Considerations of Linking WordNet with MRD

Changhua Yang
Department of Computer and Information Science, Soochow University
Taipei, Taiwan 100
ms8903@cis.scu.edu.tw

Sue J. Ker
Department of Computer and Information Science, Soochow University
Taipei, Taiwan 100
ksj@cis.scu.edu.tw

Abstract

This paper introduces a bilingual MRD (English-Chinese LDOCE) to help with the construction of a Chinese WordNet. Considerable linking strategies are discussed and proper translations are attached to 28,388 noun synsets and 10,380 verb synsets of WN 1.7 automatically. The steady precision (92~94%) shows the corresponding headwords and matching keywords provide considerable guarantee for linking. The Chinese phrases need to be processed further to match the correct synset due to the fine-grained senses of WordNet.

Introduction

Princeton WordNet[1] was developed by a group led by George A. Miller (Miller 1993). WordNet contains information about the nouns, verbs, adjectives and adverbs in English. Words of the same part-of-speech are organized as *synset* to reflect their synonymous relationship. The main semantic relations among synsets are *hypernymy*, *hyponymy*, *meronymy* and *antonymy*. As researchers who are highly interested in discovering ontology information within a lexical knowledge base, currently we focus on the network organized by the *hypernymy/hyponymy* relationship.

Inspired by English WordNet, researchers within other languages have been developing their own vocabulary network. For example, the EuroWordNet (Vossen, 1998) project collects many languages' vocabulary knowledge and organizes it following the construction principles of WordNet. It indeed expands the WordNet family of world languages. If we add new language networks to the family with similar structures, it will result in lower translation costs to communicate with the WordNet-like lexical knowledge base in existence.

Every language relies on previous bilingual translation knowledge to automatically construct a new network. Machine Readable Dictionaries (MRD) are representative of this kind of knowledge. In Chinese, for this study we explore the usage of the Longman English-Chinese Dictionary of Contemporary English (Proctor, 1988) (abbreviated as E-C LDOCE) to accomplish this work. The Chinese translation field in the dictionary is linked to the synset of WordNet according to the semantics clues provided by keywords within the English definition sentences.

This paper is organized as follows. Section 1 discusses some related researches. In Section 2 we observe the resources and show the linking considerations and difficulties. In Section 3 the considerations are explored and solved further with several weighting strategies. Section 4 displays the experimental results from applying these strategies. The discussion is developed in Section 5, followed by the conclusion.

1 Related Works

It requires strenuous effort to construct the Chinese WordNet. The most trivial solution is to interpret the synset as some set of Chinese words manually by referencing the English definition and example sentences. This method relies on linguistic experts' devotion and is very time-consuming.

Agirre and Rigau (1996) introduced a supervised statistical WSD method using the

[1] http://www.cogsci.princeton.edu/~wn/, June 2002.

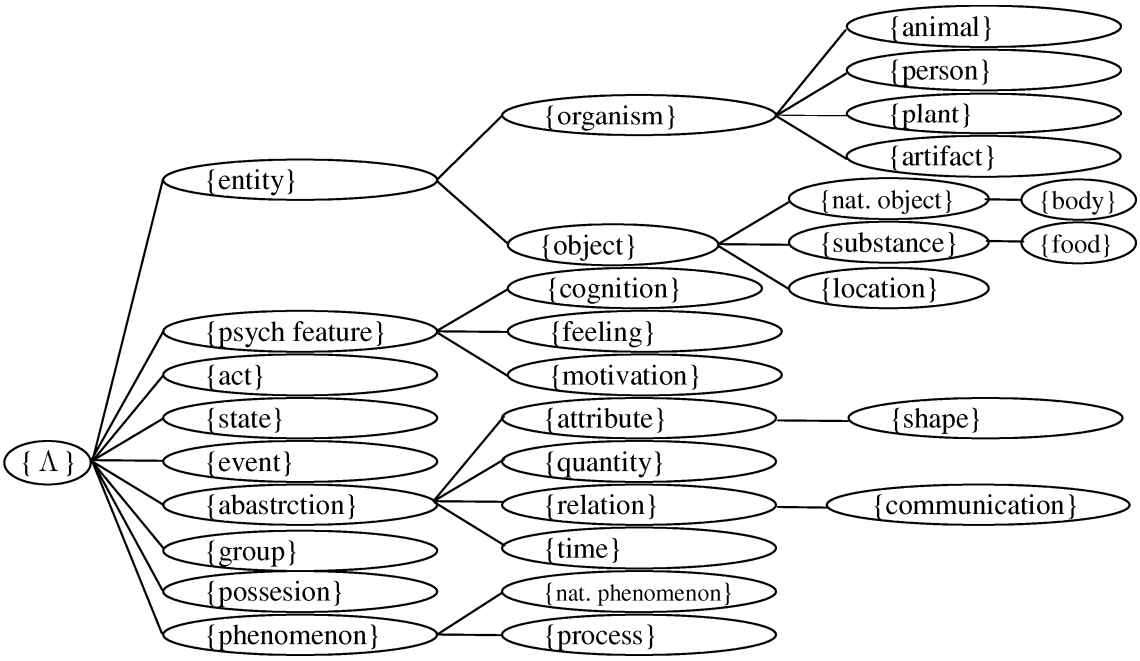

Figure 1 Beginners of noun hierarchy.

concept density of WordNet. Lee et al. (2000) explore the automatic construction of a Korean WordNet by using WSD heuristics. Bilingual MRDs are introduced in order to remove the ambiguities among mappings from Korean to English. The usage of a bilingual translation knowledge is also found in the work of Chen and Lin (2000). They map WordNet synsets to a Chinese thesaurus and tag the Chinese Corpus with reduced ambiguities.

Keywords provide the clue for linking dictionary definitions to another lexical knowledge base. Chang and Chen (1998) group the LDOCE definitions as topical clusters and link them to a thesaurus. A technique for Information Retrieval (IR) is used to complete this task. Besides, they argue that MRD's division of senses is often too fine for the task of WSD. However, we illustrate in the discussion that WordNet senses are more fine-grained by referencing the result of our linking task.

IR technique is also utilized by Carpuat et al. (2002). They align WordNet with a Chinese lexical knowledge base called HowNet (Dong and Dong, 1998). Statistical data from a bilingual corpus provides the seed vector information for translation.

Semantic relations like *hypernymy* relations provide positive sense information in language processing. Mihalcea and Moldovan (2000) propose a WSD algorithm to be used with WordNet. The indexing expanded with synonyms and hyponyms improves the efficiency of the Boolean Information Retrieval system. Chang et al. (1998) summarize the issues dealing with definitions in MRD. They elaborate the roles of genus terms building the taxonomy. We adopt their viewpoint and revise the linking by comparing the genus terms in MRD and hypernyms in WordNet.

2 Observation

2.1 Materials

As our source, we use WordNet 1.7, which contains 74,487 noun synsets and 12,753 verb synsets. Noun synsets are divided into several *hypernymy* hierarchies with 9 beginners; each of the beginners is inherited by all of its hyponyms. Figure 1 shows the top noun hierarchy of beginners in the current version. According to Miller (1998), a unique beginner {Λ} is set to be the hypernym of the 9 top beginners. It pulls all nouns into a single hierarchical structure. As for verbs, they are divided into 359 *hypernymy* hierarchies.

Table 1 Instances of E-C LDOCE for headword "club".

Sense ID	Definition	Translations
(lm.club.1)	a society of people who join together for a certain purpose, esp. sport or amusement	俱樂部 (chu4 le4 pu4)
(lm.club.2)	a building where such a society meets	俱樂部會址 (chu4 le4 pu4 hui4 chih3)
(lm.club.3)	a heavy wooden stick, thicker at one end than the other, suitable for use as a weapon	短棒 (tuan3 pang4); 棍 (kun4)
(lm.club.4)	a specially shaped stick for striking a ball in certain sports, esp. GOLF	高爾夫球棒 (kao erh3 fu chiu2 pang4); 高爾夫球桿 (kao erh3 fu chiu2 kan3)
(lm.club.5)	a playing card with one or more 3-leafed figures printed on it in black	梅花(mei2 hua)

Table 2 Concept mappings of word X.

	Sense of MRD Definitions				
Sense of Synsets	L_1	L_2	L_3	…	L_y
$S_1\{\ldots, X, \ldots\}$	$M_{1,1}$	$M_{1,2}$	$M_{1,3}$	…	$M_{1,y}$
$S_2\{\ldots, X, \ldots\}$	$M_{2,1}$	$M_{2,2}$	$M_{2,3}$	…	$M_{2,y}$
$S_3\{\ldots, X, \ldots\}$	$M_{3,1}$	$M_{3,2}$	$M_{3,3}$	…	$M_{3,y}$
…	…	…	…	…	…
$S_x\{\ldots, X, \ldots\}$	$M_{x,1}$	$M_{x,2}$	$M_{x,3}$	…	$M_{x,y}$

The E-C LDOCE uses a controlled vocabulary of some 2,000 words to define over 60,000 word senses; about 37,200 of them are nouns and 15,600 are verbs. Table 1 shows the instances of sense definitions of word "club". The dictionary definitions are converted into Chinese translations (word or phrases). These Chinese translations are the candidates for linking to WordNet.

WordNet groups the polysemous senses by looking up the synsets that share the same word. It is similar to MRD in that WordNet also gives each synset a descriptive gloss. Though the gloss is not as well-organized as a dictionary definition, it provides the first hint for linking with these two resources.

2.2 Linking with Keywords

If *X* is a polysemous word, it should be organized according to comparative senses in WordNet synsets and MRD definitions. Table 2 shows this condition. We conjecture the main meanings of word *X* spread in *x* synsets and *y* MRD definitions.

To link WordNet with E-C LDOCE, we need to assign appropriate dictionary sense of headword *X* to synset *S* that contains the same word *X*. We assume that each linking of a dictionary sense to the synset is independent; the selected sense may be the candidate of next assignment.

Table 3 shows the linking instances of synsets that contain word "club". After part-of-speech filtering, we extract the keywords of the tagged sentence from glosses, reserving nouns, verbs, adjectives and adverbs, as shown in the "Keyword" field of Table 3. We then compare the keywords (with identical part-of-speech) with MRD definitions and pick the most satisfying sense bound with its Chinese translations. In this case, the superior linking sense contains more matching keywords than other senses.

2.3 Difficulties

One problem is that one synset contains several words and these words may be polysemous in MRD definitions. We need to decide which sense of the corresponding MRD definition stands for the synset. For nouns, there are 28,388 noun synsets sharing the headwords with 34,092 definitions in E-C LDOCE. Table 4 shows the average number of senses dealt with for disambiguation processing. Seventy-eight percent of these synsets contain only one word mapping and the average number of corresponding senses in E-C LDOCE is 2.41 per word. When a synset contains more words, it is easier to find a certain one that is less ambiguous.

3 Linking Considerations

3.1 Competition Weighting

For a given synset *S*, we fetch *m* words of synonym set $\{X_1, ..., X_m\}$ and *p* keywords of

Table 3 Instances of linking result with synsets containing the word "club".

Sense ID	Synonyms	Gloss	Keyword	Sense ID	Translations
(wn.club.1)	clubhouse, club	a building occupied by a club	building/N, occupy/V, club/N	(lm.club.2)	俱樂部會址 (chu4 le4 pu4)
(wn.club.2)	club	stout stick that is larger at one end	stout/A, stick/N, large/A, end/N	(lm.club.3)	短棒 (tuan3 pang4); 棍 (kun4)
(wn.club.3)	golfclub, club	golf equipment used by a golfer to hit a golf ball	golf/N, equipment/N, use/V, hit/V, ball/N	(lm.club.4)	高爾夫球棒 (kao erh3 fu chiu2 pang4); 高爾夫球桿 (kao erh3 fu chiu2 kan3)

Table 4 The distribution of ambiguity degree.

# of Words Need to Check	# of Synsets	Percentage	Average Degree of Word Ambiguity	
			Most Polysemous	Least Polysemous
1	22105	77.8%	2.41	2.41
2	4484	15.8%	3.66	1.7
3	1177	4.1%	4.34	1.39
4	363	1.3%	5.06	1.28
5	140	0.5%	5.97	1.18
6	67	0.2%	6.15	1.16
7	34	0.1%	7.21	1.15
8~19	18	0.06%	8.44	1.06

gloss definition $\{K_1, ..., K_p\}$. Among words $\{X_1, ..., X_m\}$ we need at least one X_q, $1 \le q \le m$, that indicates the linking bound with the translations. Suppose X_q has n corresponding MRD senses $\{L_{q,1}, \ldots, L_{q,n}\}$ and for some $L_{q,r}$, $1 \le r \le n$, there are o keywords $\{T_1, \ldots, T_o\}$.

We adopt the technique of IR here to measure the basic score of the mapping between S and $L_{q,r}$. We treat the synset as a query and the MRD senses as a document. The score of mapping terms is calculated by common scheme IDF (inverse document frequency).

For S and $L_{q,r}$, we fetch mapping terms from a joint set of keywords $\{X_1, \ldots, X_m\} \cap \{T_1, \ldots, T_o\}$. The weighting score w of *term* is calculated as equation (1), where N is the number of MRD senses and df_{term} is the number of appearances of *term* in MRD senses. Equation (2) sums up the mapping score W of $(S, L_{q,r})$ from weighting scores of the mapping terms.

In the case that synset S shares only one headword X_1 with E-C LDOCE, we look it up in dictionary senses. The mapping with the maximum score δ offers the best linking ϕ as shown in equation (3) and equation (4).

$$w = idf(term) = \log\left(\frac{N}{df_{term}}\right), \quad \text{equation (1),}$$

$$W = \sum_{t=1}^{f} w_t, \quad \text{equation (2),}$$

$$\delta = \max_{1 \le r \le n}(W_r), \quad \text{equation (3),}$$

$$\phi = \arg\max_{1 \le r \le n}(W_r), \quad \text{equation (4),}$$

where $f = |\{X_1, \ldots, X_m\} \cap \{T_1, \ldots, T_o\}|$ and n is the number of comparing senses in MRD.

On the other hand, if synset S shares more than one headword with E-C LDOCE; the competition takes place. We then adopt the algorithm, as shown in Figure 2, to choose the linking with maximum score offered by one of these headwords. If all corresponding headwords are polysemous, we pick the sense is most ahead of the others. This is achieved by measuring the weighting difference between the top two scores belonging to either headword.

3.2 Enhanced Weighting

Some glosses of synsets don't follow the lexicographic custom that combines genus and differentiating terms to form the definitions of words. We could trace these synsets' meanings from their ancestors in the hierarchy formed by *hypernymy/hyponymy* relations. We interpret

Algorithm ***competition***
// *w-score* (weighted score) is calculated by equation (2)
  for each word X_q of synset S
    fetch the senses set SS_q in MRD with corresponding headword X_q
    for each sense $L_{q,r}$ of SS_q
      weight *w-score* W_r between S and $L_{q,r}$
  if for some X_q of synset S, there is only 1 corresponding sense $L_{q,1}$ in MRD **then**
    fetch all of these $L_{q,1}$ of X_q as the linking result
  else
    for each word X_q of synset S
      sort sense $L_{q,r}$ of SS_q by W_r in descending order
      record $L_{q,1}$, $diff_q(W_1, W_2)$ // top sense & weighting-difference between 1st and 2nd
    sort X_q of synset S by $diff_q$ in descending order
    choose $L_{1,1}$ *of* X_1 as the best linking to S

Figure 2 Algorithm for competition.

Function ***EnhancedWeighting***(S, $L_{q,r}$) // measure the weight of (S, $L_{q,r}$)
// *w-score* (weighted score) is calculated by equation (2)
  for any synset H that is a 1-level hypernym of S
    sum up to $W_{HYPERYM}$ with *w-score* between H and $L_{q,r}$
  W_{ORIGIN} = *w-score* between S and $L_{q,r}$
  fetch the ancestor sets AS of S
  if $\{ T_1, \ldots, T_o \} \cap AS \neq \varnothing$ **then**
    $W_{ORIGIN} = W_{ORIGIN} + \log(N)$
  return $W_{ORIGIN} + W_{HYPERYM}$

Figure 3 Function of enhanced weighting.

this clue in two ways. First, we may recover the missed description in the hypernyms' glosses while weighting. Second, as MRD uses the categorized genus terms, the terms should be searched out in the ancestors of the proper corresponding synsets that share the same headwords.

These two considerations modify equation (2) as function ***EnhancedWeighting*** shown in Figure 3. $W_{HYPERNYM}$ augments the weighting score from the keywords shared by the hyponyms' glosses and dictionary definitions. In WordNet it is possible that one node inherits multiple hypernyms. The scores they contribute are summed up. W_{ORIGIN} modifies the weighting to reflect the consideration of genus terms by adding a sufficiently large score. We don't parse the definition sentence to fetch the exact genus but verify if some possible terms exist that categorize the headword while they are discovered among the corresponding ancestors. We show in the experimental results that these two improvements enlarge the coverage of the linking task.

4 Experimental Results

There are 28,388 noun synsets and 10,380 verb synsets sharing the headwords that can be found in E-C LDOCE. Linking coverage and precision are the qualifying point for different linking strategies. After linking, we qualify 50 randomly chosen results and verify them with human judgment.

With nouns, we implement competition weighting and 20,714 (73.0%) synsets are solved with 92% precision. We then adopt the enhanced weighting function. Of those without consideration of genus, there are 23,186 (81.7%) synsets covered with 94% precision. 24,434 (86.1%) synsets are covered with 92% precision with full-function enhanced weighting. Table 5 shows the summary of the performance.

Table 5 Linking results of noun synsets.

Strategy	Covered Synsets	Coverage	Precision
Competition Weighting	20,714	73.0%	92%
plus W_{HYPERNYM} Consideration	23,186	81.7%	94%
Full Enhanced Weighting	24,434	86.1%	92%
# of Synsets	28,388		

Table 6 Linking results of verb synsets.

Strategy	Covered Synsets	Coverage	Precision
Competition Weighting	7,106	68.5%	94%
plus W_{HYPERNYM} Consideration	8,177	78.8%	92%
Full Enhanced Weighting	8,641	83.2%	92%
# of Synsets	10,380		

There are also *hypernymy* relations among verb synsets, so we can implement the same weighting algorithm as with nouns. Table 6 shows that with full-function enhanced weighting, 83.2% percent of verb synsets could be linked with 92% precision.

5 Discussion

WordNet collects many vocabulary items that don't appear in common dictionaries. For example, synsets {phaneromania} and {logorrhea, logomania} have no corresponding headword in E-C LDOCE. Hence we could not attach any Chinese translation to them. From our observation, there are 46,099 noun synsets that share no vocabulary with E-C LDOCE. Table 7 shows the data of headword mappings under some hierarchies of synset beginners. We find 85%~87% of the hyponymy synsets have no headword mappings under beginners {group, grouping}, {plant, flora} and {region}. Words under these categories include some artificial collocations or are too domain-specific.

While examining the Chinese translations of synsets after successful linking, we find some arguable outcomes. In the case of "digger", it was located in E-C LDOCE with the definition: "a person or machine that digs" and Chinese translations: "挖掘者(wa chueh2 che3); 挖掘機(wa chueh2 chi)". WordNet interprets the same word from a finer viewpoint. The definition for "digger" was separated into two synsets whose glosses are "a laborer who digs" and "a machine for excavating". After linking WordNet with E-C LDOCE, we find the two Chinese translation phrases "挖掘者; 挖掘機" are attached to both synsets, even though "挖掘者" indicates a person and "挖掘機" is a machine.

A learner who looks "digger" up in E-C LDOCE would be able to acquire the necessary explanation with the context of the word. But, in Chinese, "挖掘者" and "挖掘機" are not synonyms. They just embrace the significant prefix "挖掘(wa chueh2)" representing "to dig". Chinese phrases with suffix "者(che3)" represent "someone who" and those with suffix "機(chi)" are kinds of "machine, engine, or airplane". The phrases need to be processed further to match the correct synset. Ker and Song (2002) parse the Chinese translations of synsets and tag the primary semantic components of the translations with HowNet. Their work may serve as the post-processing of our methods.

Because Chinese translations are additionally bound to WordNet synsets, they violate nothing of the original hierarchy. Yu (2002) define this kind of hierarchy as WordNet-like lexicon. And further, Liu et al. (2002) explore the construction criteria of its tree structure. However, the Chinese translations could be processed further to extract the semantic elements, basically one-character words, forming a new Chinese word hierarchy. Wong and Pala (2002) compare Chinese characters and radicals with EuroWordNet Top Ontology. They observe Chinese characters under several radicals, which are one way that Chinese represents concepts.

Conclusion

This paper provides an entry point for the construction of Chinese WordNet. The Chinese translations of a bilingual dictionary are attached to WordNet after linking. Descriptive keywords for word senses provide semantic information and help in the disambiguation task of linking. Among these keywords, we stress the importance of categorized terms and give them added weight in the linking process. It

Table 7 Analysis of headword mappings from WordNet to E-C LDOCE under beginners.

Synset Beginner	# of H*yponymy* Synsets	with Mappings	Percentage	without Mappings	Percentage
{object}	19660	9337	47.5%	10323	52.5%
{cause}	11426	3988	34.9%	7438	65.1%
{group, grouping}	7466	1166	15.1%	6580	84.9%
{person, someone}	10224	3685	36.0%	6539	64.0%
{plant, flora}	4913	624	12.7%	4289	87.3%
{animal, creature}	4019	1015	25.3%	3004	74.7%
{region}	2904	415	14.3%	2489	85.7%
{psy. feature}	4316	1940	44.9%	2376	55.1%
{state}	3103	1086	35.0%	2017	65.0%

does indeed help to expand the coverage of the linking task.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments. This research is partially supported by the ROC NSC grants 90-2213-E-031-005.

References

Agirre, E. and G. Rigau (1996) *Word Sense Disambiguating using Conceptual Density*, In Proceedings of COLING-96, Denmark.

Carpuat, M., G. Ngai, P. Fung and K. W. Church (2002) *Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet*, In Proceedings of First International WordNet Conference, India.

Chang, J-S. and J-N. Chen (1998) *Topical Clustering of MRD Senses Based on Information Retrieval Techniques*, Computational Linguistics, 24/1, pp. 61-95.

Chang, J-S., S-J. Ker and M. H. Chen (1998) *Taxonomy and Lexical Semantics – from the Perspective of Machine Readable Dictionary*, In Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, pp. 199-212.

Chen, H-H. and C-C. Lin (2000) *Sense-Tagging Chinese Corpus*, In Proceedings of ACL-2000 Second Chinese Language Processing Workshop, pp. 7-14.

Dong, Z. and Dong Q. (1998) HowNet, http://www.keenage.com/, June 2002.

Ker, S-J. and C-S. Song (2002) *Syntactic and Semantic Analysis of Chinese Phrases*, In Proceedings of the 3rd Conference of Chinese Lexical Semantics, pp. 57-68.

Lee, C., G. Lee and J. Seo (2000) *Automatic WordNet Mapping Using Word Sense Disambiguation*, In Proceedings of the ACL-2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Copora, pp. 142-147.

Liu, Y., J. Yu and S. Yu (2002) *A Tree-structure Solution for the Development of ChineseNet*, In Proceedings of First International WordNet Conference, India.

Mihalcea, R. and D. I. Moldovan (2000) *Semantic Indexing Using WordNet Senses*, In Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, pp. 35-45.

Miller, G. A. (1993) *Five Papers on WordNet*, http://www.cogsci.princeton.edu/~wn/, June 2002.

Miller, G. A. (1998) *Nouns in WordNet*, In "WordNet: An Electronic Lexical Database", C. Fellbaum ed., The MIT Press, London, England, pp. 23-46.

Proctor, P., ed. (1988) Longman English-Chinese Dictionary of Contemporary English, Longman Group (Far East) Ltd., Hong Kong.

Vossen, P. ed. (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, London, England.

Wong, S. H. S. and K. Pala (2002) *Chinese Characters and Top Ontology in EuroWordNet*, In Proceedings of First International WordNet Conference, India.

Yu, J. (2002) *Evolution of WordNet-like Lexicon*, In Proceedings of First International WordNet Conference, India.