Abstract

This research proposed an integrate system to mining person activities from Thai newspapers. The integrate system includes Named Entity Recognition, Activity Word Extraction and Individual Activity Formation. The Named Entity Recognition is used to extract the name of individual person. The Condition Random Fields (CRF) is applied as training and extracting techniques in this part. Swath word segmentation is utilized in Activity Word Extraction to extract the activity words. Activities of an individual are formed by utilizing named entities, activity words and specific domain corpus. The proposed formation rules are applied to the test articles to create the relations among individuals and activities. The system performance was evaluated by comparing the extraction results from 30 articles between this system and the sample of five persons as the standard. The experimental result of Named Entity Extraction by individual person activities is equaled F-score 84.68%.

Keywords: Named Entity Extraction, Named Entity, Lexicon, Conditional Random Fields, Text mining