WNMS: Connecting the Distributed WordNet in the Case of Asian WordNet

Kergrit Robkop[1]
Sareewan Thoongsup[1]

[1]Thai Computational Linguistics Lab.,
NICT Asia research center, Thailand
{kergrit, sareewan, 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
Communication Technology, Japan
isahara@nict.go.jp

Abstract

This paper deals with the development of a platform for Asian WordNet (AWN) construction. Not only for the diversity of the languages using in Asia, we also need a platform that can connect the distributedly developing WordNet to establish a network for the cross language WordNet. Each WordNet is created independently by referring to the original Princeton WordNet (PWN) as the focal representation. The Asian WordNet Management System (WNMS) is proposed as a distributed management system that allows the server for each WordNet interchange requests with each other to perform a cross language WordNet interfacing, including the fundamental web service utilities for editing, visualizing and exporting.

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. Nowadays, there have still been some efforts in developing WordNets of some languages in Asia. Some of them can make a progress on their own Wordnets, for example, Japanese WordNet (Isahara and et al., 2008; Bond and et al., 2009), Chinese WordNet (Huang, 2007), Korean WordNet (KorLex, 2006), and so on. The achievement of these projects will lead to the development of linguistic database and the cooperation among languages in Asia.

However, many languages in Asia are still in the initial stage of the development for their own WordNet. Sharing the language resources among the richer and lesser resource languages can be found in many recent efforts (Virach, 2008). Starting from the seed dictionaries we proposed an efficient way to creating a WordNet from the existing bi-lingual dictionaries (Virach and et al., 2008a). The results are now extended to share among the WordNet of each language.

To facilitate the development of the WordNet for languages in Asia, the AsianWordNet Project (AWN) is initiated based on the collaboration manner in creating an interconnection among the WordNets. The goal of AWN is to provide a communication platform to realize the cross language manipulating between the WordNet of the Asian languages. The AWN is built based on the English PWN. Therefore, the original structural information is inherited to the target WordNet through its sense translation and sense ID. The AWN finally connects each WordNet to build the complete Asian WordNet via the English Princeton WordNet.

In the first stage, we adopted KUI (knowledge Unifying Initiator) for collaborative editing to review and complete the translation (Virach and et al., 2008b). We have found that KUI is suitable for building such a community, however, it fails to show the relation between senses; the translation is for word translation rather than sense translation; and the system is also not fully distributed. As a result, we propose a new system called WNMS (Asian WordNet Management System) to dedicate its features to the Asian WordNet construction and visualization.

The following section gives an overview of the tools provided in WNMS (Asian WordNet

Management System) that are Editor, Web Service API, Visualization and Exporting tool. In Section 3 the progress report of Asian WordNet development is given. Section 4 concludes our work.

2 Asian WordNet Management System (WNMS)

WNMS is a distributed management system that makes the servers interact with each other in order to construct Asian WordNet. In the Princeton WordNet database the word entry is organized by linking of semantic relation to the word meaning. It is therefore possible to provide such semantic relation for better understanding in the translation process.

To achieve the goal of AWN, providing a communication platform for finishing WordNet, WNMS has been developed to facilitate the process of the connection between the members, the database storages, and the English WordNet translation. WNMS is easily, freely and publicly available for download. The installed WNMS server will be connected to the other servers to form the AWN network.

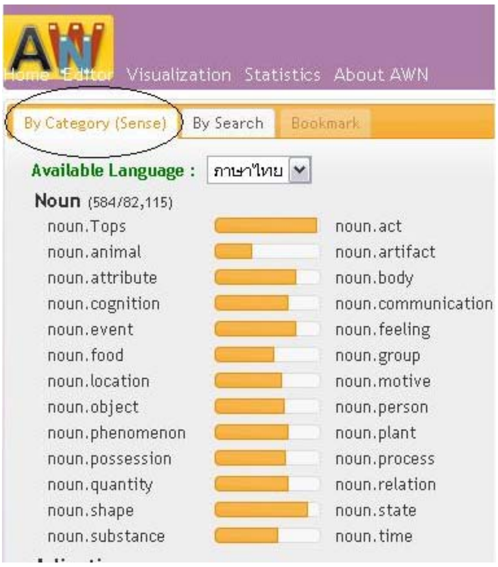
Tools in WNMS are Editor, Web Service, Visualization and Export. These tools are clearly explained in the following subsections.

2.1 Asian WordNet Editor

Asian WordNet Editor is a user-friendly tool that supports users in developing their local WordNet by using the sense translation method. This tool allows an editor or a translator to translate synsets (synonym of word) of PWN with minimal assistant from software developers or programmers.

The important features in AWN Editor are in the followings:

- **By category**: the base types of WordNet synsets has been shown in the By Category window. These base types are based on categories from PWN. An editor or translator can start to translate by searching from the base type and then go down to a synset of its. Figure 1 shows the By Category window with the base types. Base types are categorized as followings: 25 primitive groups for noun, 15 groups for verb, 3 groups for adjective and only one group for adverb.



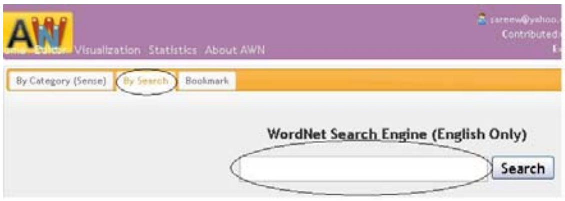
Available Language : ภาษาไทย

Noun (584/82,115)

noun.Tops	noun.act
noun.animal	noun.artifact
noun.attribute	noun.body
noun.cognition	noun.communication
noun.event	noun.feeling
noun.food	noun.group
noun.location	noun.motive
noun.object	noun.person
noun.phenomenon	noun.plant
noun.possession	noun.process
noun.quantity	noun.relation
noun.shape	noun.state
noun.substance	noun.time

Figure 1. Asian WordNet search by category

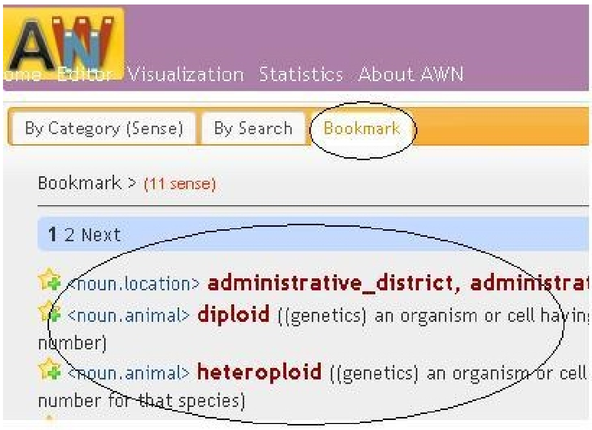
- **By Search**: By Search window is another way to start editing or translating in AWN. Figure 2 illustrates search box where an editor or translator can start creating WordNet by searching the target word on the page of WordNet Search Engine.



WordNet Search Engine (English Only)

Figure 2. Asian WordNet search box

- **Bookmark**: While working on the translation, the editor or translator can create a bookmark for placing some unclear synsets for further checking. Figure 3 shows the Bookmark page on AWN Editor.



Bookmark > (11 sense)

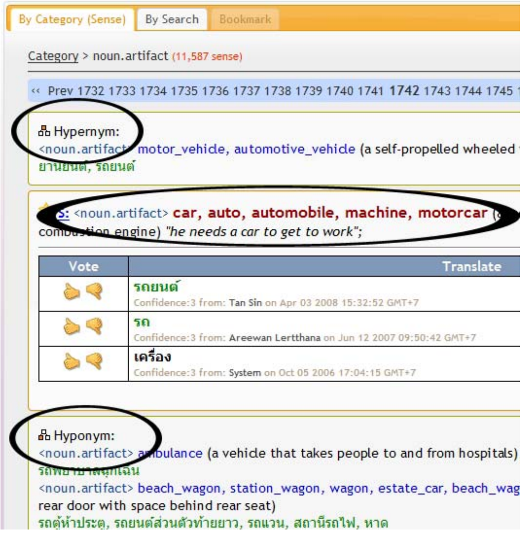
<noun.location> administrative_district, administrat

<noun.animal> diploid ((genetics) an organism or cell havin number)

<noun.animal> heteroploid ((genetics) an organism or cell number for that species)

Figure 3. Bookmark page

- **Semantic relation of a synset**: The synsets of English WordNet show the relation one upper and lower level of semantic relation. Figure 4 illustrates the synset of ‘car, auto, automobile, machine, motorcar’ with a hypernym relation and the hyponym relation. The relation of a synset that is shown by this method will help to scope the idea of word sense in translation.



By Category (Sense) | By Search | Bookmark

Category > noun.artifact (11,587 sense)

‹‹ Prev 1732 1733 1734 1735 1736 1737 1738 1739 1740 1741 **1742** 1743 1744 1745

Hypernym:
<noun.artifact> motor_vehicle, automotive_vehicle (a self-propelled wheeled
ยานยนต์, รถยนต์

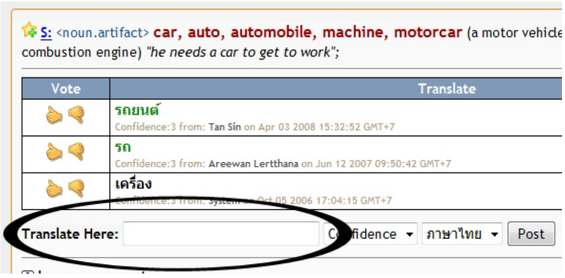
S: <noun.artifact> **car, auto, automobile, machine, motorcar** (a
combustion engine) *"he needs a car to get to work";*

Vote	Translate
👍 👎	รถยนต์ Confidence:3 from: Tan Sin on Apr 03 2008 15:32:52 GMT+7
👍 👎	รถ Confidence:3 from: Areewan Lertthana on Jun 12 2007 09:50:42 GMT+7
👍 👎	เครื่อง Confidence:3 from: System on Oct 05 2006 17:04:15 GMT+7

Hyponym:
<noun.artifact> ambulance (a vehicle that takes people to and from hospitals)
รถพยาบาลฉุกเฉิน
<noun.artifact> beach_wagon, station_wagon, wagon, estate_car, beach_wag
rear door with space behind rear seat)
รถตู้ห้าประตู, รถยนต์ส่วนตัวท้ายยาว, รถแวน, สถานีรถไฟ, หาด

Figure 4. Semantic relation of a synset

- **Insert translation**: an editor or translator can insert the translation of the synset in the translation box, as shown in figure 5.



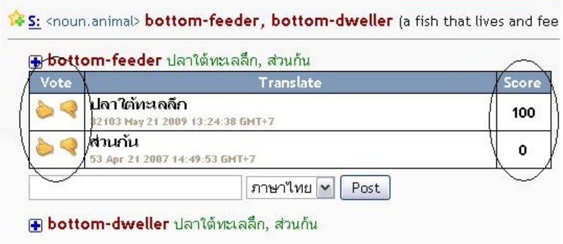
S: <noun.artifact> **car, auto, automobile, machine, motorcar** (a motor vehicle
combustion engine) *"he needs a car to get to work";*

Vote	Translate
👍 👎	รถยนต์ Confidence:3 from: Tan Sin on Apr 03 2008 15:32:52 GMT+7
👍 👎	รถ Confidence:3 from: Areewan Lertthana on Jun 12 2007 09:50:42 GMT+7
👍 👎	เครื่อง Confidence:3 from: System on Oct 05 2006 17:04:15 GMT+7

Translate Here: Confidence ▾ ภาษาไทย ▾ Post

Figure 5. Insert Translation

- **Vote for translation**: figure 6 shows voting box, the editor can verify the translation which were translated by others and vote up for the right translation or vote down for the wrong one by this voting box.



S: <noun.animal> **bottom-feeder, bottom-dweller** (a fish that lives and fee

⊞ **bottom-feeder** ปลาใต้ทะเลลึก, ส่วนก้น

Vote	Translate	Score
👍 👎	ปลาใต้ทะเลลึก 32103 May 21 2009 13:24:38 GMT+7	100
👍 👎	ส่วนก้น 53 Apr 21 2007 14:49:53 GMT+7	0

ภาษาไทย ▾ Post

⊞ **bottom-dweller** ปลาใต้ทะเลลึก, ส่วนก้น

Figure 6. Vote for translation

To reach the objective in working together in AWN construction, AWN Editor has been designed to make the user’s translation as simple and efficient as possible.

Next section is about Web Service API that functions as the connector between servers in different languages.

2.2 Web Service API

In WNMS, Web Service is designed to support machine-to-machine interaction over the network of AWN. AWN Web Service is Internet Application Programming Interfaces (API) that can be accessed over the network and executed on a remote system hosting the requested services. When running Web Service, each WordNet in AWN network can be considered autonomous. Each language WordNet in AWN network works independently. Web Service API will function as the connector among membership’s servers. By this way, data in AWN network can be exchanged among the membership.

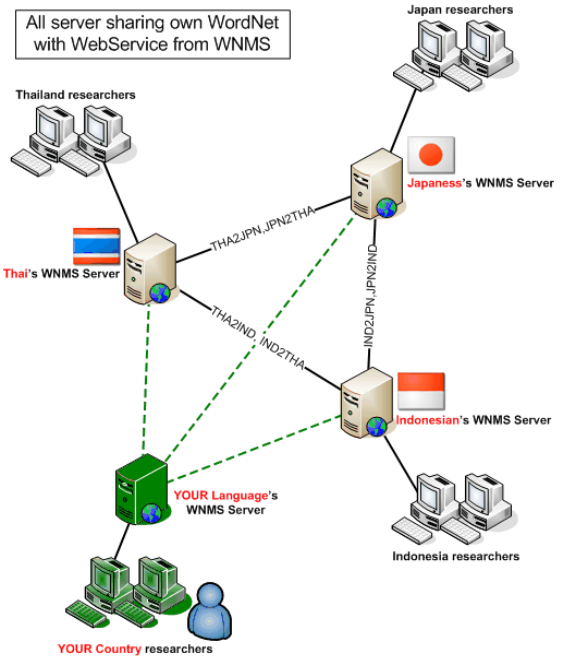
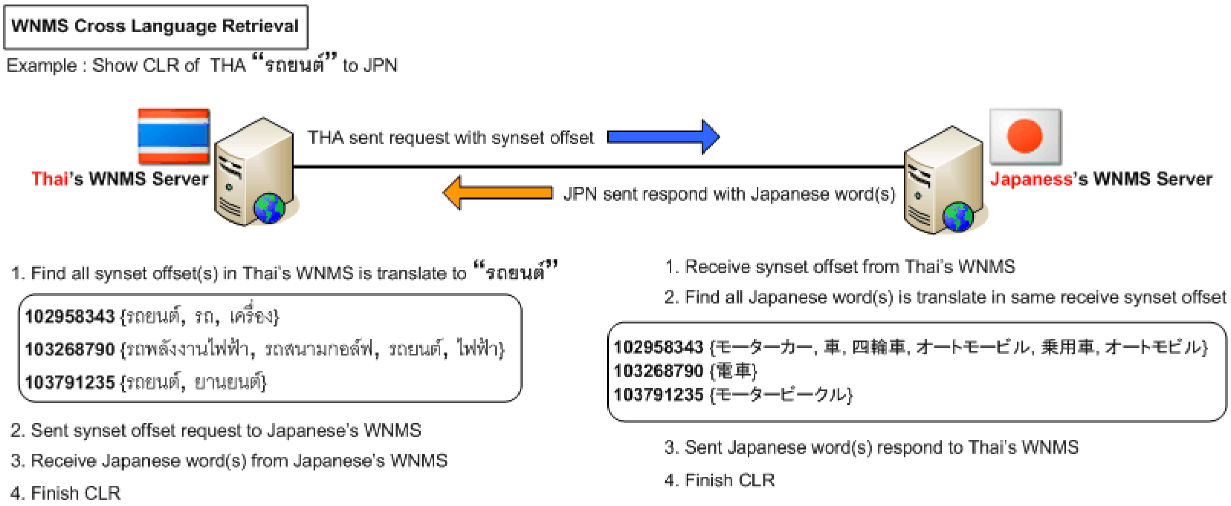


Figure 7. All server sharing own WordNet with WebService from WNMS

Figure 7 shows one-to-one connection among languages in AWN network, for example, THA2JPN is the linking between Thai and Japanese to exchange the data through English WordNet.

The different file format and data index cause some problem for information retrieval. Sometimes, an information file in each WordNet is formatted differently. So the requester needs to know how to access different file formats and to specify which file format that the WordNet local provider should use to access the data source.


WNMS Cross Language Retrieval
Example : Show CLR of THA "รถยนต์" to JPN
Thai's WNMS Server
THA sent request with synset offset
JPN sent respond with Japanese word(s)
Japaness's WNMS Server

1. Find all synset offset(s) in Thai's WNMS is translate to "รถยนต์"

102958343 {รถยนต์, รถ, เครื่อง}
103268790 {รถพลังงานไฟฟ้า, รถสนามกอล์ฟ, รถยนต์, ไฟฟ้า}
103791235 {รถยนต์, ยานยนต์}

2. Sent synset offset request to Japanese's WNMS
3. Receive Japanese word(s) from Japanese's WNMS
4. Finish CLR

1. Receive synset offset from Thai's WNMS
2. Find all Japanese word(s) is translate in same receive synset offset

102958343 {モーターカー, 車, 四輪車, オートモービル, 乗用車, オートモビル}
103268790 {電車}
103791235 {モータービークル}

3. Sent Japanese word(s) respond to Thai's WNMS
4. Finish CLR

Figure 8. WNMS Cross Language Retrieval

WNMS has been developing to adjust this problem. We attempt to set the standard of WordNet information retrieval by using WNMS in Asian WordNet.

Actually, a language database based on PWN structure is indexed by the WordNet sense index that provides for accessing synsets and word senses in the WordNet database. By using WNMS, a membership of AWN can use the WordNet sense index to retrieve synsets or other information related to a specific sense in WordNet from another.

Figure 8 shows the process of data transferring between Japanese and Thai. A Thai user needs to search Japanese words that relate with รถยนต์ rod4-yon0 'car' in Thai word. This word will be searched for synset_offsets from English WordNet by Web Service API. All of "รถยนต์" synset_offsets will be transferred to Japanese server. Then the synset_offsets of รถยนต์ rod4-yon0 'car' will be searched for the information of Japanese words. Finally they will be transferred back to Thai server.

By using Web Service API, an unfamiliar language database does not need to be stored in the server. The data of other languages will be transferred to the target server by Web Service API tool when they are required. Membership countries are therefore only responsible for their own language database.

2.3 Visualization

Visualization tool allows fast interactive viewing of WordNet structures organized in a tree. The Treebolic program (Bernard Bou, 2009) has been used for visualizing the result of WordNet structure received from Web Service API.

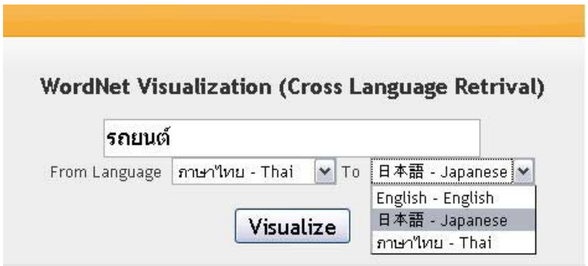
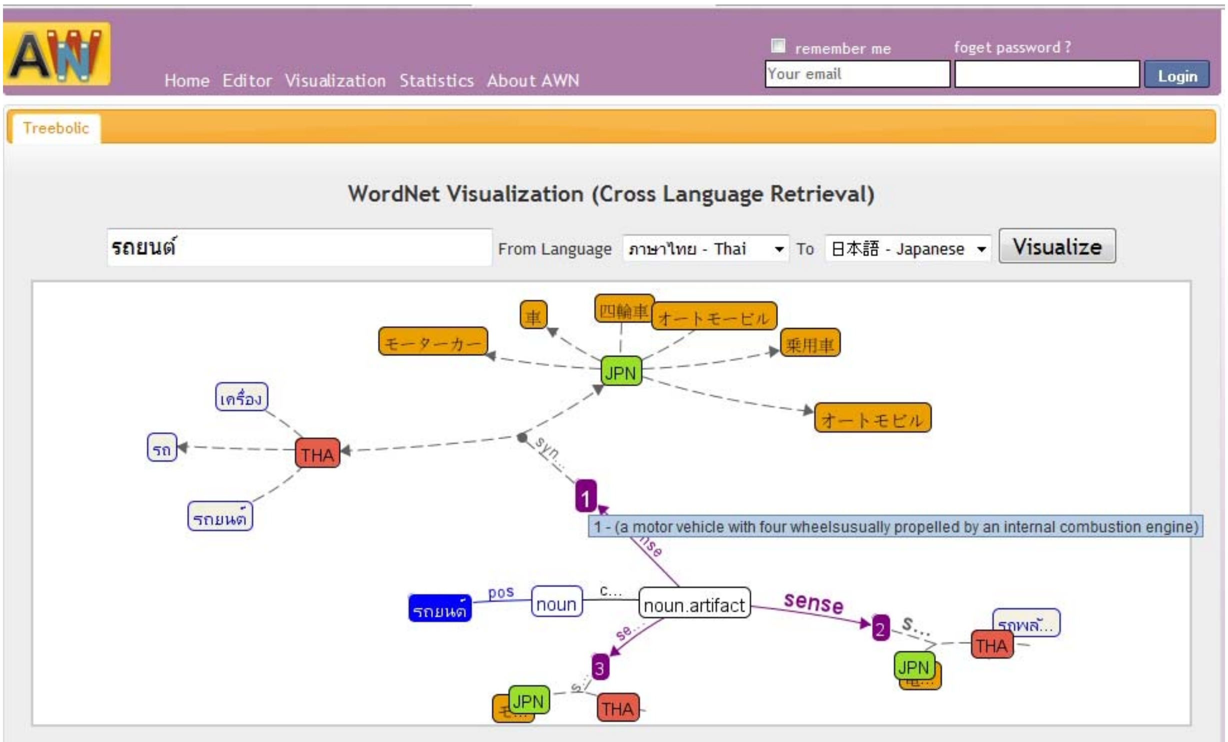


Figure 9. Asian WordNet Visualization

In the AWN Visualization page, a user can visualize the structure of WordNet by typing a word in the box and choose the source and target language, as in figure 9 and figure 10 visualizes the result of transferring data.



WordNet Visualization (Cross Language Retrieval)

รถยนต์ From Language ภาษาไทย - Thai To 日本語 - Japanese

1 - (a motor vehicle with four wheelsusually propelled by an internal combustion engine)

Figure 10. Asian WordNet Visualization

The process of transferring are:

1. When receiving the surface of word, Web Service API will search for the senses of Thai word รถยนต์ rod4-yon0 'car' in Thai WordNet database.
2. The information of Thai word รถยนต์ rod4-yon0 'car' that will be taken from the database is:
   a. Synset of Thai word
   b. Synset_offsets of English WordNet
   c. POS with category of base type and
   d. Synset of English word
3. The synset_offsets of English WordNet will be transferred to Japanese server to search for information of Japanese word.
4. The synonym sets of Japanese will be sent to Thai server.
5. The Visualization Tool visualizes the WordNet structure of Thai and Japanese word.

The following is the example of the information of Thai word รถยนต์ rod4-yon0. There are three concept for Thai word รถยนต์ rod4-yon0.

Sense 1 of Thai word รถยนต์ rod4-yon0

Synset_offset:	102958343
POS.base type:	<noun.artifact>
English synset:	car 0 auto 0 automobile 0 machine 1 motorcar 0
Gloss:	(a motor vehicle with four wheels usually propelled by an internal combustion engine)
Thai synset words:	รถยนต์, รถ, เครื่อง
Japanese synset words:	モーターカー, 車, 四輪車, オートモービル, 乗用車, オートモビル

Sense 2 of Thai word รถยนต์ rod4-yon0

Synset_offset:	103268790
POS.base type:	<noun.artifact>
English synset:	electric 0 electric_automobile 0 electric_car 0
Gloss:	(a car that is powered by electricity)
Thai synset words:	รถพลังงานไฟฟ้า, รถยนต์, รถสนามกอล์ฟ, ไฟฟ้า
Japanese synset words:	電車

Sense 3 of Thai word รถยนต์ rod4-yon0

Synset_offset:	103791235
POS.base type	<noun.artifact>
English synset:	motor_vehicle 0 automotive_vehicle 0
Gloss:	(a self-propelled wheeled vehicle that does not run on rails)
Thai synset words:	รถยนต์, ยานยนต์
Japanese synset words:	モータービークル

This information will be visualized in a tree by using Visualization in AWN, as in figure 10.

2.4 Export tool

The Export tool in WNMS allows the membership user to export data to the following format: LMF (Lexical Markup Framework) (Takenobu and et. al, 2009), XML (Extensible Markup Language), CSV text file format, and so on, so the user can use WordNet for other related projects, for example, machine translation, word sense disambiguation, and so on.

3 Progress Work on Asian WordNet

Asian WordNet has been being developed to reach the goal of the project. The success of AWN project needs to develop not only some tools for construction but also the cooperation among Asian languages. At present, there are ten Asian languages in AWN, as the following table.

Language	Synsets
Thai	80,098
Lao	72,672
Japanese	66,648
Korean	65,483
Burmese	26,033
Indonesian	21,584
Vietnamese	17,767
Mongolian	2,283
Bengali	1,775
Sinhala; Sinhalese	177

Table 1 The number of synsets in AWN

Each language has a difference in the linguistic resources, so it needs to use several methods to create and share the WordNet among Asian languages. We try to use available resources of each language to build AWN. The several methods for building are:

- Using local WordNet to link with AWN.

By supporting and cooperating from Japanese WordNet by NICT, Thai WordNet can be linked and interchange the data with Japanese WordNet by using WNMS in AWN.

- Mapping local word surface to English WordNet

Bilingual dictionary can be a resource for WordNet construction. The surface words in language have been mapped to English WordNet. However, it needs to recheck by native language. Those languages in AWN are Korean, Burmese, Indonesian, Vietnamese, Mongo, Bengali and Sinhalese.

- Using a phoneme-based transfer method for machine translation.

This method can be used for languages that are very similar in terms of grammar, lexicon, and character encoding scheme. Thai and Lao languages have these characteristics (Virach and et al., 2008).

- Using manual translation

This method has been used for Thai WordNet. We use the Editor interface on AWN to translate the concepts (synset) of English in PWN into Thai synonym sets.

The number of the synsets of English WordNet has been continuously translated into Thai synsets. From 117,659 synsets in PWN, there are:

49,514 synsets translated

40,425 translated synsets have been approved.

57,047 Thai unique lemmas

4 Conclusion

In this paper we have described tools used for Asian WordNet construction. The development of tools is to facilitate the construction and to make a better connection among WordNets of Asian languages. These tools will help extend the network of Asian WordNet. WNMS is easily, freely and publicly available for download. The installed WNMS server will be connected to the other servers in AWN network. WNMS can be downloaded at www.asianwordnet.org.

Acknowledgements

AWN project would not have been possible without our enthusiastic partners and community contributors. We gratefully thank Dr. Hitoshi Isahara and Dr. Francis Bond from NICT, Japan, Professor Chi Mai Luong from Institute of Information Technology, National Center for Science and Technology of Vietnam, Vietnam, Valaxay Dalaloy from Science Technology and Environment Agency, LAOS, Hammam Riza from Agency for the Assessment and Application of Technology, Indonesia, Professor Purev Jaimai from Center for Research on Language Processing, National University of Mongolia, Mongolia, Dr. Myint Myint Than from Myanmar Computer Federation (MCF), Myanmar, StarDict and ADD school supporters.

References

Bernard Bou. 2009. *Treebolic*. Available at http://treebolic.sourceforge.net/

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

Fellbum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. 2009. *Enhancing the Japanese WordNet*. in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009, Singapore.

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

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

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

Takenobu Tokunaga, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Yingju Xia, Chu-Ren Huang, Shu-Kai Hsieh and Kiyoaki Shirai. 2009. *Query Expansion using LMF-Compliant Lexical Resources*, Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP), Suntec, Singapore, August 6-7.

Virach Sornlertlamvanich. 2008. *Cross Language Resource Sharing*, Proceedings of Workshop on NLP for Less Privileged Languages, Hyderabad, India, January 11.

Virach Sornlertlamvanich, Chumpol Mokarat, and Hitoshi Isahara. 2008. *Thai-Lao Machine Translation based on Phoneme Transfer*, Proceedings of the 14th NLP2008, University of Tokyo, Komaba Campus, Japan, March 18-20.

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

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergrit Robkop, and Hitoshi Isahara. 2008. *KUI: Self-organizing Multi-lingual WordNet Construction Tool*, Proceedings of the Fourth Global WordNet Conference (GWC2008), Szeged, Hungary, January 22-25.