Cross Language Resource Sharing

Virach Sornlertlamvanich
Thai Computational Linguistics Lab., NICT Asia Research Center,
Pathumthani, Thailand
virach@tcllab.org

Abstract

Language resource development is crucial for language study in current approaches. Many efforts have been made to model a language on very large scaled corpora. Statistical and probabilistic approaches are playing a major role in taking the advantage of incorporating the context to improve their performance to a promising result in many areas of natural language processing such as machine translation, parsing, POS tagging, morphological analysis, etc. It is believed that if there are sufficient corpora for a language, we can develop many efficient language processing applications within an expectable period. However, corpus development is a labor intensive task and requires a continuous effort in maintaining the result to such a qualified level. The problem is magnified when we need to deal with the less computerized languages. The availability of the computerized language data can be varied by the availability of the standard of language encoding, number of speakers, economic scale of the speakers, and the language supporting tools. As a result, the technology gap between languages are widened as we can see in the evidence of online language populations and web contents which are mainly occupied by English, and others major languages distributed in Chinese, Spanish, Japanese, German and French. The major concern in the less computerized languages is how to leverage the technology for those languages which will result in scaling up the number of online language populations. Cross language resource sharing is one of the efforts to increase the opportunity for the access to those languages. We are expecting that a language may utilize the resource from other similar languages in terms of computational approaches and corpora. To relate the language resources among the less computerized languages has brought us to the following open questions:

1. Is there any intermediate representation that can efficiently relate among the languages? Will it be an approach of meaning representation such as conceptual unit, WordNet, or etymological word form representation such as Pali, Sanskrit, Chinese character?

2. Can a shallow language processing approach be used to increase the resources, namely orthographic conversion, transliteration?

3. Will English be a good intermediate language? This is because of the availability of the language pairing resources with the English language.

As a platform for cross language resource development, we have developed KUI (Knowledge Unifying Initiator: http://www.tcllab.org/kui) equipped with a voting function to measure for the most reliable translation. The English language is not only a possible intermediate representation for languages. Other appropriate approaches could be considered to maximize the resource sharing among the less computerized languages if we can determine a better common feature among those languages.