Enhanced Tools for Online Collaborative Language Resource Development

Virach Sornlertlamvanich
Thatsanee Charoenporn
Suphanut Thayaboon
Chumpol Mokarat
Thai Computational Linguistics Lab.
NICT Asia Research Center,
Pathumthani, Thailand
{virach, thatsanee, suphanut, 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

Abstract

This paper reports our recent work of tool development for language resource construction. To make a revision of Asian WordNet which is automatically generated by using the existing English translation dictionary, we propose an online collaborative tool which can organize multiple translations. To support the work of syntactic dependency tree annotation, we develop an editing suite which integrates the utilities for word segmentation, POS tagging and dependency tree into a sequence of editing.

1 Introduction

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.

This paper presents an online collaborative tool particularly to facilitate the construction of the Asian WordNet which is automatically generated by using the existing resources having only English equivalents and the lexical synonyms.

In addition, to support the work of syntactic dependency tree annotation, we develop an editing suite which integrates the utilities for word segmentation, POS tagging and dependency tree. The tool is organized in 4 steps, namely, sentence selection, word segmentation, POS tagging, and syntactic dependency tree annotation.

The rest of this paper is organized as follows: Section 2 describes the collaborative interface for revising the result of synset translation. Section 3 describes the tool for annotating Thai syntactic dependency tree corpus. And, Section 4 concludes our work.

2 Collaborative Tools for Asian WordNet Construction

There are some efforts in developing Wordnets for some of Asian languages, e.g. Chinese, Japanese, Korean, and Hindi. The number of languages that have been successfully developed their Wordnets is still limited to some active research in the area. However, the extensive development of Wordnet in other languages is of the efforts to support the NLP research and implementation. It is not only to facilitate the implementation of NLP applications for the language, but also provide an inter-linkage among the Wordnets for different languages to develop multi-lingual applications.

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 introduce KUI (Knowledge Unifying Initiator), (Sornlertlamvanich et al., 2007) to establish an online collaborative work in refining the WorNets.

KUI is a community software tool which allows registered members including language experts to 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-based 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. Through this effort, a translated version of Asian WordNet can be established.

Table 1 shows a record of WordNet displayed for translation in KUI interface. English entry together with its part-of-speech, synset, and gloss are provided if exists. The members will examine the assigned lexical entry and decide whether to vote for it or propose a new translation.

Car
[Options] POS : NOUN Synset : auto, automobile, machine, motorcar Gloss : a motor vehicle with four wheels; usually propelled by an internal combustion engine;

Table 1. A record for a synset

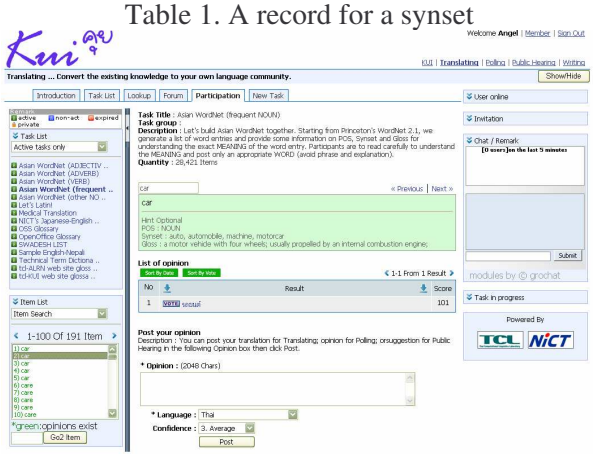


Figure 1. KUI interface (www.tcllab.org/kui)

Figure 1 illustrates the translation page of KUI In the working area, the login member can participate in proposing a new translation or voting for the preferred translation to revise the synset assignment. Statistics of the progress as well as many useful functions such as item search, chat, and list of online participants are also provided to understand the progress of work and to work online with other members.

3 Tool for Constructing a Syntactic Dependency Tree Annotated Corpus

The tool is organized in 4 steps, namely, sentence selection, word segmentation, POS tagging, and syntactic dependency tree annotation, shown in Figure 2. Sentence segmentation is yet another crucial issue for the Thai language. We, however, will not discuss about the issue in this work. The input is already a list of sentences provided for annotator to select.

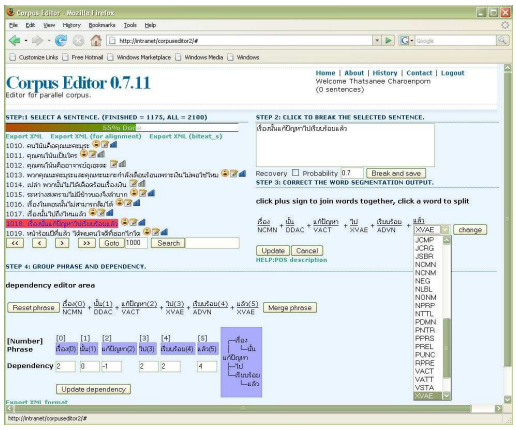
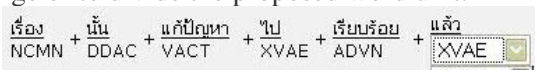


Figure 2. Syntactic Dependency Tree Annotation

3.1 POS Annotation

The result from the automatic word segmentation and POS tagging program is generated with alternative POSs for revision. A dropdown list of POSs is provided for annotator to correct the POS. Since word segmentation is processed together with POS tagging, the annotator is also provided a GUI to merge or to divide the proposed word unit.



3.2 Syntactic Dependency Tree Annotation

The result from POS annotation in Section 3.1 is passed to define the syntactic dependency between words. The dependency is assigned to form a phrase and a sentence respectively. The final output will be marked in the XML manner and shown as a tree for confirmation.

4 Conclusion

Our current work on the web-based collaborative tool for Asian WordNet construction and tool for Syntactic Dependency Tree Annotation are developed as an open platform for online contribution. A user-friendly interface and self-organizing utilities are intentionally prepared to support the online collaborative work.

References

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergit Robkop, and Hitoshi Isahara. 2007. *Collaborative Platform for Multilingual Resource Development and Intercultural Communication*, IWIC2007, Springer, LNCS4568:91-102.