Chinese Core Ontology Construction from a Bilingual Term Bank

Chen Yirong, Lu Qin, Li Wenjie, Cui Gaoying
Department of Computing, the Hong Kong Polytechnic University
E-mail: csyrchen@comp.polyu.edu.hk, csluqin@comp.polyu.edu.hk, cswjli@comp.polyu.edu.hk, csgycui@comp.polyu.edu.hk

Abstract

A core ontology is a mid-level ontology which bridges the gap between an upper ontology and a domain ontology. Automatic Chinese core ontology construction can help quickly model domain knowledge. A graph based core ontology construction algorithm (COCA) is proposed to automatically construct a core ontology from an English-Chinese bilingual term bank. This algorithm computes the mapping strength from a selected Chinese term to WordNet synset with association to an upper-level SUMO concept. The strength is measured using a graph model integrated with several mapping features from multiple information sources. The features include multiple translation feature between Chinese core term and WordNet, extended string feature and Part-of-Speech feature. Evaluation of COCA repeated on an English-Chinese bilingual Term bank with more than 130K entries shows that the algorithm is improved in performance compared with our previous research and can better serve the semi-automatic construction of mid-level ontology.

1. Introduction and Related Works

A core ontology, in a three level ontology hierarchy, is the mid-level ontology which models the most fundamental domain concepts according to the original backbones established by an upper ontology (Navigli, 2004). An upper ontology is a general ontology to ensure reusability across different domains. A domain ontology conceptualizes a specific domain. A core ontology bridges the gap between a specific domain and the abstract concepts in an upper ontology. The concept included in the core ontology is named as ***core concept***. The most representative lexical term for a given concept is referred to as the ***core term*** of core concept (Ji, 2007). In order to support the construction of different application dependent domain ontology, it is important to obtain a mid-level ontology which represents the core concepts in a specific domain. Thus proper taxonomy and reasoning axioms in a well formed upper ontology can be inherited by the core ontology to facilitate the construction of application oriented domain ontology. For example, as shown in Figure 1, "software" is normally not defined in an upper ontology. However, once the concept "software" is identified in a mid-level ontology with mappings to the upper ontology, related domain concepts such as "free software" and "public domain software" can be easily added into the domain specific ontology and they can inherit all the features of the general concept "ComputerProgram" in the upper ontology.

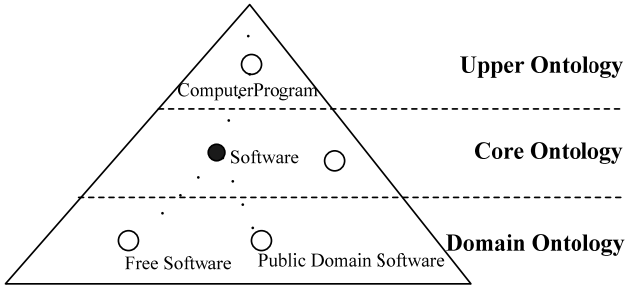

Figure 1: Three Levels of Ontology

There are not many directly related works on automatic core ontology construction. Some ontology related researches build up their concept nodes using a core lexicon (Hirst, 2004). However, many of the ontologies including Chinese ontology are manually built (Huang, 2004; Tang, 2005) and are often for a special application (Doerr, 2003). Few works are reported on automatic construction of core lexicon and core ontology. It is more difficult to automatically construct Chinese core ontology because there are relatively limited natural language resources compared to that of English. For example, HowNet, the most commonly used semantic resource for Chinese, contains about 20K number of concept nodes in its 2002 version (Dong, 2006) whereas WordNet (Fellbaum, 1998) contains close to 130K concept nodes in its version 1.6 database.

Besides core ontology construction, upper ontologies are mainly manually constructed. A widely known upper ontology is the Suggested Upper Merged Ontology (SUMO) which is a part of the IEEE Standard of the Upper Ontology Working Group by merging public available ontological content into a single, comprehensive and cohesive structure (Pease, 2002; Niles, 2001). A widely known English lexicon ontology is WordNet (Miller, 1990). An important notion in WordNet is "Synset" which is a set of synonym terms representing one unique concept. Different senses of a word are included in different synsets. From the perspective of a ontology, a synset is equivalent to a concept in an ontology. Another two famous upper level ontology works based on lexicon are CoreLex (Buitelaar, 1998) and the base synsets of EuroWordnet (Rodríguez, 1998) .

An earlier attempts on automatic Chinese core ontolog construction uses a multi-route core ontology construction algorithm (MRCOCA) (Chen, 2007) which tries to makes use of bilingual resources. In MRCOCA,

the terms representing the core concepts are referred to as core terms. For a given set of Chinese core terms, MRCOCA maps the Chinese core terms to English terms first, and then the English terms are mapped to their synsets in WordNet which are then mapped to SUMO concept nodes. But mapping to SUMO can only achieve accuracy of about 50% in MRCOCA because the features used in MRCOCA are quite primitive.

In this paper, a graph based core ontology construction algorithm (COCA) is proposed to automatically construct a core ontology from an English-Chinese bilingual term bank. This algorithm computes the mapping strength from a selected Chinese term to WordNet synset with association to an upper-level SUMO concept. The strength is measured using a graph model integrated with several mapping features from multiple information sources. The features include multiple translation feature between Chinese core terms and WordNet, extended string feature and Part-of-Speech feature. Evaluation of COCA repeated on an English-Chinese bilingual Term bank with more than 130K entries show that the algorithm is improved in performance compared to MRCOCA by about 42% accuracy improvement.

Section 2 introduces of the requirement of core ontology construction. Section 3 gives detailed algorithm description. Section 4 presents the evaluation of COCA algorithm. Section 4 concludes this paper.

2. Requirements of Core Ontology

As the purpose of a core ontology is to bridge the gap between an application oriented domain ontology to an upper ontology, it must maintain certain properties. The following gives the required properties which are considered fundamental in Chinese core ontology construction.

Firstly, the concepts represented in a core ontology must be **widely accepted and commonly referenced**. Since core ontology contains fundamental concepts in a specified domain, obviously a core concept should be widely used and commonly referenced in a consistent way in the domain. In case of polysemy, it is important to identify the correct sense. For example, the term "system" (系统) is widely used in the IT domain. But, one of its senses with the definition "*a group of physiologically or anatomically related organs or parts*" is obviously not a core concept in IT domain. So it is important to identify core terms and the appropriate sense(s) used in the domain.

Secondly, its representative core terms must be **highly used and should be productive to compose longer terms.** It is understood that a concept must presented by some lexical terms. As the realizations of the core concepts, core terms must be commonly used in a domain (e.g.: "software" in IT domain). Core terms should also have strong ability to form longer terms used in a domain ontology. In another words, core terms should have strong ability form longer terms so that the core ontology can link more domain specific concepts (often represented by longer domain terms) to upper ontology. An observation made by the study shows that the top 1,500 most productive core terms extracted can serve as suffixes to form more than 50% of the terms in a domain specific term bank (The term bank contains about 130K entries of IT terms). And since in most cases (93%), a Chinese term act as the suffix of a longer term is the head words of the longer term (Cui, 2008), the core ontology constructed using core terms can directly map most of the domain specific concepts to the upper ontology in theory.

Thirdly, the concepts/terms can be **mapped to upper ontology**. This ensures that the core ontology is not a dangling concept which will not have any relation to the upper ontology. It also ensures that all concepts in the core ontology can inherit the attributes provided by upper ontology. Upper ontologies are relatively small in size and are more carefully designed with additional more information such as axioms. Furthermore, the mapping to upper ontology can help the domain ontology to merge or interoperate with other domain specific ontology.

3. Core Ontology Construction Algorithm

The core ontology construction algorithm (COCA) is designed to construct a core ontology for Chinese. As there is not much Chinese NLP resources available, COCA is designed to make use of both a comprehensive Chinese-English term bank, and also the English WordNet and SUMO where each concept node is mapped to a synset in WordNet already. The main idea of COCA is to map each Chinese core term T_c to the most appropriate synset $Synset_c$ in WordNet first. It then make use of both the SUMO hierarchy, the synset of each SUMO object T_m , $Synset_m$ to WordNet, to build the ontology structure of the core terms. As a result, the core ontology is constructed by inheriting the hierarchical structure of SUMO extended by the hypernym structures of Wordnet. The main issue, however, is that given a Chinese core term T_c, how to map it to the appropriate synset S in the WordNet. That is

$$\arg\max_{S} P(S \mid T_C)$$

To find the appropriate mapping, two levels of ambiguity must be addressed. Firstly, a Chinese core term T_c, as a lexical item, can have multiple translations into English. Secondly, each English lexical item can have multiple senses and thus correspond to different synset in WordNet. Facing these two levels of multiplicity, the main goal of COCA is to find the most appropriate synset in WordNet for a Chinese core term.

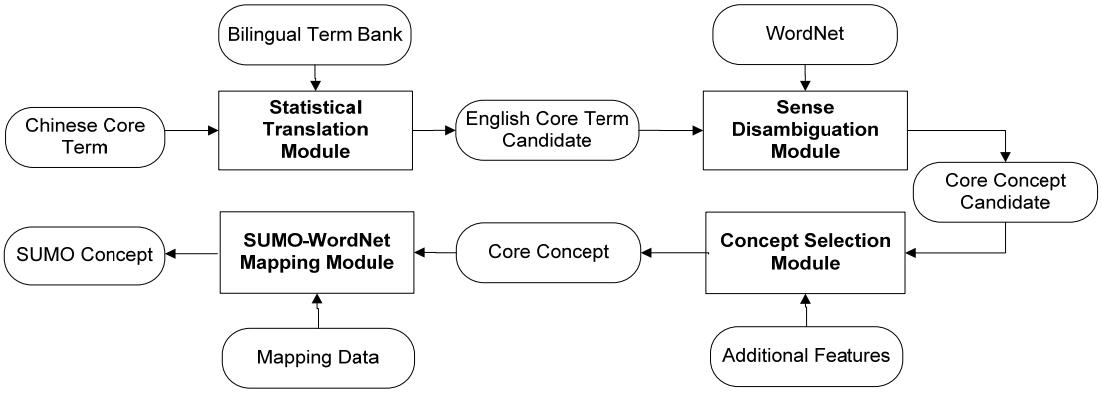

Figure 2: the Frame Work of COCA

Figure 2 shows the framework of COCA. It takes a Chinese core term as input and the corresponding core concept and a SUMO concept node as the output. The algorithm first use the bilingual term bank to map Chinese core terms into English core term candidates in the Statistical Translation Module. Second, the English core term candidates are mapped to core concept candidate (namely, synset candidate of Chinese core term) in the Sense Disambiguation Module using WordNet information. Thirdly, the core concept is finally selected among the set of candidates using multiple features (such as hyponym features) in the Concept Selection Module. Finally, using the mapping data in the SUMO-WordNet Mapping Module, the core concept is linked to a SUMO concept as the output.

3.1. Resources and Data Preparation

The COCA algorithm makes use of the following sets of resources: (1) a Chinese and English bilingual term bank, ***CETBank,*** a IT domain term bank from Institute of Computational Linguistics, Peking University which is a bilingual lexicon containing 130K Chinese general terms and IT terms, and their English translations; (2) WordNet; (3) the mappings between WordNet and SUMO nodes (Niles, 2003; SUMO-WordNet mapping data file, 2007).

The domain specific core term list, used in this work, referred to as ***ITCTerm,*** is the same as that used in (Chen, 2007). ***ITCTerm*** is extracted from the term bank ***CETBank*** by a simple segmentation algorithm which forces every entry T_i in ***CETBank*** be segmented so as to see what are the smaller component words/terms used to form T_i. (Ji ,2007; Chen, 2007). As a core term should not only be domain specific but also have strong ability to form other words, the words that are most frequently used as components of other domain terms in ***CETBank*** are selected as core terms. For example, "软件" is considered a core term because it is often used to form other terms, such as "软件工程", "软件设计", etc.. It should be pointed out that the obtained core term list ***ITCterm*** is a subset of ***CETBank***.

WordNet contains many different data. The following *synset[ ]* table lists out only five useful data fields for this work as

synset[syn_id, term_list, tag, gloss, hypernym_id].

Each data field is explained below.

a) *syn_id* is synset id unique to a synset. When used in combination with *tag*, it uniquely identifies a word sense.
b) *term_list* is a list of two tuples (*x, freq(x)*) where *x* is the term belong to this synset and *freq(x)* gives the frequency of *x* in this sense.
c) t*ag* is POS tag of the synset.
d) *gloss* is the definition of the synset in text form and examples of usage. This field is used as reference during manual evaluation..
e) *Hypernym_id* is the synset id of the hypernym of the synset. This field is used to identify hypernym relations for the core ontology.

The SUMO to WordNet mapping table *SW_Mapping[]* has three data fields for this work as

SW_Mapping[SUMO_term,syn_id, Mapping_method].

Each data field is explained below.

a) *SUMO_term* is a SUMO concept.
b) *syn_id* is the mapped synset id.
c) *mapping_method* indicates a mapping type from synset to this SUMO concept. A mapping can either be equivalent, subsumption, or instance type.

3.2. The Statistical Translation Module

For each Chinese core term T_C in ***ITCTerm***, suppose it has a translation T_E in the bilingual term bank. Because a term can have multiple translations, each T_C has a set of translations T_Set_E where $T_E \in T\_Set_E$. The objective of the Statistical Translation Module is to estimate the likelihood of every translation. $P(T_E \mid T_C)$ for all $T_E \in T\_Set_E$.

For a string s_1 and s_2, if s_1 is a substring of s_2($s_1 \subseteq s_2$), Then, s_2 is called the ***extended string*** of s_1. For example, the string "software" is the sub-string of the string "public software". Thus, "public software" is called the extended string of "software". For a translation pair $< T_C, T_E >$, if

there is a another translation pair such that $< T_{C\_e}, T_{E\_e} >$ satisfy the condition that T_{C_e} is the extended string of T_C and T_{E_e} is the extended string of T_E, respectively, the translation pair $< T_{C\_e}, T_{E\_e} >$ is called the ***extended translation pair*** of $< T_C, T_E >$. All the T_{E_e} forms an extended translation set for a given T_C, denoted as $ExtT\_Set(T_C)$. In this paper, two heuristics are considered in the probability model: (1) If the total number of extended translation pairs of a translation pair is larger, this translation should be more favorable; (2) If the difference between T_E and T_{E_e} is smaller in term of length, this translation should be more favorable. Then, a weight function $W(T_C, T_E)$ can be expressed as follows:

$$W(T_E \mid T_C) = \sum_{T_{E\_e} \in \text{ ExtT\_Set}(T_C)} \frac{len(T_E)}{len(T_{E\_e})} \quad (1)$$

where the function l*en()* returns the length of a string.

The probability of a given Chinese term T_C to be translated into T_E can then be expressed as normalized $W(T_E \mid T_C)$ given below

$$P(T_E \mid T_C) = \frac{W(T_C, T_E)}{\sum_{T_{Ei} \in T\_Set_E(T_C)}^{i} W(T_C, T_{Ei})} \quad (2)$$

3.3. The Sense Disambiguation Module

The Sense Disambiguation Module is the second step to map a given T_C to the Synset S through its translation set $T\_Set_E\ (T_C)$. Since a word has probabilities of taking different senses and the sense frequencies of a word are available in WordNet, the mapping probability from a English term T_E to a synset S is given firstly as following,

$$P(S \mid T_E) = \frac{F(T_E, S) + 1}{\sum_{x \in synset\ (T_E)} (F(T_E, x) + 1)} \quad (3)$$

where $F(T_E, S)$ the value from the field of *freq*() in the *term_list* data field of the *synset* table. If the $< T_E, S >$ pair does not occur in *synset[]*, the function $F(<T_E, S>)$ returns the value 0. So, for smoothing purpose, a value of 1 is added to each item in the formula.

Using first-order Markov chain model (Gilks, 1996) with the assumption that T_E is only influenced by T_C and S is only influenced by T_E, the synset path probability $P(S \mid T_C)$ for a given term T_C to take a particular synset S via an English translation T_E is computed in the following formula,

$$P(S \mid T_C) = P(T_E \mid T_C) * P(S \mid T_E) \quad (4)$$

Using the probability function $P(S \mid T_C)$ is enough to map a core term to a synset. However, the function $P(S|T_E)$ given in formula (4) often introduces more errors than function $P(T_E|T_C)$ given in formula (2) because the data used in formula (4) is taken from WordNet which is a general domain lexicon while the target core concepts are domain specific. To further improve performance, three additional features are introduced in Section 3.4.

3.4. The Concept Selection Module

In the Concept Selection Module, COCA takes three more features to improve performance compared to MRCOCA (Chen, 2007) including the multi-path feature, the hypernyms feature and the POS feature. Before introducing the features, a function to handle independent feature events is introduced in Section 3.4.1. Then Section 3.4.2 introduces the three features. Section 3.4.3 explains how the three features are integrated using the method discussed in Section 3.4.1.

3.4.1. Union Probability of Independent Events

The union probability of independent events comes with the assumption that features are independent events. Formula (5) is used to compute the union probability of independent events,

$$\underset{x \in E}{U}(p) = \begin{cases} 0 & |E| = 0 \\ p(x) + \underset{y \in E-\{x\}}{U}(p) - p(x) * \underset{y \in E-\{x\}}{U}(p) & |E| > 0 \end{cases} \quad (5)$$

where E is an event set, $p(x)$ is a probability function which returns the probability of event x.

If E is an empty set $\{\}$, then $U(p)$=0; if E is set $\{x\}$, then $U(p)=p(x)$; if E is set $\{x, y\}$, then $U(p)=p(x)+p(y)-p(x)p(y)$. Obviously function U returns the union probability of independent events when $|E|<=2$. For all other cases of $|E| > 2$, since events are independent, given an event x in E, the event set $Y = E-\{x\}$ can be considered a single event In these cases, y can be further processed as a new E. Then, the final value of U can be worked out recursively.

3.4.2. Three Features for Concept Selection

The features include multi-path feature, hypernyms feature and part-of-speech feature. Cases will be given for each feature to explain why the feature is introduced.

3.4.2.1. Feature 1 –Multi-Paths to Synset

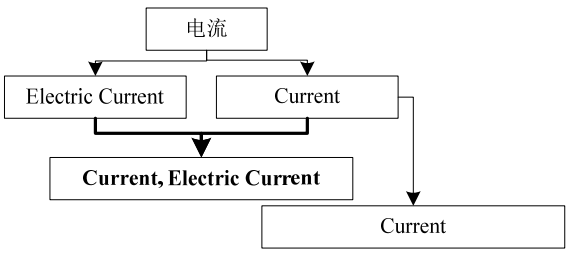

Figure 3: Case 1 – Multi-Path to Synset

Figure 3 shows a scenario of mapping of a Chinese term to a synset. The Chinese term "电流"(*current*) has two English translations "*current*" and "*electrical current*". The synset "current, electrical current" can be mapped from English term "*Current*" and "*Electrical Current*" respectively. But, the function in *formula (4)* only computes the probability via one path even when the synset can be mapped to by multi-paths from the original Chinese term. To better capture the mapping probability of between a Tc and it synset, the probability of multiple paths are merged into a single path as the Synset Probability, denoted as $SP(S \mid T_C)$, with their probabilities being added up using formula (4) as shown in the below formula.

$$SP(S \mid T_C) = \sum_{T_E \in \text{T\_Set}_E(T_C)} P(S \mid T_C) \qquad (6)$$

3.4.2.2. Feature 2 – Hyponym in domain

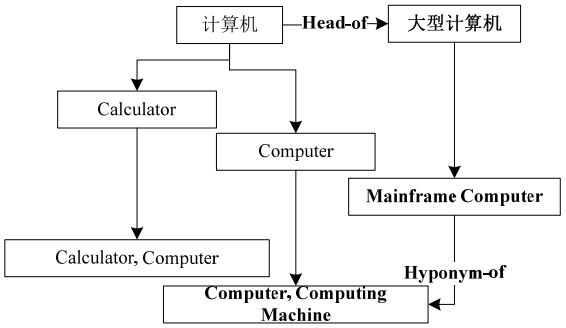

Figure 4: Case 2 – Using Hyponym in Domain

Given a core term, the first kind of additional features comes from all the extended strings which use the core term as headword. Figure 4 is an example of a core term "计算机"(*computer)* which are applied to all the extended string to map to a better concept using the hyponym relations. The Chinese term "计算机"(*computer)* can be mapped to English term *computer* and *calculator*. As the English term *computer*, when serving as a headword, has more extended terms(extended string) such as "大型计算机 "(*MainFrame Computer)* whose translation is a hypernyms of *computer*. Thus, it should give more confidence to take the translation of computer, rather than *calculator* who has no extended Chinese strings which can show any relation with *calculator.*

According to work in (Cui, 2008), about 93% extended string has their suffix term as the headword. Taking this fact as an assumption, the set of extended terms using T_C as suffix is obtained, using the function $Suffix\_Ext(T_C)$. A new measure, called the Hypernym Propability, denoted as HP($S \mid T_C$), can be formulated using the function $U(\ )$ given formula (6) with consideration of the hypernyms of those $Suffix\_Ext(T_C)$ as .

$$HP(S \mid T_C) = \underset{t \in \text{Suffix\_Etx}(T_C)}{U} \left(\sum_{h \in hyponym(S)} \frac{len(T_C)}{len(t)} SP(h \mid t) \right) \qquad (7)$$

where function *hyponym(S)* returns all the direct or indirect hyponyms of synset *S*, $Len(T_C)/ Len(t)$ is applied to compute according to the similarity between a core term T_C and its extended term *t.*

3.4.2.3. Feature 3 – Part of Speech

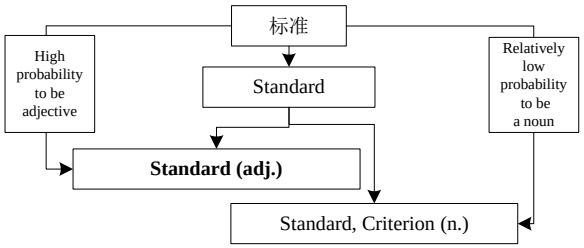

Figure 5: Case 3 – Using Part-of-Speech

Figure 5 shows a case of a better mapping when POS tag information is used. The term "标准" (standard) can be mapped to either the synset "Standard, Criterion" or "Standard". From the observation in the extended string of "standard", it can be found that "标准" (standard) has relatively low probability to be a noun while has a high probability to be an adjective. Therefore, the noun mapping should be selected. This shows that POS tagging information is also useful to select the correct meaning (or synset). However, due to the fact that the core term list nor its English translation has PoS tagging, a heuristics must be used to estimate the PoS of the Chinese core terms so its estimated PoS can match the synset with the same PoS. The estimation of PoS of the a *Tc* is done by searching *Tc* in all its extended Chinese terms. A simple heuristics is that it is more likely to be an adjective if it occurs in the beginning as a prefix , a verb if in the middle, and a noun at the end as a suffix. Then, a simple function $freq\_Pos(T_C, tag)$ is then developed which can obtain the frequency of T_C taking *tag* as its PoS by searching through the term bank ***CETBank***Then, the probability with PoS tagging being considered is denoted as $OP(S \mid Tc)$ which takes the following form:

$$OP(S \mid T_C) = \frac{freq\_pos(T_C, pos(S))}{\sum_{po \in \{noun, verb, adjective\}} freq\_pos(T_C, po)} \qquad (8)$$

where the function *pos(S)* returns the Pos tag of a synset *S.*

3.4.3. Integrate Features for Concept Selection

Finally, the probability for the mapping between T_C, and a *S* , denoted by $FP(S|T_C)$, can integrate all the features using the union probability of the these features as independent events as given below:

$$FP(T_C, S) = \max_{x \in \{SP(T_C,S), H(T_{CT},S), O(T_{CT},S)\}} U(x) \quad (9)$$

4. Evaluation

The output of COCA is pairs of <*Chinese-Core-Term, Synset*>. The evaluation takes the top 28 topped ranked nodes to manually evaluated if their mappings to Synset are the best match among all synset candidates of that Chinese core term. It should be pointed out that on average, there are 42 concept (synset) candidates for each Chinese core term evaluated. Accuracy is measured by the number of the correct mappings divided by 28 in this evaluation. The algorithm COCA achieves **71%** in accuracy. Compared to the result of MRCOCA (Chen, 2007) which achieved only 50%, the improvement is 42**%**.

Two examples of core term to syntset mapping generated by the algorithm are given in Appendix A for "软件" and "网络". Each entry shows a specific synset mapping the core terms listed in the second column. The entry in column 2 marked by the letter C indicates that it is the correct answer. The letter S indicates the choice by the COCA algorithm. The symbol "+" in the fifth column denotes that this synset is subsumed by the SUMO concept listed in column 4. In the example, for the core term" 软 件 " (software), there are 6 different synset mappings. Among them, the synset in row number 1 is the correct choice and the algorithm actually selected it as the answer. However, for the core term "网络" (net), the most appropriate choice is in row number 8. Yet, the algorithm picked row number 7 as its choice. However, by looking into more details of the SUMO objects, either one may not be appropriate. It is interesting to point out that the two SUMO objects do not have subsumption relationship.

5. Conclusion

In this paper, a graph based core ontology construction algorithm (COCA) is proposed to automatically construct a core ontology from an English-Chinese bilingual term bank. This algorithm computes the mapping strength from a selected Chinese term to WordNet synset with association to an upper-level SUMO concept. The strength is measured using a graph model integrated with several mapping features from multiple information sources. The features include multiple translation feature between Chinese core terms and WordNet, extended string feature and Part-of-Speech feature. Evaluation of COCA repeated on an English-Chinese bilingual Term bank with more than 130K entries show that the algorithm is improved in performance compared to MRCOCA by about 42% accuracy improvement.

Even though this algorithm introduced three features, analysis to these features should be conducted to see how effective they are, or how much they contribute to the performance. Other works to handle abbreviation and morphological conversion can also be evaluated.

6. Acknowledgements

This work is partially supported by CERG rant (PolyU 5225/05E: B-Q941), CERG project PolyU 5190/04E and a Hong Kong Polytechnic University funded project 4-Z08K.

7. References

Buitelaar, P., (1998). *CoreLex: Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

Chen, Y.R., Lu Q., Li, W.J., Li, W.Y., Ji, L.N., Cui, G.Y. (2007). Automatic Construction of a Chinese Core Ontology from an English-Chinese Term Bank. In *Workshop OntoLex07 – From Text to Knowledge: The Lexicon/Ontology Interface*, the 6th International Semantic Web Conference .

Cui, G.Y., Lu, Q., Li, W.J. (2008). Preliminary Chinese Term Classification for Ontology Construction. In *The 6th Workshop on Asian Language Resources*, in the Third International Joint Conference on Natural Lanugrage Processing (IJCNLP)

Doerr, M., Hunter, J., Lagoze, C., (2003). Towards a Core Ontology for Information Integration. In *Journal of Digital Information*, vol. 4, no. 1, pp. 169.

Dong, Z., Dong, Q., (2006). *HowNet and the Computation of Meaning*. World Scientific Publishing Co., Singapore

Gilks,W.R., Richardson, S. , Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Fellbaum, C., (1998). *Wordnet: an electronic lexical database*. MIT Press

Hirst, G., (2004). Ontology and the Lexicon. In *Handbook on Ontologies*, S. Staab and R. Studer: Springer, Karlsruhe, pp. 209-230.

Huang, C.N. (1997). Segmentation Problems in Chinese Processing. (In Chinese) In Applied Linguistics, 1, pp. 72-78.

Huang, C.R., Chang, R.Y., Lee, S.B., (2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 26–28.

Ji, L. N., Lu, Q., Li, Chen, Y.R. (2007). Automatic Construction of a Core Lexicon for Specific Domain. In *Proceeding of the 6th International Conference on Advanced Language Processing and Web Information Technology*. Luoyang, China

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J., (1990). Introduction to WordNet: An On-line Lexical Database. In *International Journal of Lexicography.* Oxford Univ Press, vol. 3, no. 4, pp. 235.

Navigli, R., Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. In *Computational Linguistics*, MIT Press

Cambridge, MA, USA, vol 30, no 2, pp151-179.

Niles, I., Pease, A., (2001). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems*. ACM Press New York, NY, USA. Volume 2001, pp. 2–9.

Niles, I., Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*. Las Vegas, Nevada

Pease, A., Niles, I. and Li, J. (2002). The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, Edmonton, Canada , vol. 28.

Rodríguez, H., Climent, S., Vossen, P., Bloksma L. and et al, (1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In *Computers and the Humanities*. Springer , vol. 32, no. 2, pp. 117–152.

SUMO-WordNet mapping data file, (2007). http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings

Tang, A.M., Zhen, Z., Fan, J. (2005). Thesaurus-based Approach to Build Domain Ontology. In *New Technology of Library and Information Service*, vol. 2005, no 4, pp. 1-5. (The journal is in Chinese)

Appendix: Two Examples of Synset mappings

No.	*Zh*	*En*	*SUMO Concept*		*Synset*
1	软件(SC)	Software	ComputerProgram	+	software,software_system
		(computer science) written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory			
2	软件	Facility	StationaryArtifact	+	facility,installation
		something created to provide a particular service; "the assembly plant is an enormous facility"			
3	软件	Facility	SubjectiveAssessmentAttribute	+	proficiency, facility, technique
		skillfulness in the command of fundamentals deriving from practice and familiarity; "practice greatly improves proficiency"			
4	软件	Facility	SubjectiveAssessmentAttribute	+	adeptness,adroitness,deftness,facility,quickness
		skillful performance without difficulty; "his quick adeptness was a product of good design"			
5	软件	facility	Room	+	toilet, lavatory, lav, can, facility, john, privy, bathr
		a room equipped with washing and toilet facilities			
6	软件	facility	SubjectiveAssessmentAttribute	+	facility,readiness
		a natural effortlessness; "a happy readiness of conversation"--Jane Austen			
7	网络(S)	net	Artifact	+	network,net,mesh,meshwork,reticulation
		an interconnected or intersecting configuration or system of components			
8	网络(C)	network	Collection	+	network,web
		an intricately connected system of things or people; "a network of spies" or "a web of intrigue"			
9	网络	network	SocialInteraction	+	network
		communicate with and within a group; "You have to network if you want to get a good job"			
10	网络	net	Pursuing	+	net,nett
		catch with a net; "net a fish"			
11	网络	net	Making	+	web,net
		construct or form a web, as if by weaving			
12	网络	net	SubjectiveAssessmentAttribute	+	final,last,net
		conclusive in a process or progression; "the final answer"; "a last resort"; "the net result"			
13	网络	net	CurrencyMeasure	+	net,nett
		remaining after all deductions; "net profit"			
14	网络	net	FinancialTransaction	+	net,sack,sack_up,clear
		make as a net profit; "The company cleared $1 million"			
15	网络	net	FinancialTransaction	+	net,clear
		yield as a net profit; "This sale netted me $1 million"			
16	网络	grid	Device	+	grid,gridiron
		a utensil of parallel metal bars; used to grill fish or meat			
17	网络	grid	EngineeringComponent	+	grid,control_grid
		an electrode placed between the cathode and anode of a vacuum tube to control the flow of electrons through the tube			
18	网络	grid	Abstract	+	grid,reference_grid
		a network of horizontal and vertical lines that provide coordinates for locating points on an image			
19	网络	internet	Device	+	internet,cyberspace
		worldwide network of computer computer networks that use the TCP/IP network protocols to facilitate data transmission and exchange			
20	网络	net	Fabric	+	net,mesh
		an open fabric woven together at regular intervals			
21	网络	net	Device	+	net
		a trap made of netting to catch fish or birds or insects			