Review on Development of Asian WordNet

Thai Computational Linguistics Lab., NICT Asia Research Center, and National Electronics and Computer Technology Center (NECTEC)
Virach Sornlertlamvanich

PROFILE
He received the D.Eng. degree from Tokyo Institute of Technology in 1998. He worked with NEC Corporation as a sub-project leader for Thai language processing in the Multi-lingual Machine Translation Project. He later founded the Linguistics and Knowledge Science Laboratory (LINKS) to conduct the research on Natural Language Processing (NLP) in the National Electronics and Computer Technology Center (NECTEC) of Thailand in 1992. He initiated a wide range of applied NLP projects, such as ParSit (a web-based English to Thai machine translation service), LEXiTRON (an online Thai-English corpus-based dictionary), and Sansarn (a probabilistic based Thai-English search engine). He was awarded by the National Research Council of Thailand as the Most Outstanding Researcher of the Year 2003. His research interests are machine translation, natural language processing, lexical acquisition, information retrieval and other related fields. He is currently the Co-Director of Thai Computational Linguistics Laboratory (TCL), NICT Asia Research Center, and the Assistant Executive Director of National Electronics and Computer Technology Center (NECTEC), Thailand.
virach@tcllab.org TEL

Thai Computational Linguistics Lab., NICT Asia Research Center, and National Electronics and Computer Technology Center (NECTEC)
Thatsanee Charoenporn

PROFILE
thatsanee@tcllab.org TEL

Thai Computational Linguistics Lab., NICT Asia Research Center
Kergrit Robkop

PROFILE
kergrit@tcllab.org TEL

Thai Computational Linguistics Lab., NICT Asia Research Center
Chumpol Mokarat

PROFILE
chumpol@tcllab.org TEL

National Institute of Information and Communications Technology
Hitoshi Isahara

PROFILE
isahara@nict.go.jp TEL

Abstract. This paper describes the approach we used to create Asain WordNet (AWN) from any existing bi-lingual dictionaries. We found that most of the bilingual dictionaries of a language are paired with the English language. Based on the English equivalents in the bi-lingual dictionary we estimate the WordNet synset assignment. In general, 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. The results are reviewed collaboratively online via http://www.tcllab.org and can be viewed on http://www.asianwordnet.org that connects Asian languages through the Princeton WordNet (PWN).

Keywords: WordNet, Asian language, synset assignment, visualization, collaborative tools

1 Introduction

The Princeton WordNet (PWN) [1] 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. PWN is successfully implemented in many applications, e.g., word sense disambiguation, information retrieval, text summarization, text categorization, and so on. Inspired by this success, many languages attempt to develop their own WordNets using PWN as a model, for example[1], BalkaNet (Balkans languages), DanNet (Danish), Eurowordnet (European languages such as Spanish, Italian, German, French, English), Russnet (Russian), Hindi WordNet, Arabic WordNet, Chinese WordNet, Korean 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. Spanish and Catalan Wordnets [2], for instance, are automatically constructed using hyponym relation, a monolingual dictionary, a bilingual dictionary and taxonomy [3]. Italian WordNet [4] is semi-automatically constructed from definitions in a monolingual dictionary, a bilingual dictionary, and WordNet glosses. Hungarian WordNet uses a bilingual dictionary, a monolingual explanatory dictionary, and Hungarian thesaurus in the construction [5], etc.

This paper presents a new 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 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 a 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

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

a dictionary from different sources. Section 5 exhibits the cross language visualization for Asian WordNet (AWN), 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, 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: Li denotes the lexical entry, Ej 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 confidence score of 4.

Fig. 1 simulates that a lexical entry L_0 has two English equivalents of E_{00} and E_{01}. Both E_{00} and E_{01} are included in a synset of S_1. The criterion implies that both E_{00} and E_{01} 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.

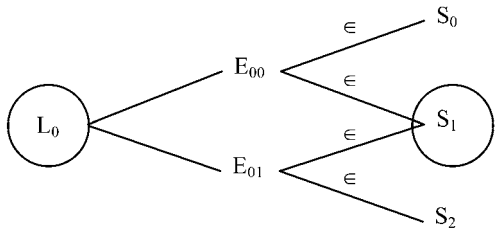

Fig. 1. Synset assignment with CS=4

Example:
L_0: เป้าหมาย
E_{00}: aim E_{01}: target
S_0: purpose, intent, intention, **aim**, design
S_1: **aim**, object, objective, **target**
S_2: **target**

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 from the synonym of the target language with a confidence score of 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. Fig. 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.

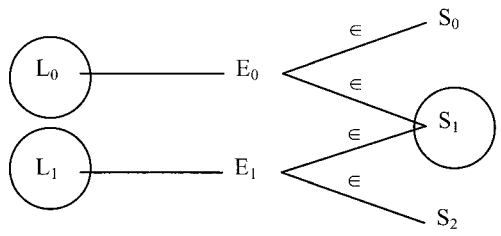

Fig. 2. Synset assignment with CS=3

Example:
L_0: จ้อง L_1: เพ่งมอง
E_0: stare E_1: gaze
S_0: **stare** S_1: **gaze, stare** S_2: **regard, gaze**

In the above example, the synset, S_1, is assigned to both lexical entries, L_0 and L1, with CS=3.

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

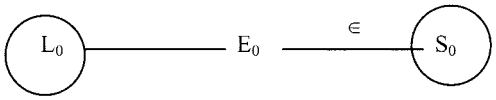

Fig. 3. Synset assignment with CS=2

Fig. 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.

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 equivalents with a confidence score of 1.

Case 4 is the most relaxed rule to provide some relation information between the lexical entry and a synset. Fig. 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.

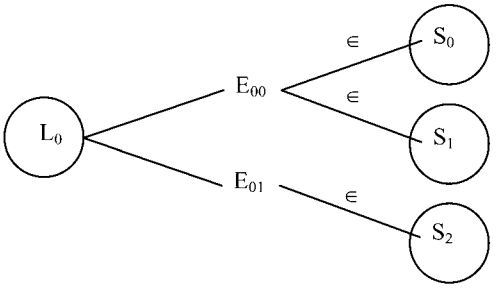

Fig. 4. Synset assignment with CS=1

Example:
L_0: ช่อง
E_{00}: hole E_{01}: canal
S_0: **hole**, hollow
S_1: **hole**, trap, cakehole, maw, yap, gap
S_2: **canal**, duct, epithelial duct, channel

In the above example, each synset, S_0, S_1, and S2 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) [6] 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.

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

1. Compound
   The English equivalent is assigned in a compound, especially in cases where there is no 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-inflected languages such as Thai and most Asian languages. For example,
   L: ร้าวระทมใจ E: grieved

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

	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 1. Synset assignment to Thai-English dictionary

We applied the same algorithm to Indonesia-English and Mongolian-English [7] 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.

	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 2. Synset assignment to Indonesian-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%)

Table 3. Synset assignment to Mongolian-English dictionary

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.

4 Evaluations

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.

It took about three years to develop the Balkan WordNet on PWN 2.0 [8], [9]. Therefore, we randomly picked up some synsets that resulted from our synset assignment algorithm. The results were manually checked and the details of synsets to be used to evaluate our algorithm are shown in Table 4.

Table 5 shows the accuracy of synset assignment by part of speech and CS. 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.

	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

	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

	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 a different source, we consulted another Thai-English dictionary (LEXiTRON) [10]. Table 6 shows the improvement of the assignment by the increased number of correct assignment in each type. We can correct more in nouns and verbs but not adjectives. Verbs and adjectives are ambiguously defined in Thai lexicon, and the number of the remaining adjectives is too few, therefore, the result should be improved regardless of 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 a 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 if a bi-lingual dictionary from different sources is available.

5 Collaborative Review and Visualization of Asian WordNet

The results of the synset assignment for each language are stored and indexed under KUI (Knowledge Unifying Initiator) environment for online collaborative review [11]. Contributors are registered to participate as a supporter of the translation by voting for the best translation or posting a better translation for each synset. From the result of the translation, a table for mapping between sense id and word entry is created. When there is a request for a pair of languages WordNet expression, the word entry of the source language will be used to retrieve the sense id, and then with the sense id the translated word entry of the target language will be obtained. Since each translated word entry is accommodated with a vote score, the word entry with the highest score will be selected to display the current best translation.

sense_id	lid	message	sense_key	synset_offset	sense_id	lid	message
28262	5865	รถราง	car%1:06:01::	102959942	28262	2549	列車
28262	5865	รถราง	car%1:06:01::	102959942	28262	2549	鉄道車両
28262	5865	รถราง	car%1:06:01::	102959942	28262	2549	貨車
177401	5865	รถราง	streetcar%1:06:00::	104335435	177401	2549	市電
177401	5865	รถราง	streetcar%1:06:00::	104335435	177401	2549	市街電車
177401	5865	รถราง	streetcar%1:06:00::	104335435	177401	2549	ストリートカー
177401	5865	รถราง	streetcar%1:06:00::	104335435	177401	2549	電車
177401	5865	รถราง	streetcar%1:06:00::	104335435	177401	2549	トロリー
177401	5865	รถราง	streetcar%1:06:00::	104335435	177401	2549	路面電車
177401	5865	รถราง	streetcar%1:06:00::	104335435	177401	2549	都電

Table 8. Result of mapped word entry between Thai and Japanese

Table 8 shows the result of mapped word entry between Thai and Japanese through the sense id when making a request for a Thai word (รถราง).

Fig. 5 shows the result of retrieving the Thai word (รถราง) for Japanese equivalents. This service can be found at http://www.asianwordnet.org/. Currently the based PWN is converted to version 3.0 for better compatibility

with other WordNets.

6 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 the same process that we implemented for the Thai language, we are expecting an acceptable result from the Indonesian, Mongolian languages and so on. Resulting from the AWN creation, the visualization of AWN across languages can efficiently serve the request for any pairs of languages through the PWN sense id.

References

1. Fellbaum, C. (ed.).: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass (1998)
2. Spanish and Catalan WordNets, http://www.lsi.upc.edu/~nlp/
3. Atserias, J., Clement, S., Farreres, X., Rigau, G., Rodr?guez, H.: Combining Multiple

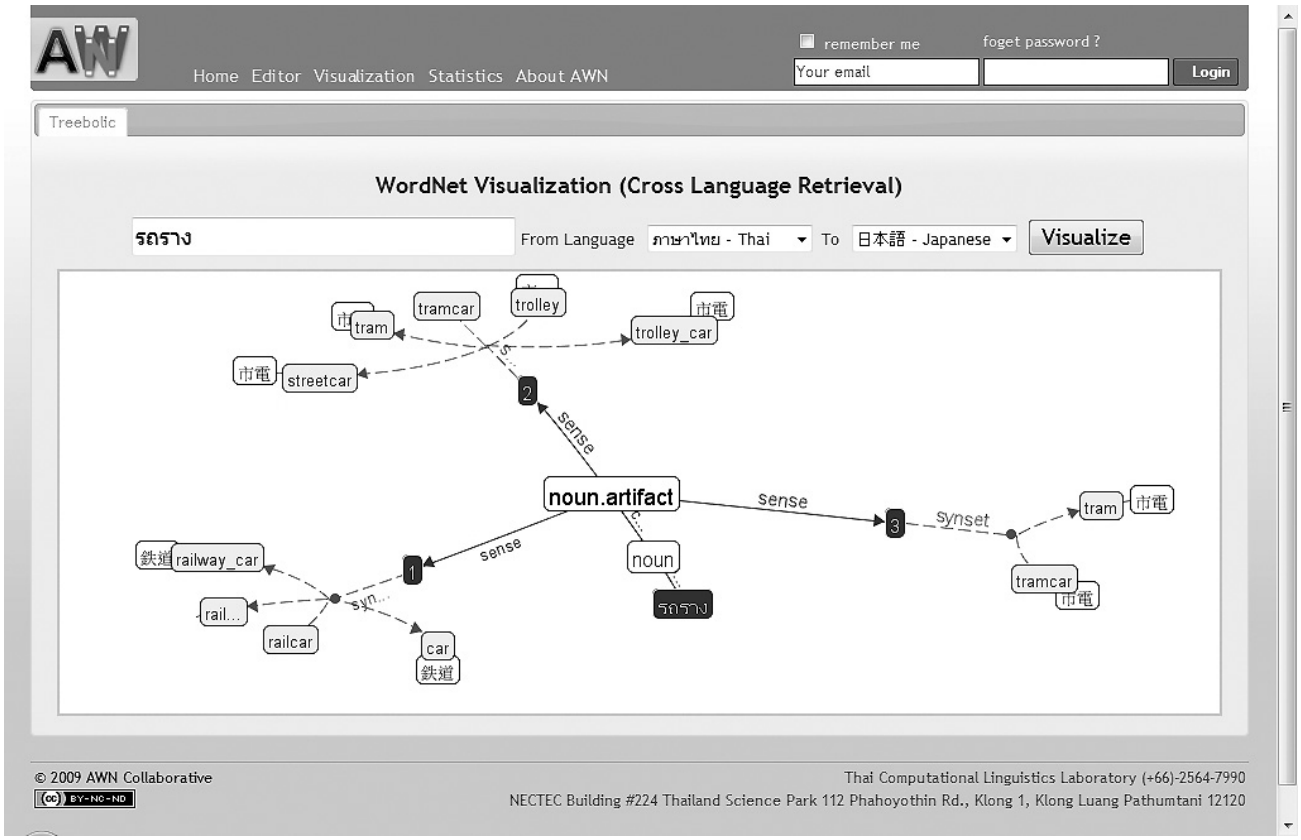

Fig. 5. Screen shot of AWN cross language visualization

Methods for the Automatic Construction of Multilingual WordNets. In: Proceedings of the International Conference on Recent Advances in Natural Language, Bulgaria. (1997)
4. Magnini, B., Strapparava, C., Ciravegna, F., Pianta, E.: A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet. IRST Technical Report # 9406-15 (1994)
5. Proszeky, G., Mihaltz, M.: Semi-Automatic Development of the Hungarian WordNet. In: Proceedings of the LREC 2002, Spain. (2002)
6. CICC.: Thai Basic Dictionary. Technical Report, Japan. (1995)
7. 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)
8. Tufis, D. (ed.).: Special Issue on the BalkaNet Project, Romanian Journal of Information Science and Technology, vol. 7, no. 1-2. (2004)
9. Barbu, E., Mititelu, V. B.: Automatic Building of Wordnets. In: Proceedings of RANLP, Bulgaria (2005)
10. NECTEC. LEXiTRON: Thai-English Dictionary, http://lexitron.nectec.or.th/
11. Sornlertlamvanich, V., Charoenporn, T., Robkop, K., and Isahara, H.: KUI: Self-organizing Multi-lingual WordNet Construction Tool. In: Proceedings of the Fourth Global WordNet Conference (GWC2008), Szeged, Hungary. (2008)

「アジアワードネットAWNの開発の省察」の概要

第一著者ウィラット・ソンラートラムワーニッチ博士は、タイ計算言語学研究所 TCL の Co- リーダであり、タイ国立電子コンピュータセンター NECTEC の参事である。20 年前、通産省 (現経産省) 所轄の CICC 近隣諸国間機械翻訳プロジェクトのタイ国チームの副代表であった。東京工業大学において博士号を取得した日本語の分かる自然言語研究者である。

1 はじめに

プリンストン大学 WordNet(PWN) は英語における意味情報を最も豊富に含んだ語彙データベースであり、語彙の知識源として広く利用されている。このデータベースは品詞（名詞、動詞、形容詞、副詞）で分類されている。その特徴は、同義語 (シノニム) をまとめて語彙の意味分類（“Synset”と呼ばれる）を与えたことである。PWN は多くの自然言語応用に利用され成功している。例えば、多義の解消、情報参照、文書要約、文書分類等である。この成功によって各国語の WN が開発されている。例えば、バルカン言語 BalkaNet、デンマーク語 DanNet、西・伊・仏・独・英語の EurowordNet、ロシア語 RussNet、Hindi WordNet、Arabic WordNet、Chinese WordNet、Korean WorddNet 等々である。（訳注：今年の 2 月に独立行政法人情報通信研究機構 NICT から日本語 WordNet が公開された。）

各国語の WordNet は PWN を初期言語資源として開発されるが、利用する言語資源によって構築方法が異なる。あるものは人手で、また、あるものは種々の言語資源を用いて開発された。スペイン語とカタール語の WordNet は全自動であり、上位下位関係、単言語辞書、対訳辞書、語彙分類を利用している。イタリア語の WordNet は半自動である。単言語辞書の語義文、対訳辞書、WordNet の語釈を利用している。ハンガリー語の WordNet は対訳辞書と説明つき単言語辞書、ハンガ

リー語のシソーラスを用いている。

この論文では、新たな WordNet 構築法を示す。利用する言語資源は、英語対訳辞書と単言語の同義語辞書である。同じ手法でアジア言語の中で言語的に異質なタイ語、インドネシア語、モンゴル語の WordNet を構築し、評価したのでここに報告する。

2 Synset の割り振り

タイ語の単語見出しに対して、上記の言語資源により Synset を対応させ、4 段階の確信度 Confidence Score を与えた。

基準 1：確信度4(高い)、1 見出し語に対し複数の英語対訳が Synset を共有しているとき。

基準2：確信度3、ターゲット言語の同義語と英語対訳の Synset を共有しているとき。

基準3：確信度2、唯一の英語対訳の場合、その英語対訳の Synset を割り振る。

基準 4：確信度 1、複数の対訳がそれぞれ異なる Synset を持っているときは全ての Synset を割り振る。

3 実験結果

Synset の総数は 207K であるが、タイ英対訳の見出語数 82K、インドネシア英対訳の見出語数 41K、モンゴル英対訳の見出語数 635 に対し、実際に Synset が割り振られたのはそれぞれ、19K(24%)、9K(24%)、410(64%) であった。

4 評価

割り振りのアルゴリズムを評価するために、1,044 語をランダムに抽出して目視チェックで評価した。先ず、サンプリングした語彙の品詞 x 確信度の表で分類し、その要素に正しかった個数を記入した。（表 5）確信度4では 80% の正解率で漸次低下して行く。新たに別のタイ英対訳辞書情報を追加して割り振ったところ、確信度4の正解率は 93% になった。（表 7）つまり、追加情報があれば、品質が向上することが分かった。

5 アジア WordNet の共同開発と可視化

アジア言語の WordNet は、知識統合支援システム（Knowledge Unifying Initiator）の下でインデックス化され、共同利用が進んでいる。図 5 はタイ語と日本語の Synset を介したクロス言語可視化の一画面である。

6 結論

英語対訳辞書と同義語辞書のみによる Synset 割り振り方式が効果的であることが分かった。さらに確信度を導入することで信頼性が数値化できることが分かった。

AWN の開発により、PWN の Synset ID を入力すると、アジア言語の単語の相互参照と表示が効率的にサービスできるようになった。

（作成：Japio 特許情報研究所）