Semi-automatic Compilation of Asian WordNet
半自動アジア・ワードネットの構築

Thatsanee Charoenporn, Virach Sornlertlamvanich, Chumpol Mokarat
Thai Computational Linguistics Lab., NICT Asia Research Center,
Pathumthani, Thailand
{thatsanee, virach, chumpol}@tcllab.org

Hitoshi Isahara
NICT, Japan
isahara@nict.go.jp

近年、単語知識として広く使われているものはワードネットがある。単語の意味を表現するために同義語群(synset)が使われるのは特徴である。英語ワードネットから始まって多数の言語ワードネットが構築されてきた。ユーロ・ワードネットの構築によってヨーロッパ言語の言語処理研究、応用などが著しく行われた。同様にアジア地区でもアジア諸言語のワードネット化の努力がされつづある。中国語をはじめ、韓国語、ヒンディー語、日本語などのワードネット化の研究、応用が見られてくる。本研究では、多様なアジア言語を如何にワードネット化し、言語間の意味的関係をワードネットの構造上の特徴によって自動的に対応付ける仕組みを提案する。基ワードネットの意味表現を活用するためには少なくとも各言語の英語への対訳辞書が必要である。我々は辞書の最小限の情報、いわゆる、見出し語と品詞情報のみで基ワードネットと対応を自動的に行う。各単語の英訳の重複程度を調べ、対応の確信度を決定する。その結果、単語のワードネット構造が形成された。また、基の同義語群(synset)を介してアジア言語間の意味的関係が得られる。あいまい性解消のため KUI(Knowledge Unifying Initiator, http://www.tcllab.org/kui)というウェブ上の社会ネットでアジア・ワードネットを提供し、誤りを集団的に修正の協力を呼びかける。

WordNet is a kind of word knowledge database which is widely used in the recent years. Basically, the word concept is defined by a set of its synonyms, called synset. English WordNet was originally proposed and developed at Princeton University. Since then, WordNet for several languages such as Euro WordNet were constructed. For Asian languages, the efforts on creating WordNet for Chinese, Korean, Hindi, and Japanese can also be found. This paper aims to create a linkage among Asian languages by adopting the concept of semantic relations and synset expressed in WordNet. Based on the Princeton WordNet (PWN), we propose a method in generating a WordNet by using an existing bi-lingual dictionary. Our algorithm is to align the PWN synset to a bi-lingual dictionary through the English equivalent and its part-of-speech (POS). Number of English equivalent of a word in a synset increases the degree of confidence in the synset assignment process. We also introduce a web-based collaborative workbench, called KUI (Knowledge Unifying Initiator), for revising the result of synset assignment and provide a framework to create Asian WordNet via the linkage through PWN synset.

1 Introduction

WordNet [1] is a kind of word knowledge database which is widely used in the recent years. The original WordNet is English WordNet proposed and developed at Princeton University. Princeton WordNet (PWN) is designed as a collection of synsets that represent synonymous English lexemes which are connected to one another with semantic relations such as hyponymy, meronymy, antonymy, and entailment. That is “synset” is used to represent “meaning” of the word entry. This structure can be mirrored in most of the WordNets developed on the basis of PWN. Inspired by the success implemented in many applications, many languages attempt to develop their own WordNets using PWN as a model, for example[1], Eurowordnet, Chinese WordNet, Korean WordNet, Japanese WordNet and so on. Though WordNet was already used as a starting resource for developing many language WordNets, the constructions of the WordNet for languages can be varied according to the availability of the language resources. Some were developed from scratch, and some were developed from the combination of various existing lexical resources.

This paper presents a method to facilitate the WordNet construction by using the existing resources having only English equivalents and the lexical synonyms. Our proposed criteria and algorithm for application are evaluated by implementing them for Asian languages which occupy quite different language phenomena in terms of grammars and word

[1] List of wordnets in the world and their information is provided at http://www.globalwordnet.org/gwa/ wordnet_table.htm

unit. The PWN version 2.1 containing 207,010 senses classified into adjective, adverb, verb, and noun are used to evaluate our criteria and algorithm. Our approach is conducted to assign a synset to a lexical entry by considering its English equivalent and lexical synonyms. The degree of reliability of the assignment is defined in terms of confidence score (CS) based on our assumption of the membership of the English equivalent in the synset.

In what follows, section 2 describes our criteria for synset assignment. Section 3 provides the results of the experiments and the evaluation. The information on KUI, the post editng tool is given in section 4. And Section 5 concludes our work.

2 Synset alignment

Under the situation of limited resources on a language, an English equivalent word in a bi-lingual dictionary is a crucial key to find an appropriate synset for the entry word in question. The synset assignment criteria described in this section relies on the information of English equivalent and synonym of a lexical entry, which is most commonly encoded in a bi-lingual dictionary.

Synset Assignment Criteria

Applying the nature of WordNet which introduces a set of synonyms to define the concept, we set up four criteria for assigning a synset to a lexical entry. The confidence score (CS) is introduced to annotate the likelihood of the assignment. The highest score, CS=4, is assigned to the synset that is evident to include more than one English equivalent of the lexical entry in question. On the contrary, the lowest score, CS=1, is assigned to any synset that occupies only one of the English equivalents of the lexical entry in question when multiple English equivalents exist.

The details of assignment criteria are: L_i denotes the lexical entry, E_j denotes the English equivalent, S_k denotes the synset, and $\in$ denotes the member of a set.

Case 1: *Accept the synset that includes more than one English equivalent with a CS of 4.*

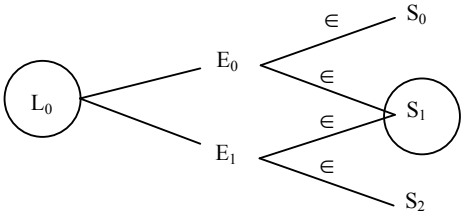

Figure 1 Synset assignment with CS=4

Example:
L_0: เป้าหมาย
E_0: aim E_1: target
S_0: purpose, intent, intention, **aim**, design
S_1: **aim**, object, objective, **target**
S_2: **aim**

In this example, the synset, S_1, is assigned to the lexical entry, L_0, with CS=4.

Figure 1 simulates that a lexical entry L_0 has two English equivalents of E_0 and E_1. Both E_0 and E_1 are included in a synset of S_1. The criterion implies that both E_0 and E_1 are the synset for L_0 which can be defined by a greater set of synonyms in S1. Therefore the relatively high confidence score, CS=4, is assigned for this synset to the lexical entry.

Case 2: *Accept the synset that includes more than one English equivalent of the synonym of the lexical entry in question with a CS of 3.*

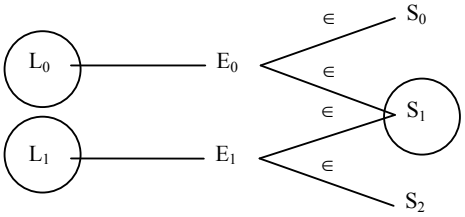

Figure 2 Synset assignment with CS=3

Example:
L_0: จ้อง L_1: เพ่งมอง
E_0: stare E_1: gaze
S_0: **gaze, stare** S_1: **stare**

In this example, the synset, S_0, is assigned to the lexical entry, L_0, with CS=3.

If Case 1 fails in finding a synset that includes more than one English equivalent, the English equivalent of a synonym of the lexical entry is picked up to investigate. Figure 2 shows an English equivalent of a lexical entry L_0 and its synonym L_1 in a synset S_1. In this case the synset S_1 is assigned to both L_0 and L_1 with CS=3. The score in this case is lower than the one assigned in Case 1 because the synonym of the English equivalent of the lexical entry is indirectly implied from the English equivalent of the synonym of the lexical entry. The newly retrieved English equivalent may not be distorted.

Case 3: *Accept the only synset that includes only one English equivalent with a CS of 2.*

Figure 3 shows the assignment of CS-2 when there is only one English equivalent and there is no synonym of the lexical entry. Though there is no English equivalent to increase the reliability of the assignment, at the same time there is no synonym of the lexical entry to distort the relation. In this case, the only English equivalent shows an uniqueness in the translation that can maintain a degree of confidence.

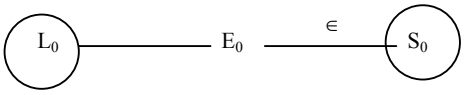

Figure 3 Synset assignment with CS=2

Example:
L_0: สูติแพทย์ E_0: obstetrician
S_0: **obstetrician**, accoucheur
In this example, the synset, S_0, is assigned to the lexical entry, L_0, with CS=2.

Case 4: *Accept more than one synset that includes each of the English equivalents with CS of 1.*

Case 4 is the most relaxed rule to provide some relation information between the lexical entry and a synset. Figure 4 shows the assignment of CS=1 to any relations that do not meet the previous criteria but the synsets include one of the English equivalents of the lexical entry.

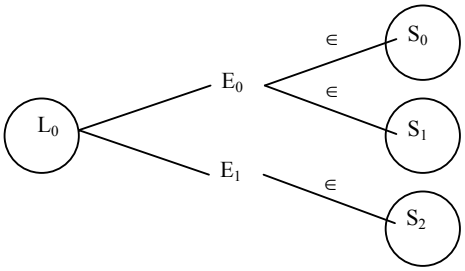

Figure 4 Synset assignment with CS=1

Example:
L_0: ช่อง
E_0: hole E_1: canal
S_0: **hole**, hollow
S_1: **hole**, trap, cakehole, maw, yap, gop
S_2: **canal**, duct, epithelial duct, channel
In this example, each synset, S_0, S_1, and S_2 is assigned to lexical entry L_0, with CS=1.

3 Experiment results and evaluation

We applied the synset assignment criteria to a Thai-English dictionary (MMT dictionary) [2] with the synset from WordNet 2.1. To compare the ratio of assignment for Thai-English dictionary, we also investigated the synset assignment of Indonesian-English and Mongolian-English dictionaries.

In our experiment, there are only 24,457 synsets from 207,010 synsets, which is 12% of the total number of the synsets that can be assigned to Thai lexical entries. Table 1 shows the successful rate in assigning synsets to the Thai-English dictionary. About 24 % of Thai lexical entries are found with the English equivalents that meet one of our criteria. We applied the same algorithm to Indonesia-English and Mongolian-English [3] dictionaries to investigate how it works with other languages in terms of the selection of English equivalents. The difference in unit of concept is basically understood to affect the assignment of English equivalents in bi-lingual dictionaries. In Table 2, the size of the Indonesian-English dictionary is about half that of the Thai-English dictionary. The success rates of assignment to the lexical entry are the same, but the rate of synset assignment of the Indonesian-English dictionary is lower than that of the Thai-English dictionary. This is because the total number of lexical entries is about in the half that of the Thai-English dictionary. A Mongolian-English dictionary is also evaluated. Table 3 shows the result of synset assignment. These experiments show the effectiveness of using English equivalents and synonym information from limited resources in assigning WordNet synsets.

Table 1 Synset assignment to Thai-English dictionary

	WordNet (synset)		TE Dict (entry)	
	total	Assigned	total	assigned
Noun	145,103	18,353 (13%)	43,072	11,867 (28%)
Verb	24,884	1,333 (5%)	17,669	2,298 (13%)
Adjective	31,302	4,034 (13%)	18,448	3,722 (20%)
Adverb	5,721	737 (13%)	3,008	1,519 (51%)
Total	207,010	24,457 (12%)	82,197	19,406 (24%)

Table 2 Synset assignment to Indonesian-English dictionary

	WordNet (synset)		IE Dict (entry)	
	total	assigned	total	assigned
Noun	145,103	4,955 (3%)	20,839	2,710 (13%)
Verb	24,884	7,841 (32%)	15,214	4,243 (28%)
Adjective	31,302	3,722 (12%)	4,837	2,463 (51%)
Adverb	5,721	381 (7%)	414	285 (69%)
total	207,010	16,899 (8%)	41,304	9,701 (24%)

Table 3 Synset assignment to Mongolian-English dictionary

	WordNet (synset)		ME Dict (entry)	
	total	assigned	total	assigned
Noun	145,103	268 (0.18%)	168	125 (74.40%)
Verb	24,884	240 (0.96%)	193	139 (72.02%)
Adjective	31,302	211 (0.67%)	232	129 (55.60%)
Adverb	5,721	35 (0.61%)	42	17 (40.48%)
total	207,010	754 (0.36%)	635	410 (64.57%)

In the evaluation of our approach for synset assignment, we randomly selected 1,044 synsets from the result of synset assignment to the Thai-English dictionary (MMT dictionary) for manually checking. The random set covers all types of part-of-speech and degrees of confidence score (CS) to confirm the approach in all possible situations. According to the supposition of our algorithm that the set of English equivalents of a word entry and its synonyms are significant information to relate to a synset of WordNet, the result of accuracy will be correspondent to the degree of CS. The results were manually checked, and it is found that a small set of adverb synsets is 100% correctly assigned irrelevant to its CS. The total number of adverbs for the evaluation could be too small. The algorithm shows a better result of 48.7% in average for noun synset assignment and 43.2% in average for all part of speech.

With the better information of English equivalents marked with CS=4, the assignment accuracy is as high as 80.0% and decreases accordingly due to the CS value. This confirms that the accuracy of synset assignment strongly relies on the number of English equivalents in the synset. The indirect information of English equivalents of the synonym of the word entry is also helpful, yielding 60.7% accuracy in synset assignment for the group of CS=3. Others are quite low, but the English equivalents are somehow useful to provide the candidates for expert revision.

4 KUI for post-editing

We also introduce a web-based collaborative workbench, called KUI (Knowledge Unifying Initiator) [4], for revising the result of synset assignment and provide a framework to create Asian WordNet via the linkage through PWN synset. KUI is an efficient online collaborative framework in producing and maintaining knowledge according to the principle of collective intelligent. KUI was designed to support an open web community by introducing a voting system and a mechanism to realize the function of selectional preference. KUI enables to connect and collaborate among individual intelligence in order to accomplish a complex mission.

Asian WordNet translation is one of the KUI-Translating rooms. At http://www.tcllab.org/kui, participants from each language can translate or revise all English words (synsets) into their own language. Online lookup, chat function, voting, invite assistants are also provided to consult a term translation.

5 Conclusion

Our synset assignment criteria were effectively applied to languages having only English equivalents and its lexical synonym. Confidence scores were proven efficiently assigned to determine the degree of reliability of the assignment which later was a key value in the revision process. Languages in Asia are significantly different from the English language in terms of grammar and lexical word units. The differences prevent us from finding the target synset by following just the English equivalent. Synonyms of the lexical entry and an additional dictionary from different sources can be complementarily used to improve the accuracy in the assignment. Applying the same criteria to other Asian languages also yielded a satisfactory result. Following this process, Asian WordNet can be constructed from existing language resources.

Reference

1. Fellbaum, C. (ed.).: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass (1998)
2. CICC.: Thai Basic Dictionary. Technical Report, Japan (1995)
3. Hangin, G., Krueger, J. R., Buell, P.D., Rozycki, W.V., Service, R.G.: A modern Mongolian-English dictionary. Indiana University, Research Institute for Inner Asian Studies (1986)
4. Charoenporn, T., Sornlertlamvanich, V., Robkop, K., and Isahara, H.: KUI: an ubiquitous tool for collective intelligence development, In: Proceedings of the Workshop on NLP for Less Privileged Languages, Hyderabad, India (2008)