| | | |
|---|---|---|
| 氏　　名 | Wattarujeekrit Tuangthong | |
| 学位（専攻分野） | 博士（情報学） | |
| 学位記番号 | 総研大甲第905号 | |
| 学位授与の日付 | 平成１７年９月３０日 | |
| 学位授与の要件 | 複合科学研究科　情報学専攻<br>学位規則第６条第１項該当 | |
| 学位論文題目 | Exploring Semantic roles for Named Entity Recognition in the Molecular biology domain | |
| 論文審査委員 | 主　査　助教授 | Nigel Collier |
| | 教授 | 佐藤　健 |
| | 教授 | 武田　英明 |
| | 教授 | 藤山　秋佐夫 |
| | 教授 | 神門　典子 |
| | 教授 | Asanee Kawtrakul<br>（Kasetsart University） |

# 論文内容の要旨

Named entity recognition (NER) in the molecular biology domain, the task of identifying and categorizing molecular entities appearing in text, is one of the most important tasks in a biological text mining engine. In general, this task is taken as the first step towards the more ambitious task of molecular event extraction (relation extraction) and, eventually, pathway discovery. However, NER in this scientific domain, which seems to be the easiest task among others in text mining, still achieves quite low performance. As can be seen from the most recent shared-task evaluations of NER in this domain (JNLPBA-2004), the best performance in terms of F1-score is only 72.6. This result is far below what is achieved by NER system in newswire domain (F1-score of about 96%) which is near the human level of performance. At present, most NER systems employ term internal features (e.g., lexical and morphology) and co-occurrence information as term external features. Due to the lack of molecular naming convention, which leads to the difficulty of terminological variations as well as the difficulty of polysemy (i.e. the sharing of names between different entities), such features are insufficient to handle the difficulties for NER in the molecular biology domain. To obtain a complete set of rules for lexical patterns of molecular names seem impossible, thus to use term external features other than co-occurrence information is of interest.

In this thesis, the semantic relationships between a predicate and its arguments in terms of semantic roles are proposed to enhance NER system in the molecular biology domain. The semantic role information is derived from a predicate-argument structure (PAS) which is a higher sentence representation level than syntactic relation and surface form levels. Thus, the use of semantic roles is more consistent than co-occurrence information derived from a surface level. To employ the semantic role for NER system, it is realized in various sets of syntactic features which were used by a machine learning model to explore the most efficient way in allowing this knowledge to provide the highest positive effect on the NER.

As a result, the best feature set composed of the 6 lexical features (i.e., *surface word*, *lemma form*, *orthographic feature*, *part-of-speech*, *phrase-chunk* and *head word of NP-chunk*) and 4 PAS-related features for representing an argument's semantic role (i.e., *predicate's surface form*, *predicate's lemma*, *voice* and the united feature of *subject-object head's lemma* and *transitive-intransitive sense*). Moreover, the use of semantic roles can show the positive effects for only the predicates conforming to the criteria as follows. A predicate must have its arguments as both *agent* and *theme* with a higher probability of belonging to a named entity class than non-named entity class; otherwise, a predicate must have its arguments as both *agent* and *theme* with a lower probability of belonging to a named entity class than non-named entity class and the number of training examples for this predicate should be large enough (by observing from empirical evidences, at least 270 sentences). The improvement in performance obtained from the NER system using PAS-related features, compared to not using these features, affirms that the using of semantic roles can enhance NER system.

論文の審査結果の要旨

Tuangthong Wattarujeekrit's research concerns the application of predicate argument structures (PAS) for supporting text mining in the field of molecular biology. The life sciences have seen an immense growth in the volume of peer-reviewed articles which is the primary and most timely record of experimental results. Since there is significant delay in updating databases with these results (e.g. BIND, KEGG, DIP, MINT etc.) the text mining community has responded to the challenge of extracting information from unstructured texts. However the transformation of text to knowledge is not simple and even low level tasks such as identifying and classifying terminology (the named entity recognition task) perform worse than expected from our experience of using shallow lexical and syntactic information in the newswire domain. Therefore it is reasonable to suppose that deeper types of knowledge are needed to help computers recognize terms, their properties and events in biology. In her thesis, Tuangthong Wattarujeekrit explored the hypothesis that predicate argument structures and in particular the semantic roles of arguments could be one of the deep knowledge sources that would help improve text mining and she applied this to the task of named entity recognition (NER) and explore what types of predicates could provide clues to the named entity classes of their arguments.

The thesis starts by presenting an outline of the tasks in biology text mining and motivates the use of predicate argument structures. It then outlines the central hypothesis of the research which is "Can the semantic information describing the relationships in terms of semantic roles between a predicate and its arguments enhance named entity recognition?'. Chapter 2 provides a survey of the tasks in biology text mining, their significance and cites the relevant literature. In particular the chapter focuses on the task of NER and analyses the difficulties that still remain to be solved. It then proceeds to outline current methods in NER and their empirical performance in gold-standard shared tasks such as JNLPBA, KDD cup and Bio-Creative 2004. The chapter then introduces the idea of predicate argument structures and the standards that exist in general language resources such as VerbNet, PropBank and FrameNet. It then argues that PAS frames are a natural candidate for supporting the task of NER in the biology domain.

Chapter 3 provides one of the two major contributions of the thesis which is the extension and development of a set of PropBank-style PAS frames for the biology domain and their realization as an online database. An analysis of the PAS structures from major biology journal articles revealed significant differences in numbers and types of arguments in the biology domain compared to general language. The work was published as an article in the journal BMC Bioinformatics.

Chapter 4 outlines the process by which selected arguments (agent and theme) from PAS frames can be mapped to surface sentences using syntactic relation features and establishes the second contribution of the research which is to show how PAS frames can be used to enhance the NER task in biology by mapping surface syntactic clues to the level of semantic roles. The model used is a state of the art support vector machine model which is trained on a named entity tagged corpus of MEDLINE abstracts enhanced with PAS frame information from a parser and mapping rules. Results show improved performance over a baseline lexical model for certain classes of predicates but lower performance for others. The classes of predicate behavior depend on the propensity of the semantic roles to take named entities as arguments. Tuangthong Wattarujeekrit then analyses the effects of parser error in a further series of experiments using a hand-parsed set of sentences and finds further evidence that PAS frames by themselves can benefit NER. The conclusion in chapter 5 is a natural consequence of the experiments, i.e. that NER can benefit to some extent from knowledge of semantic roles in PAS frames but only for certain classes of verbs and where parsing errors are reduced.