Thai WordNet Construction

Sareewan Thoongsup[1]
Kergrit Robkop[1]
Chumpol Mokarat[1]
Tan Sinthurahat[1]
[1] Thai Computational Linguistics Lab.
NICT Asia Research Center, Thailand
{sareewan, kergrit,
Chumpol, tan, thatsanee,
virach}@tcllab.org

Thatsanee Charoenporn [1,2]
Virach Sornlertlamvanich [1,2]
Hitoshi Isahara [3]
[2]National Electronics and Computer
Technology Center Thailand, Thailand
[3]National Institute of Information and
Communications Technology, Japan
isahara@nict.go.jp

Abstract

This paper describes semi-automatic construction of Thai WordNet and the applied method for Asian wordNet. Based on the Princeton WordNet, we develop a method in generating a WordNet by using an existing bi-lingual dictionary. We align the PWN synset to a bi-lingual dictionary through the English equivalent and its part-of-speech (POS), automatically. Manual translation is also employed after the alignment. We also develop 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

The Princeton WordNet (PWN) (Fellbuam, 1998) is one of the most semantically rich English lexical banks widely used as a resource in many research and development. WordNet is a great inspiration in the extensive development of this kind of lexical database in other languages. It is not only an important resource 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. There are some efforts in developing WordNets of some languages (Atserias and et al., 1997; Vossen, 1997; Farrers and et al., 1998; Balkova and et al., 2004; Isahara and et al., 2008). But the number of languages that have been successfully developed their WordNets is still limited to some active research in this area. This paper, however, is the one of that attempt.

This paper describes semi-automatic construction of Thai WordNet and the applied method for Asian WordNet. Based on the Princeton WordNet, we develop a method in generating a WordNet by using an existing bi-lingual dictionary. We align the PWN synset to a bi-lingual dictionary through the English equivalent and its part-of-speech (POS), automatically. Manual translation is also employed after the alignment. We also develop 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.

The rest of this paper is organized as follows: Section 2 describes how we construct the Thai WordNet, including approaches, methods, and some significant language dependent issues experienced along the construction. Section 3 provides the information on Asian WordNet construction and progress. And Section 4 concludes our work.

2 Thai WordNet Construction Procedure

Different approaches and methods have been applied in constructing WordNet of many languages according to the existing lexical resources. This section describes how Thai WordNet is constructed either approach or method.

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 139–144,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

2.1 Approaches

To build language WordNet from scratch, two approaches were brought up into the discussion: the merge approach and the expand approach.

The merge approach is to build the taxonomies of the language; synsets and relations, and then map to the PWN by using the English equivalent words from existing bilingual dictionaries.

The expand approach is to map or translate local words directly to the PWN's synsets by using the existing bilingual dictionaries.

Employing the merge approach, for Thai as an example, we will completely get synsets and relations for the Thai language. But it is time and budget consuming task and require a lot of skilled lexicographers as well, while less time and budget is used when employing the expand approach to get a translated version of WordNet. But some particular Thai concepts which do not occur in PWN will not exist in this lexicon. Comparing between these two approaches, the Thai WordNet construction intended to follow the expand approach by this following reasons;

- Many languages have developed their own WordNet using the PWN as a model, so we can link Thai lexical database to those languages.
- The interface for collaboration for other languages can be easily developed.

2.2 Methods

As presented above, we follow the expand approach to construct the Thai WordNet by translating the synsets in the PWN to the Thai language. Both automatic and manual methods are applied in the process.

2.2.1 Automatic Synset Alignment

Following the objective to translate the PWN to Thai, we attempted to use the existing lexical resources to facilitate the construction. We proposed an automatic method to assign an appropriate synset to a lexical entry by considering its English equivalent and lexical synonyms which are most commonly encoded in a bi-lingual dictionary. (Charoenporn 2008; Sornlertlamvanich, 2008).

	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 entries in Thai-English dictionary

For the result, there is only 12% of the total number of the synsets that were able to be assigned to Thai lexical entries. And about 24% of Thai lexical entries were found with the English equivalents that meet one of our criteria. Table 1 shows the successful rate in assigning synsets to the lexical entry in the Thai-English Dictionary.

Considering the list of unmapped lexical entry, the errors can be classified into three groups, as the following.

1. The English equivalent is assigned in a compound, especially in case that there is no an appropriate translation to represent exactly the same sense. For example,
   L: ร้านค้าปลีก raan3-khaa3-pleek1
   E: retail shop
2. Some particular words culturally used I 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: กรรเจียก kan0-jeak1
   E: bouquet worn over the ear
3. 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: ร้าวระทม raaw3-ra0-thom0
   E: greived

By using this method, a little part of PWN has been translated into Thai. About 88% of the total number of the synsets still cannot be assigned. Manual step is therefore applied.

2.2.2 Manual Construction

Human translation is our next step for synset translation. Two important issues were taken into discussion, when starting the translation process. Those are;

- How to assign or translate new concepts that still do not occur in the Thai lexicon. Compound word or phrase is acceptable or not.
- Which equivalent do we need to consider, synset-to-synset equivalent or word-to-word equivalent?

For the first issue, we actually intend to translate the PWN synsets into single Thai word only. But problems occurred when we faced with concept that has not its equivalent word. For example,

filly#1 -- (a young female horse under the age of four)
colt2#1 – (a young male horse under the age of four)
hog2#2, hogget#1, hogg#2 – (a sheep up to the age of one year: one yet to be sheared)

There is not any word that conveys the meaning of the above concepts. That is because of the difference of the culture. In this case, phrase or compound word will be introduced to use as the equivalent word of the concept. This phenomenon always occurs with cultural dependent concept, technical terms and new concepts.

As for the second issue, considering between (1) synset-to-synset equivalent assignment or (2) word-to-word equivalent assignment has to be discussed. Let consider the following concept of "dog" in the PWN.

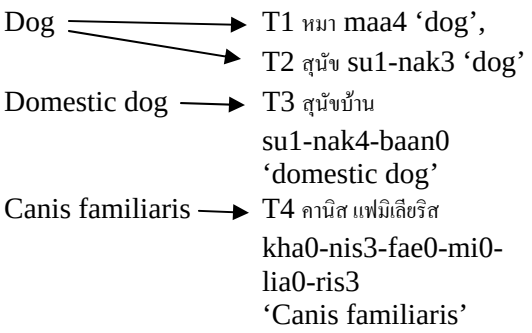
dog#1, domestic dog#1, Canis familiaris#1 -- (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; "the dog barked all night")

The above synset consists of three words; dog, domestic dog, and Canis familiaris. The set of Thai synonyms that is equivalent to this English synset is the following.

Thai synset of 'dog'
{T1 หมา maa4 'dog' (normal word),
T2 สุนัข su1-nak3 'dog' (polite word),
T3 สุนัขบ้าน su1-nak3-baan0 'domestic dog',
T4 คานิส แฟมิเลียริส kha0-nis3-fae0-mi0-lia0-ris3 'Canis familiaris'}

These words have the same concepts but are different in usage. How do we choose the right Thai word for the right equivalent English word? It is a crucial problem. In the paragraph below, three English words which represent the concept "dog" are used in the different context and cannot be interchanged. Similarly, T1, T2 and T3 cannot be used substitutionally. Because it conveys different meaning. Therefore, word-to-word is our solution.

"...Dog usually means the **domestic dog**, Canis lupus familiaris (or "**Canis familiaris**" in binomial nomenclature)...."



Consequently, word-to-word equivalent is very useful for choosing the right synonyms with the right context.

In conclusion, the main principle for the English to Thai translation includes

(1) "Single word" is lexicalized the existence of concepts in Thai.
(2) "Compound" or "Phrase" is represented some concepts that are not lexicalized in Thai.
(3) Synset-to-synset equivalent is used for finding Thai synset that is compatible with PWN synset.
(4) Word-to-word equivalent is used for finding the right Thai word that is compatible with PWN word in each synset.

2.3 Language Issues

This section describes some significant characteristics of Thai that we have to consider carefully during the translation process.

2.3.1 Out of concepts in PWN

There are some Thai words/concepts that do not exist in the PWN, especially cultural-related words. This is the major problem we have to solve during the translation.

One of our future plans is to add synsets that do not exist into the PWN.

2.3.2 Concept differentiation

Some concepts in the PWN are not equal to Thai concepts. For example, a synset {appear, come out} represents one concept "be issued or published" in English, but meanwhile, it represents two concepts in Thai, the concept of printed matter, and the concept of film or movie, respectively.

2.3.3 Concept Structure differentiation

In some cases, the level of the concept relation between English and Thai is not equal. For example, {hair} in the PWN represents a concept of "a covering for the body (or parts of it) consisting of a dense growth of threadlike structures (as on the human head); helps to prevent heat loss; …" but in Thai, it is divided into two concepts;

T1 ขน khon4 'hair'
= "hair" that cover the body
T2 ผม phom4 'hair'
= "hair" that cover on the human head

This shows the nonequivalent of concept. Moreover, it also differs in the relation of concept. In PWN "hair" is a more general concept and "body hair" is more specific concepts. But in Thai T1 ขน khon4 'hair' (hair that covers the body) is more general concept and T2 ผม phom5 'hair' (hair that covers on the human head) is more specific one.

2.3.4 Grammar and usage differentiation

- Part of speech
    - "Classifier" is one of Thai POS which indicates the semantic class to which an item belongs. It's widely use in quantitative expression. For example, 'คน knon' used with 'person', 'หลัง lang' used with house.
    - Some adjectives in English, such as 'beautiful', 'red' and so on can function as the adjective and attribute verb in Thai.
- Social factors determining language usage
    - In Thai, some social factors, such as social status, age, or sex play an important role to determine the usage of language. For example, these following three words กิน kin0, ฉัน chan4 and เสวย sa0-waey4, having the same meaning 'eat', are used for different social status of the listener or referent. These words cannot be grouped in the same synset because of their usage.

3 From Thai to Asian WordNet

AWN, or Asian WordNet, is the result of the collaborative effort in creating an interconnected WordNet for Asian languages. Starting with the automatic synset assignment as shown in section 2, we provide KUI (Knowledge Unifying Initiator) (Sornlertlamvanich, 2006), (Sornlertlamvanich et al., 2007) to establish an online collaborative work in refining the WorNets. KUI is 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.

3.1 Collaboration on Asian WordNet

Followings are our pilot partners in putting things together to make KUI work for AWN.

- Thai Computational Linguistics Laboratory TCL), Thailand
- National Institute of Information and Communications Technology (NICT), Japan
- National Electronics and Computer Technology Center (NECTEC), Thailand
- Agency for the Assessment and Application of Technology (BPPT), Indonesia

- National University of Mongolia (NUM), Mongolia
- Myanmar Computer Federation (MCF), Myanmar

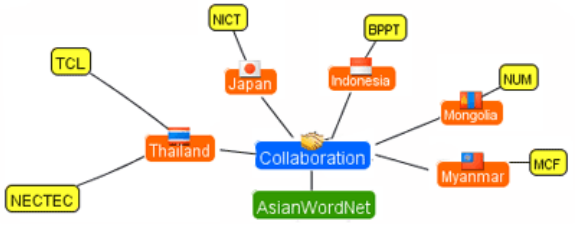


Figure 1. Collaboration on Asian WordNet

3.2 How words are linked

In our language WordNet construction, lexical entry of each language will be mapped with the PWN via its English equivalent. On the process of mapping, a unique ID will be generated for every lexical entry which contains unique sense_key and synset_offset from PWN. Examples of the generated ID show in Table 2. When a word with a unique ID is translated into any language, the same unique ID will be attached to that word automatically. By this way, the lexicon entry in the community can be linked to the each other using this unique ID.

id	sense_key	synset_offset
28259	car%1:06:00::	02929975
28260	car%1:06:01::	02931574
28261	car%1:06:02::	02931966
28262	car%1:06:03::	02932115
28263	car%1:06:04::	02906118

Table 2. Examples of the unique index with sense_key and synset_offset

3.3 Progress on Thai WordNet and Asian WordNet

This section presents the progress on Asian WordNet and Thai WordNet construction.

3.3.1 Current Asian WordNet

At present, there are ten Asian languages in the community. The amount of the translated synsets has been continuously increased. The current amount is shown in the table 3. As shown in the table, for example, 28,735 senses from 117,659 senses have been translated into Thai.

Language	Synset (s)	% of total 117,659 senses
Thai	28,735	24.422
Korean	23,411	19.897
Japanese	21,810	18.537
Myanmar	5,701	4.845
Vietnamese	3,710	3.153
Indonesian	3,522	2.993
Bengali	584	0.496
Mongolian	424	0.360
Nepali	13	0.011
Sudanese	11	0.009
Assamese	2	0.008
Khmer	2	0.002

Table 3. Amount of senses translated in each language

3.3.2 Sense Sharing

Table 4 shows the amount of senses that have been conjointly translated in the group of language. For example, there are 6 languages that found of the same 540 senses.

Language	Sense (s)	%
1-Language	27,413	55.598
2-Language	11,769	23.869
3-Language	5,903	11.972
4-Language	2,501	5.072
5-Language	1,120	2.272
6-Language	540	1.095
7-Language	53	0.107
8-Language	4	0.008
9-Language	2	0.004
10-Language	1	0.002
Total	49,306	100.000

Table 4. Amount of senses translated in each language

3.3.3 Amount of Words in Thai synsets

From the synset in Thai WordNet, there are the minimum of one word (W1) in a synset and the maximum of six words (W6) in a synset. The percentage shown in Table 5 presents that 89.78% of Thai synset contain only one word.

Amount of word in Thai Synset	Sense (s)	%
W1	19,164	89.78
W2	1,930	9.04
W3	211	0.99
W4	27	0.13
W5	4	0.02
W6	8	0.04
Total	21,344	100.00

Table 5. Amount of Word in Thai synsets

4 Conclusion

In this paper we have described the methods of Thai WordNet construction. The semi-auto alignment method constructed the database by using the electronic bilingual dictionary. The manual method has constructed by experts and the collaborative builders who works on the web interface at www.asianwordnet.org.

References

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

Xavier Farreres, German Rigau and Horacio Rodriguez. 1998. *Using WordNet for building WordNets*. In: *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki. 2008. *Development of the Japanese WordNet*. In *LREC-2008*, Marrakech.

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau and Horacio Rodriguez. 1997. Combinding multiple Methods for the automatic Construction of Multilingual WordNets. In proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP'97). Tzigov Chark, Bulgaria.

Piek Vossen, 1997. *EuroWordNet: a multilingual database for information retrieval.* In proceedings of DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.

Thatsanee Charoenporn, Virach Sornlertlamvanich, Chumpol Mokarat, and Hitoshi Isahara. 2008. *Semi-automatic Compilation of Asian WordNet*, In proceedings of the 14th NLP2008, University of Tokyo, Komaba Campus, Japan, March 18-20, 2008.

Valenina Balkova, Andrey Suhonogov, Sergey Yablonsky. 2004. *Rusian WordNet: From UML-notation to Internet/Infranet Database Implementation*. In Porceedings of the Second International WordNet Conference (GWC 2004), pp.31-38.

Virach Sornlertlamvanich, Thatsanee Charoenporn, Chumpol Mokarat, Hitoshi Isahara, Hammam Riza, and Purev Jaimai. 2008. *Synset Assignment for Bi-lingual Dictionary with Limited Resource.* In proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP2008), Hyderabad, India, January 7-12, 2008.