Thai Text Processing and Its Applications

Virach Sornlertlamvanich
Sirindhorn International Institute of Technology (SIIT),
Thammasat University, Thailand
virach@siit.tu.ac.th

Talking about the language, which has no explicit word and sentence delimiter, is extremely hard for processing on computer. The text flows as a stream of conversation with arbitrary pause of the speaker. The Thai language is one of the good examples of the languages, which need preprocessing of word and sentence segmentation in the early step of morphological analysis. The current research has reported that the accuracy of word segmentation can be achieved at the rate of higher 98% depending on the text set. However, it is still behind the accuracy of human judgement. By the way, beyond the problem of word segmentation, the needs of text processing in higher level become more significant in today's growth of data, especially in terms of text data, in the cyberspace.

The talk will be formalized into three fundamental problems in Thai text processing. Those are word segmentation, named entity recognition or keyword extraction, and semantic relation extraction. In the flood of information today, we spend most of the time to grasp the essence of the information rather than to enjoy the reading. Many approaches have been proposed to handle these fundamental issues, however, there is still much room for improvement. The introduced approach is not the best one, but it is aimed to make the problem well recognized. Mutual information and entropy are effective measures to uncover the possible word boundary for the non-segmenting languages such as the Thai language. It is remarkably to note that with the approach, the result has shown that about 30% of the extracted words are not defined in the Thai-Thai dictionary published by Thai Royal Institute in 1982. Keyword labeling is also a task that we can effectively apply a machine learning approach such as MIRA (Margin Infused Relaxed Algorithm) to capture the word context. This can be done on the result from the word segmentation task. Undoubtedly, the accuracy of the annotated tag is ranked from person (PER), date (DAT), location (LOC), and organization (ORG). This is because tag for person has the least ambiguity. The pattern for extracting the semantic relation between the type-annotated keywords is accordingly assigned to the word form of the disambiguated verb phrase. The experimental result shows that most of the distance between the keyword and the target verb phrase is not more than one word. Therefore, we can find the target verb phrase in the adjacent position or one word skipped position with the highest probability.

Based on the solution for the above NLP fundamental issues, many more tasks are made possible on the current viable Internet connection. The talk demonstrates the three constructive applications on the huge generated data i.e. linked data formation for knowledge map reasoning; keyword tracking on social media to understand the online social movement; and hyper local news publishing to fill in the information gap between urban and rural life.

The task of natural language processing today is not just only for the language itself any more, but it can bring along the possibilities on the advance of the Internet, big data, and machine learning technique.