Journal of Information Processing Vol.24 No.3 554–564 (May 2016)

[DOI: 10.2197/ipsjjip.24.554]

Regular Paper

NaDev: An Annotated Corpus to Support Information Extraction from Research Papers on Nanocrystal Devices

Thaer M. Dieb[1,a)] Masaharu Yoshioka[1] Shinjiro Hara[2]

Received: May 22, 2015, Accepted: January 12, 2016

Abstract: The process of nanocrystal device development is not well systematized. To support this process, analysis of the information produced by developmental experiments is required. In this study, we constructed an annotated corpus to support the extraction of experimental information from relevant publications. We designed the corpus-construction guidelines by cooperating with a domain expert. We evaluated these guidelines through corpus-construction experiments with graduate students from this domain, and then evaluated the corpus with the domain expert. In the corpus construction experiments, we achieved a sufficient level of Inter-Annotator Agreement by using a loose agreement measure that ignored the term-boundary mismatch problem, and made an agreement corpus that excluded annotations based on misunderstanding the guidelines. The domain expert evaluated this agreement corpus and modified the guidelines based on real examples. Using these guidelines, we finalized the corpus called "NaDev" (Nanocrystal Device development corpus). The NaDev corpus and its construction guidelines will be released via our website, http://nanoinfo.ist.hokudai.ac.jp/. The NaDev corpus aims to support automatic information extraction from publications relevant to nanocrystal device development. This information can be used to solve problems in the nanotechnology domain using the massive availability of fresh information. To the best of our knowledge, this is the first corpus constructed for the development of nanocrystal devices.

Keywords: nanoinformatics, annotation, corpus construction, information extraction, nanocrystal device development

1. Introduction

Nanoinformatics is a newly developing interdisciplinary research domain that aims to use information technology to support research in the nanoscience field [1], [2]. Nanoinformatics is 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 [3]. Alternatively, nanoinformatics has been defined as an emerging area of information technology at the intersection of bioinformatics, computational chemistry, and nanobiotechnology [4]. Nanoinformatics could play the same role in nanotechnology and nanomedicine as bioinformatics and medical informatics have played in biology and medicine [5].

Nanocrystal device development is an area of nanoscale research where nanoelectronic devices are developed for future nanoelectronic industrial applications using electronic materials such as semiconducting, insulating, and magnetic materials [6], [7], [8], [9], [10]. However, the process of nanocrystal device development is not well systematized, requiring both engineering knowledge and craftsmanship [11]. We have been conducting a project called the "Knowledge Exploratory Project for Nanodevice Design and Manufacturing" to support the nanocrystal device development process [12]. This is a joint research project between the Research Center for Integrated Quantum Electronics (RCIQE) and the Division of Computer Science at Hokkaido University. As part of this project, we want to exploit information about the development of nanocrystal devices that is reported in research publications. This information would be used to facilitate a more effective development process through various applications, including but not limited to experimental result analysis.

In this study, we have developed a method for constructing an annotated corpus of publications relevant to nanocrystal device development to support automatic information extraction. The tag set and the corpus-construction guidelines were designed in collaboration with a domain expert. We evaluated the reliability of these guidelines through corpus construction experiments with graduate students from this domain. We evaluated the constructed corpus using Inter-Annotator Agreement (IAA) and confirmed that the guidelines achieved a satisfactory IAA level. We also constructed an agreement corpus that excluded incorrect annotations based on misunderstanding the guidelines. The domain expert evaluated this agreement corpus and modified the guidelines by checking them with real annotation examples. Based on these modified guidelines, we finalized the corpus called "NaDev" (Nanocrystal Device development corpus) and its construction guidelines for an official release.

Several attempts have been conducted to extract information

[1] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan

[2] Research Center for Integrated Quantum Electronics, Hokkaido University, Sapporo, Hokkaido 060–8628, Japan

[a)] diebt@kb.ist.hokudai.ac.jp

related to nanoinformatics. However, such efforts have not focused on extracting the information necessary to analyze experimental results. Our tag set is designed to support the extraction of experimental information. To the best of our knowledge, this is the first attempt to construct a corpus for the development of nanocrystal devices.

This paper has six additional sections. Section 2 reviews related research. Section 3 contains a discussion of the corpus construction approach. In Section 4, we describe our method and experiments for constructing the corpus. Section 5 presents the corpus evaluation experiments using the domain expert. In Section 6, we discuss the corpus release and usage. Finally, Section 7 concludes this paper.

2. Related Work

One of the very important initiatives to roadmap the nanoinformatics domain was Nanoinformatics 2010 [3], a collaborative roadmapping and workshop project at which informatics experts, nanotechnology researchers, and other stakeholders and potential contributors collaborated to develop a roadmap for the domain.

There have been several attempts to learn how informatics can be used to advance nanomanufacturing. For example, the Greener Nano 2012: Nanoinformatics Tools and Resources Workshop [13] aimed at establishing a better understanding of state-of-the-art approaches to nanoinformatics and clearly defining the immediate and projected informatics infrastructure needs for the nanotechnology community. De la Iglesia et al. also discussed the needs and challenges, as well as extant initiatives and international efforts, in the field [2].

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 [14], aiming to fill the gap between materials and devices by establishing common concepts across various domains. They also aimed at building a creative design-support system using systematized knowledge. Another approach aimed at developing a nanoparticle ontology to represent knowledge underlying the preparation, chemical composition, and characterization of nanomaterials involved in cancer research [15]. This approach focused on nanoparticles related to cancer research, and is therefore insufficiently general. Other researchers are working on developing databases and repositories for nanomaterials and their related applications. For example, the Nanomaterial Registry [16] uses information about nanomaterials to support a robust data curation process that promotes integration across a diverse data set. Another example is the DaNa project [17], which provides information about products and applications related to nanomaterials, aiming to illuminate their health and environmental aspects. Based on the DaNa project, some researchers are trying to capture knowledge at a higher semantic level in a database called DaNaVis, which increases the accessibility of the DaNa project results by means of interactive visualization components [18]. However, these studies did not consider the analysis of experimental results. It is necessary to analyze such results to support the effective planning of experiments.

The use of literature in the nanotechnology domain is currently oriented toward the nanomedical field, focusing 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 [19]. This portal offers information related mainly to the biomedicine domain. However, because nanomaterials can be used in other domains such as nanoelectronics, the need for general knowledge about nanodevice development experiments is becoming more widespread.

The extraction of information from research publications using text-mining techniques is a growing trend in various areas, particularly the bioinformatics research domain. Researchers can build large-scale corpora using text-mining approaches, such as the GENIA corpus [20]. The definition of suitable corpora can help overcome problems related to the massive availability of information in fields such as molecular biology. For example, the annotation of proteins and protein-related events can help to assemble protein–protein interactions from different publications.

There are well-defined corpora that have been established in other domains, such as the BioCreative IV CHEMDNER corpus [21]. However, because such a corpus contains only chemical information, it is therefore oriented toward solving problems related to the chemistry domain.

To the best of our knowledge, there are no well-established approaches for constructing corpora related to nanotechnology research, and there have been only a few attempts to construct such corpora. For example, Garca-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 [5]. Jones et al., using a natural language processing technique, tried to extract numeric values for the biomedical properties of poly (amidoamine) dendrimers from the nanomedicine literature [22]. However, the information used to construct these corpora is insufficiently general, being oriented toward nanomedicine.

3. Corpus Construction Approach in Bioinformatics

3.1 Introduction

An approach that uses information extraction from research publications has several advantages as a data collection process for the nanoinformatics domain. It can utilize the freshness and massive availability of information in research publications, thereby facilitating 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. For example, this information could be used to find similarities between previous experiments and planned experiments to enable a more effective experiments design. A well-defined corpus is essential to support such an information-extraction process.

No previous studies have extracted information from publications on nanocrystal device development. We therefore decided to employ applicable techniques from the bioinformatics research domain. We consider the GENIA corpus as a model of such corpora in bioinformatics [23].

3.2 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 [24].
- Event annotation: GENIA corpus event annotation marks expressions describing 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, such as a theme or a cause [25].
- Relation annotation: GENIA corpus relation annotation aims to complement event annotation in the 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, particularly capitalized names and abbreviations such as NFAT, CD4, and RelB, which make it difficult to distinguish between proper nouns and common nouns.

4. Corpus Construction Process

4.1 Background and Motivation

Previously, we have built an experimental record management system to analyze the results of experiments related to the development of nanocrystal devices [12]. Using pattern-mining techniques on this system, we found that different sets of parameters were used to form the same layer structure. Parameter settings would be decided depending on experimental motivation and evaluation criteria, which were not available in the system. The information stored in the system was insufficient for detailed analysis. Therefore, additional information would be required to ensure adequate analysis.

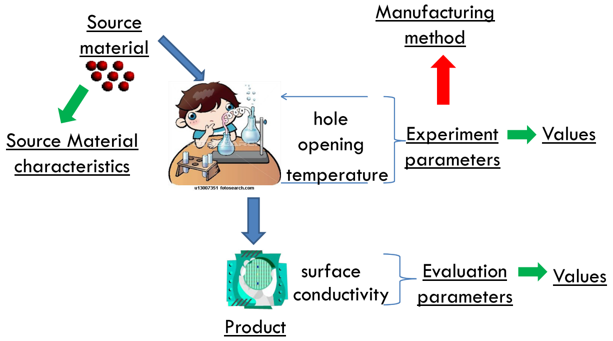


Fig. 1 Information categories used in nanocrystal device development experiments.

We conducted discussions with researchers in the field of nanocrystal devices, who suggested that two additional information sources could be used to obtain the necessary information. The first source suggested was a set of research notes related to experiments. However, this approach would require extra work by the nanotechnology researchers, who might not be available at the appropriate time (such as graduate students who had 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 the background information. The second source suggested was a set of research publications relevant to the experiments. Because research publications are written after a series of experiments, they often contain complete descriptions of the motivation, purpose, and other relevant information. Researchers in the field of nanocrystal devices recommended using these publications as a source from which to extract the necessary information. In our approach, we constructed an annotated corpus to extract the necessary information from relevant research publications.

4.2 Tag Set Design

To extract information from research publications, it is necessary to identify 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 1** shows the experimental abstract model.

In their experiments, researchers usually employ source materials such as a gas or MnAs, where each source material has specific characteristics such as the distinctive group of that material in the periodic table. The experimental conditions can be controlled by adjusting experimental parameters such as the temperature and pressure. However, because different development methods may use different sets of experimental parameters, a set of parameters may be relevant only to a particular development method. An experiment yields a final product, namely 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. Therefore, 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 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 material, 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 for 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 for 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.

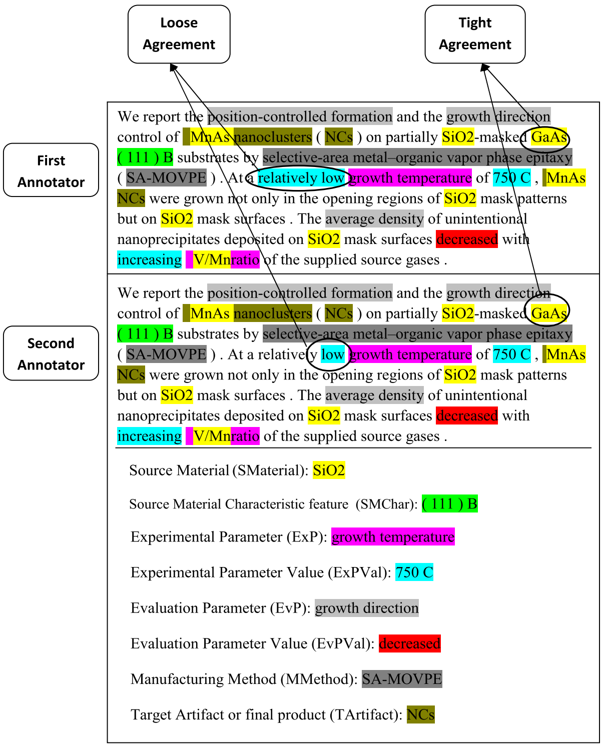
4.3 Corpus Construction Guidelines

Before we constructed the corpus, it was necessary to specify the corpus construction guidelines. To construct these guidelines, we asked two graduate students from RCIQE to annotate the same publication [26] 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 guidelines for annotating research publications. This draft was progressively improved as more papers were annotated. In addition, the guidelines were checked by an expert researcher in nanocrystal device development. The annotation was implemented 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 different ways. Therefore, it was necessary to check the reliability of the corpus construction guidelines.

4.4 Reliability Measures

To evaluate the quality of the corpus construction guidelines, we used reliability to represent the accuracy of the annotated information, which is the likelihood of extracting all of the requisite information. Therefore, reliability represents consistency in this case. We checked the reliability of the corpus using the IAA for two different annotators based on the kappa coefficient [27]. The kappa coefficient is a robust measure because it takes into consideration the possibility that the agreement may occur by chance.



We report the position-controlled formation and the growth direction control of MnAs nanoclusters (NCs) on partially SiO2-masked GaAs (111) B substrates by selective-area metal–organic vapor phase epitaxy (SA-MOVPE) . At a relatively low growth temperature of 750 C , MnAs NCs were grown not only in the opening regions of SiO2 mask patterns but on SiO2 mask surfaces . The average density of unintentional nanoprecipitates deposited on SiO2 mask surfaces decreased with increasing V/Mnratio of the supplied source gases .

We report the position-controlled formation and the growth direction control of MnAs nanoclusters (NCs) on partially SiO2-masked GaAs (111) B substrates by selective-area metal–organic vapor phase epitaxy (SA-MOVPE) . At a relatively low growth temperature of 750 C , MnAs NCs were grown not only in the opening regions of SiO2 mask patterns but on SiO2 mask surfaces . The average density of unintentional nanoprecipitates deposited on SiO2 mask surfaces decreased with increasing V/Mnratio of the supplied source gases .

Source Material (SMaterial): SiO2

Source Material Characteristic feature (SMChar): (111) B

Experimental Parameter (ExP): growth temperature

Experimental Parameter Value (ExPVal): 750 C

Evaluation Parameter (EvP): growth direction

Evaluation Parameter Value (EvPVal): decreased

Manufacturing Method (MMethod): SA-MOVPE

Target Artifact or final product (TArtifact): NCs

Fig. 2 Corpus example illustrating tight and loose agreement.

However, the annotation of a text using the proposed tag set requires some consideration of the term-boundary mismatch problem. To separate the issues of term-category selection and term-boundary identification, we developed two different evaluation criteria for the analysis. They are "tight agreement," which considers term boundaries, and "loose agreement," which ignores the term-boundary problem. **Figure 2** illustrates the difference between tight and loose agreements in a corpus example.

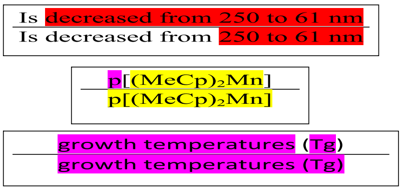
4.5 Corpus Construction Experiments

We asked the same two graduate students to annotate the same publication independently [28] according to the guidelines, 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% for tight agreement, and 74% for loose agreement. According to Green (1997) [29], high agreement (sufficiently reliable agreement) requires a kappa coefficient of ≥ 0.75. The results of the first experiment showed that the annotation was not quite sufficiently reliable for loose agreement and definitely inadequate for tight agreement. It was therefore necessary to improve the guidelines to resolve the mismatches between annotators and check the reliability again.

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, and 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** shows examples of term-boundary mismatches that occurred between the two annotators in the first experiment.

Based on these results, we revised the guidelines and conducted a second annotation experiment using four research papers [30], [31], [32], [33]. In this experiment, the corpus-annotation support tool XConc Suite [34], which was originally developed for constructing the GENIA corpus [20], 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 the IAA. In this experiment, the IAA was 0.63 for tight agreement and 0.77 for loose agreement. For this second experiment, based on Green (1997) [29], the guidelines with loose agreement now achieved sufficient reliability. **Table 1** and **Table 2** show the experimental results for the tight and loose agreement ratios, respectively.

Some disagreements were caused by careless mistakes or misunderstanding of the guidelines by one of the students and were solved after discussion with the students. We can confirm that the new guidelines and the corpus-annotation support tool improved the quality of the annotation.



Is decreased from 250 to 61 nm
Is decreased from 250 to 61 nm

p[(MeCp)$_2$Mn]
p[(MeCp)$_2$Mn]

growth temperatures (Tg)
growth temperatures (Tg)

Fig. 3 Examples of term-boundary mismatches between the two annotators. Each box is an example of two annotators' annotation of the same text.

Table 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 is the Other class of unannotated text (or terms with boundary mismatches that prevent tight agreement). T is the total.

Table 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 is the Other class of unannotated text (or terms with boundary mismatches that prevent tight agreement). T is the total.

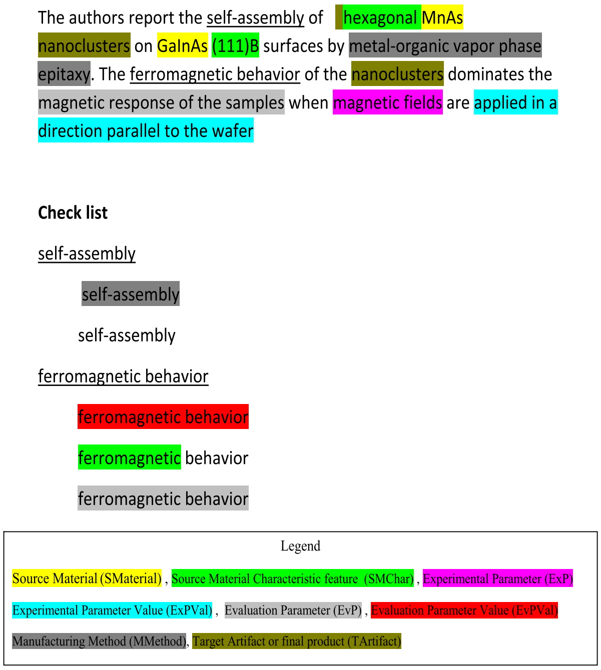
5. Corpus Evaluation Experiments with a Domain Expert

5.1 Experiment Setup

In the previous two experiments, we constructed a corpus using graduate students. Even though the corpus construction guidelines reached a reliable level in the case of loose agreement, it remained 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 guidelines.

From 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 guidelines. We therefore asked him to evaluate the results of the previous corpus-construction experiments.

The evaluation data was 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 failing to add an annotation, and typical types of disagreement, such as one of the annotators misunderstanding the guidelines. 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 as that used for representing the corpus. For the disagreed annotations, we underlined the related text and provided the students' annotation candidates to the domain expert (In some cases, we provided additional annotation candidates resulted from discussion with students). **Figure 4** shows an example of the evaluation experiment data.



The authors report the self-assembly of hexagonal MnAs nanoclusters on GaInAs (111)B surfaces by metal-organic vapor phase epitaxy. The ferromagnetic behavior of the nanoclusters dominates the magnetic response of the samples when magnetic fields are applied in a direction parallel to the wafer

Check list

self-assembly

self-assembly

self-assembly

ferromagnetic behavior

ferromagnetic behavior

ferromagnetic behavior

ferromagnetic behavior

Legend

Source Material (SMaterial), Source Material Characteristic feature (SMChar), Experimental Parameter (ExP)

Experimental Parameter Value (ExPVal), Evaluation Parameter (EvP), Evaluation Parameter Value (EvPVal)

Manufacturing Method (MMethod), Target Artifact or final product (TArtifact)

Fig. 4 Example of the evaluation experiment data.

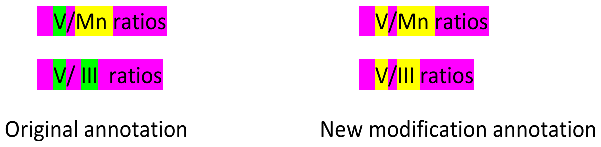


Fig. 5 Different representations of ratios between source materials.

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, suggest a new candidate.
- Annotate any terms that have not been annotated.

5.2 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 [28]. In this experiment, we spent almost one hour evaluating the annotation results for the paper, including discussion of the corpus-construction guidelines. Because there was no specific problem with the experimental setup, the second step involved an experiment that used the other four papers [30], [31], [32], [33]. This required almost two hours, again including discussion of the corpus-construction guidelines. 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 [28], [30], [32], [33] focus on the synthesis of new nanomaterials.
- Characterization papers: Paper 5 [31] 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 guidelines 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 guidelines:

- 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 to which it belongs (e.g., Group III or V), are used for representing a group of source materials. For example, the ratios among source materials and/or groups 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 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 6** shows an example of these 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 specify these characteristics as inherited from the source materials or resulting from the manufacturing process.

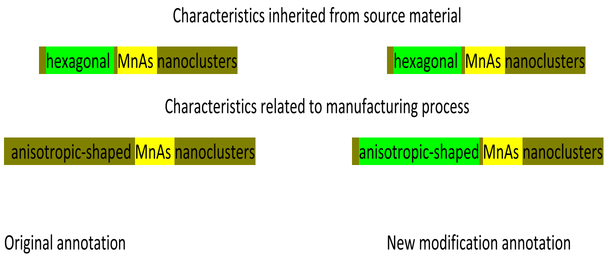

Characteristics inherited from source material
hexagonal MnAs nanoclusters
hexagonal MnAs nanoclusters
Characteristics related to manufacturing process
anisotropic-shaped MnAs nanoclusters
anisotropic-shaped MnAs nanoclusters
Original annotation
New modification annotation

Fig. 6 Different sources for the final product characteristics.

We 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 the characterization paper, we provide separate comparisons for synthesis and characterization papers to characterize the differences between these two types of paper. **Table 3** and **Table 4** show the comparison matrices between the domain-expert corpus and original corpus for synthesis papers and the characterization paper, respectively. We calculated the precision and recall for each category. We also calculated the precision and recall when excluding the effects of guidelines modifications.

Table 3 and Table 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 guidelines modifications. It is therefore important to have discussions among the annotators after the annotation process. Such discussions can resolve mismatches caused by careless mistakes or misunderstanding of the guidelines. Recall is also high (ranging between 91% and 100%). However, because disagreed annotations caused by ambiguity were separated from the agreed annotations in the original corpus (as prepared for the evaluation experiment), it was 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 annota-

Table 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 guidelines modifications. Rec: Recall. Prec: Precision (Numbers in parentheses represent recall and precision excluding mismatches caused by guidelines modifications).

Table 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 guidelines modifications. Rec: Recall. Prec: Precision (Numbers in parentheses represent recall and precision excluding mismatches caused by guidelines modifications).

Table 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 the candidate list (Numbers in parentheses represent terms and coverage when excluding mismatches caused by modifications to the guidelines).

Table 6 Analysis of disagreed annotations in the characterization paper.

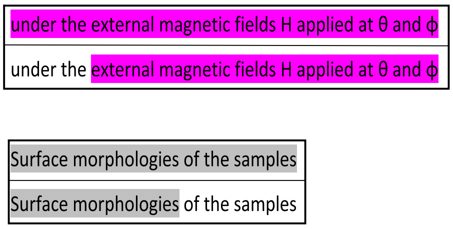
	SM	MC	MM	TA	EP	Ev	EPV	EvV	T
Total	12(9)	24(8)	2(0)	13(0)	10(0)	18(0)	2(0)	20(0)	101(17)
Candidate	3	4	1	8	1	5	0	9	31
Cov	0.25(1.0)	0.17(0.25)	0.5(0.5)	0.62(0.62)	0.10(0.10)	0.28(0.28)	0(0)	0.45(0.45)	0.31(0.37)

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 the candidate list (Numbers in parentheses represent terms and coverage when excluding mismatches caused by modifications to the guidelines).

tion candidate was provided in the candidate list. We calculated the coverage of cases, i.e., the fraction of disagreed annotation cases for which one annotator was able to provide an appropriate annotation candidate. We also calculated the coverage when excluding the effects of guidelines modifications. **Table 5** and **Table 6** summarize the analysis of disagreed annotations for the synthesis and characterization papers, respectively.

For the synthesis papers, if we exclude the effects of guidelines 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 their levels of domain knowledge, the recall for those categories is 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, Table 4, Table 5, and Table 6, we can conclude generally that information categories such as SMaterial,



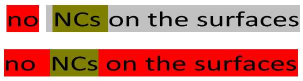
under the external magnetic fields H applied at θ and ϕ

under the external magnetic fields H applied at θ and ϕ

Surface morphologies of the samples

Surface morphologies of the samples

Fig. 7 Examples of the boundary-identification problem for terms in parameter categories between two annotators.



no NCs on the surfaces

no NCs on the surfaces

Fig. 8 Example of the boundary-identification problem for evaluation parameter values between two annotators.

MMethod, and ExPVal tend to be easier to annotate. Conversely, information categories such as the parameters ExP, EvP, and EvPVal tend to be more difficult to annotate, requiring deeper domain knowledge, particularly 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," or "increasing temperature from x to y". 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 an experiment parameter or as an evaluation parameter depending on the context. For example, "size" can be used for ExP in "mask-opening size" but for EvP in "size of nanocluster," even within the same paper. **Figure 7** shows examples of term-boundary mismatches for various 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 with longer statements that describe the final product. Furthermore, the value of the evaluation parameter can often 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 8** shows an example of boundary mismatch for the evaluation parameter value EvPVal.

6. Release of the Corpus and Its Usage

6.1 Corpus Release

From the analysis of the results of the annotation experiments, we found that the precision was high: the overall precision was 99% for synthesis papers and 97% for the characterization paper (when the effects of guidelines modifications were excluded). Recall was high for the synthesis papers (96% when excluding the guidelines-modification effects) but not high for the characterization paper (81% when excluding the guidelines-modification effects). In both cases, it is necessary to identify the appropriate annotation from the disagreed annotation results to improve the recall. The annotators' levels 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 guidelines might help the annotators. These results show that the guidelines for annotating papers related to nanocrystal device development are now suitable for release.

We plan to release the final NaDev corpus, as examined and modified by the domain expert, and its construction guidelines through our website, http://nanoinfo.ist.hokudai.ac.jp/. The corpus currently comprises five fully annotated papers, 392 sentences, and 2,870 annotated terms in eight information categories.

6.2 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 [35]. 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.

6.3 Corpus Construction Strategy in the Nanocrystal Device Domain

The proposed procedure for constructing a high-quality corpus for new research papers is as follows:

- 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 guidelines. In addition, for the disagreed annotations, the selection of one of the annotation candidates should take into account the knowledge levels of the annotators and any similarity between the annotation and examples in the guidelines. If neither of the annotators has high confidence in an annotation, it is better to check with a domain expert, given that the number of annotations requiring such checking is likely to be much smaller than for the whole corpus.

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.

6.4 Discussion

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 [20], [21]. However, as discussed in Section 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 [16], [17], [18]. The extraction of experimental information is supposed to be applied as a preprocessing step for such shared data construction in nanoinformatics. Our preliminary work [36], [37], [38] is recognized as one of the main efforts in applying natural language processing to extract such information [39].

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

- Corpus size: The NaDev corpus uses the full text of research papers instead of the abstracts that are often used in constructing corpora. Abstracts usually do not contain detailed explanations about experimental parameters in relation to the output evaluations. However, annotation of an abstract will take much less time than annotation of a full paper. We plan to increase the size of the corpus by including annotated abstracts.
- 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 [25]. 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.

7. Conclusion

In this study, we have developed a method for constructing an annotated corpus of research papers about nanocrystal device development. The method aims to support the automatic extraction of useful information for the analysis of experimental results in this field. The corpus and its construction guidelines have been examined and evaluated by a domain expert. The guidelines have now reached release level, and can be used to annotate research papers about nanocrystal device development in a consistent manner. The resulting NaDev corpus will be released soon.

In future work, we plan to increase the size of the corpus by annotating more papers for similar nanocrystal device development research areas. In addition, we plan to implement an automatic information extraction framework to support the effective collection of useful information from related publications.

Acknowledgments This research was supported partially by a grant to the Hokkaido University Global COE program, "Next-Generation Information Technology Based on Knowledge Discovery and Knowledge Federation," from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and JSPS KAKENHI Grant Number 26540165 by Japan Society for the Promotion of Science. We would like to thank Mr. Komagata, Mr. Morita, Mr. Yatago, Mr. Sakita, and Mr. Fujimagari for their contributions to the creation of the corpus. The authors wish to acknowledge the anonymous reviewers for their detailed and helpful comments to improve the manuscript.

References

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

[2] 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.*, Vol.6, 01401 (online), DOI: 10.1088/1749-4699/6/1/014011 (2013).

[3] 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 (online), available from ⟨http://eprints.internano.org/607/1/Roadmap_FINAL041311.pdf⟩ (accessed 2015-05-19), DOI: 10.4053/rp001-110413 (2011).

[4] 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*, Vol.44, pp.43–51 (2011).

[5] 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 (online), DOI: 10.1155/2013/410294 (2013).

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

[7] 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.*, Vol.5, pp.2339–2345 (online), DOI: 10.3762/bjnano.5.243 (2014).

[8] Capan, I., Carvalho, A. and Coutinho, J.: Silicon and germanium nanocrystals: Properties and characterization, *Beilstein J. Nanotechnol.*, Vol.5, pp.1787–1794 (online), DOI: 10.3762/bjnano.5.189 (2014).

[9] 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.*, Vol.58, pp.2018–2020 (1991).

[10] 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.*, Vol.95, No.4, pp.289–293 (2012).

[11] 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*, Vol.19, 265604-1-8 (2008).

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

[13] 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.*, Vol.6, 014008 (online), DOI: 10.1088/1749-4699/6/1/014008 (2013).

[14] 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, pp.63–64 (Oct. 2003).

[15] Thomas, D.G., Pappu, R.V. and Baker, N.A.: NanoParticle ontology for cancer nanotechnology research, *J Biomed Inform.*, Vol.44, No.1, pp.59–74 (2011).

[16] 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.*, Vol.6, 014007 (online), DOI: 10.1088/1749-4699/6/1/014007 (2013).

[17] DaNa project (online), available from ⟨http://www.nanoobjects.info/en/⟩ (accessed 2015-05-19).

[18] 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.*, Vol.7, 014001 (online), DOI: 10.1088/1749-4699/7/1/014001 (2014).

[19] 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.*, Vol.6, 014010 (online), DOI: 10.1088/1749-4699/6/1/014010 (2013).

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

[21] BioCreative IV CHEMDNER corpus (online), available from ⟨http://www.biocreative.org/resources/corpora/bc-iv-chemdner-corpus/⟩ (accessed 2015-05-19).

[22] 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.*, Vol.2, No.1, e83932 (2014).

[23] GENIA Project, Tsujii Laboratory, University of Tokyo (online), available from ⟨http://www.nactem.ac.uk/genia/⟩ (accessed 2015-05-19).

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

[25] Kim, J.D., Ohta, T. and Tsujii, J.: Corpus annotation for mining biomedical events from literature, *BMC Bioinformatics*, Vol.9, No.10 (online), DOI: 10.1186/1471-2105-9-10 (2008).

[26] 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.*, Vol.49, No.4, 04DH08-1-5 (2010).

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

[28] 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.*, Vol.298, pp.612–615 (2007).

[29] Green, A.M.: Kappa statistics for multiple raters using categorical classifications, *Proc. 22nd Annual SAS Users Group International Conference*, San Diego, CA, pp.1110–1115 (Mar. 1997).

[30] 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.*, Vol.89, 113111 (2006).

[31] 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.*, Vol.94, 243117 (online), DOI: 10.1063/1.3157275 (2009).

[32] 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.*, No.310, Vol.7, pp.2390–2394 (online), DOI: 10.1016/j.jcrysgro.2007.12.026 (2008).

[33] 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.*, Vol.48, 04C137 (online), DOI: 10.1143/JJAP.48.04C137 (2009).

[34] XConc Suite (online), available from ⟨http://www.nactem.ac.uk/genia/tools/xconc⟩ (accessed 2015-05-19).

[35] 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, *Proc. 27th Annual Meeting of the Japanese Society of Artificial Intelligence* (*jsai2013*), Toyama, Japan, CD-ROM 1C3-4 (2013).

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

[37] Dieb, T.M., Yoshioka, M. and Hara, S.: Automatic Information Extraction of Experiments from Nanodevices Development Papers, *Proc. 3rd IIAI International Conference on e-Services and Knowledge Management* (*IIAI ESKM2012*), Fukuoka, Japan, pp.42–47 (2012).

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

[39] Lewinski, N.A. and McInnes, B.T.: Using natural language processing techniques to inform research on nanotechnology, *Beilstein J. Nanotechnol.*, Vol.6, No.1439–1449 (online), DOI: 10.3762/bjnano.6.149 (2015).

Thaer M. Dieb received his B.S. degree in computer science from Faculty of Information Technology, Damascus University, in 2006, and M.S. degree in information science from Graduate School of Information Science and Technology, Hokkaido University, Japan, in 2011. He is now a 4th year Ph.D. student in the same school. His research theme focuses on nanoinformatics, with research interests in named entity recognition, text mining, machine learning, and chemical information. Mr. Dieb is a member of the Information Processing Society of Japan, and the Syrian Society for Scientific Research.

Masaharu Yoshioka received his B.E. and M.E. degrees of precision machinery engineering and the Ph.D. degree of precision machinery engineering from the University of Tokyo, Japan, in 1991, 1993, and 1996, respectively. From April 1996 to March 2000, he was a Research Associate of the National Center for Science and Information Systems, Japan and engaged in research for constructing engineering ontology for intelligent CAD system. From April 2000 to May 2001, he was a Research Associate of the National Institute of Informatics, Japan and started work in the field of information retrieval and natural language processing for knowledge management. From June 2001, he joined the Graduate School of Engineering as an Associate Professor and this school was reorganized as the Graduate School of Information Science and Technology in 2004. He served as a member of the Global Center Of Excellence program in the school for interdisciplinary collaboration among nanodevice researchers and computer science researchers. His research interests includes a knowledge management framework for supporting nanodevice development researchers (e.g., text mining from nanodevice related papers and experiment record management), the construction of engineering ontology for design, and the application of the knowledge processing technique to information retrieval. Prof. Yoshioka is a member of the Japanese Society for Artificial Intelligence, the Information Processing Society of Japan and the Association for Computing Machinery.

Shinjiro Hara received his B.E. and M.E. degrees of electrical engineering and the Ph.D. degree of electronics and information engineering from Hokkaido University, Sapporo, Japan, in 1993, 1995, and 1998, respectively. From 1995 to 1998, he was a Research Fellow of the JSPS, Tokyo, Japan, and engaged in the research on formation and semiconductor laser application of quantum wires utilizing multi-atomic steps grown by metal-organic vapor phase epitaxy (MOVPE). From 1998 to 2004, he was with the Optical Semiconductor Devices Laboratory of Fujitsu Laboratories Ltd, Atsugi, Japan, as a Research Scientist, and engaged in the research and development of MOVPE growth technologies of compound semiconductor and magnetic semiconductor materials for optical semiconductor devices. From 2002 to 2003, he worked at the Materials Science Center of Philipps-University Marburg, Germany, as a Visiting Research Scientist, where he was involved in the research on MOVPE growth and characterizations of magnetic semiconductor materials. In 2004, he joined the Graduate School of Information Science and Technology, and the Research Center for Integrated Quantum Electronics (RCIQE) of Hokkaido University, Sapporo, Japan, as a Research Associate (an Assistant Professor), and became an Associate Professor in 2007. Form 2007 to 2011, he was concurrently a Research Scientist in the precursory research for embryonic science and technology of the JST, Tokyo, Japan, where his research topic was Selective-area formation of nanoscale spin valve structures for spin-logic applications. His current research interests include electronic, photonic, and magnetic nanodevice applications using compound semiconductor nanowires and magnetic nanostructures grown by selective-area MOVPE. He is a member of the Japan Society of Applied Physics, the Japanese Association for Crystal Growth, the IEEE Photonics Society, the Materials Research Society, and the Magnetics Society of Japan.