From Non-Segmenting Language Processing to Web Language Engineering

Virach Sornlertlamvanich
Thai Computational Linguistics Laboratory (TCL), NICT, Thailand
virach@tcllab.org

It is interesting to look at the statistics of the online languages reported by the Global Reach (www.global-reacg.biz). In September 2004, it was reported that the top six online language populations were English 35.2%, Chinese 13.7%, Spanish 9.0%, Japanese 8.4%, German 6.9%, and French 4.2% while the web contents were English 68.4%, Japanese 5.9%, German 5.8%, Chinese 3.9%, French 3.0%, and Spanish 2.4%. There are some changes in ranking between the online language populations and the existing of the web contents. However, English is still the majority language used in the online community. Many efforts have been making to prevent the fall-off in using of other languages, especially the less computerized languages. It is said that there are about 7,000 languages using in all over the world. At the same time the less computerized languages are disappearing. The Rosetta Project (http://64.81.54.21:8080/live/) is a global collaboration to build an online archive of all documented human languages. The Language Observatory Project (www.language-observatory.org) initiated by Nagaoka University of Technology to search for the disappearing languages.

To deal with languages as many as we can find online, it is much more efficient to consider the language independent approaches. The big difference between segmenting languages (i.e. English and other European languages) and non-segmenting languages (i.e. Thai, Lao, Khmer, Japanese, Chinese and a lot of Asian languages) in the existing of word boundary marker causes the change in language processing. Most of the current approaches are based on the assumption that words are already identified disregarding the existing of the word boundary markers. The research on word boundary is separately conducted under the topic of word segmentation. On contrary, we proposed some algorithms to handle the non-segmenting languages (Virach 2005a, Virach 2005b) to establish a language independent approach.

In our recent research, we proposed a language interpretation model to deal with an input text as a byte sequence rather than a sequence of words. It is an approach to unify the language processing model to cope with the ambiguities in word determination problem. The approach takes an input text in the early stage of language processing when the exhaustive recognition of total word identity is not necessary. In our research, we present the achievements in language identification, indexing for full text retrieval, and word candidate extraction based on the unified input byte sequence. Our experiments show comparable results with the existing word-based approaches.

In our statistical-based word extraction research (Virach et al. 2000), it was reported to yield about 30% of the total word candidates being the unregistered words of a published dictionary, when the recall threshold was set to 56%. Character-based mutual information and entropy provided significant information to C4.5 algorithm for selecting appropriate candidates for words. The approach greatly supported the process of developing a dictionary, and later was extended to fulfill a dictionary-less search engine (Virach et al. 2003). The search engine had introduced a word score as a heuristic value to determine the word likelihood of a string. The word score was a normalized value of a mutual information value. The minimum score of the left and right hand side of a string in question was assigned as the word score of the string. Based on the proposed approach, we successfully implemented a multi-lingual search engine with minimum modification.

Language identification (Canasai et al. 2005) is yet another challenging task when it is done without any parsing knowledge. Byte sequence is the only magic key in our approach to determine the language of the input text. We introduced string kernel for this language identification task. We conducted experiments using 2 kernel classifiers i.e. centroid-based and support vector machine (SVM) methods. The accuracy of identification was acceptable for both methods. The accuracies reached 95 percent with only 10 training sets (2 KB per set). It was also found that the simple centroid-based classifier is comparable to the SVM classifier based on the string kernel.

Our approaches had been proven effective under the Thai language and the multi-lingual environment of 16 European and 4 Asian languages including Thai, Chinese, Japanese, Korean, English and many other European languages. We are expanding our corpus for conducting our experiments under the environment of a large number of languages.

Based on the successful results of word extraction, language identification and language independent indexing for search engine, we are conducting an experiment of the collaborative crawler on the high speed link (45 mbps) between Thailand and Japan. This collaborative work will provide an infrastructure for collecting web contents to study about the web language. The language together with its encoding of every webpage will be automatically identified and indexed to make the archive. Collaborative search engine will then go through all archives in all registered sites to present the ranked search results for any particular requests in any languages. The reports on the web languages from any perspectives can also be constructed by the proposed web language engineering.

Reference:

Virach Sornlertlamvanich. Implementations that Unify the Language Processing, Proceedings of the 9th NCSEC, University of Thai Chamber of Commerce, Bangkok, Thailand, pp. 1053-1062, 27-28 October, 2005.

Virach Sornlertlamvanich. Statistical-Based Approaches for Non-Segmenting Languages, Proceedings of Pacific Association for Computational Linguistics (PACLING), Meisei University, Tokyo, Japan, pp. 75-84, 24-27 August 2005.

Virach Sornlertlamvanich, Tanapong Potipiti and Thatsanee Charoenporn. Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm. Proceedings of the 18th International Conference on Computational Linguistics (COLING2000).

Virach Sornlertlamvanich, Pongtai Tarsaku, Prapass Srichaivattana, Thatsanee Charoenporn and Hitoshi Isahara. Dictionary-less Search Engine for the Collaborative Database, Proceedings of The Third International Symposium on Communications and Information Technologies (ISCIT-2003), Songkhla, Thailand, 3-5 September 2003.

Canasai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. Language Identification Based on String Kernels, Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT-2005), Beijing, China, October 12-14, 2005.