Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 7, July 2018, pp. 1459-1467
DOI: 10.23940/ijpe.18.07.p9.14591467

Concept Meaning Acquisition based on HowNet and Its Application in the Construction of Taxonomy

Jian Xu[a], Jianhou Gan[b], Xianming Yao[a], and Liming Zhang[c,*]

[a]*School of Information Engineering, Qujing Normal University, Qujing, 655011, China*
[b]*Key Laboratory of Educational Informatization for Nationalities, Yunnan Normal University, Kunming, 650500, China*
[c]*School of Humanities, Qujing Normal University, Qujing, 655011, China*

Abstract

In studies related to the construction of domain ontology, the acquisition of concept meaning has not received enough attention. According to the definition of the concept, the acquisition of concept meaning is a requisite task in the process of ontology construction. This paper studied the automatic acquisition of concept meaning based on HowNet and researched the problem of meaning acquisition for complex terms and synonym removal. Grounded on concept meaning, this paper put forward a sememe suffix tree algorithm and applied it to the construction of ontology taxonomy. Compared to traditional algorithms, this method is more efficient and comprehensible. This paper implemented the methods to the domain of ethnic minorities, and the experimental results showed that this paper is referable.

Keywords: concept meaning acquisition; HowNet; sememe suffix tree algorithm; taxonomy acquisition

(Submitted on April 2, 2018; Revised on May 13, 2018; Accepted on June 11, 2018)

1. Introduction

Automatic domain ontology construction has been studied for many years, and some research achievements have been obtained [10]. Traditionally, domain terms were extracted from domain corpora, and they were used to create taxonomy and none-taxonomy relationships [18]. In this process, the term was used as a concept directly, and subtle differences between them have not been taken into consideration. This difference would bring fuzziness and inconsistency to explicitly and formally defined ontology [6]. As the cornerstone of ontology, a well-defined concept is of great importance to ontology construction.

According to the definition in literature [9], a concept is composed by its word realization (term), definition, and extension (instance). Traditional ontology construction process simply included obtaining terms and ontology populations [12,13,14]. Concept meaning acquisition was barely mentioned. Based on previous work and aimed at the deficiency of concept meaning acquisition, in the review of paper [20], we studied the task of concept meaning acquisition based on HowNet, researched concept meaning acquisition for complex terms, and also provide a method for synonym removal. This paper puts forward a sememe suffix tree algorithm to construct taxonomy relationships.

Taking ethnic minority as a research area, this paper studied concept meaning acquisition and domain ontology taxonomy relation construction to validate the effectiveness of the methods proposed above. Experimental data proved that this method is referable.

2. Concept Meaning Acquisition

Studies on concept meaning acquisition are some of the most important and basic work in this paper. In this section, we studied concept meaning acquisition based on HowNet, especially terms with complex structure and those not included in HowNet. In the last section, we studied synonym removal.

* Corresponding author.
E-mail address: qjncxj@126.com

2.1. HowNet Based Concept Meaning Acquisition

So far, concept meaning acquisition was barely mentioned in the domain of ontology construction. Primary research work was focused on concept definition extraction, which is defined in natural text form. For instance: paper [8] studied the structure and classification of the definition and presents a multiple-rules-extraction method; paper [16] proposed a definition extraction method based on both hard pattern matching and soft pattern matching. The precision of these methods reached over 80% [19], and they could extract definition combined with domain knowledge effectively. Deficiency of these methods lies in that the definition is expressed in natural text and it can hardly be used to assist other tasks. Also, these research works were not based on ontology construction.

In the light of domain term meaning acquisition, paper [20] studied its acquisition method and partly resolved the task for simple and compound nouns. Paper [20] profited from an existing knowledge base, named HowNet, which defined part of speech, instances and structured definition for each entry. Definitions could be obtained from structured definition directly, and the definition of a compound term could be extracted out and recombined to a rational format. For those complex nouns, there are still no good solutions, and the taxonomy relation construction remains unresolved. This paper adopted and improved the method mentioned in paper [20]. A word sense disambiguation procedure was introduced for identity meaning of each word, so as to make the definition from HowNet more rational. Other techniques, such as term matching and part of speech matching, were used as important filtering rules to improve precision.

The philosophy of HowNet and its structure have been discussed in many other literatures [4], and there is no need to introduce here again. W_C, G_C, E_C and DEF represent term, part of speech, instances, and definition relatively. W_C represent the characters of the word itself. It is the most direct features of the word. In order to get the definition from HowNet, W_C was firstly used to retrieve the relevant concept. G_C, as part of speech, indicates the role it plays in sentences. This information is critical in most cases. E_C are instances where this concept is mostly used. This could provide relevant words for word sense disambiguation. Lastly, DEF is the definition of the word in HowNet, and this is what we want to get in this procedure.

For each term from our previous work, the POS procedure is needed to get words that compose this term. These words were sent to HowNet to retrieve concepts according to its entry. In the following, part of speech from HowNet and POS matching procedures were introduced to exclude non-relevant concepts. Lastly, the word sense disambiguation procedure was introduced to get the most similar concept. After these procedures, definitions of terms could be obtained and reassembled to form a whole definition for each term. In this process, not all definition of terms could be obtained successfully due to some words in POS that are not included in HowNet. The flow of concept meaning acquisition in this paper is depicted in Figure 1.

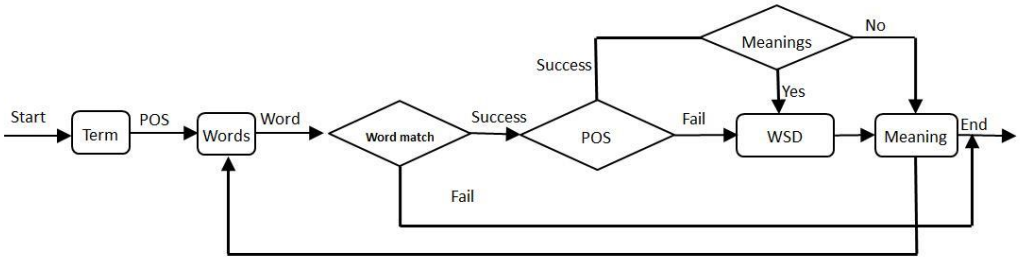

Figure 1. Concept meaning acquisition flow

1. POS. Execute part of speech for each term, so as to get its words and corresponding POS. POS procedure is done by a tool of NLPIR [4].

2. Filtering procedure. Get one word from previous steps and retrieve it in HowNet. Usually, the returned result is a set of concepts. We had to select the most similar concept as the word's meaning. Word matching procedure could be introduced here to judge whether W_C of these concepts is the same as this word; concepts that share the same W_C with this word could be kept. Subsequently, part of speech matching procedure is required too. If G_C information of concepts is matched with POS, these concepts could be kept, otherwise, they are abandoned. What needs to be noticed here is that NLPIR and HowNet use different POS sets. A conversion procedure is needed to convert POS of NLPIR to HowNet.

3. Word sense disambiguation. After previous filtering procedures, if there remains only one concept, it could be considered that the definition was acquired successfully. If there remains over one concept, and they do not share the same definition, word sense disambiguation procedure should be introduced to select the most similar concept as the target concept.

Word sense disambiguation procedure is done by a bag of words model. **Context** is words that co-occur with the word in the domain corpora. It includes other words in the same term, other terms that include the word, and words that co-occur at documents level. **Glossary** is words include sememe words, like E_C and relevant words provided by HowNet to a specific concept. These words could be obtained via access of interface provided by HowNet. The similarity of word and concept could be calculated out according to the overlap of Context words and Glossary words. Concepts that share the highest similarity will be chosen as the target definition.

For instance, "Shui3Zu2" is a polysemic word; it could be considered as a "fish" or "tribe". Context words of "Shui3Zu2" include "Zu2", "Min2Zu3", "Zhong1Guo2" "Mao3Jie2", "Shui4Shu1", and "Xin1Niang2" etc. Glossary words of "fish" are "Bu3Lao1Neng2Li4", "Yu2Shi4", "Shui3Zu2Guan3", "Shui3Chan3Ye4" etc. Glossary words of "tribe" are "Zu2Qun2", "Zhong1Guo2" etc. We could easily conclude that "Shui3Zu2" means "tribe" in our experiment, for they share the same glossary word "Zhong1Guo2".

4. Next loop for current term. If there are other words in the part of speech result to the current term, go to the second step, or else go to the next step. For example, "Min2Zu2Xiang1" is segged as "Min2Zu2 Xiang1". In the first loop, we could conclude the meaning of "Min2Zu2" is "tribe". We need to get the meaning of "Xiang1" in the next loop, so we have to go back to the second step. This processing is critical in this paper; once we stop and go to the last step, we could only acquire meanings for simple and single terms, and we could guarantee that terms with unfixed length will still get their meanings.

5. Reassemble word meaning for term. According to the discipline where core meaning is positioned left in DEF, the meaning of the core word of term positions left too. Other meanings of accessary words align from the right to left positions from left to right in meaning. This reassemble procedure is explained in detail in the next sub section.

2.2. Meaning Acquisition for Complex Term

Meaning acquisition is relatively easier for simple structure terms that have only one or two characters. It could be obtained from HowNet directly, but for other terms that are composed of several words, they are not included in HowNet, meaning the acquisition is more difficult. These terms are named complex terms. Complex terms have clear meanings and are common to a special domain. Usually, they are composed of several words, and each word has a special meaning and plays a dedicated semantic role. Actually, complex terms are composed by the meaning of these words. We could get the meaning for each word individually and reassemble these meanings together to form a complete meaning so that the complex term meaning acquisition could be obtained. This process is depicted in Figure 2, which takes the term "Fei1Wu4Zhi4Wen2Hua4Yi2Chan3" as an example.

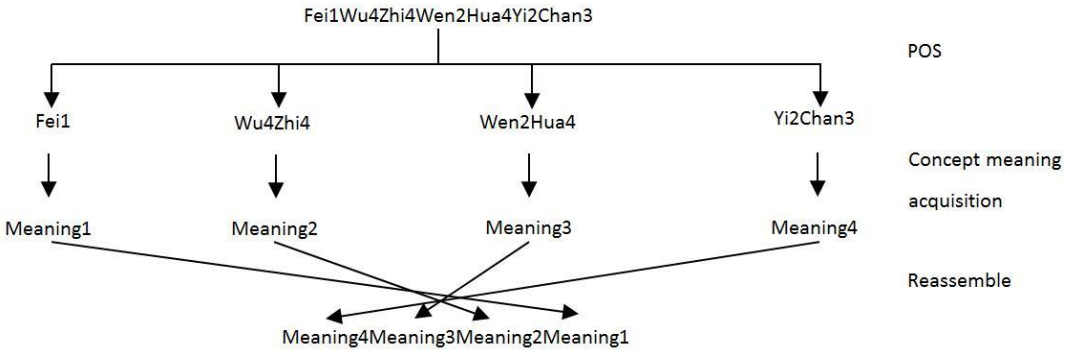

Figure 2. Complex term meaning acquisition process

As the example shown in Figure 2, "Fei1Wu4Zhi4Wen2Hua4Yi2Chan3" could not be retrieved in HowNet, but it is composed of "Fei1", "Wu4Zhi4", "Wen2Hua4", and "Yi2Chan3". These words could be separated by POS tools. The meaning of "Fei1" is "{PropertyValue|Te4Xing4Zhi2:scope={BeNot|Fei1}}"; meaning of "Wu4Zhi4" is "{physical|Wu4Zhi4}"; meaning of "Wen2Hua4" is "{knowledge|Zhi1Shi2:domain={education|Jiao4Yu4}{entertainment|Yi4}{literature|Wen2}}";

meaning of "Yi2Chan3" is "{thing|Wan4Wu4:modifier={precious|Zhen1},{PassOn|Liu2Gei3:agent={human|Ren2:modifier={forefathers|Zhu3Xian1}},possession={~}}}". The meanings of these words could be obtained from HowNet. These meanings could be reassembled as a complete meaning to describe "Fei1Wu4Zhi4Wen2Hua4Yi2Chan3". According to the discipline that the core meaning of the concept is placed at the left in HowNet, and the core word of the complex term is at the right, the meaning of the core word should be placed at the left, and other meanings of words should be placed from left to right after the core meaning. So, the meaning of "Fei1Wu4Zhi4Wen2Hua4Yi2Chan3" is "{thing|Wan4Wu4:modifier={precious|Zhen1},{PassOn|Liu2Gei3:agent={human |Ren2:modifier={forefathers|Zhu3Xian1}},possession={~}}};{knowledge|Zhi1Shi2:domain={education|Jiao4Yu4}{entertainment|Yi4}{literature|Wen2}};{physical|Wu4Zhi4};{PropertyValue|Te4Xing4Zhi2:scope={BeNot|Fei1}}". The meaning reassemble procedure is a little sophisticated and the procedure is depicted in Figure 2. Note that the final meaning order is the reverse of the sequence of the words.

This process is reflected in the fourth step of the previous section. For each complex term, it is segged into several words, and the meaning of each word is obtained separately in each loop. When one loop is finished, the other meanings will start in the next loop. In the final step, meanings will be reassembled.

2.3. Synonym Removing based on Concept Meaning

In the previous section, concept meaning acquisition has provided a prerequisite for other tasks of ontology construction. This paper researched synonym removal, which is based on concept meaning.

Currently, studies on synonym removal are relatively fewer and decentralized. The main method focuses on similarity computation according to the knowledge base or distributional features, etc. Threshold will be given out to select those above it as a synonym. The method we adopted below belongs to the knowledge base, and this paper checks whether two concepts share the same meaning acquired before.

Concept meaning in this paper has been obtained in previous sections. If two terms are synonyms, they should share the same meaning. So, synonym removal could be done by comparing term meanings. According to this discipline, a sorting to concept meaning from previous sections will put those synonyms together and select one as term, so as to achieve the goal of synonym removal. It is noteworthy that synonyms should share the same meaning rigidly, and any difference will not be taken into consideration. This method could avoid the problem of setting experimental thresholds in similarity computation. It also preserves subtle differences between similar terms; this difference would be helpful for taxonomy relation construction.

For example, "Yi1Fu2", "Fu2Shi4", and "Fu2Zhuang1" share the same meaning "{clothing| Yi1Fu2}", so they could be considered synonyms. The meaning of "Wu3Dao3" is "{fact|Shi4Qing2:CoEvent={recreation|Yu2Le4},domain={entertainment|Yi4}}", while "Yi2Shi4" has the meaning of "{fact|Shi4Qing2:modifier={formal|Zheng4Shi4}}". Two terms have the same core meaning of "fact|Shi4Qing2", but one is relevant to "Yu2Le4" and the other represents "formal" "things". They share a certain similarity while differences between them exist at the same time; therefore, they could not be considered as synonyms.

2.4. Concept Meaning Acquisition in Ethnic Minorities

This paper implemented an algorithm in concept meaning acquisition in the domain of ethnic minorities to test its effectiveness. In our previous work, we had extracted 523 terms. After the implementation of the above algorithm, we obtained 492 meanings for concepts and 31 terms failed to get its meaning. After synonym removal, we were left with 349 concepts and 68 synonyms.

A part of the complex terms meanings is listed in Table 1.

Table 1. Part of complex terms meaning

ID	Complex term	meaning
1	Chuan2Tong3Jie2Ri4	{time|Shi2Jian1:TimeFeature={festival|Jie2},{congratulate|Zhu4He4:time={~}}};{Conventionality|Zheng4Zong1Xing4:host={group|Qun2Ti3}{human|Ren2}}
2	Feng1Shu2Xi2Guan4	{Habit|Xi2Guan4:host={group| Qun2Ti3}{human| Ren2}};{Habit|Xi2Guan4:host={group| Qun2Ti3}{human|人},modifier={local|Gui1Shu3Di4Fang1}}
3	Yuan2Shi3Zong1Jiao4	{community|Tuan2Ti3:domain={religion|Zong1Jiao4}};{original|Yuan2}
4	Huo3Ba3Jie2	{time|Shi2Jian2:TimeFeature={festival|Jie2},{congratulate|Zhu4He4:time={~}}};{tool|Yong4Ju4:{illuminate|Zhao4She4:instrument={~}},{lighting|Dian3Ran2:instrument={~}}}
5	Ren2Jun1Cun3Shou1Ru4	{wealth|Qian2Cai2:quantity={net|Jin4E2},{earn|Zhuan4:possession={~}}};{average|Ping2Jun2:scope={human|Ren2}}
6	Dai3Wen2	{character|Wen2Zi4};{tribe|Zu2Qun2:belong="China|Zhong2Guo2"}

Statistics show that the precision rate of concept meaning acquisition reached 88%. This mainly profited from a large domain text, which included 4242 files with a size of 14M. There are enough context words to describe terms.

A part of synonyms and their meanings are listed in Table 2.

Table 2. Part of synonyms and their meaning

ID	term	meaning
1	La3Ma1, Huo2Fo2, Seng1Lv3	{human|Ren2:belong="Buddhism|Fo2Jiao4",domain={religion|Zong1Jiao4}}
2	Shi4Zu2, Min2Zu2, Shao3Shu4Min2Zu2	{tribe|Zu2Qun2}
3	Si4Yuan4, Ta3, Si4, Qin1Zhen1Si4, Fo2Si4, Fo2Ta3	{facilities|She4Shi1:domain={religion|Zong1Jiao4}}
4	Yin1Yue4, Min2Ge1, Chang4Qiang1	{music|Yin1Yue4}
5	Ge1, Shan1Ge1, Qu3	{music| Yin1Yue4:{sing|Chang4:content={~}}}
6	Cun1Zhai4, Cun1, Cun1Zi3	{place|Di4Fang1:PlaceSect={village|Xiang1}}

Statistics show that the precision rate of synonym removal reached 78%. The main reason for the error is that concepts with the same type that share the same meaning in HowNet are considered synonyms. For example, "Nan2Fang1" and "Nv3Fang1" have the same meaning "{part|Bu4Jian4:PartPosition={aspect|Ce4},RelateTo={GetMarried|Jie2Hun1},whole={community| Tuan2Ti3}}", and so they were merged as synonyms. But, they are not equal to each other. In addition, errors that occurred in concept meaning acquisition impact synonym removal as well.

3. Concept Taxonomy Relation Construction based on Meaning

Taxonomy relation refers to the hierarchical relationship between concepts, including part-whole relation, inheritance relation, concept-instance relation, and concept-feature relation [5]. Establishment of taxonomy relations has always been difficult in the construction of domain ontology. Paper [7] adopted a pattern based method to extract relation by constructing a hypernym pattern and appositive pattern, and achieved good effect. However, the ability of transplanting to other fields is poor, and the recall rate may decline greatly. The most frequently used methods for construct taxonomy include top-down [3] or bottom-up [1] hierarchical clustering algorithms [21] by calculating the similarity of concepts. It still has problems of sparse data and tightness between classes is difficult [17].

Based on concept meaning from the previous section, this paper finds that structured concept meaning could be very useful in concept taxonomy construction. Taxonomy could be constructed by comparing meanings of concepts. This paper divided the construction process into two steps. Firstly, group concepts share the same sememe and construct a local taxonomy tree. Secondly, append local taxonomy trees into an "Entity" tree, which is obtained from HowNet according to core sememe, so that a global taxonomy tree could be obtained and a lightweight ontology could be constructed.

3.1. Local Taxonomy Relation Tree Construction based on Sememe Suffix Tree

This paper put forward a sememe suffix tree algorithm, which is based on concept meaning obtained in the previous section. Sememe is the smallest unit in DEF [4], and the difference between concepts is embodied at the difference of sememe. For example, "Di4Qu1" has meaning "{place|Di4Fang1}", "Shi4" has meaning "{place|Di4Fang1:PlaceSect={city|Shi4}}", and they are similar to each other, for they share the same core sememe "Di4Fang1". But, "Di4Qu1" is an upper concept to "Shi4" because "Shi4" has a more detailed description of "city|Shi4". Concepts that share the same core sememe could be organized to a local taxonomy relation tree.

According to our analysis, we find four rules: firstly, if a concept has only one core sememe, it could be considered as the root concept to other concepts; secondly, if two concepts share the same sememe, they could be considered as one concept, so that they are synonyms; thirdly, if two concepts share the same core sememe but other sememes are wholly different from each other, they could be considered as brother concepts; fourthly, if sememes of one concept are included in another concept, it could be concluded that the previous concept may be an upper concept of the latter one, and vice versa.

Take data in Table 3 as an example, all concepts that share the core sememe "place|Di4Fang1" in our experimental data have been listed out.

The column headed "Terms" in Table 3 represents domain terms extracted from free text. They share the same core sememe "place|Di4Fang1"; "Meaning" represent concept meaning obtained from HowNet, they were reassembled, such as "Min2Zu2Xiang1"; The "Sememes" column represents sememes extracted out from the Meaning column.

The term “Di4Qu1” has only one sememe “place|Di4Fang1”, and it is the core sememe to other terms. According to the first rule, it is the root node of the local taxonomy relation tree. “Guo2Jia1”, “Zhou1”, “Shi4”, “Xian4”, “Cun1Zhai4” are brother concepts according to the third rule. “Shi4” is an upper concept of “Gu3Cheng2” according to the fourth rule. Other relations could be constructed according to these rules.

Table 3. All the concepts which share core sememe “place|Di4Fang1” in our experimental data

ID	Terms	Meaning	Sememes
1	Di4Qu1	{place|Di4Fang1}	place|Di4Fang1
2	Shi4	{place| Di4Fang1:PlaceSect={city|Shi4}}	place|Di4Fang1 city|Shi4
3	Gu3Cheng2	{place| Di4Fang1:PlaceSect={city| Shi4},modifier={past|Guo4Qu4}}	place|Di4Fang1 city| Shi4 past|Guo4Qu4
4	Guo2Jia1	{place| Di4Fang1:PlaceSect={country|Guo2Jia1},domain={politics|Zheng4}}	place|Di4Fang1 country|Guo2Jia1 politics|Zheng4
5	Wo3Guo2	{place| Di4Fang1:PlaceSect={country| Guo2Jia1},domain={politics| Zheng4},modifier={self|Ji3}}	place|Di4Fang1 country| Guo2Jia1 politics| Zheng4 self|Ji3
6	Xian4	{place| Di4Fang1:PlaceSect={county|Xian4}}	place|Di4Fang1 county|Xian4
7	Zhou1	{place| Di4Fang1:PlaceSect={provincial|Sheng3}}	place|Di4Fang1 provincial|Sheng3
8	Xing2Shen3	{place| Di4Fang1:PlaceSect={provincial| Sheng3}};{able|Neng2}	place|Di4Fang1 provincial| Sheng3 able|Neng2
9	Cun1Zhai4, Cun1, Cun1Zi3	{place| Di4Fang1:PlaceSect={village|Xiang1}}	place|Di4Fang1 village|Xiang1
10	Min2Zu2Xiang1	{place| Di4Fang1:PlaceSect={village|Xiang1}};{tribe|Zu2Qun2}	place|Di4Fang1 village|Xiang1 tribe|Zu2Qun2
11	Zhai4Zi3	{place| Di4Fang1:{defend|Fang2Shou3:patient={~}}}	place|Di4Fang1 defend|Fang2Shou3

Although these rules look a little sophisticated, in our experiment, we found that there exists a simple method to construct these relations more quickly. A sorting process of sememes could group similar sememes together. The “Sememes” column in Table 2 is sorted. We could note that sememes close to each other have a common prefix. This prefix is an important indicator to infer their relation. A comparison to the string of them could obtain their relationship. This process is just like a suffix algorithm. So, we call this algorithm a sememe suffix tree algorithm.

According to data in Table 3, the local concept taxonomy relation tree shown in Figure 3 could be constructed based on a sememe suffix tree algorithm. Figure 3 is a graph that takes “Di4Qu1” as the root node and where other leaves and branches are its hyponym concepts. So, it is named the local taxonomy relation tree in this paper.

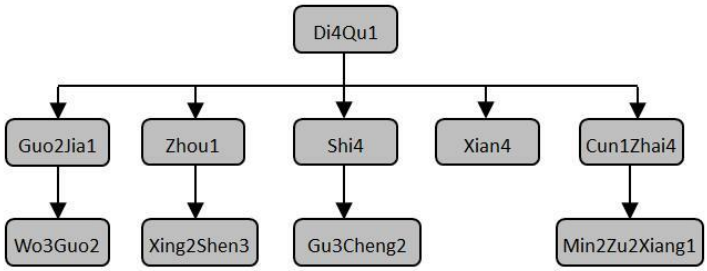

Figure 3. Local taxonomy relation graph

Sememe suffix tree based concept taxonomy relation construction is a simple and effective algorithm. Based on the general knowledge base, it could create a taxonomy relation for concepts that share the same core sememe, and precision could also be guaranteed because semantic information has been taken into consideration.

3.2. Global Taxonomy Relation Tree Construction based on Local Tree

After construction for all the local taxonomy relation trees, concepts will be divided into different groups according to its core sememe. In order to construct a global taxonomy relation tree, local trees should be integrated into a whole one. Sememe tree named “Entity” in HowNet is a preferable choice as a global skeleton for ontology of ethnics, for meaning of concept in this paper is based on HowNet, and as one of 8 kinds of sememe trees [11], the “Entity” tree had given out a global view to our world. Another reason lies that concept extraction was mainly focused on noun phrases. Borrowing “Entity” tree as a global taxonomy skeleton is reasonable, as other kinds of sememe trees in HowNet are not in consideration now.

The process of global taxonomy relation construction is as follows:

Step1. Acquire sememe tree of "Entity" from HowNet.

Step2. Acquire the root node of the local taxonomy relation tree and search its core sememe in "Entity" tree. Append the local tree to global "Entity" according to the root node of the local tree.

Step3. Cut branches in the "Entity" tree that shares no local graphs. The global tree left is a domain taxonomy relation tree, or a lightweight ontology.

3.3. Domain Taxonomy Relation Construction for Ethnic Minorities

Based on a sememe suffix tree algorithm, this paper studied concept taxonomy relation construction. We constructed 144 local taxonomy relation graphs based on 349 concepts. After analysis, 288 concepts could be merged to an "Entity" tree; other concepts belong to other sememe trees in HowNet. The final global taxonomy relation graph included 345 nodes and the depth of the graph is 9. Part of the global graph is depicted in Figure 3.

Figure 4 sketched out the global taxonomy tree, in which rounded rectangles with the white background are from the "Entity" tree in HowNet. Part of the data comes from the experiment; the rounded rectangle with grey background outlined data from the experiment. They were appended to the "Entity" tree according to the local taxonomy relation graph. This graph demonstrated a local taxonomy relation graph that shared a core sememe "tool|Yong4Ju4". It has over fourteen hyponyms, and it was appended to the root node "Qi4Ju4". With the limitation of space, other local taxonomy relation graphs were not depicted here.

3.4. Evaluation to Ontology of Ethnics

Automatic construction of ontology has always been a difficult task. This paper studied the related theory, put forward our own methods, and finally realized the domain of ethnics. This work will fill the blanks in this area. Other tasks will benefit from this work too. Whether this ontology is effective to our interested domain or could assist other tasks correctly still remains to be validated.

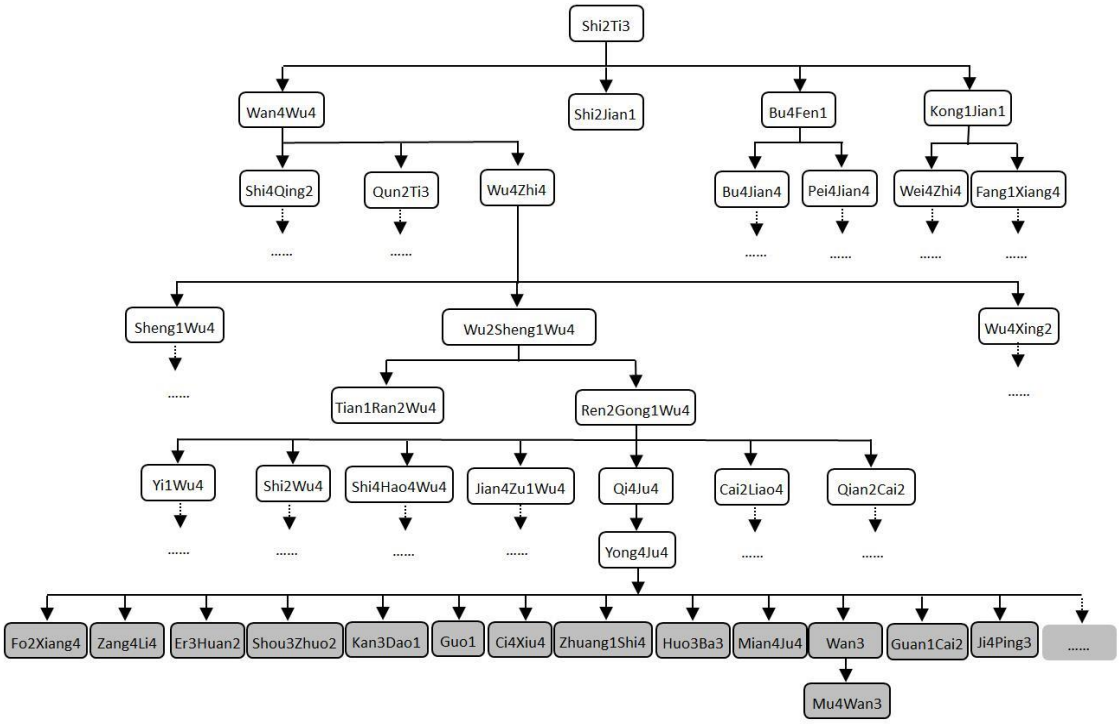

Figure 4. Part of global taxonomy tree

Evaluation to ontology is not an easy task for ontology restricted to theory, method, and standard. Fewer evaluation works were mentioned in previous work. This paper departed the evaluation task into two different subtasks: evaluation of concepts and taxonomy relations. In each subtask, the golden standards of precision rate and recall rate in information retrieval were adopted.

Concept evaluation, this task focused on validating whether concepts belonged to our interested domain. Rigidly speaking, all concepts hidden in text that belonged to this domain could hardly be recognized by humans. This problem brings great difficulty. In this paper, we adopted the methods mentioned in paper [2]. Firstly, a set of concepts was given out according to our interest in ethnics. The number of concepts reached 387, which were extracted from domain text based on word co-occurrence and selected artificially like "Min2Ge2", "Xi2Shu2" and "Fu2Shi3", etc. Secondly, we listed out all terms of concept in ontology built by us. This set of terms is augmented, for one concept corresponds to one term; this augmented term set is more tolerant to reality. Thirdly, we obtained the precision rate and recall rate. Precision rate is the ratio of overlapped terms to all terms in our ontology; recall rate is the ratio of overlapped terms to artificially checked concept sets. Statistics in Table 4 show the evaluation result. There are 403 terms in ontology constructed by us and 387 terms were provided by us as a reference. There were 343 overlapped terms.

Table 4 shows that the result is promising. But, there remains a problem where terms from HowNet are not in the given concept set. This problem leads to the decline of the precision rate. Concepts like "Wu2Sheng1Wu4", "Tian1Ran2Wu4" etc. did not occur in the domain text but are in one category in HowNet.

Table 4. Evaluation of concept in ontology

Terms in ontology	Terms given artificially	Overlapped terms	Precision rate	Recall rate	F-score
403	387	343	85%	87%	86%

Taxonomy evaluation: this task checks whether the hyponym relation in ontology is right. So far, there is no good method to check these relationships. In this paper, we assume that the hyponym relation between two concepts is hidden in the text. Lots of clues could check its relationship, such as "li4Ru2", "Bao1Kuo4", "Shu3Yu2" etc. Firstly, this paper picked up all concepts with a hypernym relation and formed a tuple (like concept1, concept2). Secondly, we retrieved concept strings across all documents to find sentences that included both strings of concept in a tuple. This could alleviate human work greatly. Finally, a series of searching, sorting, visualization, and pattern matching tools were adopted to check concept hypernym relation by the human. Table 5 shows the detailed evaluation result.

Table 5. Evaluation of taxonomy relation in ontology

Tuples	Sentences returned	Precision rate
402	5281	73%

In ontology built by this paper, there are 402 pairs of tuples, and 5281 sentences were extracted from the domain text. This is still a very large data set. Results show that the precision rate reached 73%, which is also promising. Although hypernym relation is relatively stable in ontology, experimental data shows that there are other kinds of relations (non-taxonomy relation). Hypernym relation in the text is relatively rare. Another problem that needs to be explained is that relations spring from HowNet are considered as right even though they do not occur in domain text.

4. Conclusions

This paper studied concept meaning acquisition based on the general knowledge base and discussed complex term meaning acquisition and synonym removal. This paper put forward a sememe suffix tree algorithm and its application in concept taxonomy relation construction. We implemented the algorithm in the domain of ethnic minorities and reached our expected results, which proved its reference value.

Although this paper partly realized concept meaning acquisition and taxonomy creation, its effectiveness is still waiting to be validated. In the future, we would like to try other general knowledge bases like the Chinese Knowledge Base [11], a large-scale knowledge base built from the Chinese Wiki Encyclopedia.

Acknowledgements

This research is supported by the National Nature Science Fund Project (61562093) and the Key Project of Applied Basic Research Program of Yunnan Province (2016FA024).

References

1. S. Abeyruwan, U. Visser, V. Lemmon, and S Rer. "PrOntoLearn: Unsupervised Lexico-Semantic Ontology Generation Using Probabilistic Methods," International Conference on Uncertainty Reasoning for the Semantic Web, vol. 74, no.9, pp. 25-36, 2010
2. J. Brank, M. Grobelnik, and D. Mladenić. "A Survey of Ontology Evaluation Techniques," SIKDD at Multiconference 2005.

3. P. Cimiano, and S. Staab, "Learning by Googling," ACM SIGKDD Explorations, vol. 6, no.2, pp. 24-33, 2004
4. Z. D. Dong, and Q. Dong. "HowNet Knowledge Database," http://www.keenage.com
5. Z. H. DENG, S. W. TANG, M. ZHANG, D. Q. YANG, and J. CHEN. "Overview of Ontology," Acta Scientiarum Naturalium Universitatis Pekinensis, vol. 38, no.5, pp. 730−738, 2002
6. T. R. Gruber, "A Translation Approach to Portable Ontology Specifications," Technical Report, KSL 92·7 1, Knowledge System Laboratory, 1993
7. Hearst, and A. Marti, "Automatic Acquisition of Hyponyms from Large Text Corpora" Conference on Computational, Linguistics, pp. 539-545, 1992
8. B. L. Hua, Y. L. Liu, and Y. N. Zheng, "Studies on Methods of Formulating Rules for Academic Definition Extraction," Information Studies: Theory & Application, vol. 34, no.12, pp. 5-9, 2011
9. International Organization for Standardization – ISO. (2000b). ISO 1087-1, "Terminology Work, Vocabulary - Part 1: Theory and Application," Geneva: ISO.
10. R. Iqbal, "An Analysis of Ontology Engineering Methodologies: A Literature Review," Research Journal of Applied Sciences Engineering and Technology, vol. 6, no. 16, pp. 2993–3000, 2013
11. Knowledge Engineering Group, Tsinghua University, "A Large Scale Knowledge Base Built from Chinese Wiki Encyclopedia," http://keg.cs.tsinghua.edu.cn/project/ChineseKB/zhan_dian/Chinese_Knowldge_Base.html
12. D. Maynard, Y. Li, and W. Peters, "NLP Techniques for Term Extraction and Ontology Population," Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text, IOS Press, 2008
13. L. K. Mcdowell, and M. Cafarella, "Ontology-driven, Unsupervised Instance Population," Web Semantics Science Services & Agents on the World Wide Web, vol. 6, no.3, pp. 218-236, 2008
14. G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos, "Ontology Population and Enrichment: State of the Art," Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Springer Berlin Heidelberg, pp. 134-166, 2011
15. K. Purabi, and K. B. Anup, "Word Sense Disambiguation: A Survey," International Journal Of Engineering And Computer Science, vol. 4, no.5, pp. 11743-11748, 2015
16. F. QIAN, and C. F. YUAN, "A Definition Extraction Algorithm Combining Hard Pattern Matching and Soft Pattern Matching" , COMPUTER TECHNOLOGY AND DEVELOPMENT, vol. 22, no.9, pp. 32-36, 2012
17. C. Wen, Z.X. Shi, and X Zhang, "A Survey on Ontology Concept Hierarchy Acquisition," Computer Applications and Software, vol. 27, no.9, pp. 103-107, 2010
18. W. Wong, W. Liu, and M. Bennamoun, "Ontology Learning from Text: A Look Back and into The Future," , ACM Computing Surveys, vol. 44, no.4, pp. 20–36, 2012
19. E. D. Xun, and C. Li, "Applying Terminology Definition Pattern and Multiple Features to Identify Technical New Term and Its Definition," Journal of Computer Research and Development, vol. 46, no.1, pp. 62-69, 2009
20. X. M. Yao, "Studies on Automatic Domain Concept Extraction," , Kunming University of Science and Technology, 2010
21. A. Zouaq, and R. Nkambou, "A Survey of Domain Ontology Engineering: Methods and Tools," In Advances in Intelligent Tutoring Systems, Springer-Verlag, Berlin Heidelberg, pp. 103-119, 2010

Jian Xu borns in 1977. Master, associate professor. Member of the Teacher Education Branch of the National Institute of Computer Education, Executive Director of the Yunnan Computer Teaching and Research Association. His main research interests include knowledge graph, machine learning and natural language processing.

Jianhou Gan borns in 1976. PhD, professor, Master Tutor. Academic and Technical Leader in Yunnan Province, he was selected as the "Light of the West" talent training program. His main research interests include intelligent information processing and database technology.

Xianming Yao borns in 1984. Master, lecturer. His main research interests include ontology, information extraction and web search engine.

Liming Zhang borns in 1978. Master, lecturer. Her main research interests include knowledge graph and News communication.