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

Thaer M. Dieb

Graduate School of Information Science and Technology
Hokkaido University

November 2015

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Thaer M. Dieb
November 2015

Acknowledgements

I would like to express my deepest gratitude to my PhD advisor, Professor Masaharu Yoshioka. Prof. Yoshioka always provided insightful discussions about the research to deepen my reasoning behind research results. Prof. Yoshioka extensive, immediate, and patient support, not only in academic but also in personal issues related to my life in Japan is invaluable. In addition, he was kind and patient enough to answer my questions on Japan culture. Without his guidance and patience, this dissertation would not have been possible.

My gratitude is extended to Professors Makoto Haraguchi, and Yoshiaki Okubo for their valuable discussions to improve my research. Prof. Haraguchi comments were always challenging to have a deeper understanding on my research. Prof. Okubo support for lab activities and working environment makes it possible to conduct my research.

In the same manner, I would like to deeply thank Professor Hiroki Arimura for his discussion on the future of my research and academic career.

Similarly, I would like to extend my deep thanks to Professor Shinjiro Hara for his cooperation to conduct this interdisciplinary research. Without his support, this project would not have seen the light.

In addition, I am very thankful to Dr. Marcus Newton of Southampton University, U.K for his cooperation on developing the corpus.

I would like to thank all members of knowledge base lab, present and past ones for being friendly during my stay here.

Special acknowledgment to members of Research Center for Integrated Quantum Electronics for their contribution in constructing the corpus.

Additionally, a special acknowledgment to the Japanese government for funding my research and life in Japan via the MEXT scholarship.

Dedication

To my country on its way to freedom and dignity.

To my family for their unconditional support.

A special feeling of gratitude to my mother Ayda, my sister Rana, and my uncle Khaled for their support in the difficult times.

Deep gratitude for those few around me who provided comfort along the way till the finish line.

Abstract

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.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background and motivation

Nanocrystal device development is an area of nanoscale research where nanoelectronic devices are developed for future nanoelectronic industry applications using electronic materials, such as semiconducting, insulating, and magnetic materials [1–6]. This development process is not well systematized, and requires both engineering knowledge and craftsmanship skills [7]. Researchers have to conduct several experiments before reaching the appropriate manufacturing process to produce the desired output. Skilled engineers can make the development process more efficient by well planning of the manufacturing experiments. However, knowledge about this planning is difficult to transfer from skilled engineers to novices.

In order to support this process, analysis of experiments' results is necessary. Domain researchers recommended using related research publications as a source to extract experiment-related information. These publications usually include detailed discussion about experiments including motivation and evaluation criteria. We propose a framework to exploit experimental information reported in research publications on the development of nanocrystal devices using machine-learning and natural language processing techniques based on an annotated corpus approach. This is a joint research project between the Research Center for Integrated Quantum Electronics (RCIQE) and the Division of Computer Science at Hokkaido University. This interdisciplinary research, where information technology is used to support nanoscale research is associated with a newly emerging domain known as nanoinformatics [8, 9].

Information extraction from research publications approach has several advantages. It can utilize the freshness and massive availability of information in research publications, thus facilitate collaboration among researchers in the areas of nanocrystal device development, computer science, and natural language processing, which can overcome problems related

to the excess of information in the nanotechnology domain. This Information can be used -for example- to find similarities between previous experiments and planned experiments for a more effective experiments' design. A well-defined corpus is essential to support this information extraction process.

In this chapter, we overview the nanocrystal device development experiment, and propose our approach to support this process.

1.2 Nanocrystal device development experiments

In RCIQE, researchers are developing various kinds of nanodevices using selective-area metal-organic vapor phase epitaxy (SA-MOVPE) method. SA-MOVPE is a chemical vapor deposition method of epitaxial growth of materials, especially compound semiconductors from the surface reaction of organic compounds or metalorganics and metal hydrides containing the required chemical elements. Figure 1.1 shows an illustration of SA-MOVPE. Even

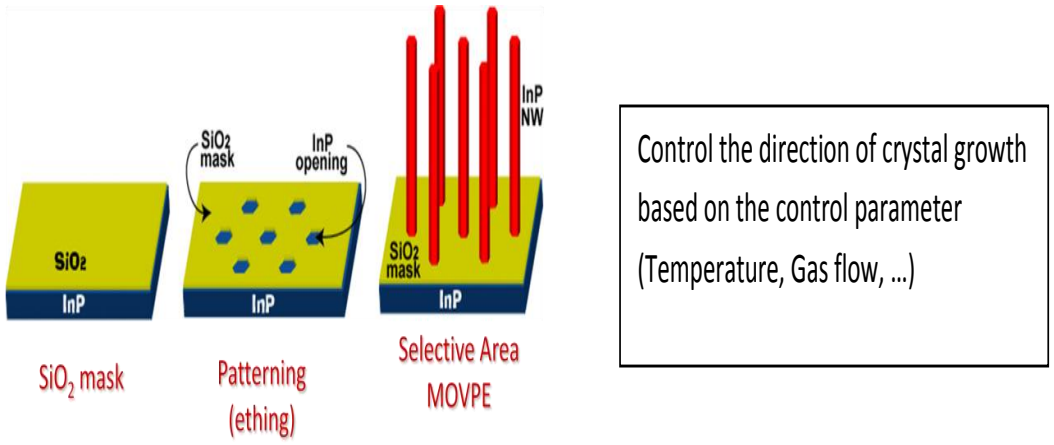


Figure provided by Prof. FUKUI Takashi (RCIQE, Hokkaido University)

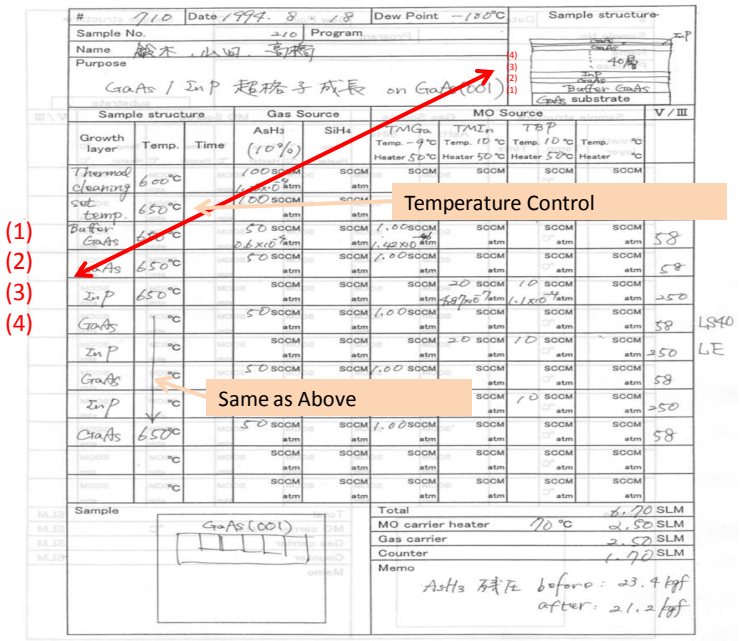
Fig. 1.1 Selective-area metal-organic vapor phase epitaxy (SA-MOVPE)

though SA-MOVPE is a good method to control the quality of the device, researchers still have to go through many trial-and-error experiments to arrive at the final process. Researchers use the SA-MOVPE growth parameter record sheet to keep record of each experiment. These sheets have the following types of information.

- Background information: ID, Date, name of the experimenter, Purpose

- Growth layers information : Growth layers with parameter settings used to control operation for each growth layer (Gas source, metal organic, gas temperature, pressure, mixture ...).
- Memo...

Figure 1.2 shows an example of experiment record sheet.



Values described in this sheet are compiled only for demonstration purpose, and are not used in real experiments

Fig. 1.2 MOVPE growth parameter record sheet

After series of experiments, research publications are written by domain researcher. These publications contain complete descriptions of the motivation, purpose, and other related experimental information.

1.3 Contribution of the thesis

The major contribution of this thesis is to provide framework to extract experimental information in nanocrystal device development related papers based on an annotated corpus approach. We can divide our contribution as follows:

- Construction of the corpus: We designed a tag set in collaboration with a domain expert. We constructed the annotated corpus with domain graduate students [10, 11].

- Development of an automatic information extraction framework based the constructed corpus using machine-learning and natural language processing techniques [12–14].
- Propose a method to utilize the extracted information to cluster research papers based on different similarity metrics. This method will provide a quicker access for researcher to relevant information when planning their experiments [15].
- Since there is a considerable amount of chemical information in nanocrystal device related publication represented as experiment source materials, we develop a chemical entity recognition system based on a ensemble-learning approach to support the extraction of source material information [16, 17].

1.4 Thesis organization

This thesis has additional five chapters. In Chapter 2, we overview the related work in nanoinformatics. We also discuss efforts to utilize information in research papers including those in bioinformatics domain. Chapter 3 presents our approach to construct an annotated corpus of research papers related to nanocrystal device development to support data collection process in this domain. In Chapter 4, we discuss the utilization of our constructed corpus to develop an automatic information extraction framework to automatically extract the desired information categories from research papers using machine learning and natural language processing techniques. Since there is considerable amount of chemical information in nanocrystal device development research papers, we also discuss the development of an automatic chemical information extraction framework using ensemble-learning approach. Chapter 5 introduces our preliminary work to utilize extracted information to support the nanocrystal development process. Finally, Chapter 6 concludes this study and proposes future development.

Chapter 2

Related works

2.1 Introduction

In this chapter, we discuss the related works dividing them into two main research streams: first, we review the efforts in utilization of research papers' information using text mining in different domains, specially, in bioinformatics, where there are well-established projects such as GENIA project [18]. Second, we review the efforts in the nanoinformatics domain, including collection of experimental data.

2.2 Utilization of research papers' information using text mining in different domains

2.2.1 Overview

A large number of research papers is being available that contain massive amount of information written by domain specialists in different domains. The necessity to use the freshness and availably of such information is growing to help reduce information overload on researchers. Several efforts have been conducted to achieve this purpose using text-mining techniques. These efforts can be mainly categorized into 2 categories: dictionary-based, and machine-learning based systems. Due to the large number of varieties of information in research papers, the dictionary-based systems are not efficient enough. However, for the machine-learning based systems, it is necessary to have a corpus of research papers annotated with the desired information. For new domains, where well-defined corpora do not yet exist, the construction of such corpora is crucial. Several attempts have been adopted in different domains. In bioinformatics for example, researchers can build large-scale corpora using

text-mining approaches to support research in the field of molecular biology. GENIA corpus [19] was the first attempt for constructing large corpus to overcome the bottleneck problem for applying NLP techniques in biological domain. GENIA corpus version 3.0 consists of 2000 MEDLINE abstracts with more than 400,000 words and almost 100,000 annotations for biological terms (more details of the corpus will be explained in Section 2.2.2). By using this corpus, many researchers from NLP can participate in the research on automatic information extraction from research papers in biological domain [20–22], and as a result, several new tools and techniques were developed in these tasks [23, 24]. The most common approach in this task is modeling the information extraction task as a sequence labeling task on the morphological analysis results. In this approach, the system breaks one sentence into a sequence of morpheme. After that, the system tries to identify positions where target terms start and end by using Support Vector Machine (SVM) [25] or Conditional Random Field (CRF) [26] as a machine-learning system. In this system, researchers use several features that include linguistic feature and domain knowledge related features. They also proposed a new framework to utilize the extracted information. For example, one of the important utilization of the information extracted from the papers is protein-protein interaction information. By integrating such fragment interactions, they proposed a framework to identify the pathway that represents the biological systems [27]. This is one of the good applications for utilizing extracting information.

Based on the success in biological domain, there were several attempts for constructing corpora in other domains. Corpus construction for chemical named entity recognition is one of the well-known examples for this process. At first, SCAI corpus [28] that includes 100 abstracts for recognizing chemical named entities described in IUPAC [29] (International Union of Pure and Applied Chemistry) style was constructed. Several tools [30, 31] were developed based on this corpus. Most of the machine-learning based system used similar techniques to those in the biological domain. However, since those tools were not good enough to extract chemical related information in biological/biomedical domain, CHEMD-NER corpus [32] that contains 10,000 abstracts for recognizing chemical and drug named entities was developed, and different approaches compete to extract chemical entities and drug names automatically based on this corpus [33].

There are also several projects in nanoinformatics domain, those projects are discussed in Section 2.3.2 in detail.

2.2.2 GENIA Corpus development

The GENIA corpus was created to support the development and evaluation of information extraction and text-mining systems in the domain of molecular biology. GENIA employs multilayer annotation, which encompasses both syntactic and semantic annotation, as follows:

- Part-of-speech (POS) annotation: In general, GENIA POS annotation follows the Penn Treebank POS tagging scheme.
- Constituency (phrase structure) syntactic annotation
- Term annotation: This refers to the identification of linguistic expressions that relate to entities of interest in molecular biology, such as proteins, genes, and cells [34].
- Event annotation: GENIA corpus event annotation marks expressions stating biomedical events, or changes in the states or properties of physical entities. Event annotations are text-based associations of arbitrary numbers of entities in specific roles (e.g., a theme or a cause) [35].
- Relation annotation: GENIA corpus relation annotation aims to complement event annotation in a corpus by capturing (primarily) static relations, i.e., relations between entities, such as "part of," that do not necessarily involve changes.
- Co-reference annotation: This refers to identifying expressions in texts that relate to the same thing.

The GENIA term corpus is available in an XML format, which is described in the GENIA corpus manual.

During the construction of the GENIA corpus, several problems had to be overcome that originated from the nature of biomedical research abstracts. Unlike everyday English text, the research abstracts used in the molecular biology domain include the following items:

- Nonproper names and abbreviations that begin with capital letters.
- Chemical and numeric expressions that include nonalphanumeric characters such as commas, parentheses, and hyphens.
- Participles of unfamiliar verbs that describe domain-specific events.
- Fragments of words, especially names and abbreviations, that begin with capital letters (e.g., NFAT, CD4, and RelB), which makes it difficult to distinguish between proper nouns and common nouns.

2.3 Nanoinformatics

2.3.1 Nanoinformatics roadmap 2020

Nanoinformatics is gaining more attention recently because of the diverse application domains of nanotechnology, and the need to get use of the massive information available. Nanoinformatics can be defined as the science and practice of determining which information is relevant to the nanoscale science and engineering community, and then developing and implementing effective mechanisms for collecting, validating, storing, sharing, analyzing, modeling, and applying this information [36]. Alternatively, nanoinformatics can be defined as an emerging area of information technology at the intersection of bioinformatics, computational chemistry, and nanobiotechnology [37]. Nanoinformatics would allow researchers to leverage the findings of other efforts in support of their own investigations and to broaden the impact of their research. For example, using mapping, visualization, and advanced analytical tools, a researcher may uncover important information, which points research in new directions. Such cyber-enabled discoveries can quickly advance the exploration and application of systems too complex to be understood solely from first-principles science. Nanoinformatics could play the same role in nanotechnology and nanomedicine as bioinformatics and medical informatics in biology and medicine [38].

There have been several attempts to learn how informatics is used to advance nanomanufacturing. For example, The Greener Nano 2012 (GN12): Nanoinformatics Tools and Resources Workshop [39] aimed at establishing a better understanding of state-of-the-art approaches to nanoinformatics and clearly define immediate and projected informatics infrastructure needs for the nanotechnology community. De la Iglesia et al. also discuss the needs and challenges, as well as the extant initiatives and international efforts in the field [9]. One of the very important initiatives to roadmap the nanoinformatics domain was Nanoinformatics 2010 [36], a collaborative road mapping and workshop project at which informatics experts, nanotechnology researchers, and other stakeholders and potential contributors collaborated to develop a roadmap for the area.

There are three main research themes in nanoinformatics:

- Data collection and curation
- Tools for innovation, analysis and simulations
- Data accessibility and information sharing.

Data collection process is considered a very essential step towards developing computational frameworks to utilize information in nanoinformatics domain. Some researchers have

focused on assembling fundamental knowledge related to the development of nanodevices to support nanotechnology research. For example, Kozaki et al. systematized fundamental nanotechnology knowledge through ontology engineering [40] to fill the gap between materials and devices by establishing common concepts across various domains. They also aimed to build a creative design support system using systematized knowledge. Another approach aimed at developing a NanoParticle Ontology (NPO) to represent knowledge underlying the preparation, chemical composition, and characterization of nanomaterials involved in cancer research [41]. Several other approaches have been conducted to manage and share data related to nanoscale, including construction of databases of nanomaterials [42], and setting up portals for sharing useful information [43–45]. Other researchers are working on the DaNa project [46] to provide information on products and applications of nanomaterials, and illuminate health and environmental aspects. Based on the DaNa project, researchers are trying to capture knowledge on a semantically higher level in a database called DaNaVis to increase the accessibility of the DaNa project results by means of interactive visualization components [47]. The major focus of such projects are applications related to health and environment.

2.3.2 Extraction of research paper information in nanoinformatics domain

The use of literature in the nanotechnology domain is still in its early stage. Few efforts have been conducted; however, they focus on the study of nanoparticles and nanomaterials and their potential use and side effects in medical applications. For example, Gaheen et al. are working on a data-sharing portal called caNanoLab, which provides access to experimental and literature-curated data from the NCI Nanotechnology Characterization Laboratory, the Alliance and the greater cancer nanotechnology community [48]. This portal offers information related mainly to the biomedicine domain. Some researchers try to extract information from full-text nanotoxicity related publications [49]. García-Remesal et al. developed a method for the automatic identification of relevant nanotoxicology entities in published studies using a text-mining approach, and they constructed a corpus for this purpose [38]. Jones et al., using a natural language-processing technique, tried to extract numeric values of the biomedical property terms of poly (amidoamine) dendrimers from the nanomedicine literature [50]. However, nanomaterials can be used in other domains such as nanoelectronics; hence, the need for general knowledge about nanodevice development experiments is growing, and these efforts are not sufficient.

2.3.3 Information collection from experimental record sheets

Since it is expensive to conduct new experiments in nanotechnology to obtain new experimental data, it is desirable to collect and share such information. One possible resource could be used to obtain experimental information is the experimental record sheets as in 1.2.

Yoshioka et. al. have been conducting the project "Knowledge exploratory project for nanodevice design and manufacturing" to collect data from experiment record sheets related to nanocrystal device development [51]. They have implemented a prototype for the SAMOVPE experiment record management system. This system stores each sheet as an XML semi structured data, and use structured queries based on the XML data. We constructed a database of real experiment records from 2005 to 2008, and provide the system with the following supporting functions.

- Data record retrieval with structured query (e.g., name, layer structure)
- Frequent pattern mining for understanding the parameter commonly used...

Based on the analysis of frequent pattern mining results from that system, researchers found that different sets of parameters are used to make same layer structure. Figure 2.1 shows the frequency of parameter settings used for making different layers. For example magnesium arsenide (MnAs) layers represented by filled circles, can be produced at different temperature and gas flow rate.

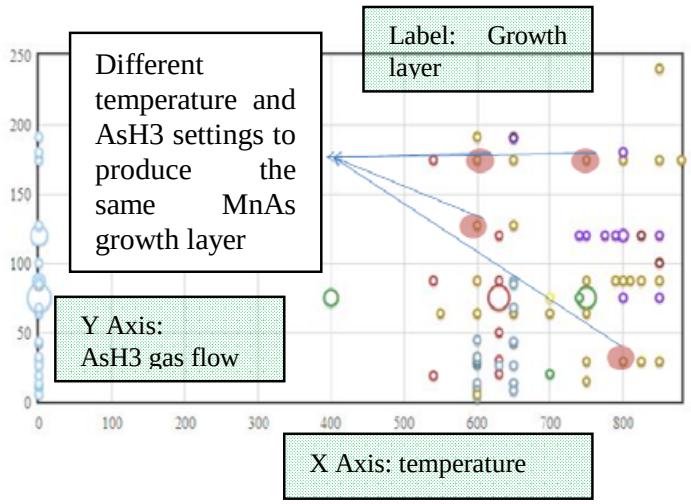


Fig. 2.1 Different parameter settings for making same layers

Novice engineers may have difficulties on selecting the appropriate parameter settings for their task. On the contrary, skilled engineers can understand the difference of the parameter

settings, and select the settings used in the experiments for similar cases. For supporting such a selection process, information described in the sheets is not enough for a detailed analysis, and it is necessary to collect related information (such as purpose, evaluation criteria) from other resources.

One way to obtain the necessary additional information is the research notes related to experiments. However, this approach would require extra work by the nanocrystal researchers, who might not be available at the time of the project (e.g., graduate students who completed their courses). Furthermore, the research notes related to experiments might not include comprehensive information about a series of related experiments, such as the evaluation criteria used and background information.

2.4 Summary

As discussed in Section 2.2, in order to utilize research papers' information, it is crucial to have a well-defined corpus that represents the target information to be extracted. Even though it is not so easy to construct good corpora, those corpora attract NLP researchers to develop new information extraction tools. In addition, from the experience in the chemical named entity recognition task, it is not necessary to start with a large size corpus (more than 1,000 abstracts). It is important to construct certain amount of the corpus and develop tools for extracting such information to attract other researchers to expand this research activity.

In the nanoinformatics domain, there are very few attempts to construct corpora related to nanotechnology. However, these efforts do not aim to extract comprehensive experimental information (e.g., parameters and their values), such information is useful to analyze information in the experiential record sheets and research notes. This information is also useful to find out papers that contain information to be shared in portals related to the nanoinformatics domain.

Chapter 3

NaDev corpus: An annotated corpus for nanocrystal device research papers

3.1 Introduction

As discussed in Chapter 2, to the best of our knowledge, there is no well-established corpus that tries to annotate comprehensive experimental information from research papers related to nanotechnology. These types of information are useful to support experimental results analysis of nanocrystal device development and find out papers that contains information to be shared at portals related the nanoinformatics domain. In this chapter, we discuss the development of a method for constructing an annotated corpus of publications related to nanocrystal device development to support automatic information extraction. This is a first step to attract other researchers to expand this research activity.

The corpus-construction guideline was designed in collaboration with a domain expert. We evaluated the reliability of this guideline through corpus construction experiments with graduate course students in this domain. We evaluated the constructed corpus using Inter-Annotator Agreement (IAA) and confirmed the guideline achieved a satisfactory level of IAA. We also constructed an agreement corpus that excludes wrong annotation based on the misunderstanding of the guideline. A domain expert evaluated this agreement corpus and modified the guideline by checking the real annotation example. Based on this modified guideline, he finalized the corpus called "NaDev" (Nanocrystal Device Development corpus) and its construction guideline for the official release.

3.2 Corpus construction process

3.2.1 Tag Set Design

To extract information from research publications, it is necessary to identify the information categories and to understand why these categories are needed to analyze the experiments. We conducted interviews with researchers in the field of nanocrystal devices at RCIQE, Hokkaido University. In collaboration with these researchers, we built an abstract model for experiments in nanocrystal device development. Figure 3.1 shows the experimental abstract model.

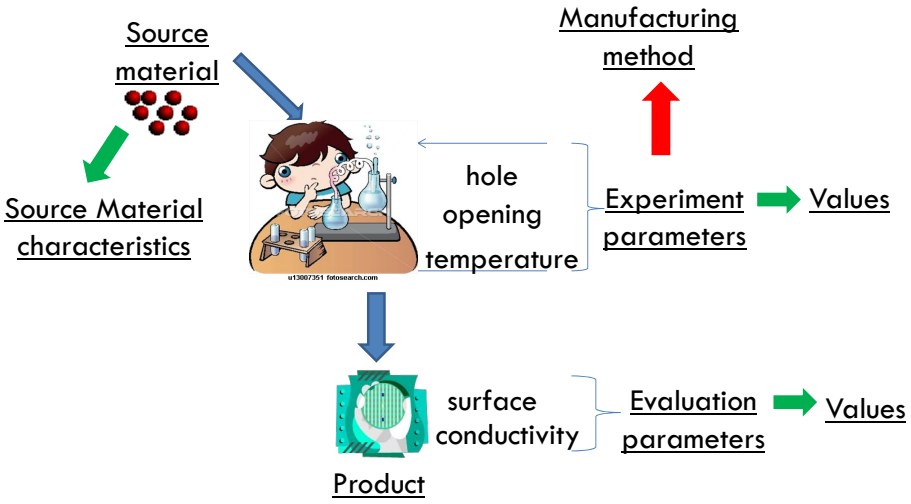


Fig. 3.1 Information categories used in nanocrystal device development experiments

In experiments, researchers usually employ source materials such as gases or MnAs, where each source material has specific characteristics; e.g., the distinctive group of that material in the Periodic Table. The experimental conditions can be controlled by adjusting specific experimental parameters such as the temperature and pressure. However, different development methods may use different sets of experimental parameters, so a set of parameters may be relevant only to a particular development method. An experiment yields a final product; i.e., the target artifact. To evaluate the success of an experiment, it is important to understand the type of device for which the target product is designed. Thus, researchers use evaluation criteria to evaluate the suitability of the final product based on its intended purpose, such as the smoothness of a semiconducting nanocrystal surface or its electrical conductivity. These evaluation criteria are measured using relative values.

Based on discussions with the researchers in the nanocrystal device field, we developed a candidate tag set for annotating research publications, which categorizes the information in the experimental abstract as follows:

- Source material (SMaterial): Source material employed in the experiment, such as As or InGaAs.
- Source material characteristic feature (SMChar): Characteristic feature of the source materials, such as (111) B, hexagonal.
- Experimental parameter (ExP): Control parameter for adjusting experimental conditions, such as diameter or total pressure.
- Experimental parameter value (ExPVal): Value of an experimental parameter, such as 50 nm or 10 atoms.
- Evaluation parameter (EvP): Parameter that is used to evaluate the output of the experiment, such as peak energy.
- Evaluation parameter value (EvPVal): Value of an evaluation parameter, such as 1.22 eV.
- Manufacturing method (MMethod): Method used in the experiment to achieve the desired product, such as selective-area metalorganic vapor-phase epitaxy.
- Target artifact or final product (TArtifact): Final output of the experiment, such as semiconductor nanowires.

3.2.2 Construction guideline

Before we constructed the corpus, it was necessary to specify the corpus construction guideline. To construct these guideline, we asked two graduate students from RCIQE to annotate the same publication [52] independently. Next, we compared both sets of annotations and discussed the disparities. Based on this discussion, we prepared a first draft of the corpus construction guideline for annotating research publications. This draft has been progressively improving as more papers were annotated. Additionally, the guideline was checked by an expert researcher in nanocrystal device development. The annotation was performed by assigning different colors to the information categories that we wanted to extract.

Computer scientists might find it difficult to define clearly what needs to be extracted and the method of extraction, because of a lack of experience in the nanotechnology domain.

This means that annotators might interpret and annotate the same text in a different manner. Thus, it was necessary to check the reliability of the corpus construction guideline.

3.2.3 Reliability measures

To evaluate the quality of the corpus construction guideline, we used reliability to represent the accuracy of the annotated information; i.e., the likelihood of extracting all of the requisite information. Thus, reliability represents consistency in this case. We checked the reliability of the corpus using the IAA between two different annotators based on the kappa coefficient [53]. The kappa coefficient is a robust measure because it takes into consideration the agreement occurring by chance. For more information on calculating IAA, please refer to Appendix B.

However, the annotation of a text using the proposed tag set requires some consideration of the term boundary mismatch problem. Thus, to separate the issues of term category selection and term boundary identification, we developed two different evaluation criteria for the analysis. The first criterion is tight agreement, which considers the term boundary, whereas the other is loose agreement, which ignores the term boundary problem. Figure 3.2 illustrates the difference between tight and loose agreement in a corpus sample.

3.2.4 Corpus construction experiments

We asked the same two graduate students to annotate the same publication independently [54] according to the guideline, and we calculated the IAA using the kappa coefficient. The annotation was performed manually by highlighting each information category with the corresponding color. The kappa coefficient was 41% in case of tight agreement, and 74% in case of loose agreement. According to Green (1997) [55], high agreement (i.e., sufficiently reliable agreement) requires a kappa coefficient of ≥ 0.75. The results of the first experiment showed that the annotation was sufficiently reliable for cases with loose agreement, but inadequate for cases with tight agreement.

Two types of mismatches were observed: term category and term boundary mismatches. Fewer problems were related to term category mismatches, and most of these were mismatches between SMChar and TArtifact. This was because the characteristics of the source materials were also the characteristics of the final product in some cases, so the annotators confused these two categories. For the term boundary mismatches, most of the common errors occurred in the EvPVal and ExP categories. Figure 3.3 shows examples of term boundary mismatches that occurred between the two annotators in the first experiment.

Based on these results, we revised the guideline and conducted a second annotation experiment using four research papers [56–59]. In this experiment, the corpus-annotation support tool XConc Suite [60], which was originally developed for constructing the GENIA corpus [19], was used for the annotation. We asked two graduate students (different from the first experiment) to annotate these papers independently, and evaluated the annotation results using IAA. In this experiment, the IAA was 0.63 for tight agreement and 0.77 for loose agreement. Table 3.1 and table 3.2 show the experimental results for the tight and loose agreement ratios, respectively.

Table 3.1 Tight agreement ratio, kappa coefficient = 0.63

	SM	SMC	EP	EPV	Ev	EvV	MM	TA	O	T
SM	95	1								96
SMC		32						4	15	51
EP			24						3	27
EPV		1		14					6	21
Ev					38	2			18	58
EvV						25			17	42
MM							18	1		19
TA							3	45	6	54
O		23	4	6	9	14	1	5		62
T	95	57	28	20	47	41	22	55	65	430

SM: SMaterial, SMC: SMChar, EP: ExP, EPV: ExPVal, Ev: EvP, EvV: EvPVal, MM: MMethod, and TA: TArtifact are from the tag set. O: Other class of unannotated text (or terms with boundary mismatches that prevent tight agreement). T: Total.

Table 3.2 Loose agreement ratio, kappa coefficient = 0.77

	SM	SMC	EP	EPV	Ev	EvV	MM	TA	O	T
SM	95	1								96
SMC		44				4		6	6	60
EP		1	27						3	31
EPV		1		18					2	21
Ev		2			40	6			12	60
EvV		1				36			5	42
MM							18	1		19
TA		5					3	47	4	59
O		3	1	2	6	3	1	1		17
T	95	58	28	20	46	49	22	55	32	405

SM: SMaterial, SMC: SMChar, EP: ExP, EPV: ExPVal, Ev: EvP, EvV: EvPVal, MM: MMethod, and TA: TArtifact are from the tag set. O: Other class of unannotated text (or terms with boundary mismatches that prevent tight agreement). T: Total.

Some disagreements were caused by careless mistakes or misunderstanding of the guideline by one of the students and were solved after discussion with the students. We could

confirm that the new guideline and the corpus-annotation support tool improved the quality of the annotation.

3.3 Corpus evaluation with a domain expert

3.3.1 Experiment setup

In the previous two experiments, we had constructed a corpus using graduate students. Even though the corpus construction guideline reach a reliable level in case of loose agreement. It is necessary to evaluate this corpus and finalize it with a domain expert researcher to ensure reliability. Therefore, we asked Prof. Hara (the domain expert involved in the design of the tag set) to evaluate the quality of the corpus and its construction guideline.

From the previous annotation experiments, we found that it requires more than 10 hours to annotate a single research paper from scratch (i.e., with no annotation information). It would be onerous for the domain expert to annotate five full corpus papers based on the guideline. We therefore asked him to evaluate the results of the previous corpus-construction experiments.

The evaluation data were assembled as follows. First, we classified the annotation results into two categories: agreed and disagreed. In the annotation experiments, there can be careless mistakes, such as one annotator missing to add an annotation, and typical types of disagreement, such as one of the annotators misunderstanding the guideline. These kinds of disagreements were easily checked in the discussion after each annotation experiment. To reduce the time required to evaluate the corpus, we considered these cases as part of the agreed annotations. For the agreed annotations, we used the same style for representing the corpus. For the disagreed annotations, we underline the related text and provide the students' annotation candidates to the domain expert. Figure 3.4 shows a sample of the evaluation-experiment data.

Using this information, we asked the domain expert to perform the following three tasks:

- Consider the appropriateness of the agreed annotations and identify any problematic annotation cases.

- Choose the appropriate annotation for each disagreed-annotation case. If none is appropriate, he should suggest a new candidate.

- Annotate any terms that had not been annotated.

3.3.2 Experimental results and discussion

We conducted the evaluation experiment in two steps. In the first step, we checked the validity of the experimental setup by using a single research paper [54]. In this experiment, we spent almost one hour evaluating the annotation results for the paper, including discussion of the corpus-construction guideline. Because there was no specific problem with the experimental setup, we conducted a second experiment that used the other four papers [56–59] as a second step. This required almost two hours, again including discussion of the corpus-construction guideline. The examination of the corpus during this evaluation experiment revealed that there are two types of papers in the corpus:

- Synthesis papers: Papers 1, 2, 3, 4 [54, 56, 58, 59] focus on the synthesis of new nanomaterials.

- Characterization papers: Paper 5 [57] focuses on the analysis and characterization of nanomaterials.

For each type of paper, there are specific statements that only apply to that type. The first synthesis paper required about one hour for its evaluation, because we needed to discuss necessary guideline modifications. The remaining synthesis papers were evaluated much more quickly, because the writing style of those papers was similar to the first. The characterization paper also required about one hour, including discussion related to the specific style of writing for this type of paper.

To improve the consistency of the annotation, and to overcome problems found by examining the corpus, the domain expert proposed two major modifications to the corpus-construction guideline:

- The intrinsic characteristics of a source material should be treated as SMaterial.
  In many cases, the intrinsic characteristics of a source material such as the distinctive group in the periodic table (e.g., Groups III or V) are used for representing a group of source materials. For example, the ratios among source materials and (or) group of source materials are sometimes represented as V/Mn or V/III. To maintain consistency among these descriptions, the intrinsic characteristics of a source material should be treated as SMaterial. Figure 3.5 shows an example of such cases from the corpus.

- Substitute MChar for SMChar.
  In some cases, the characteristics of the final product result from the manufacturing process instead of being inherited from the source materials. Figure 3.6 shows an example of two sources for the final product characteristics. Even if the final product

characteristics appear during the manufacturing process, they are as important as those inherited from the source materials. Therefore, it is not necessary to identify these characteristics as inherited from the source materials or resulting from the manufacturing process.

We have constructed a final version of the corpus to reflect all the corrections and modifications suggested by the domain expert. We compared this corpus with the original corpus constructed for the evaluation experiment, to analyze the quality of the original. Because there are different types of error for synthesis papers and characterization papers, we provide separate comparisons for synthesis and characterization papers to characterize the differences between these two types of paper. Table 3.3 and table 3.4 show the comparison matrices between the domain-expert corpus and original corpus for synthesis papers and characterization papers, respectively. We calculate the precision and recall for each category. We also calculate the precision and recall when excluding the effects of guideline modifications.

Table 3.3 Comparison of annotation results for the domain-expert corpus and the original corpus for synthesis papers

		Domain expert										
		SM	MC	MM	TA	EP	Ev	EPV	EvV	O	T	Prec
Original	SM	558								15(0)	573(0)	0.97(0.97)
	MC	11(11)	247							10(0)	268(11)	0.92(0.96)
	MM			109						0(0)	109(0)	1.0(1.0)
	TA				300					0(0)	300(0)	1.0(1.0)
	EP					225				1(0)	226(0)	1.0(1.0)
	Ev						281			3(0)	284(0)	0.99(0.99)
	EPV							195		0(0)	195(0)	1.0(1.0)
	EvV								209	0(0)	209(0)	1.0(1.0)
	O	137(136)	36(27)	11(0)	26(0)	5(0)	11(0)	3(0)	21(0)		250(163)	
	T	706(147)	283(27)	120(0)	326(0)	230(0)	292(0)	198(0)	230(0)	29(0)	2414(174)	0.98(0.99)
	Rec	0.79(1.0)	0.87(0.96)	0.91(0.91)	0.92(0.92)	0.98(0.98)	0.96(0.96)	0.98(0.98)	0.91(0.91)		0.89(0.96)	

SM: SMaterial, MC: MChar, MM: MMethod, TA: TArtifact, EP: ExP, Ev: EvP, EPV: ExPVal, and EvV: EvPVal are from the tag set. O: Other class of unannotated text (or terms with boundary mismatches that prevent tight agreement). T: Total. Numbers in parentheses represent mismatches caused by guideline modifications. Rec: Recall. Prec: Precision (Numbers in parentheses represent recall and precision excluding mismatches caused by guideline modifications).

Table 3.3 and table 3.4 show that, for synthesis papers, the agreed-annotation results obtained through discussion after the annotation experiments have high precision for all information categories (ranging between 96% and 100%), when we exclude the effects of guideline modifications. It is important to have discussions between annotators after the annotation process. Such discussions can resolve mismatches caused by careless mistakes or misunderstanding of the guideline. Recall is also high (ranging between 91% and 100%). However, because disagreed annotations caused by ambiguity were separated from the agreed

Table 3.4 Comparison of annotation results for the domain-expert corpus and the original corpus for the characterization paper

		Domain expert										
		SM	MC	MM	TA	EP	Ev	EPV	EvV	O	T	Prec
Original	SM	58								4(0)	62(0)	0.94(0.94)
	MC		67							3(0)	70(0)	0.96(0.96)
	MM			14						0(0)	14(0)	1.0(1.0)
	TA				77					2(0)	79(0)	0.97(0.97)
	EP					20				0(0)	20(0)	1.0(1.0)
	Ev						55			2(0)	57(0)	0.96(0.96)
	EPV							34		1(0)	35(0)	0.97(0.97)
	EvV								46	0(0)	46(0)	1.0(1.0)
	O	16(13)	31(13)	2(0)	13(0)	12(0)	18(0)	2(0)	20(0)		114(26)	
	T	74(13)	98(13)	16(0)	90(0)	32(0)	73(0)	36(0)	66(0)	12(0)	497(26)	0.97(0.97)
	Rec	0.78(0.95)	0.68(0.79)	0.88(0.88)	0.86(0.86)	0.63(0.63)	0.75(0.75)	0.94(0.94)	0.70(0.70)		0.76(0.81)	

SM: SMaterial, MC: MChar, MM: MMethod, TA: TArtifact, EP: ExP, Ev: EvP, EPV: ExPVal, and EvV: EvPVal are from the tag set. O: Other class of unannotated text (or terms with boundary mismatches that prevent tight agreement). T: Total. Numbers in parentheses represent mismatches caused by guideline modifications. Rec: Recall. Prec: Precision (Numbers in parentheses represent recall and precision excluding mismatches caused by guideline modifications).

annotations in the original corpus (as prepared for the evaluation experiment), it is necessary to analyze in detail the quality of the disagreed annotations in the original corpus. For the characterization paper, the precision is high (ranging between 94% and 100%), but the recall is low because of the larger number of disagreed annotations in this case. The students' lack of deep domain knowledge for the characterization paper seems to have had a considerable effect on the quality of its annotation.

To investigate the recall problem in detail, we analyzed the evaluation results for disagreed annotations in the original corpus. There were several cases involving different levels of domain knowledge for which the students could not reach confident agreement. In such cases, one of the annotators was able to make an appropriate annotation and the other could not. If both annotators had insufficient domain knowledge, no appropriate annotation candidate was provided in the candidate list. We calculated the coverage of cases where one annotator was able to provide an appropriate annotation candidate as a function of the total number of disagreed annotations. We also calculated the coverage when excluding the effects of guideline modifications. Table 3.5 and table 3.6 reflect the analysis of disagreed annotations for synthesis and characterization papers, respectively.

In the synthesis papers, if we exclude the effects of guideline modifications, it seems that the coverage is high, particularly for SMaterial, TArtifact, ExP, and ExPVal. For those categories, whenever we can select the appropriate annotation from the candidates by considering differences in level of domain knowledge, the recall for those categories is

Table 3.5 Analysis of disagreed annotations in synthesis papers

	SM	MC	MM	TA	EP	Ev	EPV	EvV	T
Total	29(26)	18(9)	9(0)	24(0)	5(0)	11(0)	3(0)	20(0)	119(35)
Candidate	3	8	7	23	5	9	3	16	74
Cov	0.1(1.0)	0.44(0.89)	0.78(0.78)	0.96(0.96)	1.0(1.0)	0.82(0.82)	1.0(1.0)	0.80(0.80)	0.62(0.88)

SM: SMaterial, MC: MChar, MM: MMethod, TA: TArtifact, EP: ExP, Ev: EvP, EPV: ExPVal, and EvV: EvPVal are from the tag set. T: Total number of disagreed annotations. Candidate: Number of selections of disagreed annotations by the domain expert from annotation candidates. Cov: Coverage of terms that were selected from candidate list. (Numbers in parentheses represent terms and coverage when excluding mismatches caused by modifications to the guideline).

Table 3.6 Analysis of disagreed annotations in the characterization paper

	SM	MC	MM	TA	EP	Ev	EPV	EvV	T
Total	12(9)	24(8)	2(0)	13(0)	10(0)	18(0)	2(0)	20(0)	101(17)
Candidate	3	4	1	8	1	5	0	9	31
Cov	0.25(1.0)	0.17(0.25)	0.5(0.5)	0.62(0.62)	0.10(0.10)	0.28(0.28)	0(0)	0.45(0.45)	0.31(0.37)

SM: SMaterial, MC: MChar, MM: MMethod, TA: TArtifact, EP: ExP, Ev: EvP, EPV: ExPVal, and EvV: EvPVal are from the tag set. T: Total. Candidate: Number of selections of disagreed annotations by the domain expert from annotation candidates. Cov: Coverage of terms that were selected from candidate list. (Numbers in parentheses represent terms and coverage when excluding mismatches caused by modifications to the guideline).

higher. However, for the characterization paper, the coverage level is not high. Information categories such as EvP and EvPVal seem to have a lower coverage, particularly for the characterization paper.

From table 3.3, table 3.4, table 3.5, and table 3.6, we can conclude generally that information categories such as SMaterial, MMethod, and ExPVal tend to be easier to annotate. Conversely, information categories such as the parameters, ExP and EvP, and EvPVal tend to be more difficult to annotate, requiring deeper domain knowledge, in particular for the characterization paper. Most of the disagreed annotations in these categories resulted from difficulties in setting correct boundaries for these information categories. Boundary-identification problems can have a number of causes, as we describe below.

Parameters usually have basic keywords with variations that depend on context. For example, "temperature" is a parameter that can appear variously as "growth temperature," "at room temperature," "increasing temperature from x to y," and so on. Such variations make it difficult for annotators to define clear boundaries for the same parameter. Furthermore, parameters can be highly context dependent. The same parameter can be used either as experiment parameter or as evaluation parameter depending on the context. For example,

Table 3.7 Number of categorized terms in NaDev corpus

Information category	SMaterial	MMethod	MChar	TArtifact	ExP	EvP	ExPVal	EvPVal	Total
Terms	780	136	381	416	262	365	234	296	2870
Of total	27%	5%	13%	15%	9%	13%	8%	10%	

"size" can be used for ExP in "mask-opening size," and for EvP in "size of nanocluster," even in the same paper. Figure 3.7 shows examples of term boundary mismatches for parameters.

In addition, the evaluation of the final product is not only expressed with quantitative values such as numbers. In many cases, the evaluation can be expressed in longer statements that describe the final product. In many cases, the value of the evaluation parameter can also exist without the explicit appearance of the parameter itself in the same sentence. This can sometimes cause an annotator to confuse the evaluation parameter with its value. Such cases can make it difficult to identify the correct boundary for the evaluation statement. Figure 3.8 shows an example of boundary mismatch for the evaluation parameter value EvPVal.

3.4 Release of the corpus and its usage

3.4.1 Corpus Release

From the analysis of the results of the annotation experiments, we found that precision was high; the total precision was 99% for synthesis papers, and was 97% for the characterization paper (when the effects of guideline modifications were excluded). Recall was high for the synthesis papers (96% when excluding the guideline-modification effects), but not high for the characterization paper (81% when excluding guideline-modifications effects). However, in both cases, it is necessary to identify the appropriate annotation from the disagreed annotation results to obtain an increased recall. The level of knowledge about the subject domain should be a candidate criterion for such an evaluation process. In addition, for the boundary-identification problem, adding examples of appropriate annotations for ambiguous cases to the guideline may help the annotators. These results show that the guideline for annotating papers related to nanocrystal device development is now reliable to be used. For more information on corpus construction guideline, please refer to Appendix A.

We have released the corpus construction guideline. NaDev corpus can be also distributed upon request [11]. The corpus currently comprises five fully annotated papers, 392 sentences, and 2,870 annotated terms in eight information categories. Table 3.7 shows the number of categorized terms in NaDev corpus.

3.4.2 NaDev Usage

By using this corpus as training data, we plan to implement an automatic annotation framework to extract experimental information from research papers related to nanocrystal device development. The annotation results of this framework can be used as keywords with semantic category information for the papers. We will be able to construct a paper-retrieval system for a nanocrystal device development portal by using these information categories. For example, the user could find papers that involve MnAs as a source material in developing nanoclusters as a target artifact. Information such as this would be helpful in finding research papers that contain the results of recent analyses of particular types of experiments and would support the data collection process. In addition, these annotation results can be used to find similarities between research papers based on different similarity metrics [15]. For example, similarity metrics can be focused on certain information categories of interest for the researchers (such as source material or final product) rather than overall similarity based on the general content of the paper. Such flexible similarity metrics can help researchers plan experiments more efficiently by using insights from similar experimental settings reported in research papers.

3.4.3 Corpus Construction Strategy in the Nanocrystal Device Domain

This is the proposed procedure for constructing a high-quality corpus for new research papers:

- Conduct an independent annotation with two annotators. It is preferable to have at least one annotator who is familiar with the subject domain of the paper.
- Discuss the results after the annotation process. This is necessary to exclude both careless mistakes and errors based on misunderstanding the guideline. In addition, for the disagreed annotations, selection of one of the annotation candidates should take into account the knowledge level of the annotator and any similarity between the annotation and examples in the guideline. If none of the annotators has high confidence in an annotation, it is better to check with a domain expert. However, the number of annotations requiring such checking is likely to be much smaller than for the whole corpus.

3.5 Summary

In this chapter, we have developed a method for constructing an annotated corpus of research papers on nanocrystal device development, which aims to support the automatic extraction

of useful information for the analysis of experiments' results in this field. The corpus and its construction guideline were examined and evaluated by a domain expert. The corpus called "NaDev" (Nanocrystal Device Development corpus), and its guideline is now released, and can be used to annotate research papers about nanocrystal device development in a consistent manner.

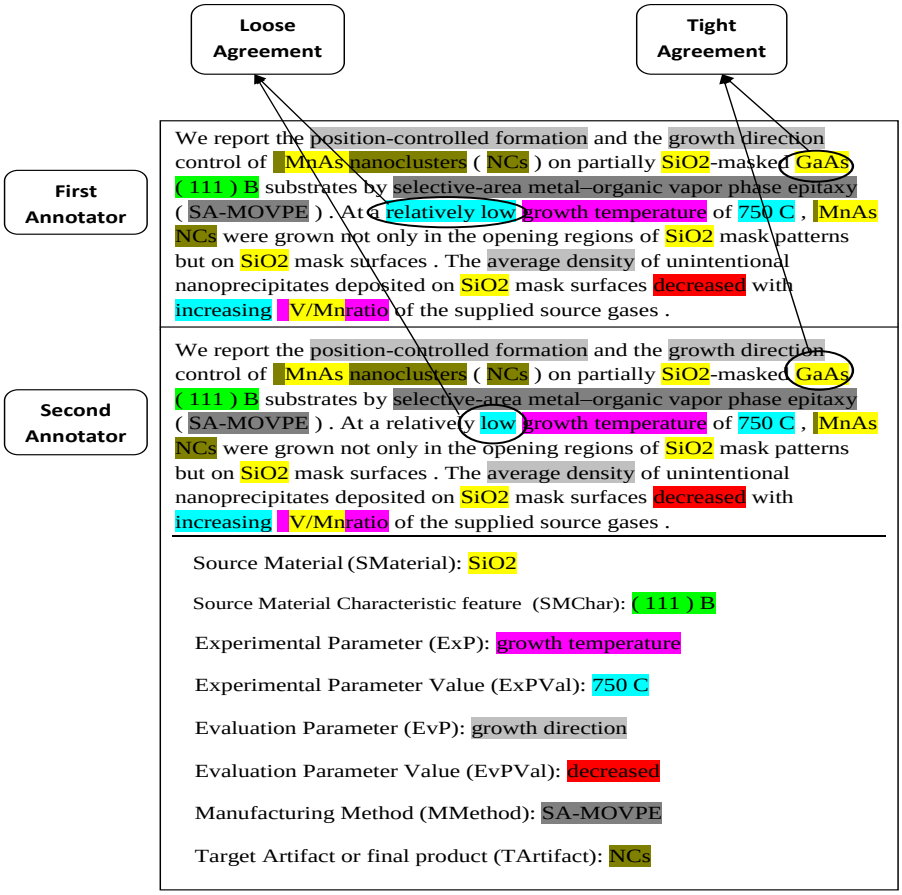


Fig. 3.2 Corpus sample illustrating tight and loose agreement.

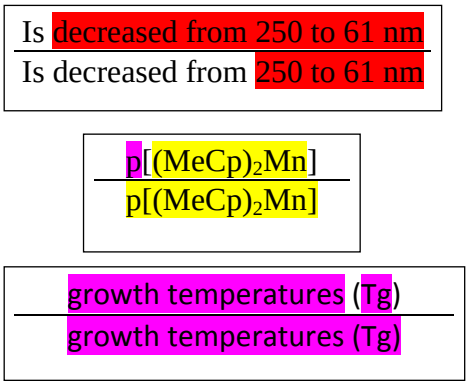


Fig. 3.3 Examples of term boundary mismatches between the first annotator (above) and the second annotator (below).

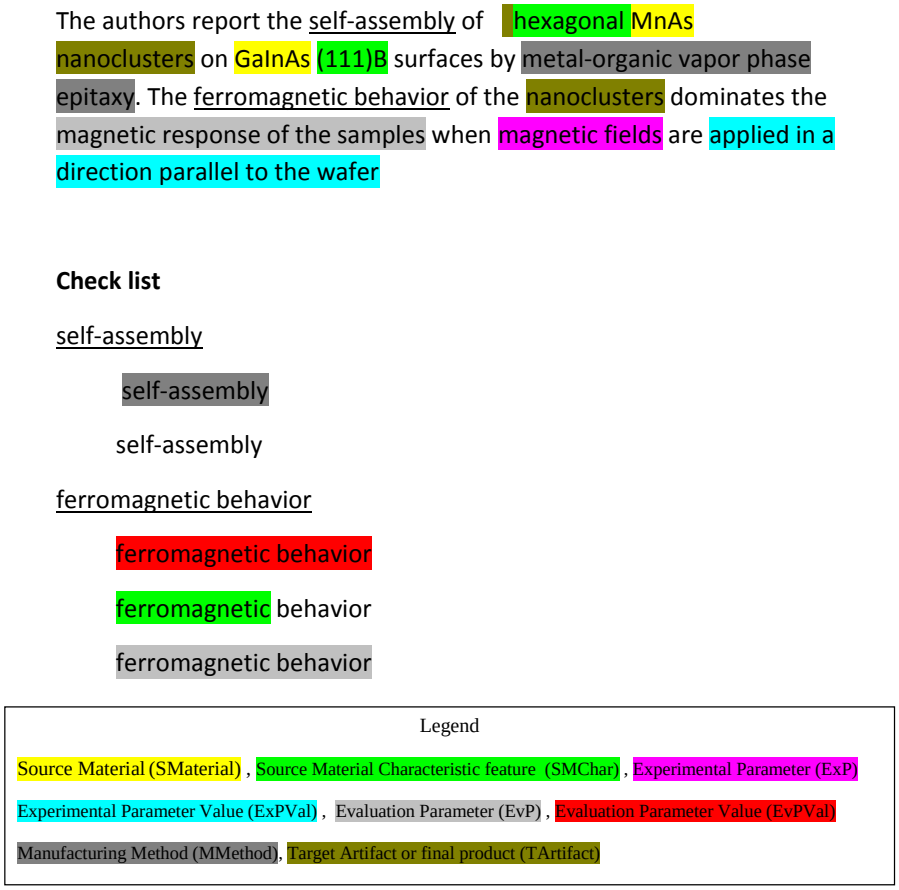


Fig. 3.4 Sample of the evaluation-experiment data

Fig. 3.5 Different representations of ratios between source materials

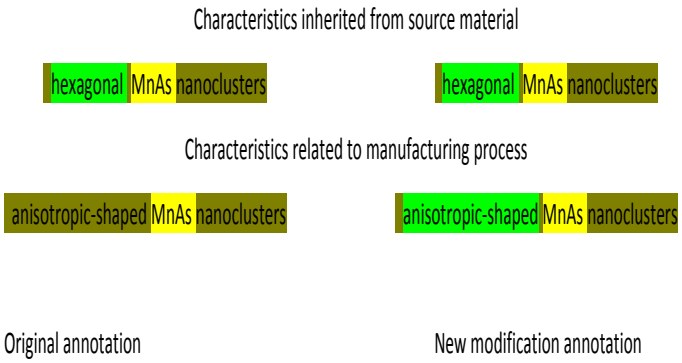


Fig. 3.6 Different sources for the final product characteristics

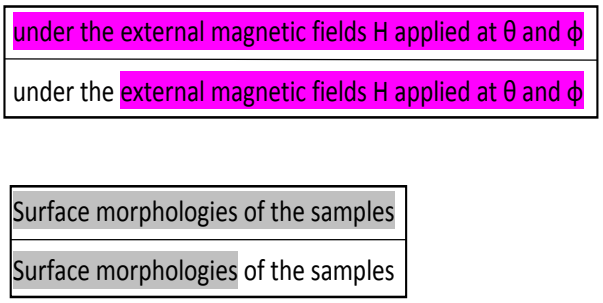


Fig. 3.7 Examples of the boundary-identification problem for terms in parameter categories

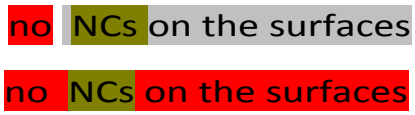


Fig. 3.8 Example of the boundary-identification problem for terms in evaluation parameter values

Chapter 4

NaDevEx: Automatic annotation framework for nanocrystal device research papers

4.1 Introduction

In this chapter, we propose a framework for automatic information extraction, NaDevEx (Nanocrystal Device Automatic Information Extraction Framework) from research papers on nanocrystal devices and evaluate the system using the NaDev corpus we discussed in chapter 3. Our system uses a step-by-step (cascading style) extraction based on machine-learning and natural language processing techniques. Linguistic and domain knowledge features are used to support information extraction. We discuss domain-related issues that reflect the nature of entities in nanocrystal devices development domain when designing the system. We also discuss the quality of automatic information extraction compared with that from human annotators and conduct failure analysis to identify future research issues. Additionally, we compare system performance based on paper type, and analyze the effect of domain knowledge features on system performance.

Since there is significant amount of chemical information in nanocrystal device development publication, we discuss a chemical entity recognition system using machine-leaning techniques based on an ensemble learning approach.

4.2 Automatic information extraction

4.2.1 System design

4.2.1.1 Chemical entity recognition

In literature related to nanocrystal device development, most of the source material entities are chemical compounds. We assume identifying chemical entities (e.g., As) is helpful to identify source material entities.

We have developed a new chemical entity recognizer called SERB-CNER (Syntactically Enhanced Rule-Based CNER) enhance the identification of source material entities [12]. SERB-CNER is a rule-based chemical entity recognizer that uses regular expressions to identify chemical compounds. In addition to that, SERB-CNER uses syntactic rules to eliminate some mismatches that might occur between chemical entities and general text.

4.2.1.2 Cascading style information extraction

In nanocrystal device development domain, entities sometimes overlap within each other, and not always simple. Figure 4.1 shows an example of overlapping entities. Because of

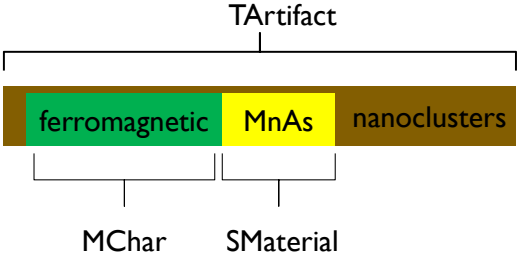


Fig. 4.1 Overlapped entities

this overlapping, same chunk of text might have information related to more than one entity at the same time. That makes it difficult for a machine-learning system to learn to set the correct entity information all at once. To tackle this issue, we have separated overlapped information categories into 5 groups where entities of these information categories do not overlap with other entities of information categories of the same group.

- group 1: SMaterial, and MMethod.
- group 2: MChar.
- group 3: TArtifact.

- group 4: ExP and EvP.
- group 5: ExPVal, and EvPVal.

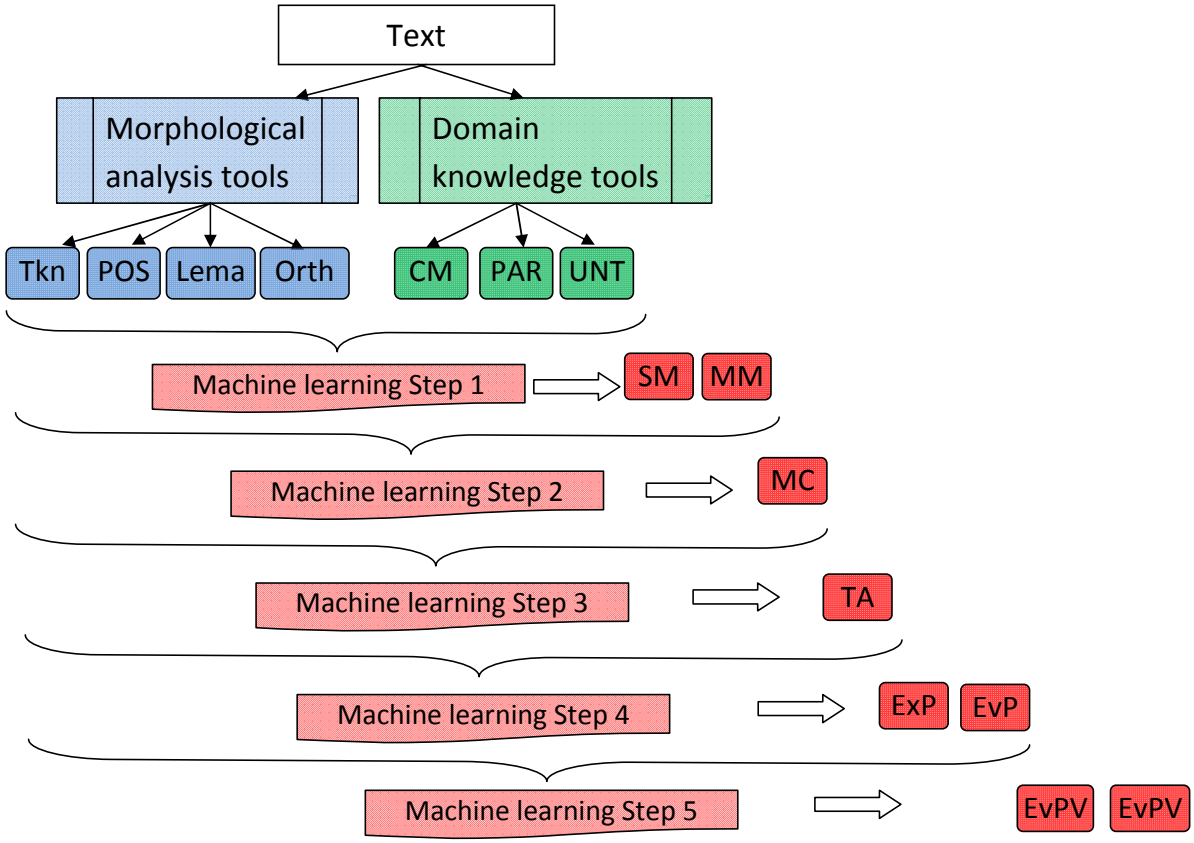
Using the overlapping structure between entities, we can use information of entities of one group to support identification of entities from another group. For example, identification of source material (e.g., As) is useful for identifying term boundaries of experimental parameters (e.g., pressure of AsH_3 gas). The order of information categories for extraction was designed by using the overlapping structure between information categories. For example, for experimental parameters and source materials (e.g., pressure of AsH_3 gas), the extraction of source material should be prior to extraction of experimental parameters. Based on that, we divided the machine-learning process into 5 step-by-step (cascading) levels i.e. cascading named entity recognition [61].

4.2.1.3 Physical quantities list

There are two types of parameters exist in this nanocrystal device development papers corpus; Experiment parameter (ExP), and evaluation parameter (EvP). Since (ExP) represents a control parameter for the experimental equipment and (EvP) represents ones measured by measuring devices, most of the entities of these two information categories are associated with physical quantities and they contains (a) term(s) that represent(s) its characteristics. For example, “density of the nanoclusters” and “height of the nanoclusters” contains physical quantity term “density” and “height” respectively. Based on [62], most of the compound nouns that are constructed from basic head noun, identification of head noun contributes to make up compound noun is useful to extract the whole compound noun. Identification of physical quantities may support to extract parameters. For example, identification of "size" as a physical quantity might support identification of “size of nanoclusters” as a parameter. We have constructed a list of physical quantities [13] to support the identification of parameters in nanocrystal device development papers.

4.2.2 System layout

Figure 4.2 shows an outline for our system to extract the desired information categories step-by-step. First, linguistic features such as part-of-speech (POS) tags, orthogonal features, and lemmatization features are generated using the results from a morphological analysis tool [63]. Second, we use domain knowledge tools (i.e., the output of a chemical named entity recognition tool [12], matching results from a physical quantities vocabulary list, and a list of common measurement units [13]) to generate domain knowledge-related features (CNER,



Legend: Tkn: token, POS: part of speech, Lema: lemmatization, Orth: orthogonal, CM: chemical named entity, PAR: physical quantity matching, UNT: measurement unit list matching, SM: SMaterial, MM: MMethod, MC: MChar, TA: TArtifact, ExP: ExP, EvP: EvP, ExPV: ExPVal, and EvPV: EvPVal

Fig. 4.2 Outline of our automatic information extraction system

PAR, and UNT, respectively). For the latter step, we used CRF++ [64], an implementation of conditional random field (CRF) [26] as a machine learning system that uses part of the corpus as training data for information extraction. In each step, we use all the features generated by the tools, including linguistic features and domain knowledge-related features.

4.2.3 System implementation

The NaDevEx system accepts plain text as input and adds annotations to the terms in the text that belong to the information categories defined in the NaDev corpus construction guideline.

Information about the most recent version of the system, which was used for these experiments, is as follows.

- Linguistic features: GPostLL tagger (ver. 0.9.3) [63].
- An orthogonal feature was added using regular expressions based on the definition in [23].
- Domain knowledge-based features:
  - A chemical named entity feature was added using SERB-CNER (Syntactically Enhanced Rule-Based Chemical Named Entity Recognition System) that we developed to annotate chemical entities in nanocrystal device papers.
  - A parameter identification feature was added based on a list of physical quantities: we compiled a list that contains physical properties of matter (e.g., density, concentration), common parameters found in nanocrystal device papers (e.g., height, conductivity), and several keywords that usually correlate with parameters (e.g., ratio, rate). The list was checked by nanocrystal device researchers as a basic list for physical quantities.
  - A parameter value identification feature was added based on a list of common measurement units.
- CRF tool: CRF++ (ver.0.58)

The input for the CRF++ tool is in IOB format, which identifies the position (beginning, inside, out of) of a token of text related to a term. Figure 4.3 shows an example of input data for the CRF++ tool.

For the training, NaDevEx first added linguistic features and results of the domain knowledge-based systems to the original texts. Then information about correct annotations was used to train the machine learning system CRF++ in cascading style. For the information

Tkn	POS	Lema	Orth	CM/PAR/ UNT	SM/MM	MC	TA	ExP/EvP	ExPV/EvPV
MnAs	NP	mnas	TwoCaps	B-CM	B-SMaterial	O	B-TArtifact	O	O
thin	JJ	thin	Lowercase	O	O	O	I-TArtifact	O	O
films	NNS	film	Lowercase	O	O	O	I-TArtifact	O	O
have	VHP	have	Lowercase	O	O	O	O	O	O
shown	VVN	show	Lowercase	O	O	O	O	O	O
ferromagnetic	JJ	ferromagnetic	Lowercase	O	O	O	O	B-EvP	O

Legend: Tkn: token, POS: part of speech, Lema: lemmatization, Orth: orthogonal, CM: chemical named entity, PAR: physical quantity matching, UNT: measurement unit list matching, SM: SMaterial, MM: MMethod, MC: MChar, TA: TArtifact, ExP: ExP, EvP: EvP, ExPV: ExPVal, and EvPV: EvPVal

Fig. 4.3 Example of CRF++ input data

extraction, the system used the same tools to add linguistic features and results of domain knowledge and used the learning results of CRF++ in cascading style to generate the final answer.

4.2.4 Experiment plan

In this paper, we evaluate our automatic information extraction system (NaDevEx) and discuss the characteristics of this system by using the NaDev corpus. We design an experiment plan to address the following three main issues:

- System performance analysis compared with human annotators
- System performance analysis for each type of corpus paper (synthesis or characterization)
- Effect of domain knowledge features on system performance.

4.2.4.1 System performance analysis compared with human annotators

We evaluated our system performance using NaDev corpus. We used five fold cross validation and calculated precision, recall, and F-score. In each fold, we trained the system using four of the five papers as training data and evaluated its performance using the fifth paper. Because NaDev gold standards are based on the domain expert's annotation, those results represent

Table 4.1 Average performance of NaDevEx and the human annotation results compared with the domain expert's annotation

	Human			**NaDevEx**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
SMaterial	0.97	0.79	0.87	0.95	0.94	0.94
MMethod	1.00	0.91	0.95	0.97	0.73	0.82
MChar	0.93	0.84	0.88	0.94	0.67	0.75
TArtifact	0.99	0.90	0.94	0.88	0.73	0.80
ExP	1.00	0.91	0.94	0.93	0.68	0.76
EvP	0.98	0.91	0.94	0.78	0.55	0.64
ExPVal	0.99	0.97	0.98	0.80	0.53	0.64
EvPVal	1.00	0.86	0.92	0.75	0.39	0.51
Total	0.98	0.86	0.91	0.89	0.69	0.77

the comparison between NaDevEx performance and the domain expert's annotation. Because NaDevEx is built using machine-learning techniques, deep domain knowledge is difficult to acquire using NaDevEx. Therefore, we contrast NaDevEx performance with that based on agreement between two novice annotators, as discussed previously. These comparison results represent the ideal level of annotation without deep domain knowledge.

Table 4.1 contrasts the average performance for each information category between NaDevEx and the human annotation results compared with the domain expert's annotation. Underlining indicates that the difference between NaDevEx performance and the human annotation results is statistically insignificant at the 5% level ($P >= 0.05$). The human annotations were made prior to the released version of the guideline [11]. Recall of categories that were subject to new definitions (SMaterial and MChar) is underestimated. If we assume that all the new added annotations based on the released guideline were identified by human annotators, recall of SMaterial and MChar is increased to 0.99 and 0.93, respectively.

From Table 4.1, the performance of NaDevEx on the SMaterial category is almost comparable with human annotation. For MMethod, MChar, and ExP, performance is comparatively good for precision but not so good for recall. For the other categories, the system performance is not so good for precision and worse for recall. Based on the nature of the machine learning system, it is easier to extract the terms that appear in the training data than ones that are unique in the test data. However, if there are similar terms (e.g., a term that overlap with one in the training data or terms used in similar context) in the training data, the system can extract such terms.

Table 4.2 Average performance of NaDevEx and the human annotation results for loose agreement compared with the domain expert's annotation

	Human			**NaDevEx**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
SMaterial	0.99	0.81	0.89	0.98	0.97	0.97
MMethod	1.00	0.91	0.95	0.98	0.73	0.83
MChar	0.94	0.85	0.89	0.96	0.68	0.77
TArtifact	1.00	0.90	0.95	0.96	0.79	0.86
ExP	1.00	0.91	0.95	0.97	0.71	0.79
EvP	0.99	0.92	0.95	0.86	0.60	0.71
ExPVal	1.00	0.97	0.99	0.92	0.62	0.74
EvPVal	1.00	0.86	0.92	0.88	0.46	0.60
Total	0.99	0.87	0.92	0.95	0.74	0.83

There are several cases that show the term boundary identification problem, especially for unique compound terms. To check the effect of such problems, we used the loose agreement metric as illustrated in figure 3.2.

For human annotators, even though there were many cases of loose agreement between the two annotators, discussion after annotation experiments generally resolved these boundary mismatch issues. Table 4.2 contrasts the average performance for each information category for NaDevEx and the human annotation results for loose agreement compared with the domain expert's annotation. Underlining indicates that the difference between NaDevEx performance and the human annotation results is statistically insignificant at the 5% level ($P >= 0.05$).

The differences between the evaluation results of Table 4.1 and Table 4.2 reflect the difficulty of identifying term boundaries. For NaDevEx, performance for loose agreement improves for all information categories in precision and recall, especially for TArtifact, EvP, ExPVal, and EvPVal. This shows that these categories have many problems related to identifying term boundaries. If we accept loose agreement as correct (in most cases we can find appropriate head nouns such as temperature, or pressure in loose matching terms), TArtifact and EvPVal also become almost comparable with human annotation for precision.

In general, Table 4.1 and Table 4.2 show that NaDevEx has problems in identifying term boundaries in categories where human annotators have the same difficulty. However, discussion between the annotators after each annotation experiment helped to reduce these difficulties.

In addition, recall of the categories MChar, ExP, EvP, ExPVal, and EvPVal is comparatively worse than that made by the human agreement. For these categories, there are varieties

Table 4.3 NaDevEx average performance on synthesis and characterization papers using five-fold cross validation

	Average synthesis papers						Characterization paper					
	Prec	**Rec**	**F**	**L-Prec**	**L-Rec**	**F**	**Prec**	**Rec**	**F**	**L-Prec**	**L-Rec**	**F**
SMaterial	0.95	0.94	0.94	0.98	0.97	0.97	0.93	0.96	0.95	0.96	0.99	0.97
MMethod	0.97	0.75	0.84	0.98	0.76	0.85	1.00	0.63	0.77	1.00	0.63	0.77
MChar	0.94	0.78	0.85	0.96	0.79	0.86	0.92	0.22	0.36	1.00	0.24	0.39
TArtifact	0.93	0.79	0.85	0.95	0.81	0.87	0.69	0.49	0.57	1.00	0.71	0.83
ExP	0.91	0.77	0.83	0.96	0.81	0.87	1.00	0.31	0.48	1.00	0.31	0.48
EvP	0.80	0.57	0.66	0.88	0.62	0.73	0.73	0.48	0.58	0.77	0.51	0.61
ExPVal	0.81	0.57	0.66	0.95	0.67	0.78	0.76	0.41	0.53	0.82	0.44	0.57
EvPVal	0.74	0.41	0.53	0.87	0.48	0.62	0.79	0.33	0.46	0.90	0.37	0.53
Total	0.90	0.75	0.82	0.95	0.79	0.86	0.82	0.47	0.60	0.93	0.53	0.68

Prec: precision, Rec: recall, L-Prec: loose precision, L-Rec: loose recall, F: F-score

of compound terms that usually contain characteristic technical terms within their boundaries. However, because of the variability in using these technical terms for constructing compound terms, NaDevEx cannot extract such terms appropriately. We discuss this issue in detail in section 4.2.4.3.

4.2.4.2 System performance analysis based on type of paper

System performance differs between synthesis papers and characterization papers. Table 4.3 shows the average performance of NaDevEx for four synthesis papers and one characterization paper including loose agreement cases using five-fold cross validation. One reason for the lower performance with the characterization paper is a lack of examples of sentences and terms that are frequently used in characterization papers and not in synthesis papers. To discuss this effect, we conducted a 10-fold cross validation that uses four papers and half of the fifth paper as training data, evaluated on the other half of the fifth paper. Table 4.4 shows the average performance of NaDevEx on four synthesis papers and one characterization paper using 10-fold cross validation including loose agreement. In this case, because we can use one-half of a paper as training data, the number of terms that are unique to the test data decreased. The performance for 10-fold cross validation is slightly better than that for five-fold cross validation. However, in total, the increased ratio for characterization with loose recall was slightly better than that for synthesis papers.

Table 4.4 NaDevEx average performance on synthesis and characterization papers using 10-fold cross validation

	Average synthesis papers						Average characterization paper					
	Prec	**Rec**	**F**	**L-Prec**	**L-Rec**	**F**	**Prec**	**Rec**	**F**	**L-Prec**	**L-Rec**	**F**
SMaterial	0.95	0.94	0.94	0.98	0.97	0.97	0.96	0.97	0.96	0.97	0.99	0.98
MMethod	0.96	0.81	0.87	0.96	0.81	0.87	1.00	0.63	0.77	1.00	0.63	0.77
MChar	0.95	0.83	0.89	0.97	0.84	0.90	0.84	0.35	0.46	0.87	0.37	0.49
TArtifact	0.95	0.85	0.90	0.96	0.87	0.91	0.71	0.53	0.61	0.98	0.75	0.85
ExP	0.93	0.81	0.86	0.98	0.86	0.91	0.59	0.33	0.42	0.88	0.46	0.61
EvP	0.80	0.63	0.70	0.88	0.69	0.77	0.77	0.47	0.58	0.87	0.53	0.66
ExPVal	0.81	0.67	0.73	0.93	0.77	0.83	0.69	0.46	0.55	0.78	0.51	0.61
EvPVal	0.75	0.48	0.58	0.88	0.56	0.68	0.78	0.35	0.48	0.93	0.41	0.57
Total	0.91	0.79	0.84	0.96	0.83	0.89	0.80	0.51	0.62	0.93	0.59	0.72

Prec: precision, Rec: recall, L-Prec: loose precision, L-Rec: loose recall, F: F-score

4.2.4.3 Effect of domain knowledge features on system performance

As we have already discussed, it is difficult for the machine learning system to find terms that are unique to the test data. Table 4.5 shows the number of unique terms in each paper and the system performance for extracting such terms.

For SMaterial, even though there are many terms that are unique to the test data, the system can identify such terms with a considerably higher coverage ratio than is obtained for other information categories. In most cases, those terms are identified as Chemical Named Entities and the system can generalize the training data by using the information that has been provided by the CNER tool, discussed earlier. For the parameters ExP and EvP, precision is good when the system can use parameter list to identify parameter-related terms. However, because of the insufficient coverage of parameter-related terms used in nanocrystal device development, recall of these parameters is worse than human annotators' results.

These results show that preprocessing annotation based on domain knowledge is generally promising, but coverage of the parameter information based on a list of physical quantities is not enough for nanocrystal device papers. As we have already discussed in section 4.2.4.1, there are many compound terms that contain particular domain-specific terms within their boundaries for characterizing categories. Figure 4.4 shows an example of such domain-specific terms. Human annotators might be able to recognize such domain-specific terms with their domain knowledge; however, NaDevEx lacks such ability, especially with small training examples. It is necessary to evaluate the effectiveness of such a list by using a larger corpus.

Table 4.5 Unique term analysis for each paper

	Synthesis papers												Characterization paper			
	Paper 1			Paper 2			Paper 3			Paper 4			Paper 5			Corpus average coverage
	Uniq	Extracted	Coverage	Uniq	Extracted	Coverage	Uniq	Extracted	Coverage	Uniq	Extracted	Coverage	Uniq	Extracted	Coverage	
SMaterial	15	8	0.53	6	5	0.83	16	10	0.63	12	0	0.00	7	6	0.86	0.57
MMethod	0	0	NA	0	0	NA	14	4	0.29	10	2	0.20	7	2	0.29	NA
MChar	6	2	0.33	23	7	0.30	25	14	0.56	10	1	0.10	68	3	0.04	0.27
TArtifact	11	3	0.27	12	4	0.33	17	9	0.53	13	2	0.15	46	4	0.09	0.28
ExP	8	5	0.63	10	0	0.00	7	3	0.43	11	1	0.09	22	0	0.00	0.23
EvP	11	3	0.27	27	2	0.07	21	4	0.19	52	11	0.21	49	17	0.35	0.22
ExPVal	26	10	0.38	13	5	0.38	20	6	0.30	38	11	0.29	23	8	0.35	0.34
EvPVal	29	13	0.45	33	10	0.30	39	15	0.38	44	10	0.23	52	9	0.17	0.31
Total	106	44	0.42	124	33	0.27	159	65	0.41	190	38	0.20	274	49	0.18	0.29

Uniq: number of unique terms in each paper; Extracted: number of terms identified by NaDevEx; Coverage: coverage percentage of unique terms identified.

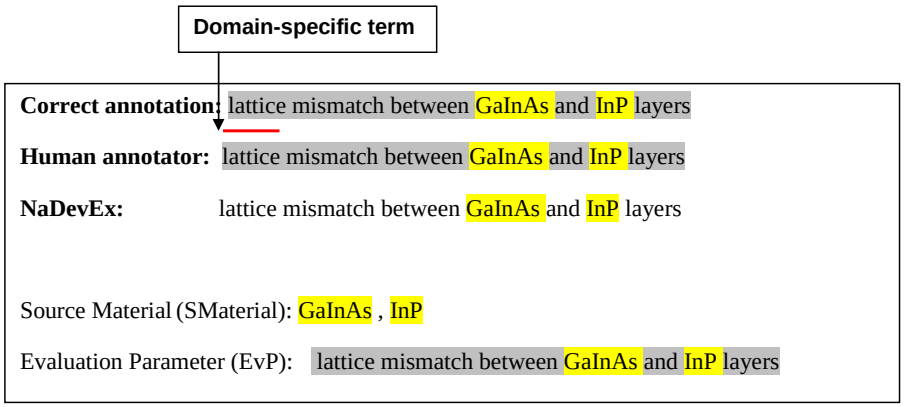


Fig. 4.4 Domain-specific terms in NaDev corpus

4.2.4.4 Discussion

The performance of NaDevEx is good for precision (95% for loose agreement overall), but is not good for recall (74% for loose agreement in total) at present. For the information category with rich domain knowledge information (SMaterial), our system performance is almost comparable with that of human annotators. Precision of the system output is generally high: it is good (more than 95%) for MMethod, MChar, TArtifact and ExP but modest (more than 85%), for other categories (EvP, ExPVal, and EvPVal) with loose agreement. In contrast, recall of the system is low (46%–73%), even with loose agreement.

It is necessary to take into account the effect of the corpus size. As we discussed in Table 3.6, it is difficult to extract unique terms that do not exist in the training data (percentage of the unique terms among total terms is almost 30% (853/2870)). It is better to check the percentage of the unique terms among total terms when the size of the corpus increases. On the contrary, identification of non-unique terms is comparatively easier for such a small size corpus.

There are two possible research approaches to increase recall of the system output. One approach is to increase the corpus size. It is good to use one whole paper for clear understanding of the role of the terms in the paper, but the varieties of terms are not greatly increased because of the repetitive mention of terms. For the next step, it may be better to construct an abstract-based corpus to increase the variety of terms. It is also preferable to have a balanced mixture of synthesis and characterization papers. Another approach is to construct resources for representing domain knowledge. A list of terms that are frequently used in nanocrystal device papers is helpful to extract related terms that are in the list and variations of the terms based on the head terms in the list. There are physical parameters that cannot be extracted using the general physical quantities list (e.g., lattice, (1 1 1)B surface), so it is better to use vocabulary lists that include the parameters in this domain.

NaDevEx can be used as a preprocessor to find research papers that contain recent analysis results on nanocrystal devices to support the data collection process. Because NaDevEx is good at identifying source material, we can construct appropriate queries to restrict the output to papers that discuss a particular type of source material. Usage of other information categories may work well for finding related papers in a precision oriented manner, but it may miss papers because of the bad recall performance. A possible solution to this problem is implementing a framework that utilizes user-defined keyword lists as a knowledge resource for extracting such information. Another is using simple keyword search to find more papers that may contain such information.

4.2.5 Summary

We introduced NaDevEx, which automatically extracts useful information from nanocrystal device research papers based on the information categories defined in the NaDev corpus. NaDevEx uses a cascading style machine-learning based approach with support of domain knowledge features. This system's performance was not defeated much by the human annotators for source material information, because of the good performance of the chemical named entity recognition system. For other categories, the precision is good (better than 85% in case of loose agreement), but there is a problem with recall because of the lack of examples, especially for characterization papers.

4.3 Extraction of chemical entities by ensemble Learning of different characteristics Chemical NER tools

4.3.1 Introduction

Recently, we have become able to use large quantities of textual data for extracting useful information. As an example, we can use a research article database as "big data" for understanding research trends and new research results. This is a new frontier for utilizing machine-learning techniques. There are two main approaches in this domain, namely analyzing bibliographic information to identify research trends [65, 66] and extracting useful information by using text-mining techniques.

Chemical named-entity recognition (chemical NER) is an application domain for extracting chemical information from text. Extraction of all the chemical named entities from a paper is desirable in finding articles that are related to particular chemical named entities. In addition, chemical information plays a significant role in a variety of related disciplines such as bioinformatics [33] and nanoinformatics [36]. For example, chemical information could help detect drug-protein interaction in the bioinformatics domain or source materials in nanodevice-development experiments within the nanoinformatics discipline.

Chemical NER tasks began with extracting general chemical named-entity information and expanded to meet the demand for extracting chemical-related named entities (such as drugs) that are used in particular research domains. In such an expansion, new guidelines were created to include the new types of entities. Because these guidelines also aim to extract general chemical named entities, they are similar to those used for the general chemical NER task but also include guidelines for the extraction of new types of chemical-related named entities. At an early stage, the SCAI corpus [28] of general chemical named entities

was created to identify the International Union of Pure and Applied Chemistry (IUPAC) entities [29]. Another approach was to create a chemical named-entity dictionary such as the Chemical Entities of Biological Interest (ChEBI) [67].

To support these chemical NER tasks, several chemical NER tools were developed to extract chemical named entities from text. Because there was no standard guideline of the chemical NER task, these tools were developed based on one of the guidelines. At an early stage of chemical NER tool development, most tools were evaluated using the SCAI corpus, which was a large corpus that was freely available for such tasks. This means that most chemical NER tools, such as OSCAR4 [30], and ChemSpot [31], were developed primarily for the chemical NER task defined by the SCAI corpus.

Additionally, these tools have different characteristics. For example, ChemSpot [31] uses a machine-learning approach that uses several linguistic features such as POS, lemmatization form, and orthogonal features to identify chemical named entities. It also uses dictionary-based features. On the other hand, OSCAR4 [30] uses a rule based approach, with a chemical dictionary and syntactic patterns to represent chemical named entities via regular expressions (OSCAR4 also uses machine-learning-based methods in the form of a maximum-entropy Markov model). Other tools also use a hybrid approach, combining rule-based and machine-learning-based methods.

We propose a method for applying these chemical NER tools to the BioCreative IV, CHEMDNER task [33] based on an ensemble-learning technique. This task aims to extract drug names in addition to general chemical named entities. BioCreative IV, CHEMDNER corpus uses the abstracts of chemical-related papers in MEDLINE. Chemical-related terms are identified as the offset information of such terms from the beginning position of the text. Left side of Figure 4.5 shows an example of those terms' information. Right side of the figure represents illustrated interpretation using abstract text information.

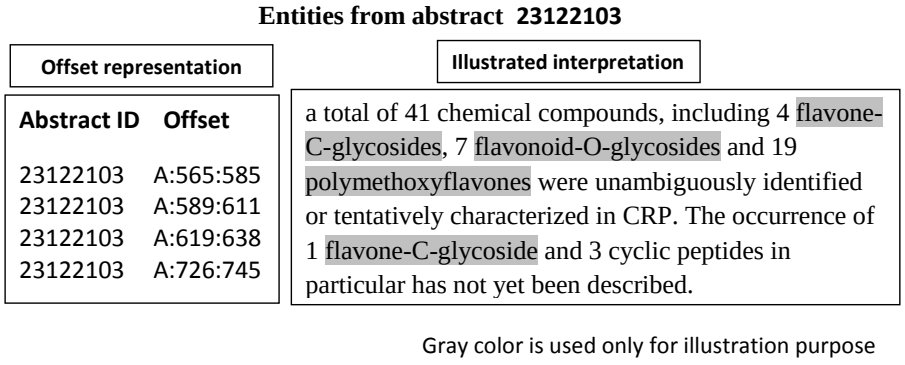


Fig. 4.5 BioCreative IV, CHEMDNER corpus data snapshot

For this very recent task, a simple ensemble-learning approach based on voting [68] is not appropriate. Therefore, we use all the system outputs of these chemical NER tools as features of a conditional random field (CRF) model [26], in addition to linguistic features, such as lexical and orthogonal features, that are widely used for chemical NER tasks. This approach is similar to the concept of domain adaptation [69] in natural language processing (NLP), which uses machine-learning results from corpora of a variety of domains to analyze texts in a new domain.

Because different chemical NER tools use different tokenization schemas for text tokenization, it is necessary that the ensemble-learning approach should handle any inconsistency between the outputs of the chemical NER tools. We apply a post-tokenization mechanism to generate more flexible tokenization schema that can adapt to a variety of chemical NER-tool tokenization schemas.

Since there are many research domains that use chemical information, these chemical NER tools have been applied to a variety of research domains. However, because of the variations in chemical-related entities across domains, these tools may not be sufficient to extract all chemical-related information in a particular domain. For example, drug names are chemical-related named entities, but general chemical NER tools cannot recognize all of them. To meet these new demands, a new corpus for extracting chemical and drug names was developed in the BioCreative IV, CHEMDNER task [33]. Even though the corpus guidelines share certain chemical entities with general chemical NER guidelines, several differences remain. Because of these differences, chemical NER tools that were developed using general chemical NER guidelines, such as ChemSpot and OSCAR4, might not perform well when tested with the new task.

Some implementations for named-entity recognition have adopted an ensemble-learning approach. For example, Dimililer et al. [70] describe classifier subset selection for biomedical named-entity recognition. In this work, a vote-based classifier selection scheme has an intermediate level of search complexity between static classifier selections and real-value and class-dependent weighting approaches. Zhou et al. [71] describe voting-based ensemble classifiers to detect hedges and their scopes. Another ensemble-learning approach assumes that instead of searching for the best-fitting feature set for a particular classifier, an ensemble of several classifiers that are trained using different feature representations could be a more fruitful approach. For example, Ekbal et al. [72] apply this approach to named-entity recognition. However, all of these approaches assume that all machine-learning systems are constructed for the same task.

4.3.2 Framework for Ensemble-learning Approach

4.3.2.1 Framework Architecture

The ensemble approach we are proposing uses CRF to fuse several chemical NER tools that use different recognition schemas. This framework decomposes input text into a sequence of tokens (tokenization), generates characteristic features for each token, namely linguistic features, and the results of chemical NER tools for this token, and then uses CRF to predict the label of the token. Based on CRF results, the system can identify chemical entities and drug names in a text.

A General text tokenizer (e.g., POS tagger) might not be good enough to adapt to multiple tokenization schemas applied by different chemical NER tools. Our system implements a post-tokenization mechanism to overcome such problems. First, we discuss the tools we are using in more detail.

- Chemical NER Tools We have used the following chemical NER tools:
  - **SERB-CNER (Syntactically enhanced rule-based chemical NER)** is a rule-based chemical-entity recognizer that uses regular expressions to identify chemical compounds [12].
  - **ChemSpot** is a named-entity recognition tool for identifying mentions of chemicals in natural language texts, including trivial names, drugs, abbreviations, molecular formulas and IUPAC entities. ChemSpot uses a hybrid approach that combines a CRF with a dictionary. ChemSpot is trained by using SCAI corpora [73] annotated mainly with IUPAC [29] entities.
  - **OSCAR4** is an open-source extensible system for the automated annotation of chemistry in scientific articles. It uses a rule-based approach, in addition to machine-learning-based methods in the form of a maximum-entropy Markov model, to identify chemical entities.
- Linguistic Features We have used GPoSTTL as a basic text tokenizer and part-of-speech tagger to define the basic type of each token. GPoSTTL is an enhanced version of Eric Brill's rule-based tagger. In addition to the POS tag, GPoSTTL generates a lemmatization feature. Based on GPoSTTL results, we use regular expressions to generate orthogonal features as defined in [23]. An orthogonal feature is a symbol that represents various styles of surface symbols (such as all capitals, lowercase, or digits).
- Conditional Random Field (CRF) A CRF [26] is a probabilistic sequence-labeling model commonly used in NER tasks. In such a task, a CRF model takes an input of

a text token sequence and seeks to assign a categorical label for each member of a sequence relying on statistical inference. Because a named entity might span over multiple tokens, IOB format is used to define entity boundaries, where "B" identifies the beginning of a named entity, "I" declares that the token is inside the named entity and "O" means that the token is outside the named entity. To label the token sequence, a CRF model builds a set of inference rules using a training dataset in which each token is attached to a feature set and labeled correctly. As noted linguistic features such as token surface, POS tag, lemmatization, and orthogonal features, are commonly used in NER tasks.

The inference rules take into consideration the target label of a token in relation to both its own feature set and also the feature sets of neighboring tokens within a certain feature window size. The feature window is defined as a function of the target label to n-gram feature combination. For example, in a bigram, the current target label is defined as a function of the combination of two features, one from the current token's feature set plus another one from a neighboring token's feature set. This makes CRF well suited for natural language processing applications [74, 75]. Figure 4.6 shows an outline of the CRF model.

4.3.2.2 System Implementation

Figure 4.7 shows an overall activity diagram for the system. In our system, in addition to the linguistic features, we use the results of the chemical NER tools for the CRF feature set. For the feature template, we use a template that is compatible with the CoNLL 2000 shared task and the Base-NP chunking task [76]. We use unigram, bigram, and trigram feature combinations. This template can handle a large number of features for one element. Table 4.6 shows an example of training data for CRF. The system uses CRF++ (Version 0.58) [64], an implementation of CRF, as a tool for the sequence-labeling task. The features of CRF++ are:

- Surface symbol: symbol used to represent a term.
- Part-of-speech (POS) tag: result from the GPoSTTL tagger (Version 0.9.3) [63].
- Lemmatization: symbol that is a result from the POS tagger.
- Orthogonal feature: identified using regular expressions based on the POS tag.
- SERB-CNER tag: output of the SERB-CNER system in IOB format.
- ChemSpot tag: output of ChemSpot (Version 1.5) [77] in IOB format.

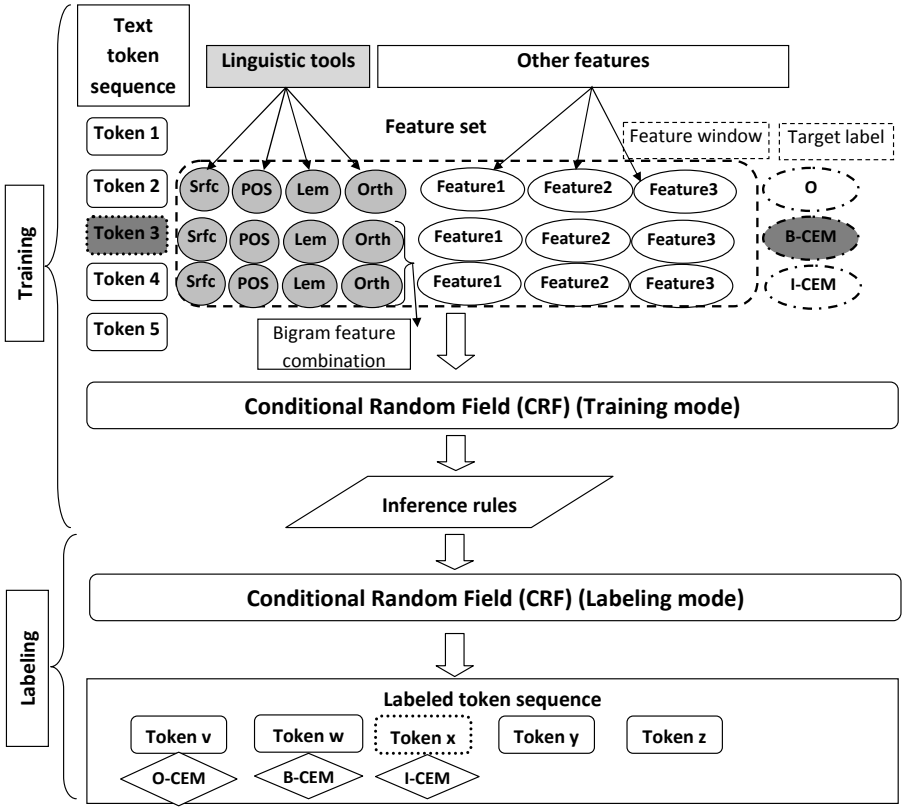


Fig. 4.6 Outline of the CRF model.
Srfc=token surface, POS=POS tag, Orth=orthogonal feature, Lem=lemmatization, CEM is target label.

– OSCAR4 tag: output of OSCAR4 in IOB format. (We use the output of OSCAR4-related ChemicalTagger (Version 1.3) [78].)

Orthogonal features are defined in [23]. The tokenization mechanism and merging chemical NER results are discussed in detail in Section 4.3.2.3. Confidence values for the extracted terms are calculated based on the CRF output. The confidence values for multiple terms are calculated by multiplying of confidence values for all values of "B" and "I".

4.3.2.3 Tokenization Mechanism

In a sequence-label task, a certain tool returns the labeling result as a feature of each token, word boundaries of the recognized named entities are defined by using tokenization results. In chemical NER, parts of long complex terms are often annotated as chemical named entities. However, because it is usually not necessary to analyze the inner structure of a term in a

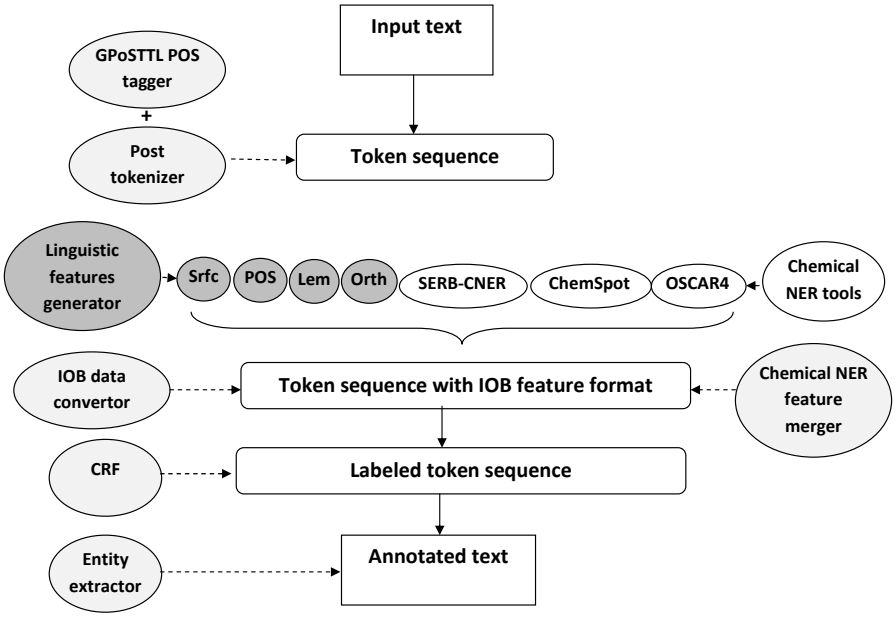


Fig. 4.7 A system overall activity diagram
POS=POS tag, Orth=orthogonal feature, Lem=lemmatization.

general POS tagging task, a general POS tagger (such as the GPoSTTL tagger) tends to treat such a long complex term as one token. Similar problems in the biomedical domain have already been discussed [79].

In this study, we aim to aggregate the results of different chemical NER tools. Depending on recognition schema, each chemical NER tool has its own text tokenizer. In many cases, these tokenization schemas are inconsistent. For example, in Abstract 23122060 of the BioCreative IV, CHEMDNER corpus, “d-glucose” is tokenized as one entity by the POS tagger and OSCAR4 tokenizer and labeled as a chemical by OSCAR4, whereas the ChemSpot tokenizer considers only “glucose” as a chemical entity. Because of this discrepancy, this result from ChemSpot cannot be matched with the POS tagger tokenization. Figure 4.8 illustrates this case of inconsistent tokenization.

To solve this problem, it is necessary to apply particular tokenization techniques to generate a greater number of tokens. This will achieve better labeling results, and has the advantage of being highly consistent [80].

We have analyzed the matching ratio between the tokenization of the text by GPoSTTL and chemical entities, and the drug-name boundaries in the annotation results of other chemical NER tools, including the “gold standard” manual annotation of the BioCreative IV, CHEMDNER corpus. The tokenization of GPoSTTL did not achieve a high matching ratio, particularly with the “gold standard” annotation of the BioCreative IV, CHEMDNER

Table 4.6 A sample training data for CRF

Token	POS	Lem	Orth	SERB-CNER	ChemSpot	OSCAR4	CEM
chemical	NN	chemical	Lowercase	O	O	O	O
compounds	NNS	compound	Lowercase	O	O	O	O
,	,	,	Comma	O	O	O	O
including	VVG	include	Lowercase	O	O	O	O
4	CD	4	DigitNumber	O	O	O	O
flavone	NNS	flavone	OtherHyphon	O	B	B	B-CEM
-	NNS	-	OtherHyphon	O	I	I	I-CEM
C	NNS	C	OtherHyphon	O	I	I	I-CEM
-	NNS	-	OtherHyphon	O	I	I	I-CEM
glycosides	NNS	glycosides	OtherHyphon	O	I	I	I-CEM

POS=POS tag, Orth=orthogonal feature, Lem=lemmatization, CEM=target label.

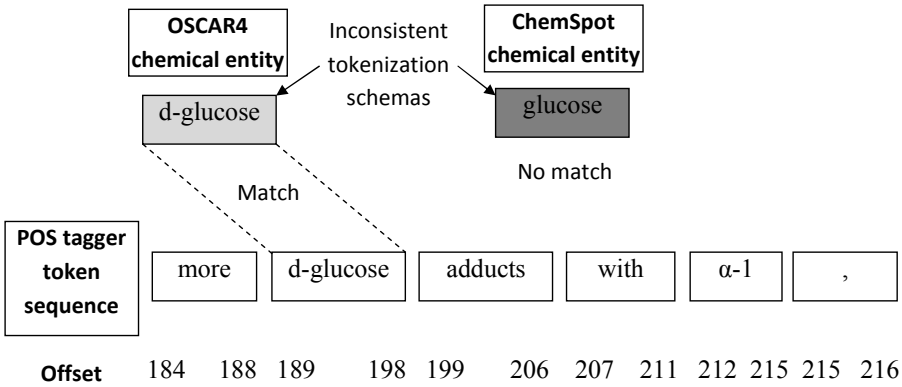


Fig. 4.8 Inconsistent tokenization schemas.

corpus. A low matching ratio between POS tagger tokens on the one hand and results from chemical NER tools and "gold standard" annotation on the other will cause inappropriately noisy training data. For example, unmatched results from chemical NER tools will not be labeled correctly (either unlabeled or loosely labeled as a chemical entity). Therefore, the performance will not be satisfactory [16].

To handle this issue, we analyzed the tokenization schemas of chemical entities and drug names annotated by the chemical NER tools including the "gold standard" annotation of the BioCreative IV, CHEMDNER corpus. We found that ChemSpot and the "gold standard" annotation of the BioCreative IV, CHEMDNER corpus tend to tokenize chunks of text containing "–", "+" and " / " into multiple tokens using these elements as separators. Because GPoSTTL does not decompose such chunks, the boundaries of chemical entities and drug names in the output of other chemical NER tools cannot be matched correctly. To solve this problem, it is necessary to generate a greater number of smaller tokens to adapt to such mismatches in tokenization schemas and to achieve better labeling results.

We implemented a post-tokenization mechanism for the POS tagger by adding three new tokenization rules for post-tokenization processing using the GPoSTTL tagger. We partitioned chunks containing "–", "+" and " / " into different tokens as for other chemical NER tools. We then checked the matching ratio after this post-tokenization process on the three datasets of the BioCreative IV, CHEMDNER corpus (training dataset, development dataset and test dataset). Table 4.7 shows the results of this analysis.

Table 4.7 Tokenization matching ratio analysis

	Training dataset		**Development dataset**		**Test dataset**	
	GPoSTTL	Post-tokenization	GPoSTTL	Post-tokenization	GPoSTTL	Post-tokenization
ChemSpot	0.93	1	0.92	1	0.92	0.99
OSCAR4	0.99	0.99	0.99	0.99	0.98	0.98
Gold standard	0.87	0.99	0.88	0.99	0.88	0.99

It is clear from table 4.7 that the matching ratio between the POS tagger tokens and the chemical entities and drug names boundaries has increased considerably. For the "gold standard" annotation data, we achieved a matching ratio of 0.99. This matching ratio has increased the performance of the overall system.

4.3.3 Experiments and Discussion

4.3.3.1 First experiment: Evaluation of the ensemble-learning approach and post-tokenization mechanism

The goal of the first experiment was to compare the performance of the ensemble-learning approach with a simple domain-adaptation approach that used only one chemical NER tool combined with CRF, on the BioCreative IV, CHEMDNER corpus. In addition, we wanted to check the effectiveness of post-tokenization on the performance. The BioCreative IV, CHEMDNER corpus contains three datasets (training, development and test). Each of the training and development datasets contains 3,500 abstracts, and the test dataset contains 3,000 abstracts.

We compared the system performance of the ensemble-learning approach before and after post-tokenization to evaluate the effectiveness of the post-tokenization process. We also compared the ensemble-learning approach with the results of a simple domain-adaptation approach that used CRF plus one of the chemical NER tools at a time (ChemSpot, OSCAR4 and SERB-CNER), together with post-tokenization. We used ten fold cross-validation on the combined three sets of the BioCreative IV, CHEMDNER corpus (training, development and test). In each fold, we used 90% of each of the three sets as training data and the remaining

10% as test data. We measured the performance using both macro- and micro-averages for precision, recall and F-score. The macro-average uses the performance of each abstract in the test dataset for calculating the average for all test data, whereas the micro-average uses all abstracts as one document for calculating the performance. Table 4.8 shows the macro-average and micro-average results for the ten fold cross-validation.

Table 4.8 Average system performance on the BioCreative IV, CHEMDNER corpus

	Macro-average			**Micro-average**		
	Precision	Recall	F-score	Precision	Recall	F-score
SERB-CNER+CRF	85.31	69.52	74.23	89.24	68.15	77.28
ChemSpot+CRF	85.26	76.77	78.84	88.10	76.21	81.72
OSCAR4+CRF	86.00	76.41	78.88	88.65	74.67	81.06
Ensemble	78.72	70.83	72.72	82.26	70.86	76.13
Ensemble/p.tok	86.62 $*	79.46 $*#	81.13 $*#	88.76 *	78.60 $*#	83.37 $*#

CRF: Conditional Random Field.
Ensemble = (SERB-CNER+ChemSpot+OSCAR4+CRF) without post-tokenization.
Ensemble/p.tok = (SERB-CNER+ChemSpot+OSCAR4+CRF) with post-tokenization.
Underlining indicates significant values for the ensemble system compared with the performance before post-tokenization. A dollar sign ($) indicates a significant value compared with SERB-CNER combined with CRF. An asterisk (*) indicates a significant value compared with ChemSpot combined with CRF. A hash (#) indicates a significant value compared with OSCAR4 combined with CRF. All significance measures were at the 0.05 level ($P < 0.05$).

Considering table 4.8, we can observe the following:

- Tokenization considerably affects the performance. Comparing the performance of the system before and after the post-tokenization process, it is clear that tokenization of text by the POS tagger can significantly affect the annotation of chemical entities and drug names. The use of a chemical-oriented POS tagger can improve the system performance because it can overcome some of the tokenization mismatches that can occur between normal text and chemical entities.

- Our system (the ensemble-learning approach with CRF) has, in general, obtained better F-scores than any of the simple domain-adaptation approaches. The system clearly outperforms the original chemical NER tools. There might be some discrepancies between the definitions of what is considered a chemical entity by different recognizers. However, we find that the use of orthogonal features has helped to reduce the effect of this problem by enabling the CRF system to learn rules that include both lexical and chemical tags. We found that ensembling different chemical NER tools with different

characteristics and different annotation criteria could leverage the performance because each tool can add new information to the system.

4.3.3.2 Second experiment: Use of the ensemble-learning approach for a well-tuned rule-based chemical NER

The goal of the second experiment was to check the ability of the ensemble-learning approach to leverage the performance of a well-tuned rule-based system for a specific task.

To investigate this effectiveness, we used one of the best performing rule-based chemical NER systems in the official BioCreative IV, CHEMDNER task, namely LeadMine [81]. LeadMine is a grammar- and dictionary-driven approach to chemical entity recognition. We asked the developer of LeadMine to provide the results data officially used for the BioCreative IV, CHEMDNER task and used this data for this experiment. In the experiment, we added the output of LeadMine as a feature, in addition to the features of the other chemical NER tools we discussed before. We used a ten fold cross-validation test on the BioCreative IV, CHEMDNER corpus. Because LeadMine was tuned using the training and development datasets of the corpus, it was not appropriate to use these datasets in the evaluation. Therefore, in each fold, we trained both systems (ensemble and LeadMine with CRF) on a combination of the full training and development datasets and 90% of the test dataset. We then tested the systems in each fold on 10% of the test dataset. Table 4.9 shows the macro-average and micro-average results for the ten fold cross-validation.

Table 4.9 Average system performance including LeadMine on the BioCreative IV, CHEMD-NER test dataset

	Macro-average			**Micro-average**		
	Precision	Recall	F-score	Precision	Recall	F-score
LeadMine+CRF	90.34	85.88	86.88	91.46	85.42	88.33
Ensemble/LeMi	<u>90.67</u>	85.97	<u>87.14</u>	<u>91.91</u>	85.63	<u>88.65</u>

CRF: Conditional Random Field.
Ensemble/LeMi = (SERB-CNER+ChemSpot+OSCAR4+LeadMine+CRF) with post-tokenization.
Underlining indicates significant values at the 0.05 level ($P < 0.05$).

From Table 4.9, it is clear that the ensemble-learning approach slightly leverages the performance of a rule-based system tuned for a specific task. Even though the improvement is small, it is statistically significant for precision and F-scores.

It is also clear that the ensemble-learning approach can help find new rules by checking terms that can only be extracted by the CRF. Analyzing the performance of LeadMine (a rule-

based system, and one of the best systems in the BioCreative IV, CHEMDNER task), we find that approximately 6% of the "gold standard" entities were recalled by the ensemble-learning approach with CRF but not with LeadMine. However, because we apparently also lost a different 6% of the "gold standard" entities, the recall stayed almost the same, while the precision improved. The CRF could identify some entities that would not be identified by the rule-based system. For example, in Abstract 22173956 in the chunk, "heterocyclic amines" is an entity in the "gold standard" annotation. However, the word " heterocyclic" was not identified by any chemical NER tool as a chemical entity or drug name, whereas "amines" was identified as such by all rule-based tools. The CRF enables us to identify such cases by learning them from the training dataset. Table 4.10 illustrates this case in IOB format.

Table 4.10 Gold standard entity recognized by CRF.

Tkn	B-POS	E-POS	POS	Lem	Orth	CNER	ChemSpot	OSCAR4	Lead	CEM
heterocyclic	469	481	JJ	heterocyclic	Lowercase	O	O	O	O	B-CEM
amines	482	488	NNS	amine	Lowercase	B	O	B	B	B-CEM

Tkn=token, B-pos=beginning of position, E-pos=end of position, POS=part-of-speech tag, Lem=lemmatization, Orth=orthogonal feature, CNER=SERB-CNER, Lead=LeadMine, CEM=gold standard.

4.3.3.3 Third experiment: System evaluation using the official BioCreative IV, CHEMD-NER test dataset

The goal of this experiment was to evaluate our final system in terms of the official test dataset of the BioCreative IV, CHEMDNER task. We also evaluated each chemical NER tool performance. The results of this experiment can be used as a reference for comparison between our system and other systems.

We trained the system on a combination of the training dataset and the development dataset provided by BioCreative IV, CHEMDNER corpus. We tested the system using different combinations of chemical NER tools with the layouts described above (linguistic features + chemical NER-tool combinations) using the official test dataset of the BioCreative IV, CHEMDNER corpus. Table 4.11 shows the performance of the various chemical NER systems on the official test dataset. For the LeadMine system, because we could only obtain the final output of the system, we show the performance as reported in the official BioCreative IV, CHEMDNER task [33]. For SERB-CNER, the performance, particularly for recall, is low because this tool uses only very simple rules to identify chemicals. These simple rules fail to generalize towards more chemical entities.

Table 4.11 Performance of different chemical NER systems for the official test dataset

System	Macro-average			Micro-average		
	Precision	Recall	F-score	Precision	Recall	F-score
SERB-CNER	23.79	11.37	13.42	43.95	11.26	17.93
ChemSpot	66.92	57.59	58.52	72.94	58.87	65.15
OSCAR4	42.71	62.88	47.34	40.66	62.08	49.13
LeadMine	87.25	81.41	82.72	89.25	81.48	85.19
Ensemble	87.36	78.17	80.68	89.41	77.47	83.01
Ensemble/LeMi	**90.09**	**85.09**	**86.34**	**91.52**	**84.85**	**88.06**

CRF: Conditional Random Field.
Ensemble = SERB-CNER+ChemSpot+OSCAR4+CRF.
Ensemble/LeMi = SERB-CNER+ChemSpot+OSCAR4+LedMine+CRF.

4.3.3.4 Discussion

We confirmed that a simple domain-adaptation method that uses linguistic features and one tool output for learning improves the performance of the automatic extraction. This result shows that it is better to use such a domain-adaptation method for chemical NER tasks that aim to extract new chemical-related entities.

We also confirmed that the ensemble-learning approach that uses multiple outputs of chemical NER tools further improves the performance and that the improvement is statistically significant. This result shows that consistent differences between target task guideline and each chemical tool can be used to construct new inference rules to add more precise annotation. In addition, the ensemble-learning approach can find new entities that a rule-based system tuned for a specific task cannot. Therefore, the ensemble-learning approach can be used to construct new rules that can be added to the rule-based system.

This approach can also be used in an expansion toward different kinds of chemical-entity-related domains in the future. For example, in the nanoinformatics domain, researchers use chemicals as source materials for their experiments. It is necessary to extract chemical entities in this domain when analyzing experimental results.

4.3.4 Summary

We have discussed an ensemble-learning approach that aggregates different chemical NER tools with different characteristics and different annotation criteria. This approach combines simple domain adaption and general ensemble-learning features. We confirmed that this approach is generally promising, because each chemical NER tool can contribute some unique new findings, thereby leveraging the performance. This approach can also be used in enhancing the performance of a well-tuned rule-based chemical NER system by providing

information to enable the creation of new rules. Finally, we have found that the text-tokenization method considerably affects the performance of the system.

Chapter 5

Utilization of the corpus information to support nanocrystal device development

5.1 Introduction

Several approaches can be conducted to utilize the extracted information by NaDevEx (discussed in chapter 4) to support nanocrystal device development. In this chapter, we discuss our preliminary experiments as an approach to fulfill this purpose. This work aims at clustering research papers based on different similarity metrics using NaDevEx. However, at the time of conducting these experiments, we used the latest version available at that time of NaDevEx [12]. We discuss other approaches to utilize the NaDevEx extracted information in section 6.2.

Usually researchers have to go through a trial and error process, modifying parameter settings for experiment several times before they can reach the most convenient settings that can yield to the desired final product. This trial and error process is time and money consuming. Research papers contain summarization for several nanocrystal device development experiments and evaluation about experiment results. Research papers can be used by novices to find similar experiments done before to help them plan their new experiment more effectively.

Different similarity measures can be used to study the similarity between papers. Clustering research papers based on the similarity of their content can allow us to study similarity measures and their effect on papers' similarity. We propose to use the automatic information extraction framework [12] in the process of calculating similarity between papers. Automatic information extraction system can provide additional similarity measures, hence deeper analysis on the study of similarity, especially when categories of information are playing

different roles in determining the similarity between research papers. Automatic information extraction system can tell which information category is more important in finding similarity.

In order to discuss the effectiveness of automatic information extraction on the similarity measure analysis, we are conducting two clustering experiments on the same set of papers, one without automatic information extraction, and the other with automatic information extraction using the framework we proposed.

There are different ways to cluster research papers based on similarity; However, in this chapter, we study one way using bag-of-words approach. Other ways of clustering might also be studied and analyzed in the future.

5.2 Papers similarity

There are many ways to find similarity between two papers. However, in any way, we need to transform the paper into representative model to be able to calculate similarity between it and another paper. We use bag-of-words approach as a model for papers to find similarity. Bag-of-words approach is a simplifying representation of text documents as unordered vector of words and their frequencies.

As we have mentioned in section 5.1, we have conducted two experiments, the first one is on non-annotated papers (base system), and in that case, we transformed each research paper (non-annotated) into a bag-of-words model. The other one is on annotated papers, and in this case, we transformed each paper (annotated) into an array of vectors; each vector contains a bag-of-words representation of all chunks annotated under certain information category. We have eight categories of information that has been annotated. In addition to that, we added extra category called "Other", that is non-annotated chunks of text within the annotated paper. In total, we have nine categories of information. Each annotated paper is transformed into an array of nine vectors; each one is a bag-of-words representation of tags under certain information category.

In this study, we use weighted cosine similarity metric. Cosine similarity metric is given by equation 5.1

$$similarity = \frac{a \cdot b}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} (a_i \times b_i)}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \times \sqrt{\sum_{i=1}^{n} (b_i)^2}} \tag{5.1}$$

Where
a, b are document vectors.

In the case of non-annotated paper, we use whole bag-of-words vector as a document vector for calculating similarity. However; in case of annotated paper, where paper is segmented into 9 vectors of bag-of-words, each one has different weight, there are different ways to calculate weighted cosine similarity. In this study, we have encoded 2 ways to calculate weighted cosine similarity:

- Long vector encoding: we construct a long vector that concatenates 9 vectors with weight that represents importance of each information category
  $a = (\alpha_1 a_1, \alpha_2 a_2, ..., \alpha_9 a_9)$
  Where
  a_i and α_i represent bag-of-words vector of a document and weight for i-th category (SMaterial, SMChar, MMethod, TArtifact, ExP, EvP, ExPVal, EvPVal, and Other). Similarity is calculated using equation 5.1.

- All sum encoding: we calculate cosine similarity for bag-of-words vectors for every category including "Other" and summed all similarity with weight as follows:

$$similarity = \sum_{i=1}^{n} \alpha_i \frac{a_i \cdot b_i}{\|\mathbf{a_i}\| \cdot \|\mathbf{b_i}\|} \tag{5.2}$$

Weighted cosine similarity can allow us to neglect or emphasize certain category of information based on the importance it plays in determining similarity between papers.

5.3 Experiments

5.3.1 Experiment setup

We conducted nanodevice paper clustering experiments by using conference proceedings [82], and session categories for these papers are used for the correct category labels (classes). We pick up 5 sessions (A-E) and select 32 papers from each session (total 160 papers). All papers are annotated using our automatic annotation framework. We have used hierarchal clustering technique [83] with R language [84] to perform both experiments. Hierarchal clustering seeks to build a hierarchy of clusters, each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. There are different methods to merge the observations on the way up depending on the distance between them. We have tested them, and "complete linkage" method, that uses the maximum distance seems to perform best. We cut the clustering result tree into 5 levels representing 5 clusters, and then we compare the 5 resulting clusters with the original 5 classes to determine the quality

of clustering. We evaluate the clustering quality using the entropy and purity measures. Entropy measures how the conceptual classes are distributed within each cluster. The smaller the entropy the better the clustering is. Purity is the fraction of the cluster size that the largest class of chunks assigned to that cluster represents. The larger the purity, the better the clustering is.

5.3.2 Base system (non-annotated paper clustering)

We performed the first experiment using non-annotated papers. Figure 5.1 shows the resulting clusters. Entropy and purity were 0.28 and 0.38 respectively.

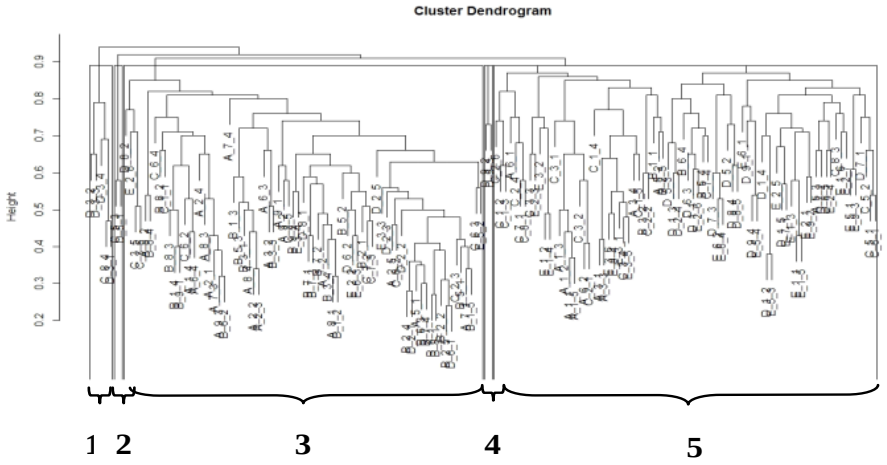


Fig. 5.1 Hierarchal clustering result for non-annotated papers

Analyzing figure 5.1, we find that, it is unbalanced structure of 5 clusters. There are 2 big clusters and 3 very small ones. However; deep look onto the distribution of papers within clusters, we found that A and B session papers are mostly clustered in cluster 3, and session E papers are mostly clustered in cluster 5. C and D session papers are distributed mostly between cluster 3 and 5. Because of the varieties of papers in both cluster 3 and 5, entropy and purity are not good enough.

5.3.3 Annotated paper clustering

As we discussed in section 5.2, we have 2 ways to encode the paper vector for calculating similarity, which are long vector and all sum. In each way of encoding, we can apply different weights for different information categories, and observe the quality of the clustering to find the best clustering strategy.

Table 5.1 shows the entropy and purity for different weighting strategies in both long vector and all sum encoding.

Table 5.1 Clustering performance for annotated papers

[SM, SMC, MM, TA, EP, Ev, EPV, EvV, O]	**Entropy**		**Purity**	
	Long vector	All sum	Long vector	All sum
Base system (non-annotated papers)	**0.28**		**0.38**	
[1,1,1,1,1,1,1,1,1]	0.27	0.28	0.39	0.31
[1,1,1,1,1,1,1,1,0]	0.30	0.29	0.27	0.32
[1,1,1,1,1,1,0,0,1]	0.27	0.30	0.37	0.26
[10,10,10,10,10,10,10,10,1]	0.28	0.29	0.34	0.31
[1,1,1,1,10,10,0,0,1]	0.27	0.29	0.40	0.33
[1,10,1,1,10,10,0,0,1]	**0.26**	0.29	**0.40**	0.29
[1,20,1,1,20,20,0,0,1]	0.28	0.30	0.34	0.28

SM: SMaterial, SMC: SMChar, MM: MMethod, TA: TArtifact, EP: ExP, Ev: EvP, EPV: ExPVal, and EvV: EvPVal are from the tag set. O: Other class.

5.3.4 Results analysis

We have conducted various weight settings to find out effective categories to calculate similarity. Since the clustering result of [10,10,10,10,10,10,10,10,1] is worse than one of the base system, we assume it is better to select useful categories for similarity calculation. We conducted several experiments and found following three categories are more effective to calculate similarity; Characteristic Feature of Material (SMChar), Experiment Parameter (ExP), and Evaluation Parameter (EvP). In addition, it is better to ignore values of parameters (ExPVal and EvPVal). Figure 5.2 shows the hierarchal clustering of best performed weight setting.

Analyzing best performance hierarchal clustering after annotation, we found it has almost the same structure as non-annotated papers clusters, keeping also 2 big clusters: 4, and 5, where cluster 4 has the bulk of sessions A and B papers, and cluster 5 has the bulk of E session papers. However, there are some documents moved to smaller clusters making 2 small clusters bigger in size and almost pure. Even though the clustering structure and clustering performance were not much better than the basic system, but this means the automatic annotation might have some good effect on the quality of clustering. If the automatic annotation quality increases, it might increase the quality of the clustering; However, we cannot confirm that based only on these results. It is necessary to do more experiments with larger data size and more balanced clustering structure.

Considering the entropy and purity values from table 5.1, we can find the following notes:

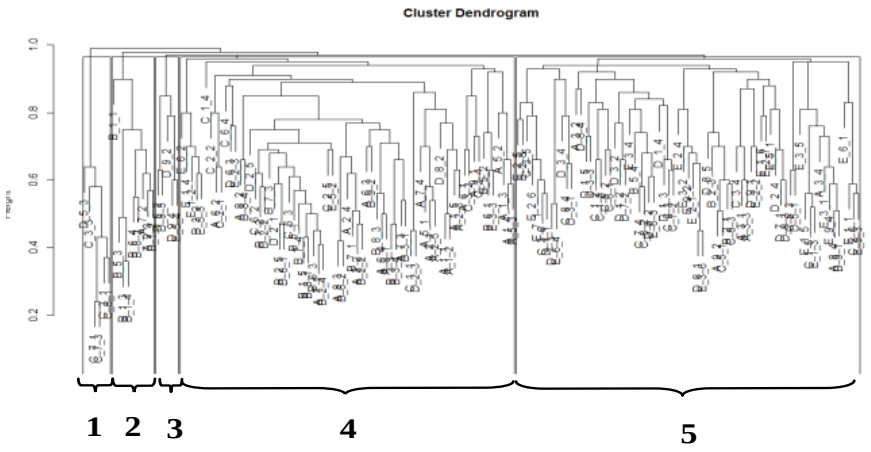
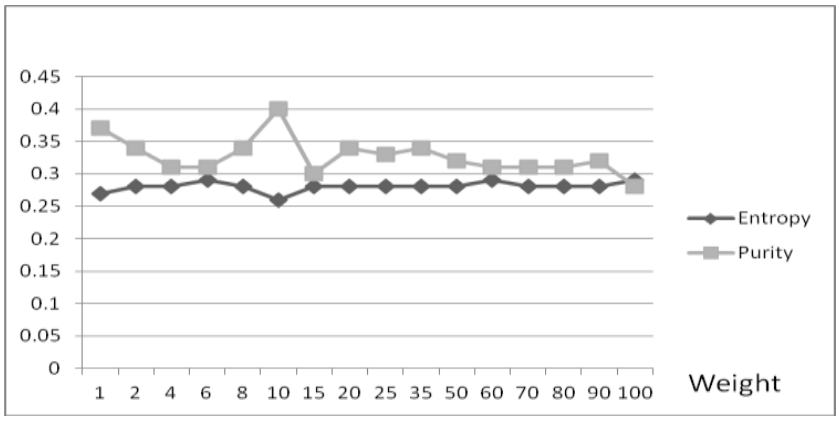


Fig. 5.2 hierarchal clustering results for [1,10,1,1,10,10,0,0,1]

- Encoding method considerably affect the clustering quality, noticing that the best performed weight setting in long vector encoding did not do well in the all sum encoding. Long vector encoding generally performed better.
- The "Other" category seems to play significant role in similarity, and that is because automatic annotation quality is still not good enough.
- Increasing weights of effective information categories does not always increase the quality of the clustering. Figure 5.3 shows the relation between clustering quality and weight in long vector encoding best performance weights array.



Weight for SMChar, ExP, EvP in [1,x,1,1,x,x,0,0,1] weight array

Fig. 5.3 Weight vs. performance in long vector encoding

5.4 Summary

In this chapter, we have conducted nanocrystal device research paper clustering experiments based on similarity of the content. We have conducted 2 experiments, one before annotating the papers using automatic information extraction framework, and the other one after the annotation. We discussed the effect of automatic annotation on similarity measures analysis.

This technique, can be used in the future to discover relations between different parameters, specially, experiment and evaluation parameters. Such relations can be very useful in determining the effect of a certain parameter change on the quality of the final product. In addition to that, different clustering approaches can be discussed.

Chapter 6

Conclusion and future works

6.1 Conclusion

In this study, we have developed a framework for experimental information extraction from research papers to support nanocrystal device development. We have constructed an annotated corpus called "NaDev" (Nanocrystal Device Development corpus) in collaboration with a domain expert and domain graduate students. This corpus contains entities annotated in eight information categories related to nanocrystal device development experiments. NaDev then used to build a cascading style automatic information extraction system using machine learning and natural language processing techniques. Automatic information extraction system was finalized as NaDevEx. NaDevEx uses linguistic and domain knowledge features, such as chemical entity recognition, and physical quantities list. We found that, NaDevEx has almost not much defeated by the human annotators for source material information, because of the good performance of the chemical named entity recognition system. The NaDev construction guideline was released, NaDev can be distributed upon request [11].

Since there is a significant amount of chemical information in publication related to nanocrystal device development represented as source material information for experiments, recognition of chemical entities might be supportive to identify source material information. Therefore, we have developed a chemical named entity recognition (chemical NER) system using ensemble learning approach.

Extracted information using this framework can be used to analyze experiments results to provide insight for novice researchers to achieve a more efficient planning for their experiments.

To discuss the novelty and appropriateness of our designed tag set, it is necessary to consider related efforts to extract information from research papers. Nanoinformatics is considered to be at the intersection of bioinformatics, computational chemistry, and

nanobiotechnology. There have been several attempts to extract chemical or nanomedicine-related information from research papers [19, 32]. However, as discussed in Chapter 2, these efforts have not focused on extracting the information necessary to analyze experimental results. By contrast, our tag set is designed in collaboration with a domain expert to support the extraction of experimental information.

Because it is costly to conduct new experiments to obtain new experimental data in nanotechnology, several approaches tried to share such information [42, 46, 47]. The extraction of experimental information is supposed to be applied as a preprocessing step for such shared data construction in nanoinformatics. Our preliminary work [10, 12, 13] is recognized as one of the main efforts in applying natural language processing to extract such information [85].

During our corpus-construction process, we found that it is not easy to design the tag set before conducting actual annotation experiments. To overcome reliability-related issues, we have developed the two-step annotation method. This method can support the construction of new corpora in new domain.

6.2 Future work

- Corpus development: Several issues might be considered for further development of the corpus, as follows:

  - Corpus size: NaDev corpus uses full text of research papers instead of abstracts that are commonly used for constructing such corpora. Abstracts usually do not contain detailed explanation about experiments' parameters in relation with output evaluation. It was necessary to extract such information to analyze experiments' results adequately. However, abstract can offer wider variety of experimental information. I plan to extend the corpus using large number of abstracts. The abstract should also be diverse and represent different sub-domains in nanocrystal research. Additionally, abstract should include the two types of research text: synthesis and characterization. The increase in size and diversity will add more reliability to the corpus that can support a wider range of nanocrystal research.

  - Inter-entity relations: In bioinformatics, the relation between entities such as event annotations is considered important. For example, in the GENIA corpus, such information is well represented [35]. By contrast, the annotation of relations between entities such as parameters and their values is not a requirement for this particular domain. Such annotation can be handled as a general task. The

NaDev construction focused on the identification of basic entities to simplify the annotation process. However, it might be preferable to add the annotation of relations in its future development.

- Automatic information extraction development: Experiments showed that preprocessing annotation based on domain knowledge is generally promising, but coverage of the parameter information based on a list of physical quantities was not enough for nanocrystal device papers. There are many compound terms that contain particular domain-specific terms within their boundaries for characterizing categories. I plan to construct resources for representing domain knowledge, such as a list of common final products and list of common parameters in nanocrystal device development. These resources will enhance the performance of the automatic information extraction system.

- Utilization of extracted information The information of this framework can be utilized to support nanocrystal development in several ways:

  - Paper retrieval system: We will be able to construct a paper-retrieval system for a nanocrystal device development portal by using these information categories. For example, the user could find papers that involve MnAs as a source material in developing nanoclusters as a target artifact. Information such as this would be helpful in finding research papers that contain the results of recent analyses of particular types of experiments and would support the data-collection process.

  - Papers' clustering: These annotation results can be further used to find similarity between research papers based on different similarity metrics. For example, similarity biased towards certain information categories of interest for researchers (e.g. source material, final product, and so on) rather than similarity based on the general content of the paper (e.g. bag of words approach). Such flexible similarity can support researchers towards a more efficient experiment planning based on insight from similar experiment settings reported in research papers.

  - Graph-based parameter change detection: Figures represent important resources for researchers in nanocrystal devices. Researchers usually start checking figures when studying a paper to find relations between parameters. It is helpful for researchers in nanocrystal domain to find figures from several research papers that discuss relations between certain parameters in a certain experiment settings. For example, density and height of the nanocluster. Figures captions contain descriptive text to describe parameters represented in the figure. Using the annotation framework we have developed, we will be able to annotate captions

of figures. This annotation can be used in a figure retrieval system to find similar figures discuss relation between same parameters.

References

[1] Kriegel, I., and Scotognella, F.:Tunable light filtering by a Bragg mirror/heavily doped semiconducting nanocrystal composite. *Beilstein J. Nanotechnol.*, 6, 193-200. DOI:10.3762/bjnano.6.18. (2015).

[2] Davydova, M., Kulha, P., Laposa, A., Hruska, K., Demo, P., and Kromka, A.:Gas sensing properties of nanocrystalline diamond at room temperature. *Beilstein J. Nanotechnol.*, 5, 2339-2345. DOI:10.3762/bjnano.5.243. (2014).

[3] Capan, I., Carvalho, A., and Coutinho,J.:Silicon and germanium nanocrystals: properties and characterization. *Beilstein J. Nanotechnol.*, 5, 1787-1794. DOI:10.3762/bjnano.5.189. (2014).

[4] Fukui, T., Ando, S., Tokura, Y., and Toriyama, T.:GaAs tetrahedral quantum dot structures fabricated using selective area metalorganic chemical vapor-deposition. *Appl. Phys. Lett.*, 58, 2018-2020. (1991).

[5] Sasa, S., Yano, M., Maemoto, T., Koike, k., and Ogata, K.:High-performance ZnO-based FETs and growing applications of oxide semiconductor devices. *J. IEICE.*, 95, 4, pp.289–293. (2012).

[6] Yatsui, T., Morigaki, F., and Kawazoe, T.: *Beilstein J. Nanotechnol.*, 5, 1767-1773. Doi:10.3762/bjnano.5.187. (2014).

[7] Ikejiri, K., Sato, T., Yoshida, H., Hiruma, K., Motohisa, J., Hara, S., and Fukui, T.:Growth characteristics of GaAs nanowires obtained by selective area metal–organic vapour-phase epitaxy. *Nanotechnology*, 19: 265604-1-8. (2008).

[8] Ruping,K., and Sherman, B. W.: Nanoinformatics: Emerging computational tools in nanoscale research. *Proceedings of NSTI-Nanotech*, Boston, Massachusetts, USA, Mar. 2004, ,Volume 3, pp. 525-528 (2004).

[9] De la Iglesia, D., Cachau, R. E., Garcia-Remesal, M., and Maojo, V.:Nanoinformatics knowledge infrastructures: bringing efficient information management to nanomedical research. *Comput. Sci. Disc.*, 6 01401. DOI:10.1088/1749-4699/6/1/014011. (2013).

[10] Dieb, T.M., Yoshioka, M., and Hara, S.: Construction of tagged corpus for Nanodevices development papers. *Proceedings of International Conference on Granular Computing (GrC)*, Kaohsiung, Taiwan; Nov., pp 167-170. (2011).

[11] Dieb, T.M., Yoshioka, M., and Hara, S.: NaDev (Nanocrystal Device development) Corpus Annotation Guideline. *TCS Technical Reports*, TCS-TR-B-15-12., July 2015. Hokkaido University, Division of Computer Science. (2015).

[12] Dieb, T.M.; Yoshioka, M. ; Hara, S. Automatic Information Extraction of Experiments from Nanodevices Development Papers. *Proceedings of International Conference on Advanced Applied Informatics (IIAIAAI)*, Fukuoka, Japan; 20-22 Sep. 2012, pp. 42-47. (2012).

[13] Dieb, T.M.; Yoshioka, M. ; Hara, S.; Newton, M. C.: Automatic Annotation of Parameters from Nanodevice Development Research Papers. *Proceedings of the 4th International Workshop on Computational Terminology Computerm 2014*, Dublin, Ireland; 23 Aug. 2014, pp. 77-85. (2014).

[14] Dieb, T.M.; Yoshioka, M. ; Hara, S.; Newton, M. C.:Framework for automatic information extraction from research papers on nanocrystal devices. *Beilstein J. Nanotechnol.*, 6, 1872–1882. (2015).

[15] Dieb, T. M., Yoshioka, M., and Hara, S.: Knowledge Exploratory Project for Nanodevice Design and Manufacturing: Knowledge Discovery from Experimental Records (3rd Report) -Nanodevice Research Papers Clustering based on Automatic Paper Annotation. *Proceedings of 27th annual meeting of the Japanese society of artificial intelligence (jsai2013)*, Toyama, Japan. CD-ROM 1C3-4.(2013).

[16] Yoshioka, M., and Dieb, T.M. :Ensemble Approach to Extract Chemical Named Entity by Using Results of Multiple CNER Systems with Different Characteristic. *Proceeding of the fourth BioCreative challenge evaluation workshop*, vol. 2, (2013).

[17] Dieb, T. M., Yoshioka, M.: Extraction of Chemical and Drug Named Entities by Ensemble Learning Using Chemical NER Tools Based on Different Extraction Guidelines. *Trans. on machine learning and data mining*. Vol. 8, No. 2 pp. 61-76. (2015).

[18] GENIA Project, Tsujii Laboratory, University of Tokyo. Available from (http://www.nactem.ac.uk/genia/).

[19] Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J.:GENIA Corpus-semantically annotated corpus for bio-textmining. *Bioinformatics*, 19, i180-i182. (2003).

[20] Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N.:Introduction to the Bio-Entity Recognition Task at JNLPBA.*Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04)*, pp.70-75, (2004).

[21] Kim, J. D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J.: Overview of BioNLP'09 Shared Task on Event Extraction. *Proceedings of the Workshop on BioNLP: Shared Task*, pages 1–9. (2009).

[22] Kim, J. D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J.: Extracting Bio-Molecular Events From Literature – The BioNLP'09 Shared Task. *Computational Intelligence*, Vol. 27, No. 4, pages 513-540. (2011).

[23] Takeuchi, K. and Collier, N.: Bio-medical entity extraction using support vector machines. *Artif. Intell. Med.*, 33, 2, 125-137. (2005).

[24] Gaizauskas, R., Demetriou, G., Artymiuk, P. J., and Willett, P.:Protein Structures and Information Extraction from Biological Texts: The PASTA System. *J. Bioinformatics.*, 19, 1, 135-143. (2003).

[25] Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning* 20 (3): 273. (1995).

[26] Lafferty, J.D., McCallum, A., and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01*, San Francisco, CA, USA; 2001, pp. 282-289. (2001).

[27] Kim, J. D., Ohta, T., Oda, k., and Tsujii, J.: From text to pathway: corpus annotation for knowledge acquisition from biomedical literature. *Proceedings of the 6th Asia Pacific Bioinformatics Conference, Series on Advances in Bioinformatics and Computational Biology*, Vol. 6, pp. 165-176. (2008).

[28] SCAI corpora: http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/research-development/ information-extraction-semantic-text-analysis/named-entity-recognition/chem-corpora.html.

[29] IUPAC http://www.iupac.org/.

[30] Jessop, D., Adams, S., Willighagen, E., Hawizy, L.,Murray-Rust, P. :OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics* 3, 41, (2011).

[31] Rockt aschel, T., Weidlich, M., and Leser, U.:ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28, 1633–1640, (2012).

[32] BioCreative IV CHEMDNER corpus. Available from (http://www.biocreative.org/resources/corpora/bc-iv-chemdner-corpus/).

[33] Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M, Sayle, R.A., Batista-Navarro, R.T., Rak, R., Huber, T., Rocktaschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K.H., Ramanan, S.V, Nathan, S., zitnik, S., Bajec, M,. Weber, L., Irmer, M., Akhondi, S.A., Kors, J.A., Xu, S., An, X., Sikdar, U.K., Ekbal, A., Yoshioka, M., Dieb, T.M., Choi, M., Verspoor, K., Khabsa, M., Giles, C.L., Liu, H., Ravikumar, K.E., Lamurias, A., Couto, F.M., Dai, H., Tsai, R.T., Ata, C., Can, T., Usie, A., Alves, R., Segura-Bedmar, I., Martinez, P., Oyarzaba, J., Valencia, A. J. :The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(Suppl 1):S2. Doi:10.1186/1758-2946-7-S1-S2. (2015).

[34] Tomoko, O., Tateisi, Y., Mima, H., and Tsujii, J.:The GENIA corpus: an annotated research abstract corpus in molecular biology domain. *Proceeding of The Human Language Technology Conference (HLT 2002)*, San Diego, USA, Mar. 2002, pp. 82-86. (2002).

[35] Kim, J. D., Ohta, T., and Tsujii, J.:Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10, 2008. DOI:10.1186/1471-2105-9-10. (2008).

[36] De la Iglesia,D., Harper,S., Hoover, M. D., Klaessig, F., Lippell, P., Maddux, B., Morse, J., Nel, A, Rajan, K., Reznik-Zellen, R., and Tuominen,M.: Nanoinformatics 2020 roadmap. National Nanomanufacturing Network. Amherst, MA 01003. Available from (http://eprints.internano.org/607/1/Roadmap_FINAL041311.pdf) DOI: 10.4053/rp001-110413. (2011)

[37] Gonzalez-Nilo, F., Perez-Acle, T., Guinez-Molinos, S., Geraldo, D. A., Sandoval, C., Yevenes, A., Santos,L. S., Laurie, V. F., Mendoza,H., and Cachau, R. E.:Nanoinformatics: An emerging area of information technology at the intersection of bioinformatics, computational chemistry and nanobiotechnology. *Biol. Res.* 44, 43-51, (2011).

[38] Garcia-Remesal, M., Garcia-Ruiz, A., Perez-Rey, D., De la Iglesia, D., and Maojo, V.:Using nanoinformatics methods for automatically identifying relevant nanotoxicology entities from the literature. *Biomed. Res. Int.*, article ID 410294. DOI: 10.1155/2013/410294 (2013).

[39] Harper, S. L., Hutchison, J. E., Baker, N., Ostraat, M., Tinkle, S., Steevens, J., Hoover, M. D., Adamick, J., Rajan, K., Gaheen, S., Cohen, Y., Nel, A., Cachau, R. E., and Tuominen, M.:Nanoinformatics workshop report: current resources, community needs and the proposal of a collaborative framework for data sharing and information integration. *Comput. Sci. Disc.*, 6 014008. DOI:10.1088/1749-4699/6/1/014008. (2013).

[40] Kozaki, K., Kitamura, Y., and Mizoguchi, R.: Systematization of nanotechnology knowledge through ontology engineering - A trial development of idea creation support system for materials design based on functional ontology. *Poster notes of ISWC2003*, Sanibel Island, Florida, USA, Oct. 2003, pp. 63-64. (2003).

[41] Thomas, D. G., Pappu, R. V., and Baker, N. A.:NanoParticle ontology for cancer nanotechnology research. *J Biomed Inform.*, 44(1): 59-74. (2011).

[42] Guzan, K. A., Mills, K. C., Gupta, V., Murry, D., Scheier, C. N., Willis, D. A., and Ostraat, M. L., Integration of data: the Nanomaterial Registry project and data curation. *Comput. Sci. Disc.*, 6, 014007. DOI:10.1088/1749-4699/6/1/014007. (2013).

[43] Madhavan, K., Zentner, L., Farnsworth, V., Shivarajapura, S., Zentner, M., Denny, N. and Klimeck, G.:nanoHUB.org: Cloud-based Services for Nanoscale Modeling, Simulation, and Education. *Nanotechnology Reviews.*, 2, 1, 107-117. (2013).

[44] Integrated Nanoinformatics Platform for Environmental Impact Assessment of Engineered Nanomaterials. Available from (http://nanoinfo.org/).

[45] Liu, R., Hassan, T., Rallo, R. and Yoram, C.:HDAT: web-based high-throughput screening data analysis tools. *Comput. Sci. Disc.*, 6, 014006. (2013).

[46] DaNa project. Available from (http://www.nanoobjects.info/en/).

[47] Kimmig, D., Marquardt, C., Nau, K., Schmidt, A., and Dickerhof, M.:Considerations about the implementation of a public knowledge base regarding nanotechnology. *Comput. Sci. Disc.*, 7, 014001. DOI:10.1088/1749-4699/7/1/014001. (2014).

[48] Gaheen, S., Hinkal, G. W., Morris, S. A., Lijowski, M, Heiskanen, M., and Klemm, J. D.:caNanoLab: data sharing to expedite the use of nanotechnology in biomedicine *Comput. Sci. Disc.*, 6, 014010. (2013).

[49] Xiao, L., Tang, K., Liu, X., Yang, H., Chen, Z. and Xu, R.: Information extraction from nanotoxicity related publications. *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 25-30. (2013).

[50] Jones, D. E., Igo, S., Hurdle, J., and Facelli, J. C.:Automatic extraction of nanoparticle properties using natural language processing: NanoSifter an application to acquire PAMAM dendrimer properties. *PLoS One.*, 2; 9(1):e83932. (2014).

[51] Yoshioka, M., Tomioka, K., Hara, S., and Fukui, T.: Knowledge exploratory project for nanodevice design and manufacturing. *Proceedings of iiWAS 10*, Paris, France, Nov.2010, pp. 869-872. (2010).

[52] Yoshimura, M., Tomioka, K., Hiruma, K., Hara, S., Motohisa, J., and Fukui, T.:Growth and characterization of InGaAs nanowires formed on GaAs (111) B by selective-area metal organic vapor phase epitaxy. *Jpn. J. Appl. Phys.*, 49 (4): 04DH08-1-5. (2010).

[53] Di Eugenio, B., and Glass, M.:The Kappa Statistic: A Second Look. *Computational Linguistics*, Vol. 30, No. 1, pp.95-101. (2004).

[54] Hara, S., Motohisa, J., and Fukui, T.:Self-assembled formation of ferromagnetic MnAs nanoclusters on GaInAs/InP (1 1 1) B layers by metal-organic vapor phase epitaxy. *J. Cryst. Growth.*, 298: 612-615. (2007).

[55] Green, A. M.: Kappa statistics for multiple raters using categorical classifications. *Proceedings of the 22ed Annual SAS Users Group International Conference*, San Diego, CA, Mar. 1997. pp 1110-1115. (1997).

[56] Hara, S., and Fukui, T.:Hexagonal ferromagnetic MnAs nanocluster formation on GaInAs/InP (111) B layers by metal–organic vapor phase epitaxy. *Appl. Phys. Lett.*, 89: 113111. (2006).

[57] Ito, S., Hara, S., Wakatsuki, T., and Fukui, T.:Magnetic domain characterizations of anisotropic-shaped MnAs nanoclusters position-controlled by selective-area metal-organic vapor phase epitaxy. *Appl. Phys. Lett.*,94,243117. DOI: 10.1063/1.3157275. (2009).

[58] Hara, S., Kawamura, D., Iguchi, H., Motohisa, J., and Fukui, T.:Self-assembly and selective-area formation of ferromagnetic MnAs nanoclusters on lattice-mismatched semiconductor surfaces by MOVPE. *J. Cryst. Growth.*, 310, 7, 2390-2394. DOI: 10.1016/j.jcrysgro.2007.12.026. (2008).

[59] Wakatsuki, T., Hara, S., Ito, S., Kawamura, D., and Fukui, T.:Growth Direction Control of Ferromagnetic MnAs Grown by Selective-Area Metal–Organic Vapor Phase Epitaxy. *Jpn. J. Appl. Phys.*, 48, 04C137. (online) DOI:10.1143/JJAP.48.04C137. (2009).

[60] XConc Suite. Available from (http://www.nactem.ac.uk/genia/tools/xconc).

[61] Kano, Y., Miwa, M., Cohen, K. B., Hunter, L. E., Ananiadou, S., and Tsujii, J.: U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, vol. 55, no. 3, pp. 11:1-11:10, (2011).

[62] Nakagawa, H., and Mori, T.: Automatic term recognition based on statistics of compound nouns and their components, *Terminology*, Vol. 9, No. 2, pp. 201-219, (2003).

[63] GPoSTTL: http://gposttl.sourceforge.net.

[64] CRFpp: http://taku910.github.io/crfpp/.

[65] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Yon Rhee, S. : Big data: The future of biocuration. *Nature*, 455, 47-50, (2008).

[66] Otsuki, A., Kawamura, M. :The Study of the Role Analysis Method of Key Papers in the Academic Networks. *Trans. on machine learning and data mining*. Vol. 6, No.1, 3-18 (2013).

[67] Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M.:ChEBI: a database and ontology for chemical entities of biological interest. *Nucl. Acids Res.* 36 (suppl 1): D344-D350. Doi:10.1093/nar/gkm791. (2008).

[68] Zhou, G., Shen, D., Zhang, J., Su, J., and Tan, S. :Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*, 6, 1–7, (2005).

[69] Blitzer, J., McDonald, R., and Pereira, F. : Domain adaptation with structural correspondence learning. Proc. the 2006 Conference on Empirical Methods in Natural Language. *Processing (EMNLP '06), Association for Computational Linguistics*, Stroudsburg, PA, USA, 120-128, (2006).

[70] Dimililer, N., Varoğlu, E., and Altınçay, H. :Classifier subset selection for biomedical named entity recognition. *Applied Intelligence*, 31, 3, pp 267-2825, (2009).

[71] Zhou, H., Li, X., Huang, D., Yang, Y., and Ren, F. :Voting-Based Ensemble Classifiers to Detect Hedges and Their Scopes in Biomedical Texts. *IEICE TRANSACTIONS on Information and Systems*,E94-D,10, pp.1989-1997, (2011).

[72] Ekbal, A., and Saha, S. :Multiobjective optimization for classifier ensemble and feature selection: an application to named entity recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 15, 2, pp 143-166, (2012).

[73] Kolarik, C., Klinger, R., Friedrich, C.M., Hofmann-Apitius, M., and Fluck, J. :Chemical names: terminological resources and corpora annotation. *Proceeding of the Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Marrakech, Morocco, pp. 51–58, (2008).

[74] Klinger, R., and Tomanek, K.: Classical probabilistic models and conditional random fields. *Technical Report TR07-2-013. Department of Computer Science, Dortmund University of Technology*; ISSN 1864-4503, (2007).

[75] McDonald, R., and Pereira, F.:Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics.* 6(Suppl. 1):S6, (2005).

[76] CoNLL 2000 http://www.cnts.ua.ac.be/conll2000/chunking/.

[77] Chemspot1.5: http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/chemspot/chemspot/.

[78] chemicalTagger-1.3: http://chemicaltagger.ch.cam.ac.uk/.

[79] Yeh, A., Morgan, A., Colosimo, M., and Hirschman, L. :BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2. Doi:10.1186/1471-2105-6-S1-S2. (2005).

[80] Leaman, R., and Gonzalez, G. :BANNER: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput*, 2008:652-63. (2008).

[81] Lowe, D.M., and Sayle, R.A. :LeadMine: A grammar and dictionary driven approach to chemical entity recognition. *Proceeding of the fourth BioCreative challenge evaluation workshop*, vol. 2, (2013).

[82] Extended Abstract of the 2008 International Conference on Solid State Devices and Materials, Tsukuba, (2008).

[83] Gothai, E.: Performance evaluation of hierarchical clustering algorithms, *Proceedings of 2010 International Conference on Communication and Computational Intelligence (INCOCCI)*, pp.457-460, (2010).

[84] R-project: http://www.r-project.org/

[85] Lewinski, N. A., and McInnes, B. T.: Using natural language processing techniques to inform research on nanotechnology. *Beilstein J. Nanotechnol.*, 6, 1439-1449. DOI:10.3762/bjnano.6.149 (2015).

Appendix A

NaDev corpus constructing guideline

INTRODUCTION

This guideline aims to help manual annotating of NaDev (Nanocrystal Device development) corpus by using nanocrystal device development papers. Extracted information will be used for analyzing contents of the paper as index terms with specific semantic category by using this corpus as training data for machine learning automatic annotation tool.

Followings are categories annotated in this corpus.

A) Source Material Information (SMaterial)
Material Information: Example: As, InGaAs, TMG, …
B) Material Characteristics (MChar)
Information about the feature characteristics features of the materials:
Example: (111) B, minor axes,...
C) Experimental parameters (ExP)
Control parameter to characterize the attribute value: Example: total pressure …
D) Value of experimental parameters (ExPVal)
The specific value of the above: Example: 50 to 200nm,
E) Evaluation parameters (EvP)
Attribute value used when in the analysis process: Example: PL peak energy, FWHMs, …
F) Value of the evaluation parameter (EvPVal)
Attribute value used in the analysis process: example: 1.22-1 .25 eV, ...
G) Manufacturing method (Mmethod)
Techniques and method for creating nanostructures: Example: SA-MOVPE, VLS, …
H) Final product or Artifact (TArtifact)
final product: semiconductor nanowires, metal-semiconductor field-effect transistors, …

Annotation

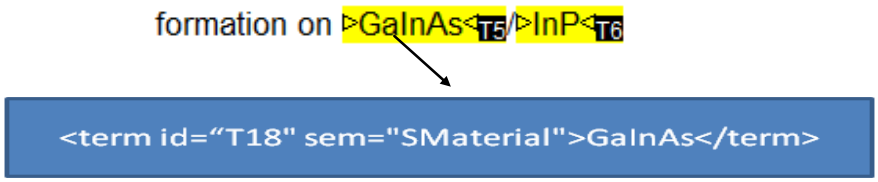
The annotator can use annotation support tool Xconc suite [1] that was originally developed originally to annotate biomedical entities in GENIA project [2]. XCnoc uses XML to represent the annotation.

Below is an example of the annotation

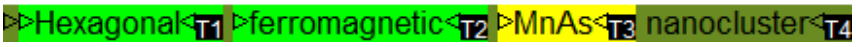
Unannotated text

the growth of GaInAs and InP layers

Annotated text



In case of overlapped terms (multiple layers of marking), the annotation will look like below



Detailed explanation about each category

➢ Material Information (SMaterial)

Defines source input materials of experiments. Below are some notes need to be considered when annotating Material Information.

Example

Hydrogen

- Source materials of current experiment should always be annotated as material information even if they are the results of a previous experiment discussed in different paper.

- When there is a compound material of 2 or more materials the compound should be annotated as material
  Compound materials such GaInAs should be annotated as one material although it is a mix of Ga, In, and As

Example

the growth of GaInAs

- In some cases, experiments start with input material then develop these materials into different materials, new one will be used to develop the final experiment output in such a case, these materials developed in the middle are still considered as source materials.

Example

starting with Zn and then during the experiment we get ZnO and use it to achieve some nanowire

- Sometimes source material falls within some parameter and can't be separated from it. In that case, the name of the material is annotated in a nested way inside the parameter.

Example

estimated p[$(MeCp)_2Mn$]
V/Mn = p[AsH_3]/p[$(MeCp)_2Mn$].

- In some cases the source material comes in a certain form like wafer or fine particle, in this case this form should be annotated as part of the input material

Example

InP (001) wafers

- Groups of source materials that classify source materials based on some attributes like groups V and III-V are classified as source materials

Example

on the III–V compound

group III

➢ Material Characteristics (MChar)

Defines electro-chemical characteristics of material.

- Features describing electro-chemical characteristics of material that is used to make certain final product should be annotated as electro-chemical characteristics

Example

ferromagnetic MnAs nanoclusters

hexagonal NCs

➢ Experiment Parameter and Experiment Parameter Value (ExP,ExPVal)

- Sometimes the parameter value is not stated concretely, however it is referenced to some value that not stated concretely within the text, in that case we also consider it as parameter value

Example

at room temperature

- If experiment parameter value and the experiment parameter can be separated from each other, we should simply annotate them as separate terms

Example

low growth temperatures
increasing p[(CH3C5H4)2Mn]

However when the experiment parameter comes within its value and can't be separated, it should be annotated as inner term within the value

Example

increasing V/Mn ratios in a supply gas from 60 to 750

with increasing growth temperatures from 550 to 700 °C

increasing the MnAs growth time from 3 to 30 min on

applied magnetic fields in a direction perpendicular to the wafer planes (out-of-plane),

➢ Evaluation Parameter and Evaluation Parameter Value (EvP.EvPVal)

- Things that help determine whether the results were satisfying or not like atomically flat crystal facets are considered as evaluation parameters.
- Evaluation parameter usually related to the analysis process of the final product, and of general criteria of interest or related to the purpose of the experiment.

- Common examples: facet, magnetic, direction, diameter, height, crystal structure, nanocluster shape, surface, crystallographic structure….etc.

Example

top surfaces of the MnAs NCs are atomically flat

However, in some cases these common examples might not refer to an evaluation parameter

Example

self-assembled on planar GaInAs surfaces

- if evaluation parameter value and the evaluation parameter can be separated from each other, we should simply annotate them as separate terms

Example

strong ferromagnetic coupling

- Modifiers of parameter value (much, less…) should be included in the same term as the value

much smaller numbers of the MnAs nanoclusters were formed on the surfaces when the V/Mn ratios were low.

➢ Method (Mmethod)

- A method term is a full name of a manufacturing method or an abbreviation of that method

Example

metal-organic vapor phase epitaxy

MOVPE

buildup fabrication and magnetic domain characterizations

- We should separate the method name from the experiment parameters used in it if possible

Example

The MOVPE growth conditions.

- Varieties of a method (sub method) should be annotated as method

Example

SA-MOVPE, LT-MBE

- Usually a method is a way of manufacturing that the experimenter chooses to apply for certain reasons, and not a natural outcome of an experiment.

Example

ZnO fine particles deposited on silica surfaces

deposited is NOT a method here

➢ Final product (TArtifact)

- When a final product is combined with the name of some material, the product name should be annotated as product; however the

material name within that product should also be annotated as overlapped term inside the first term.
Example
We fabricated InGaAs nanowires (NWs)

- Layers are not considered as final products

Example

GaInAs/InP (1 1 1) B layers,

➢ General comments

- In some cases, the paper discuses some previous experiments, these discussion should also be annotated based on that experiment perspective (input materials for that experiment are annotated as input materials, final products for that experiment are annotated as final products and so on)

- An abbreviation of a text should be annotated as the same class of the text itself

Example

of the NWs increased when the growth temperature was

- Words like above, below, under, around, at, from, to, between, increased from, decreased from… that helps indicating parameter values should be included as part of this value if they proceed the value.

Example

above 675 ºC

from 300 to 50 nm

whereas their density is increased from 10^7 to 10^8 cm^{-2}. It is increased from 66 to 319nm

- "The" should not be included in the annotation of the term

Example

The heights

- When a text is followed by an abbreviation of that text, we annotate the text separately from the abbreviation

Example

metal-organic vapor phase epitaxy (MOVPE),

tri-methyl-gallium (TMGa)

- We don't include parenthesis () in the annotation

Example

(TMGa)

- In case of [A of C], [A and B of C] or [A for B and C] and such cases, the general idea is to simply annotate separately each one of A, B and C if they can be separated from each other semantically and not necessary to be compound.

  Example

  magnetic domain characterizations of anisotropic-shaped MnAs nanoclusters

  The temperature of $(MeCp)_2Mn$

averaged height and density of the nanoclusters on the V/Mn ratios.

The estimated partial pressures for TMIn and TMGa

However, If A and C used to describe terms of the same type, or it is not possible to separate them, we don't break them and annotate them as one term

Example

aspect ratio of the initial mask openings

numbers of the MnAs nanoclusters

References

[1] XConc Suite. http://www.nactem.ac.uk/genia/tools/xconc\

[2] GENIA Corpus. http://www.nactem.ac.uk/aNT/genia.html

Appendix B

Inter Annotators Agreement Calculation

Calculating Inter Annotators Agreement (IAA)

Please refer to R. Artstein, Quality control of corpus annotation through reliability measures, ACL-2007 tutorial [http://ron.artstein.org/publications/2007-acl-t5-slides.pdf] for more details.

In all of the following example, we use A, B and C to refer to term category.

- Observed agreement:

Proportion of items on which 2 annotators agree.

Example

	A	B	Total
A	41	3	44
B	9	47	56
Total	50	50	100

Agreement: $\frac{41+47}{100} = 0.88$

- Agreement above chance

Some agreement is expected by chance alone.

- Two coders randomly assigning “A” and “B” labels will agree half of the time.
- The amount expected by chance varies depending on the annotation scheme and on the annotated data.

Meaningful agreement is the agreement above chance.

Example

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

Agreement=88/100, due to chance: 12/100, above chance: 76/100

The meaningful agreement ratio (agreement above chance) formula will be

$$\frac{A_0 - A_e}{1 - A_e} \qquad (1)$$

Where A_0: is the observed agreement

A_e: agreement by chance

- ✓ How to calculate the amount of agreement expected by chance (A_e)?

We use the following formula

$$A_e^S = q.(\frac{1}{q})^2 = \frac{1}{q} \qquad (2)$$

Where q: number of category labels.

As an example

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

A_0=0.88, Ae=0.5, $S = \frac{0.88-0.5}{1-0.5}$=0.76

- ✓ Since not all categories are equally alike, we use this formula -that takes into consideration different chance for different categories- to calculate agreement by chance

$$A_e^{\pi} = \sum_q (\frac{n_q}{N})^2 = \frac{1}{N^2} \sum_q n_q^2 \quad (3)$$

Where

N: Total number of judgments

nqa /N : Probability of one coder picking a particular category qa.

And the agreement above chance based on the new formula will be

	A	B	C	Total
A	44	6	0	50
B	6	44	0	50
C	0	0	0	0
Total	50	50	0	100

A_0=0.88, $S = \frac{0.88-1/3}{1-1/3}$=0.82, π=$\frac{0.88-0.5}{1-0.5}$=0.76

- ✓ Different annotators have different interpretations of the instructions, so we update the formula to calculate agreement by chance to be as follow

$$A_e^k = \sum_q \frac{n_{c_1 q}}{i} \frac{n_{c_2 q}}{i} = \frac{1}{i^2} \sum_q n_{c_1 q} n_{c2q} \quad (4)$$

Where
i: Total number of items.

$\frac{n_{c_x q_a}}{i}$: Probability of annotator c_x picking a category q_a.

$\frac{n_{c_1 q_a}}{i} \frac{n_{c_2 q_a}}{i}$: Probability of both annotators picking a category q_a.

And the agreement above chance will be

	A	B	C	Total
A	38	0	12	50
B	0	12	0	12
C	0	0	38	38
Total	38	12	50	100

$A0=0.88, \pi \approx 0.7995, K \approx 0.8018$

K measure is called Kappa statistics coefficient, and it is the measurement we used to calculate Inter Annotators Agreement.