Language Resource Management System for Asian WordNet Collaboration and Its Web Service Application

Virach Sornlertlamvanich[1,2], Thatsanee Charoenporn[1,2], Hitoshi Isahara[3]

[1]Thai Computational Linguistics Laboratory, NICT Asia Research Center, Thailand

[2]National Electronics and Computer Technology Center, Pathumthani, Thailand

[3]National Institute of Information and Communications Technology, Japan

E-mail: virach@tcllab.org, thatsanee@tcllab.org, isahara@nict.go.jp

Abstract

This paper presents the language resource management system for the development and dissemination of Asian WordNet (AWN) and its web service application. We develop the platform to establish a network for the cross language WordNet development. Each node of the network is designed for maintaining the WordNet for a language. Via the table that maps between each language WordNet and the Princeton WordNet (PWN), the Asian WordNet is realized to visualize the cross language WordNet between the Asian languages. We propose a language resource management system, called WordNet Management System (WNMS), as a distributed management system that allows the server to perform the cross language WordNet retrieval, including the fundamental web service applications for editing, visualizing and language processing. The WNMS is implemented on a web service protocol therefore each node can be independently maintained, and the service of each language WordNet can be called directly through the web service API. In case of cross language implementation, the synset ID (or synset offset) defined by PWN is used to determined the linkage between the languages.

1. Introduction

Language resource becomes crucial in the recent needs for cross cultural information exchange. To manage the language resource efficiently, there are many attempts to collect the resource from several languages to a single network. The Princeton WordNet (PWN) (Fellbuam, 1998) is one of the most semantically rich English lexical banks and widely used as a resource in many aspects of 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 (Korex, 2006), and Hindi WordNet (Hindi WordNet, 2007). 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, Virach and et al., 2008a). Starting from the seed dictionaries, we proposed an efficient way to creating a WordNet from the existing bi-lingual dictionaries (Virach and et al., 2008b). 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 Asian WordNet 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., 2008c). 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 the system that are Editor, and Visualization. Section 3 discusses the Web Service API that connects each service node to form a network for Asian WordNet collaboration, and the last section is the conclusion.

2. Asian WordNet Management System

Asian WordNet management system 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 a set of semantic relations linked to the word meaning. It is therefore possible to provide such semantic relations for better understanding in the translation process.

To achieve the goal of AWN by 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 explained in detail in the following subsections.

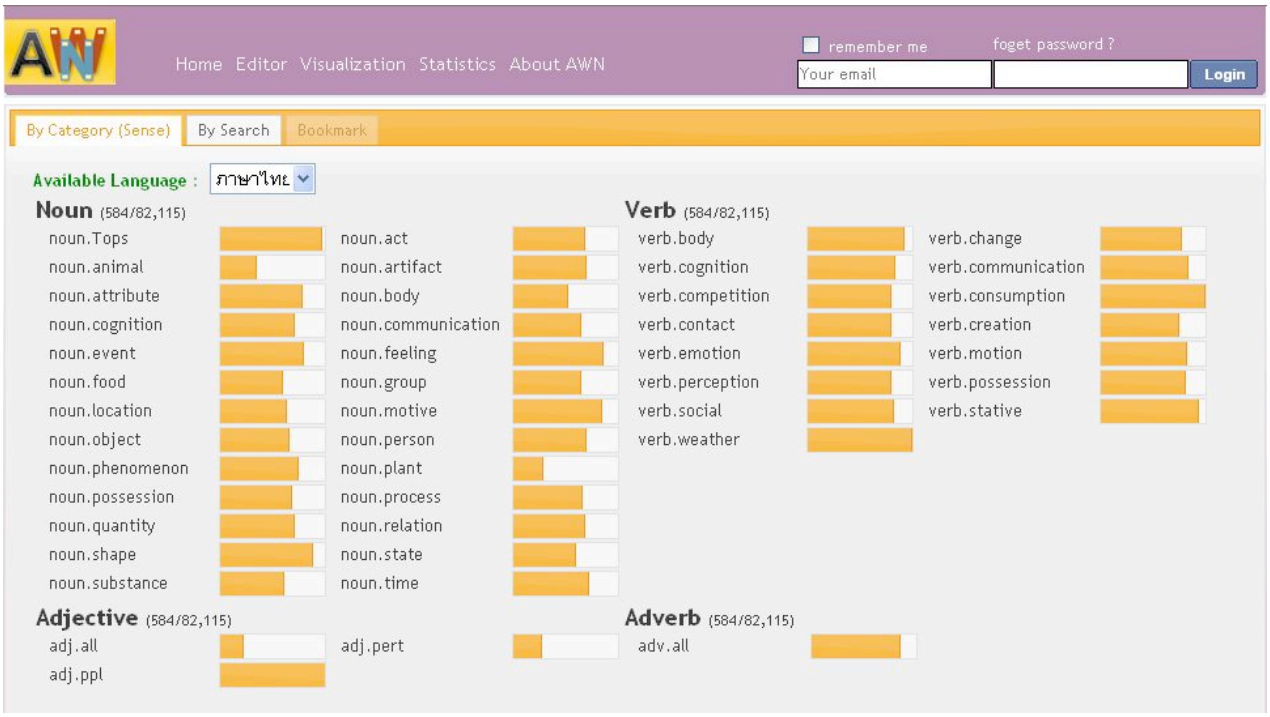

Figure 1: Asian WordNet classification by category

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 Asian WordNet can be edited in the following manners.

Category: The base types of WordNet synsets are shown in By Category window. These base types are classified on categories from PWN. An editor or translator can start to translate by searching from the base type and then go down to its synset. The base types are categorized into 25 groups of noun, 15 groups of verb, 3 groups of adjective and 1 group of adverb as shown in Figure 1. The color bar attaching to each category exhibits the ratio of the translated senses.

Searching: The specific word can be searched and directly jumped to edit or translate in AWN. An editor or translator can start the translation by inserting the target word in the WordNet Search box. An editor can insert the new translation of the synset in the translation box. If the translations are provided, the editor can verify the translation by voting. For example, the translations of the word "car" are shown in Figure 2. The editor can vote up for the right translation or vote down for the wrong one for each sense. The translation words are listed in descending order according to the voting score.

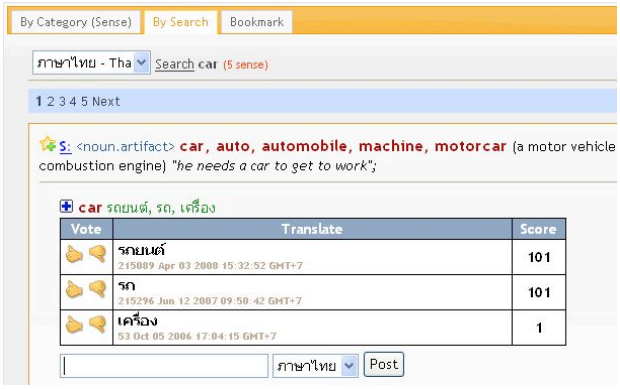

Figure 2: Asian WordNet editing and voting

2.2 Visualization

Visualization tool is an implementation of interactive tree browser for viewing the WordNet information in tree structure manner. Treebolic program (Bernard Bou, 2009) has been adopted for visualizing the result of WordNet structure. It is the result of the transferred data explained in the section of Web Service API.

In the AWN Visualization, a user can visualize a structure of WordNet by typing a word in the query box and choose the source and target languages. An example

of the visualization of the result of looking up a Thai WordNet for the word "รถยนต์" (/rod4-yon0/, car) to find the correspondent Japanese WordNet is depicted in Figure 3.

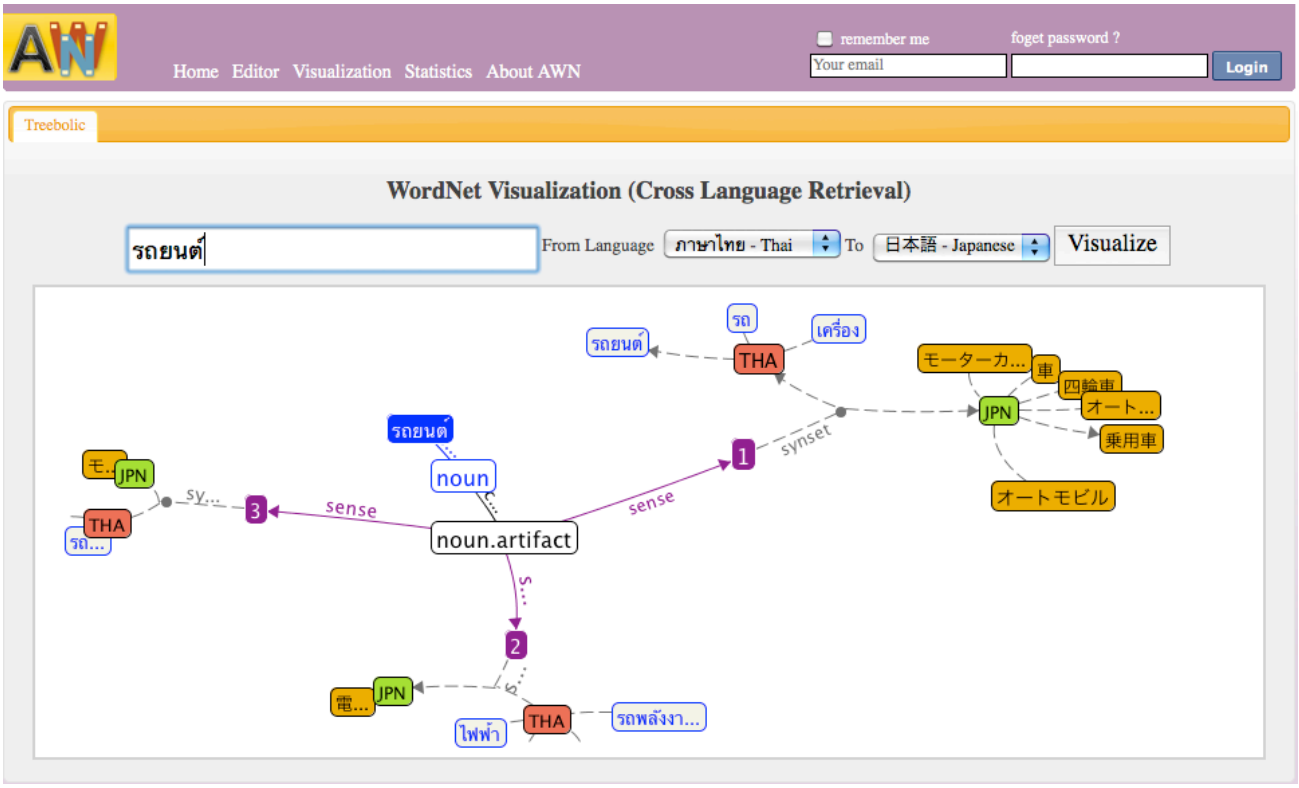

Figure 3: Asian WordNet Visualization

The process of looking up a cross language Asian WordNet from Thai to Japanese is explained in the following steps:

1. When receiving the word entry, Web Service API will look for the attaching senses of Thai word in Thai WordNet database.
2. Following information of Thai word entry will be retrieved from the database.
   a. Synset of the Thai word
   b. Synset offsets of English WordNet
   c. POS with category of base type
   d. Synset of English word
3. The synset offsets of English WordNet will be submitted to Japanese server to look for the information of Japanese WordNet.
4. The corresponding synsets of Japanese word will be sent back to Thai server.
5. The visualization tool presents the WordNet structure of the retrieved Thai and Japanese WordNet.

This information will be visualized in a tree manner by using AWN Visualization as presented in the Figure 3. Each word or synset can be visualized flexibly using the interface control such as mouse. It is beneficial to user to find the meaning of word based on its synset.

The next section will explain about Web Service API that functions as the connector between servers of different languages.

3. Web Service API

In WNMS, Web Service is designed to support machine-to-machine interaction over the network of AWN. AWN Web Service is the 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 server in AWN network can be considered autonomous. The user has no control over this service. Web Service API will function as the connector among the member's servers. By this way, the members of AWN network can exchange their WordNet databases.

Figure 4 shows one-to-one connection among languages in AWN network, for example, THA2JPN is the linkage between Thai and Japanese WordNet to exchange the data through English WordNet.

The different file format and data index are the annoying problems in the information retrieval work. 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 has been developing to reduce the problem. We attempt to set up a standard for WordNet information retrieval by using WNMS in Asian WordNet.

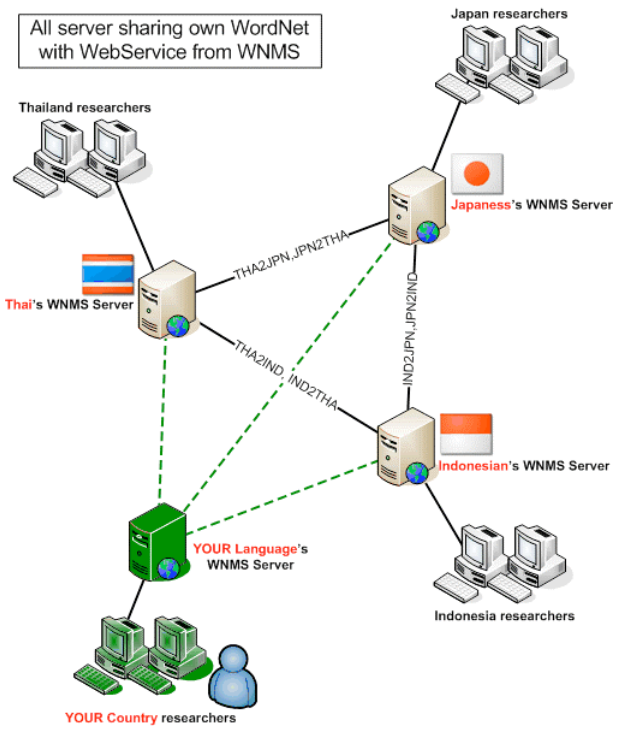

Figure 4: WordNet database sharing through Web Service

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. Language resources can be linked based on their servers from various countries as the language resource network. Member countries can use the language translation system via the Web Service API. They are only responsible for their own language database. It is so convenient for the network members to maintain their own language resources and make use of others' language resources via a standard Web Service API. In case of Asian WordNet, it is already explained in Section 2 about how to implement for the cross language Asian WordNet visualization.

4. Conclusion

In this paper, we have described the language resource management system based on the implementation for Asian WordNet development and dissemination. The development of tools is to facilitate the construction and to make a better connection among WordNets of Asian languages. We hope that these tools can help to fulfill the network of Asian WordNet. Web service API is also beneficial for cross language WordNet implementation. It links the language resource from each node and can be maintained independently as the self-maintaining system. Currently, we have the active AWN system at www.asianwordnet.org that can be used freely. In addition, the service of language resources such as Asian WordNet is compiled to the Open API. Any applications can get access to the language resources through this Open API.

5. References

Bou, B. (2009). Treebolic. Available at http://treebolic.sourceforge.net/.

Huang, C.R. (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.

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., Kanzaki, K. (2009). Enhancing the Japanese WordNet. *In Proceedings of 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/.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K. (2008). Development of the Japanese WordNet. *In Proceedings of LREC-2008*, Marrakech.

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

Sornlertlamvanich, V. (2008). Cross Language Resource Sharing. *In Proceedings of Workshop on NLP for Less Privileged Languages*, IJCNLP-2008, Hyderabad, India.

Sornlertlamvanich, V., Mokarat, C., Isahara, H. (2008). Thai-Lao Machine Translation based on Phoneme Transfer. *In Proceedings of the 14th NLP-2008*, University of Tokyo, Komaba Campus, Japan.

Sornlertlamvanich, V., Charoenporn, T., Mokarat, C., Isahara, I., Riza, H., Jaimai, P. (2008). Synset Assignment for Bi-lingual Dictionary with Limited Resource. *In Proceedings of the Third International Joint Conference on Natural Language Processing* (IJCNLP-2008), Hyderabad, India.

Sornlertlamvanich, V., Charoenporn, T., Robkop, K., Isahara, H. (2008). KUI: Self-organizing Multi-lingual WordNet Construction Tool. *In Proceedings of the Fourth Global WordNet Conference* (GWC-2008), Szeged, Hungary.