nucleus of a word or concept set refers to the primitive with minimal common meaning in the set. Therefore, such clustering nucleus can be picked out by sets computation. Procedures of identifying clustering nucleus are described as follows:

1) Group words that hold the same conceptual primitives;

2) Cluster words with the same conceptual primitives in such groups;

3) Select a word of which all its primitives are also held by other words in each word cluster, as the clustering nucleus for such word cluster, and remove the word from the word cluster, if possible;

4) Otherwise, select a word whose pragmatical primitives are mostly held by other words in a word cluster, for example, only one pragmatical primitive of the word is not held by all the other words in the word cluster.

4 Automatic Ontology Construction

In our annotation scheme, the set of primitives is pre-defined, that is, all primitives are collected as a finite set, and classified into four types: core concept, class-attributive, general pragmatical and discriminative pragmatical type. Moreover, primitives of each word are regarded as labels of semantic relations on such word, thus recognizing clustering nucleus of primitives is treated as multiple label learning, a semi-supervised machine learning task. More specifically, given a training set labeling primitive structure for each word, recognizing clustering nucleus is to find an approach that propagate labels of primitive of words in the training set to those, which have similar semantic relations with such words, in the test set.

A multi-label learning approach, namely correlated label propagation proposed by Kang *et al* [15], is used to recognize clustering nucleus. Different from first-order label propagation, correlated label propagation considers that labels propagate not only among words, but also among the labels themselves. Figure 3 shows the propagation procedure of correlated labels.

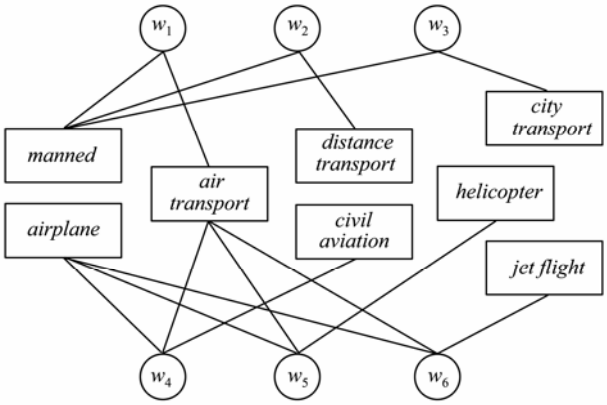

Fig. 3 Correlated label propagation procedure

In Fig. 3, $w_i(i=1,2,\cdots,6)$ in each circle denotes a word with labels, a.k.a, primitives, in each rectangle. Assume that *w* and w_i are in a same word cluster, and *w* denotes a word unlabeled with primitive. According to first-order label propagation approaches, a conclusion can be drawn that the primitive *air transport* is the most probable label assigned to *w*, as such primitive has four connections, while the primitives *manned* and *airplane* are the equally probable labels assigned to *w*, as such two primitives each has three connections. If multiple labels are able to propagate, however, the primitive *airplane* rather than *manned* is preferred to be assigned to *w* first, because the co-occurrence frequency of the primitive *air transport* and *airplane* is higher than that of *air transport* and *manned*. In a word, correlated label propagation takes relations between samples as well as relations between labels into account, and it is fit for recognizing clustering nucleus, which also needs to consider relations between words and relations between primitives.

The goal of correlated label propagation for recognizing primitive is formally described as follows: given a training set $T=\{(w_1,\boldsymbol{s}(w_1)),\cdots,(w_n,\boldsymbol{s}(w_n))\}$ labeled with primitives and an unlabeled word w, the algorithm needs to get a confidence vector $\boldsymbol{C}=\{c_1,\cdots,c_m\}$, where c_k denotes a confidence value that the *k* th label, a.k.a., primitive, is assigned to w. $\boldsymbol{s}(w_i)=\{v_{w_i,1},\cdots,v_{w_i,m}\}$ is a binary vector, where $v_{w_i,j}$ is a binary value that represents if the word w_i holds the *j* th label, *n* denotes training data size and *m* is the amount of conceptual primitives. Here, w is represented by a *d*-dimension vector, where each dimension denotes a feature from a feature space based on lexical knowledge. Such knowledge comes from thesauruses and lexicons such as WordNet or Wiki infobox, and hypernym, hyponym, sibling or path to the root of a word can be employed as a feature to profile a word.

On the other hand, assigning a label to w needs to satisfy constraints according to such training set and label propagation rules and an optimal confidence vector $\boldsymbol{C}$ from all results that mostly satisfy the constraint is finally picked out as the output for labeling conceptual primitive. This is an optimization problem for $\boldsymbol{C}$:

$$\max_{\boldsymbol{C}\in\mathbf{R}^m}\sum_{k=1}^{m}\alpha_k c_k$$

$$\text{s.t.}\quad \forall \boldsymbol{t}\in\{0,1\}^m, \boldsymbol{C}^{\mathrm{T}}\boldsymbol{t}\leqslant\sum_{i=1}^{m}K(\boldsymbol{w},\boldsymbol{w}_i)F(\boldsymbol{t}^{\mathrm{T}}\boldsymbol{s}(w_i))$$

$$\boldsymbol{C}\succeq 0$$

Where, α_k is the weight of the *k* th primitive label, which can be valued as the occurrence frequency of the *k* th primitive label in training data, *K* is a kernel function measuring the similarity of any feature vectors of two words, e.g., inner product, and *F* is a kernel function measuring two label vectors, e.g., sigmoid value. This is a linear programming problem, for which greedy algorithms can be employed to solve it.

The advantage of relevant label propagation approach is to consider semantic relations between labeled

words and unlabeled ones as well as relations between primitives, which benefits the recognition of clustering nucleus of unlabeled words given a small amount of labeled samples.

After getting optimal confidence vector $\boldsymbol{C}$, primitives that w' hold are acquired exceeding some threshold. Then, a heuristics defined in Section 3.2 is employed to construct the ontology by extracting clustering nucleuses from such primitives with a bottom-up way. More specifically, a clustering nucleus is extracted from a word group, then an ontological structure is built, in which the leaves are words and the upper node is the clustering nucleus. In the next step, a clustering nucleus is extracted from all upper nodes, a.k.a, clustering nucleuses, to update the ontological structure. This procedure is iterative until there is only one clustering nucleus in the highest layer or no clustering nucleus is extracted. In post processing, those clustering nucleuses are picked out in the last step of identifying clustering nucleus that are replaced with more fit words by manual work.

5 Conclusion

This paper gives a new approach, clustering nucleus, for ontological knowledge construction. Using this approach, semantic relations can be observed in a more elaborated view: conceptual primitives and their structures, which benefit the analysis of semantic relations in texts. This paper also introduces an approach to recognize clustering nucleus, which helps to recognize such elaborated descriptions for semantic relations automatically. By this approach, more precise ontological relations can be discovered automatically.

References

[1] Dnyanesh G R. An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain [J]. *Computers in Industry*, 2013, **64**(5): 565-580.

[2] Cimiano P, Luker J, Nagel D, *et al.* Exploiting ontology lexica for generating natural language texts from RDF data [EB/OL]. [2014-07-30]. *http://www.aclweb.org/anthology/W*13-2102.

[3] Huang X P, Du Y J, Ren Y Y. Research on automatic user ontology construction [J]. *Journal of Information and Computational Science*, 2012, **9**(2): 285-292.

[4] Subramaniyaswamy V. Automatic topic ontology construction using semantic relations from WordNet and Wikipedia [J]. *International Journal of Intelligent Information Technologies*, 2013, **9**(3):61-89.

[5] Li J, Ding Y D, Shi Y Y, *et al*. Building a large annotation ontology for movie video retrieval [EB/OL]. [2014-08-01]. *http://www.aicit.org/jdcta/ppl/*08.%20*JDCTA2*-390081.*pdf*.

[6] Xiao G Z, Wang X L. The text deduction and model realization of the lexical meanings in dictionaries based on "synset-lexeme anamorphosis" and "basic semantic elements and their structures" [C]*//*13*th Chinese Lexical Semantics Workshop*. Berlin: Springer-Verlag, 2013: 774-783.

[7] Hu D. *Primitive Structure of Mandarin in Natural Language Processing* [M]. Guangzhou: World Publishing Corporation, 2014: 3-25(Ch).

[8] Abbas J. *Structures for Organizing Knowledge*: *Exploring Taxonomies*, *Ontologies*, *and Other Schemas* [M]. New York: Neal-Schuman Publishers Inc, 2010: 5-20.

[9] Noy N F, Hafner C D. The state of the art in ontology design: A survey and comparative review [J]. *AI magazine*, 1997, **18**: 53-74.

[10] Pazzaglia J R, Embury S M. Bottom-up integration of ontologies in a database context [EB/OL]. [2014-08-01]. *http*: *//sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol*-10*/paper*7*.ps.*

[11] Wache H, Vögele T, Visser U, *et al.* Ontology-based integration of information—A survey of existing approaches [EB/OL]. [2014-08-01].*http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol*-47*/wache.pdf.*

[12] Huang X P, Du Y J, Ren Y Y. Research on automatic user ontology construction [J]. *Journal of Information and Computational Science*, 2012, **9**(2): 285-292.

[13] Cui G Y, Lu Q, Li W J, *et al*. Automatic acquisition of attributes for ontology construction [C]*//Proceedings of Computer Processing of Oriental Languages* (*ICCPOL* 2009). Berlin: Springer-Verlag, 2009: 248-259.

[14] Rosch E. Congitive representations of semantic categories [J]. *Experimental Psychology*, 1975, **104**: 192-233.

[15] Kang F, Jin R, Rahul S. Correlated label propagation with application to multi-label learning [EB/OL]. [2014-08-01]. *http:// www.cs.cmu.edu/~rahuls/pub/cvpr*2006-*labprop-rahuls.pdf.*

□