Synset Assignment for Bi-lingual Dictionary with Limited Resource

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

Hitoshi Isahara
National Institute of Information
and Communications Technology
3-5 Hikaridai, Seika-cho, soraku-gaun,
Kyoto, Japan 619-0289
isahara@nict.go.jp

Hammam Riza
IPTEKNET, Agency for the Assessment and Application of Technology,
Jakarta Pusat 10340, Indonesia
hammam@iptek.net.id

Purev Jaimai
Center for Research on Language
Processing, National University of
Mongolia, Ulaanbaatar, Mongolia
purev@num.edu.mn

Abstract

This paper explores an automatic WordNet synset assignment to the bi-lingual dictionaries of languages having limited lexicon information. Generally, a term in a bi-lingual dictionary is provided with very limited information such as part-of-speech, a set of synonyms, and a set of English equivalents. This type of dictionary is comparatively reliable and can be found in an electronic form from various publishers. In this paper, we propose an algorithm for applying a set of criteria to assign a synset with an appropriate degree of confidence to the existing bi-lingual dictionary. We show the efficiency in nominating the synset candidate by using the most common lexical information. The algorithm is evaluated against the implementation of Thai-English, Indonesian-English, and Mongolian-English bi-lingual dictionaries. The experiment also shows the effectiveness of using the same type of dictionary from different sources.

1 Introduction

The Princeton WordNet (PWN) (Fellbaum, 1998) is one of the most semantically rich English lexical databases that are widely used as a lexical knowledge resource in many research and development topics. The database is divided by part of speech into noun, verb, adjective and adverb, organized in sets of synonyms, called synset, each of which represents "meaning" of the word entry.

Though WordNet was already used as a starting resource for developing many language WordNets, the construction of the WordNet for any 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. Spanish and Catalan WordNets, for instance, are automatically constructed using hyponym relation, monolingual dictionary, bilingual dictionary and taxonomy (Atserias et al., 1997). Italian WordNet (Magnini et al., 1994) is semi-automatically constructed from definition in monolingual dictionary, bilingual dictionary, and WordNet glosses. Hungarian WordNet uses bilingual dictionary, monolingual explanatory dictionary, and Hungarian thesaurus in the construction (Proszeky et al., 2002), etc.

This paper presents a new method particularly 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 to Asian languages which occupy quite different language phenomena in terms of grammars and word unit.

To evaluate our criteria and algorithm, we use the PWN version 2.1 containing 207,010 senses classified into adjective, adverb, verb, and noun. The basic building block is a "synset" which is essentially a context-sensitive grouping of synonyms which are linked by various types of relation such as hyponym, hypernymy, meronymy, antonym, attributes, and modification. 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. A dictionary from different source is also a reliable source to increase the accuracy of the assignment because it can fulfill the thoroughness of the list of English equivalent and the lexical synonyms.

The rest of this paper is organized as follows: Section 2 describes our criteria for synset assignment. Section 3 provides the results of the experiments and error analysis on Thai, Indonesian, and Mongolian. Section 4 evaluates the accuracy of the assignment result, and the effectiveness of the complimentary use of a dictionary from different sources. Section 5 shows a collaborative interface for revising the result of synset assignment. And Section 6 concludes our work.

2 Synset Assignment

A set of synonyms determines the meaning of a concept. Under the situation of limited resources on a language, 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 elaborated as in the followings. 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 confidence score of 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 S_1. Therefore the relatively high confidence score, CS=4, is assigned for this synset to the lexical entry.

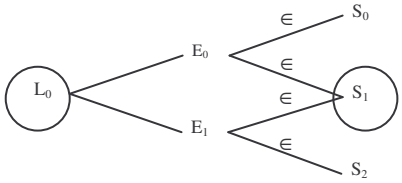

Figure 1. Synset assignment with SC=4

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

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

In case that 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.

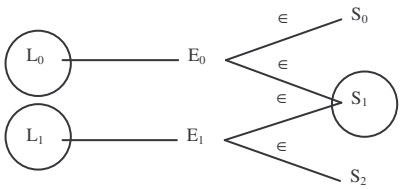

Figure 2. Synset assignment with SC=3

Figure 2 simulates that an English equivalent of a lexical entry L_0 and its synonym L_1 are included 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.

Example:

L_0: จ้อง L_1: เพ่งมอง

E_0: stare E_1: gaze

S_0: **gaze, stare** S_1: **stare**

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

Case 3: Accept the only synset that includes the only one English equivalent with confidence score of 2.

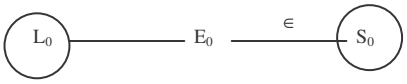

Figure 3. Synset assignment with SC=2

Figure 3 simulates 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 any English equivalent to increase the reliability of the assignment, in the same time there is no synonym of the lexical entry to distort the relation. In this case, the only one English equivalent shows it uniqueness in the translation that can maintain a degree of the confidence.

Example:

L_0: สูติแพทย์ E_0: obstetrician

S_0: **obstetrician**, accoucheur

In the above 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 Equivalent with confidence score of 1.

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

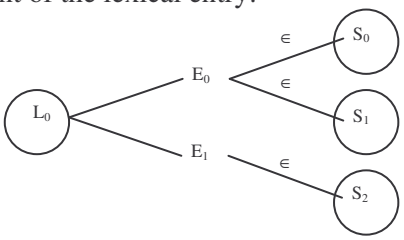

Figure 4. Synset assignment with SC=1

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 the above example, each synset, S_0, S_1, and S_2 is assigned to lexical entry L_0, with CS=1.

3 Experiment results

We applied the synset assignment criteria to a Thai-English dictionary (MMT dictionary) (CICC, 1995) with the synset from WordNet 2.1. To compare the ratio of assignment for Thai-English dictionary, we also investigate the synset assignment of Indonesian-English and Mongolian-English dictionaries.

	WordNet (synset)		T-E 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 1. Synset assignment to T-E dictionary

In our experiment, there are only 24,457 synsets from 207,010 synsets, which is 12% of the total number of the synset that can be assigned to Thai lexical entries. Table 1 shows the successful rate in assigning synset to Thai-English dictionary. About 24 % of Thai lexical entries are found with the English equivalents that meet one of our criteria.

Going through the list of unmapped lexical entry, we can classify the errors into three groups:-

1. Compound

   The English equivalent is assigned in a com-

pound, especially in case that there is no an appropriate translation to represent exactly the same sense. For example,

L: ร้านค้าปลีก E: retail shop

L: กระชาก E: pull sharply

2. Phrase
   Some particular words culturally used in one language may not be simply translated into one single word sense in English. In this case, we found it explained in a phrase. For example,

   L: ร้านสวด

   E: small pavilion for monks to sit on to chant

   L: กรรเจียก

   E: bouquet worn over the ear

3. Word form
   Inflected forms i.e. plural, past participle, are used to express an appropriate sense of a lexical entry. This can be found in non-inflection languages such as Thai and most of Asian languages. For example,

   L: ร้าวระทมใจ E: grieved

The above English expressions cause an error in find an appropriate synset.

	WordNet (synset)		I-E 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 2. Synset assignment to I-E dictionary

We applied the same algorithm to Indonesia-English and Mongolian-English (Hangin, 1986) 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 effect the assignment of English equivalents in bi-lingual dictionaries. In Table 2, the size of Indonesian-English dictionary is about half of Thai-English dictionary. The success rates of assignment to the lexical entry are the same but the rate of synset assignment of Indonesian-English dictionary is lower than one of Thai-English dictionary. This is because the total number of lexical entry is almost in the half size.

	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%)

Table 3. Synset assignment to M-E dictionary

A small set of Mongolian-English dictionary is also evaluated. Table 3 shows the result of synset assignment.

These experiments show the effectiveness of using English equivalents and synonyms information from limited resources in assigning WordNet synsets.

4 Evaluations

In the evaluation of our approach for synset assignment, we randomly selected 1,044 synsets from the result of synset assignment to 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 detail number of synsets to be used in the evaluation is shown in Table 4.

	CS=4	CS=3	CS=2	CS=1	total
Noun	7	479	64	272	822
Verb		44	75	29	148
Adjective	1	25		32	58
Adverb	7	4	4	1	16
total	15	552	143	334	1044

Table 4. Random set of synset assignment

Table 5 shows the accuracy of synset assignment by part-of-speech and CS. A small set of adverb synsets are 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.

	CS=4	CS=3	CS=2	CS=1	total
Noun	5 (71.4%)	306 (63.9%)	34 (53.1%)	55 (20.2%)	400 (48.7%)
Verb		23 (52.3%)	6 (8.0%)	4 (13.8%)	33 (22.3%)
Adjective		2 (8.0%)			2 (3.4%)
Adverb	7 (100%)	4 (100%)	4 (100%)	1 (100%)	16 (100%)
total	12 (80.0%)	335 (60.7%)	44 (30.8%)	60 (18%)	451 (43.2%)

Table 5. Accuracy of synset assignment

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. It yields 60.7% of 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.

	CS=4	CS=3	CS=2	CS=1	total
Noun	2		22	29	53
Verb		2	6	4	12
Adjective					
Adverb					
total	2	2	28	33	65

Table 6. Additional correct synset assignment by other dictionary (LEXiTRON)

To examine the effectiveness of English equivalent and synonym information from different source, we consulted another Thai-English dictionary (LEXiTRON). Table 6 shows the improvement of the assignment by the increased number of correct assignment in each type. We can correct more in noun and verb but not adjective. Verb and adjective are ambiguously defined in Thai lexicon, and the number of the remained adjective is too few, therefore, the result should be improved unconcerned with the type.

	CS=4	CS=3	CS=2	CS=1	total
total	14 (93.3%)	337 (61.1%)	72 (50.3%)	93 (27.8%)	516 (49.4%)

Table 7. Improved correct synset assignment by additional bi-lingual dictionary (LEXiTRON)

Table 7 shows the total improvement of the assignment accuracy when we integrated English equivalent and synonym information from different source. The accuracy for synsets marked with CS=4 is improved from 80.0% to 93.3% and the average accuracy is also significantly improved from 43.2% to 49.4%. All types of synset are significantly improved only if a bi-lingual dictionary from different sources is available.

5 Collaborative Work on Asian WordNet

There are some efforts in developing WordNets of some Asian languages, e.g. Chinese, Japanese, Korean (Choi, 2003), (Choi et al., 2004), (Kaji et al., 2006), (KorLex, 2006), (Huang, 2007) and Hindi (Hindi Wordnet, 2007). The number of languages that have been successfully developed their WordNets is still limited to some active research in this area. However, the extensive development of WordNet in other languages is important, not only to help in implementing NLP applications in each language, but also in inter-linking WordNets of different languages to develop multi-lingual applications to overcome the language barrier.

We adopt the proposed criteria for automatic synset assignment for Asian languages which has limited language resources. Based on the result from the above synset assignment algorithm, we provide KUI (Knowledge Unifying Initiator) (Sornlertlamvanich, 2006), (Sornlertlamvanich et al., 2007) to establish an online collaborative work in refining the WorNets.

KUI is a community software which allows registered members including language experts revise and vote for the synset assignment. The system manages the synset assignment according to the preferred score obtained from the revision process. As a result, the community WordNets will be accomplished and exported into the original form of WordNet database. Via the synset ID assigned in the WordNet, the system can generate a cross language WordNet result. Through this effort, an initial version of Asian WordNet can be fulfilled.

Figure 5 illustrates the translation page of KUI[1]. In the working area, the login member can participate in proposing a new translation or vote for the preferred translation to revise the synset assignment. Statistics of the progress as well as many useful functions such as item search, record jump, chat, list of online participants are also provided.

[1] http://www.tcllab.org/kui

KUI is actively facilitating members in revising the Asian WordNet database.

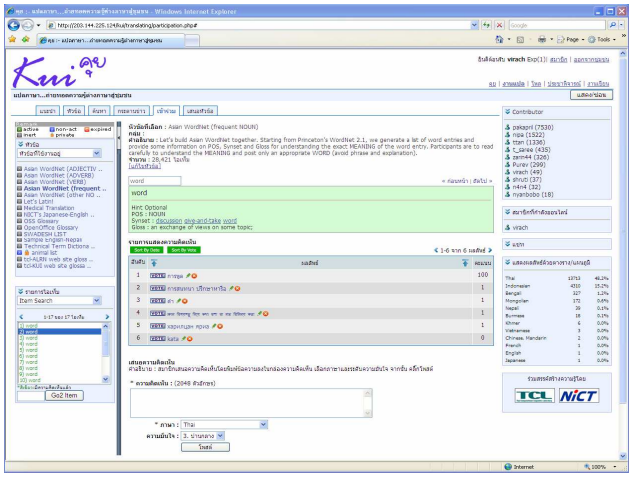

Figure 5. Sample of KUI interface

6 Conclusion

Our synset assignment criteria were effectively applied to languages having only English equivalents and its lexical synonym. Confidence score was proved 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 unit. The differences prevent us from finding the target synset by following just the English equivalent. Synonyms of the lexical entry and additional dictionary from different sources can be complimentarily used to improve the accuracy in the assignment. Applying the same criteria to other Asian languages also yielded a satisfactory result. Following the same process that we had implemented to the Thai language, we are expecting an acceptable result from the Indonesian, Mongolian languages and so on. After the revision at KUI, the initial stage of Asian WordNet will be referable through the assigned synset ID.

References

Bernardo Magnini, Carlo Strapparava, Fabio Ciravegna and Emanuele Pianta, 1994. *A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet*, IRST Technical Report # 9406-15.

Chu-Ren Huang, 2007. *Chinese Wordnet*, Academica Sinica, Available at http://bow.sinica.edu.tw/wn/

CICC. 1995. *Thai Basic Dictionary: Technical Report*, Japan.

Fellbaum, Christiane (ed.), 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Gabor Proszeky, Marton Mihaltz, 2002. *Semi-Automatic Development of the Hungarian WordNet*, Proceedings of the LREC 2002, Spain

Gombojab Hangin with John R.Krueger and Paul D.Buell, William V.Rozycki, Robert G.Service, 1986. *A modern Mongolian-English dictionary*. Indiana University, Research Institute for Inner Asian Studies.

Hindi Wordnet, 2007. Available at http://www.cfilt.iitb.ac.in/wordnet/webhwn/

Hiroyuki Kaji and Mariko Watanabe, 2006. *Automatic Construction of Japanese WordNet*, Proceedings of LREC2006, Italy.

J. Atserias, S. Clement, X. Farreres, German Rigau, H. Rodríguez, 1997. *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets*, Proceedings of the International Conference on Recent Advances in Natural Language, Bulgaria.

K.S. Choi, H.S. Bae, W.Kang, J. Lee, E. Kim, H. Kim, D. Kim, Y. Song1, and H. Shin, 2004. *Korean-Chinese-Japanese Multilingual Wordnet with Shared Semantic Hierarchy*, Proceediongs of LREC 2004, Portugal.

Key-Sun Choi, 2003. *CoreNet: Chinese-Japanese-Korean wordnet with shared semantic hierarchy*, Proceedings of Natural Language Processing and Knowledge Engineering, Beijing.

Korlex, 2006. *Korean WordNet*, Korean Language Processing Lab, Pusan National University, 2007. Available at http://164.125.65.68/

NECTEC, 2006. *LEXiTRON: Thai-English Dictionary*, Available at http://lexitron.nectec.or.th/

Spanish and Catalan WordNets, 2006. Available at http://www.lsi.upc.edu/~nlp/

Virach Sornlertlamvanich, 2006. *KUI: The OSS-Styled Knowledge Development System*, Proceedings of The 7th AOSS Symposium, Malaysia.

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergit Robkop, and Hitoshi Isahara. *Collaborative Platform for Multilingual Resource Development and Intercultural Communication*, Proceedings of the First International Workshop on Intercultural Collaboration (IWIC2007), Japan.