Participation in Language Resource Development and Sharing

Virach Sornlertlamvanich
National Electronics and Computer Technology Center
Pathumthani, Thailand
virach.sornlertlamvanich@nectec.or.th

Abstract

Language resources are really much required for understanding and modeling the language in the present approaches. The language that has a rich language resource gains a big benefit in making a big advance in language processing. On the other hand, the less resource language is struggling with preparing a large enough language resource such as raw text or annotated corpora. It is a labor intensive and time consuming task. Moreover, computerization of the text is another non-trivial effort. There needs a supportive computing environment in inputting, encoding, retrieving, analysis, etc.. Learning from the rich resource languages, we gradually collecting the resource and preparing the necessary tools. Through many efforts in the recent years, we can see some significant outcomes from PAN localization project (2004-2007, 2007-2101, http://www.panl10n.net/), ADD (2006-2010, http://www.tcllab.org/), Asian WordNet (http://asianwordnet.org/), Hindi WordNet (http://www.cfilt.iitb.ac.in/wordnet/webhwn/), BEST (since 2009, Thai Word Segmentation Software Contest, http://thailang.nectec.or.th/best/) and many NLP summer schools. The activities gain a big potential in leveraging the NLP tools development and research personnel development. It results in a big growth of Asian language resource development and research. With the spirit of sharing on social networking, the resources can efficiently be developed to a satisfied amount in a reasonable time scale. Asian WordNet is an example of developing a set of 13 languages of Wordnet connected via Princeton WordNet. Thai WordNet is open for online collaborative development. About 70K synsets and 80K words of Thai WordNet are available online. Thai-Lao conversion is an approach to exhibit the advantage in utilization of language similarity to increase the other language resource. Lao WordNet is created by converting from Thai WordNet by using the phoneme transfer approach. Taking the advantage of language similarity, the language corpus can be obtained by a quick conversion rule. In this case, the study of direct transfer is much more efficient than creating from the scratch. Currently, most of the above mentioned results are open to public for at least research purpose. However, more and more language resources are still needed to improve the language processing. The possible of online collaborative development and sharing is a key factor in the language resource development.

Proceedings of the 9th Workshop on Asian Language Resources, page 1,
Chiang Mai, Thailand, November 12 and 13, 2011.