学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　Moustafa Dieb, Thaer

学 位 論 文 題 名

Framework for Experimental Information Extraction from Research Papers to Support Nanocrystal Device Development

（ナノ結晶デバイス開発支援のための論文からの実験情報抽出フレームワーク）

Nanocrystal device development is a nanoscale research domain, where researchers produce nanocrystals for electronic and optoelectronic devices (e.g., in solar cells, light-emitting devices, and memory component). This process requires both engineering knowledge and craftsmanship skills. Since there is no well-systematized process to develop new nanocrystal devices, researchers have to conduct several experiments before reaching the appropriate manufacturing process to produce the desired output. In order to support this process, analysis of development experiments' results is necessary. Such analysis can provide insights on experiment planning leading to a quicker and less costly development process. In this study, we discuss our approach to extract experimental information related to nanocrystal devices from research papers using machine-learning techniques based on an annotated corpus approach. We defined the necessary information and designed an annotation guideline in collaboration with a domain expert. We checked the reliability of this guideline through corpus construction experiments with graduate students of this domain, and then evaluated the corpus with a domain expert. The finalized corpus called "NaDev" (Nanocrystal Device Development corpus) then has been used to build an automatic information extraction system called "NaDevEx" (Nanocrystal Device Automatic Information Extraction Framework) to automatically extract the desired information from research papers on nanocrystal devices using machine learning and natural language processing techniques.

This thesis is divided into 6 chapters. Chapter 1 introduces the nanocrystal device development process and experiments, and discusses the motivation of the study. Chapter 2 overviews the efforts in nanoinformatics, where information technology is used to support nanoscale research. This chapter discusses other efforts for extracting information from nanoscale research papers. We also review the information extraction from research papers in bioinformatics. In Chapter 3, we discuss in detail our methodology to construct the annotated corpus (NaDev). A tag set was designed in collaboration with a domain expert to annotate the desired information categories such as source material information, experimental parameters, evaluation parameters, final product, and so on. Preliminary annotation experiments were conducted with two graduate students of nanocrystal device development domain; the results of these experiments were used to build a corpus construction guideline that contains detailed definition of the desired information categories and how to annotate them with several real examples to avoid mismatches between different annotators. The reliability of this guideline was checked with corpus construction experiments using inter-annotator agreement (IAA) between two different annotators. Even though the corpus construction guideline reached a reliable level with loose agreement (where two entities agrees on information categories but disagree on the boundary, in many cases we can find

appropriate head nouns in loose matching terms), it was necessary to evaluate this corpus and finalize it with a domain expert to ensure reliability. The corpus was finalized as NaDev corpus, which includes 392 sentences, and 2870 terms annotated using eight information categories. In chapter 4, we discuss the development of the automatic information extraction framework (NaDevEx) using machine-learning techniques. Since entities from different information categories are overlapped within each other in the nanocrystal device development domain, we use a step-by-step (cascading style) information extraction system. In each step, NaDevEx extracts a group of information categories that do not overlap within each other using tagging results from previous steps as clues for information extraction. We found that, for the information category with rich domain knowledge information (source material); the system performance is almost not defeated by that of human annotators. NaDevEx also uses domain knowledge features like chemical entity recognition, and physical quantities list to support extraction of material information and parameter information respectively. The evaluation of NaDevEx using NaDev corpus is also discussed in detail regarding comparison with human annotators, paper type effect on the system performance, and domain knowledge features effect. Since there is a considerable amount of chemical entities exists in research papers related to nanocrystal devices, chemical named entity recognition is supportive for NaDevEx. We discuss in further detail a chemical named entity recognition system using ensemble-learning approach. In chapter 5, we present our preliminary efforts to utilize the information extracted to support nanocrystal device development. Finally, chapter 6 concludes the study and discusses future work.