氏名	魏 琪 (Qi WEI)			
学位（専攻分野）	博士（情報学）			
学位記番号	総研大甲第1511号			
学位授与の日付	平成２４年３月２３日			
学位授与の要件	複合科学研究科 情報学専攻 学位規則第６条第１項該当			
学位論文題目	Classifying scientific texts in biology for focus species			
論文審査委員	主 査	准教授	COLLIER Nigel	
		教授	武田 英明	
		教授	山田 誠二	
		准教授	北本 朝展	
		准教授	古山 宣洋	
		准教授	Fabio Rinaldi	University of Zurich

論文内容の要旨

In recent years high throughput methods have led to a massive expansion in the free text literature on molecular biology. Automated text mining has developed as an application technology to organize this wealth of published results into structured database entries. Presently, there are more than 10,000 species and taking the marbled lungfish (Protopterus aethiopicus) as an example, there are 132.8 billion base pairs in this fish genome. In a typical systems biology abstract, there are 4-5 genes mentioned on average. Thus, recording and encoding them manually would take prohibitive amounts of time and human resources. Building intelligent tools to help authors and database curators integrate published results into databases has therefore become a major goal of research in biomedical natural language processing. However, the multiplicity of interpretations of meanings makes the specification of the author's intended meaning extremely challenging for automated natural language processing.

In this dissertation, the contribution is presented through a series of three experiments for identifying the focus species in biological papers as an aid to classifying and summarizing the experimental result. The focus species presents the author's major claim in reporting their own results. I present a new method to identify focus species with novel features in full-text papers and abstracts. I present a new knowledge model for species citations in biomedical papers. With this scheme, I developed a tool to provide authors and curators with a high-throughput method capable of determining the focus species in experimental papers. Unlike previous studies my approach does not consider target documents in isolation but makes use of a network of citation relationships, amplifying information which is implicit in the target document. The various features explored in the thesis questions are evaluated on gold standard data sets that have been constructed by external groups for community evaluation exercises.

In the experiments, 3 model organisms are classified in full papers selected based on the BioCreative 1b dataset and 4 model organisms are classified in abstracts selected from the DECA corpus. With three experiments, I showed a best F-score of 90.7% for classifying the full papers by using internal features. I also showed that when only using internal features, full papers perform much better than abstracts. By using external features from related publications, I demonstrated a best F-score of 91.14% for classifying abstracts. Finally I developed a new typed citation scheme and showed that among the four citation classes of background, method, results and data, the strongest relation for aiding the focus species classification was the one relating author results to the target paper.

The thesis explores the general question "What features are most effective for resolving conflicting evidence about focus organism in biomedical abstract and full text?" Since the question is potentially open-ended, I break this down into three specific sub-questions.

1. What level of classification performance is achievable using state-of-the-art lexical semantic features for focus species in full papers and abstracts?
2. Of the abstracts which are cited or archived in the PubMed database, do bibliographic features provide enhanced classification accuracy?
3. Of the abstracts which are cited does a typed citation function provide enhanced classification accuracy? Also what citation types prove the most useful?

This Ph.D. dissertation presents a method for identifying the focus species of full-text papers and abstracts and a new citation scheme for biomedical papers. This dissertation consists of seven chapters. Chapter 1 gives the introduction, and Chapter 2 presents the related work. Chapter 3 describes the first experiment on focus species classification for full-text papers. Chapters 4 describes the second experiment on focus species classification for abstracts. Chapter 5 discusses the new citation scheme for biomedical papers and its application to focus species classification. And chapter 6 discusses the difficult cases for the task and online tools. Chapter 7 concludes this dissertation and discusses future work.

There is one set of experiments for each thesis question. Hypothesis 1 is explored in a series of experiments in chapter 3. Based on the findings of this experiment which showed the relative merits of various in document lexical semantic features, I conducted Hypothesis 2 experiments which are reported in chapter 4. Based on the findings of experiments in chapter 4 that showed the effectiveness of bibliographic features, I conducted Hypothesis 3 experiments which are reported in chapter 5.

博士論文の審査結果の要旨

Wei Qi gave a 45 minute presentation to the examination committee in English. The major points of her presentation were as follows:

- Understanding the focus species of a scientific text in molecular biology is an important subtask in classifying texts by lifescience database curators. With thousands of model organisms, tens of thousands of genes, wide-scale ambiguity and multiple gene mentions in each text this is a highly challenging task. Therefore automated classification of texts for the focus species is necessary in this domain.
- Qi's research aims to find ways to assess and suggest effective features for classifying biological texts for their focal species. She has compared linguistic features, text types (abstracts and full texts) and suggested a novel citation scheme to harness the information present in citation networks.
- In the first part of the method Qi presented a variety of 8 supervised learning algorithms (Naïve Bayes, Conditional Random Fields, AdaBoost, Bagging, Decision Table, Decision Tree, Logistic Regression, Support Vector Machine) to assess the contribution of various linguistic and domain features in molecular biology full texts using a gold standard data set derived from the BioCreative I task. The features used included word level features such as the gene names, curator assigned keywords (MeSH headings) as well as novel combinations of features such as journal title, gene synonyms, organism frequency, gene-species relations etc.
- The second part of the method aimed to focus on the use of external knowledge sources through direct citations and the PubMed search engine. Qi tested the assumption that the linked publications implicitly discuss the same model organism. Qi also looked at the contribution made by the gene name detection task to this process, breaking down the performance by 5 model organisms.
- Based on the results of the second method Qi proposed in the third method to create a novel typology of citation functions based on a study of earlier work by Weinstock, Oppenheim and Penn, and Teufel et al. This novel scheme was then implemented using heuristic rules and evaluated for classification performance. The citation classification was then applied to classifying focus species using papers that cite the target paper and performance compared.
- Experiments on Method 1 showed high levels of F-score performance across 3 model organisms for Naïve Bayes and that full texts provided higher levels of performance compared to abstracts alone. A combination of all features overall provided the highest level of F-score (10% higher F-score performance than a basic set of word level features) for 2 model organisms (fly and mouse) but yeast showed better performance without organism frequency, additional gene name

term and intra-sentential term-species relations.

- Experiments on Method 2 showed that bibliographic features from related papers found through a search engine provided improved levels of F-score performance. When comparing untyped citing papers and related papers the related papers proved to be substantially more informative for the classification model. Drill down analysis showed that citing papers were on average substantially later than the associated papers (6 years compared to 3 years). The experiments also showed the importance of learner selection for gene name tagging with CRFs providing state of the art performance for all species except the least represented one in the data set.
- Method 3 developed a 4 class citation scheme consisting of Background, Data, Experiment and Result classes. Experiments on classifying citations in citing texts using this schema yielded high levels of F-score (0.94 to 0.96). By applying this classification of citations and then using them as features in the focal species classifier Qi showed the effectiveness of different types of citations function. The most important citation function appears to be the Result class followed by Methods, Data and Background.

There followed an open and closed question and answer session at which the committee members asked a number of questions:
(a) The committee asked why some organisms are more difficult to classify than others;
(b) The committee requested clarification about the contribution of the three sets of experiments to the overall thesis;
(c) Clarification was asked about how generalizable the conclusions are to the general case of thousands of species;
(d) The committee asked why earlier works were not used as part of Experiment 3;

The committee members expressed their satisfaction with the quality of the answers, the work reported in the thesis, its novelty and the overall approach. The committee recognized the original contribution made by the research in the area of automated classification and knowledge modeling as well as the achievements reported in the international journal publication in BMC Research Notes.
The committee members all agreed that Qi Wei had successfully passed the thesis examination.