SUMMARY OF Ph.D. DISSERTATION

School Science for Open and Environmental Systems	Student Identification Number	SURNAME, First name SANO, Tomohisa

Title

Property Estimation of Named Entities Based on Surface Statistics

Abstract

This research is for the proactive solution of the problem of unknown word processing in the field of natural language processing, by estimating properties of named entities according to their surface information. Embedded foreign named entities have been mostly treated as unknown words by dictionary-based approaches. Area estimation should be considered for named entities rather than language estimation to extract their properties. It is a task to estimate the area where a named entity may belong. This research concentrates on area estimation of toponyms as a part of toponym resolution task. Researches have been targeted on disambiguation based on gazetteers and document contexts, and they do not have enough flexibility and robustness. This research proposes an automatic area estimation approach regarding flexibility and robustness without gazetteers and/or contexts.

First, a basic approach has been proposed based on the survey regarding the capability of the surface statistics of toponyms. This approach succeeded in leveraging the surface data by considering the correlation between areas and languages, and it gained .74 in F-measure with 94.05% recall rate for a 100,000-toponym experiment containing 10 areas. Second, a new processing unit called a block has been proposed for a better precision rate. Third, the problem of multiple area candidates caused by the linguistic and/or historical relations among areas has been resolved. Finally, an area estimation system, which possesses both robustness by omitting heuristics and flexibility in linguistic relations among areas, is constructed by integrating these methodologies. This system succeeded in obtaining .91 F-measure by 88.30% precision rate and 94.03% recall rate with a 10-area experiment, and .82 F-measure with a 20-area experiment.

This research enables to provide useful properties to embedded foreign named entities to be meaningful words for latter processing, not cumbersome unknown words. This proactive approach can be expected to accelerate the researches such as machine translation and information retrieval and other researches regarding semantic understanding in the field of natural language processing.