Construction of Tagged Corpus for Nanodevices Development Papers

Thaer M. Dieb Masaharu Yoshioka
Department of Computer Science
Faculty of Information Science and Technology
Hokkaido University, Japan

{diebt, yoshioka}@ist.hokudai.ac.jp

1 Introduction

"Nanoinformatics" is one of the emerging research fields in developing a computational framework to support nanoscale research [1]. A wide variety of research studies are conducted in Nanoinformatics [2]. Especially in the field of nanomedicine development, some integrated computational frameworks have already been proposed (e.g., [3]).

We have been working on the project "Knowledge exploratory project for nanodevice design and manufacturing" [4] that aims to propose a framework to extract useful information from the nanodevice experimental records. In this project, we found that information described on the sheet is not enough for understanding the difference of parameter settings. So we decide to use research papers associated with these experiments as resources to extract background information (such as purpose, evaluation criteria).

In this paper, we propose a framework to annotate useful information (metadata) from Nanodevices development papers that help us analyzing the experimental results.

2 An Experiment Record Management System

The nanodevice development process is not well systematized, and it requires both engineering knowledge and craftsmanship skills [5, 6]. For example, nanodevice design based on knowledge of first principles, such as atomic physics, does not mean the end of the development process. Because the manufacturing process may affect the quality of the nanodevice, much trial and error is required before the final product can be realized. Skilled engineers can conduct this experiment planning more effectively than novices.

Since knowledge about this planning process is difficult to transfer from skilled engineers to novices. Novice engineers can only acquire the knowledge required for their planning through the guidance of skilled colleagues. To accelerate this nanodevice development process, it is better to make this tacit knowledge explicit.

In the Research Center for Integrated Quantum Electronics, Hokkaido University, researchers are developing various kinds of nanodevices (e.g., nanowires on Si) by using a selective-area metal-organic vapor phase epitaxy (SA-MOVPE) method [7].

Even though SA-MOVPE is a good method that can control the quality of the device, it requires many trial and error processes to arrive at the final process. To keep records about these process, researchers use the MOVPE growth parameter record sheet for each experiment.

For supporting knowledge transfer process, we have already proposed an experiment record management system that can store the information of the original sheet. This system has following 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, we found that varieties sets of parameters are used for making a same layer structure. In order to understand the difference among these sets, it is necessary to check the background information for further understanding.

However, the experimental record sheets is not enough for understanding such information, it is necessary to check related research papers for clarifying the difference.

3 Tagged Corpus for Nanodevices Development Papers

3.1 Design of Tag Set for Annotation

In order to extract useful information from the papers, it is necessary to understand how nanodevice researchers extract and use such information. We conduct interview with nanodevice researchers of the Research Center for Integrated Quantum Electronics, Hokkaido University and found that it is important to clarify how the authors would like to use a proposed device.

This kind of information is described in following format.

- What kind of final product the authors would like to use the device for
- Evaluation criteria for the device

In addition, the research papers have lot of information about experimental settings. For example, material and parameter used in the experiment is described in the paper. That kind of information is also useful to understanding the relationship between research paper and experiments.

Based on this discussion with nanodevice researchers, we propose a candidate tag set for annotating the research papers as follows.

Material(SMaterial) Information about raw material (e.g., As, InGaAs)

Characteristic Feature of Material(SMChar) Characteristic feature about raw material (e.g., (111)B, III-V)

Experiment Parameter(ExP) Parameter that are controlled during the experiments (e.g., diameter, total pressure)

Value of the Experiment Parameter(ExPVal) Value of the experiment parameter (e.g., 50nm, 10atom)

Evaluation Parameter(EvP) Parameter that are used for evaluating the device (e.g., peak energy, FWHMs)

Value of the Evaluation Parameter(EvPVal) Value of the evaluation parameter (e.g., 1.22eV)

Manufacturing Method(Mmethod) Method for manufacturing device (e.g., SA-MOVPE)

Final Product(TArtifact) Information of final product (e.g., semiconductor nanowires)

3.2 Corpus Construction Guideline

It is not so easy to construct good tagged corpus without corpus construction guideline. In order to construct such guideline, we ask two master course students to annotate the same paper [8] without negotiation. Then we compared the both annotation results and discussed the reason why they made different annotation. Through this discussion, we made corpus construction guideline for annotating research papers.

3.3 Analysis of Corpus Construction Guideline

In order to analyze the quality of corpus construction guideline, we also ask same two master course students to conduct annotation on a same paper [9] that is different from previous one based on the guideline and measure Inter Annotation Agreement(IAA) using the Kappa statistics coefficient.

The Kappa statistics coefficient is given by the following formula:

$$k = \frac{A_0 - A_e^k}{1 - A_e^k} \tag{1}$$

Where
A_0: Observed agreement (proportion of actual agreement).
A_e^k: Probability of both annotators picking the same category, and it is given by the formula:

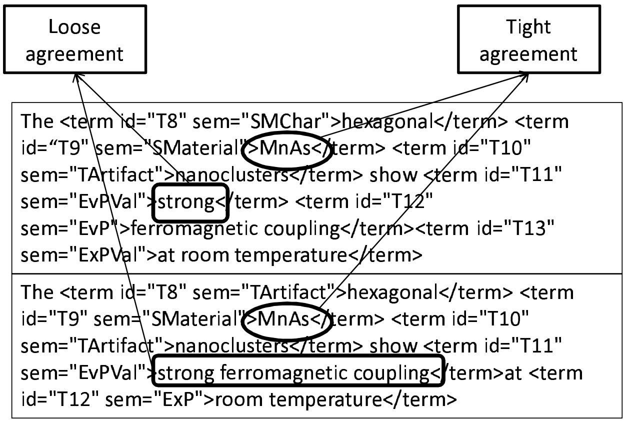
$$A_e^k = \sum_q \frac{n_{c_1 q}}{i}\frac{n_{c_2 q}}{i} = \frac{1}{i^2}\sum_q n_{c_1 q} n_{c_2 q} \tag{2}$$

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 annotator picking a category q_a.
However, for annotating text using the suggested tag set, it is necessary to deal with mismatch of term boundary problem. In order to separate the issue of term category selection and term boundary identification, we introduce two different evaluation metrics for analysis. One is tight agreement that takes term boundary into consideration, i.e.,two corresponding annotated terms (a word or a sequence of words)have the exact same boundary, and same category. The other is loose agreement that ignores term boundary problem, i.e., two corresponding terms have the same category, and overlapping boundaries.
Figure 1 shows an example of tight and loose agreement.

3.4 Experimental results

We compared the two different annotated texts on the same paper.
Tables 1 and 2 show the agreement numbers for all categories between the two annotators using tight and loose agreement metrics respectively.

Note: SM:SMaterial, SMC:SMChar, EP:ExP, EPV:ExPVal, Ev:EvP, EvV:EvPVal, MM:MMethod, and TA:TArtifact are for tag set. O:Other class is either unclassified text (or terms with boundary mismatch for tight agreement). T is for Total.



The <term id="T8" sem="SMChar">hexagonal</term> <term id="T9" sem="SMaterial">MnAs</term> <term id="T10" sem="TArtifact">nanoclusters</term> show <term id="T11" sem="EvPVal">strong</term> <term id="T12" sem="EvP">ferromagnetic coupling</term><term id="T13" sem="ExPVal">at room temperature</term>

The <term id="T8" sem="TArtifact">hexagonal</term> <term id="T9" sem="SMaterial">MnAs</term> <term id="T10" sem="TArtifact">nanoclusters</term> show <term id="T11" sem="EvPVal">strong ferromagnetic coupling</term>at <term id="T12" sem="ExP">room temperature</term>

Figure 1: Tight and Loose Agreement.

Table 1: Tight Agreement between two annotators

	SM	SMC	EP	EPV	Ev	EvV	MM	TA	O	T
SM	91								11	102
SMC		30						4	11	45
EP			32						28	60
EPV			1	17					17	35
Ev					23	2			14	39
EvV						4			34	38
MM							9		4	13
TA								44	3	47
O	12	5	24	15	10	23	10	23		122
T	103	35	57	32	33	29	19	71	122	501

Kappa Coefficient=0.41

3.5 Discussion

We notice the low agreement ratio based on the Kappa statistics as in tight agreement. We believe that is due to term boundary identification problem, since in case of loose agreement the agreement ratio is considerably higher. That means different annotators considering different chunk of texts as terms due to unclarity of the guideline.

There are two types of errors, i.e., term category mismatch and term boundary mismatch. We found that there are few problems (most of them are mismatch between SMC and TA) for selecting different categories for same term. For checking these cases, one annotator annotate the characteristics of target artifact as a characteristics of material. It is necessary to make guideline for dealing with such inconsistency.

If we have a close look on the tight agreement table, we can notice that the most common errors about term boundary mismatch with the EvPVal and ExP tags.

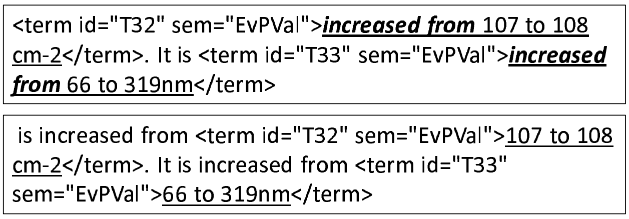
For EvPVal errors mainly come from term boundary identification of the evaluation parameter value, one annotator considers only the value of this parameter as a term while the other one considers the type of change in this parameter is also included in the parameter value hence in the term like (increasing, decreasing,...).

Table 2: Loose Agreement between two annotators

	SM	SMC	EP	EPV	Ev	EvV	MM	TA	O	T
SM	105								1	106
SMC		36						10	4	50
EP	1		53	2			8		1	65
EPV			2	33		1			2	38
Ev					32	7			3	42
EvV		1			1	24		2	12	40
MM							14			14
TA								47	1	48
O	6	2	7	2	2		1	18		38
T	112	39	62	37	35	32	23	77	24	441

Kappa Coefficient=0.74

Figure 2 shows an example of this difference.

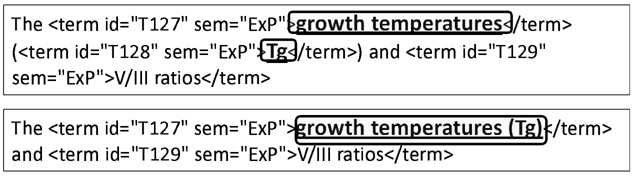


<term id="T32" sem="EvPVal">***increased from*** 107 to 108 cm-2</term>. It is <term id="T33" sem="EvPVal">***increased from*** 66 to 319nm</term>

is increased from <term id="T32" sem="EvPVal">107 to 108 cm-2</term>. It is increased from <term id="T33" sem="EvPVal">66 to 319nm</term>

Figure 2: EvPVal Term boundary mismatch

For ExP term there are two main reasons for the boundary mismatch, first one is that one annotator considers two terms as one when they come next to each other and of the same category, while the other annotator separates them as two or more terms(Figure 3).



The <term id="T127" sem="ExP">growth temperatures</term> (<term id="T128" sem="ExP">Tg</term>) and <term id="T129" sem="ExP">V/III ratios</term>

The <term id="T127" sem="ExP">growth temperatures (Tg)</term> and <term id="T129" sem="ExP">V/III ratios</term>

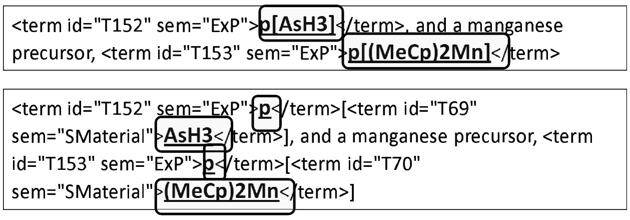
Figure 3: Boundary mismatch of ExP Term 1

The second reason is that some terms are ignored to be marked when they come within another outer terms for the first annotator,while they marked as two separate terms in another annotation(Figure 4).

In order to improve the quality of the corpus, it is better to use corpus annotation tool. We plan to use Xconc Suite[10] that is a tool originally developed for annotating biomedical information to construct GENIA corpus [11].

4 Conclusion

In this paper, we proposed an approach to build a tagged corpus for extracting useful background in-



```
<term id="T152" sem="ExP">p[AsH3]</term>, and a manganese
precursor, <term id="T153" sem="ExP">p[(MeCp)2Mn]</term>
```

```
<term id="T152" sem="ExP">p</term>[<term id="T69"
sem="SMaterial">AsH3</term>], and a manganese precursor, <term
id="T153" sem="ExP">p</term>[<term id="T70"
sem="SMaterial">(MeCp)2Mn</term>]
```

Figure 4: Boundary mismatch of ExP Term 2

formation from the nanodevice research papers. We also proposed a candidate tag set for annotation and confirm that we can construct such corpus at certain consistency level.

However, we found that the term boundary identification problem can affect notably the inter annotator agreement ratio, so we suggest to modify the guideline for the corpus standards.

As a next step we would like to modify the guideline and do the measurements again hoping to increase the agreement ratio, until we can have the standard guideline for tagging the corpus.

Acknowledgment

This research was partially supported by a grant for 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. We also would like to thank Prof. Fukui, Prof. Hara, Dr. Tomioka, Mr. Komagata and Mr. Morita for their contribution to make this corpus.

References

[1] K. Ruping and B.W. Sherman. Nanoinformatics: Emerging computational tools in nanoscale research. In *Technical Proceedings of the 2004 NSTI Nanotechnology Conference and Trade Show, Volume 3*, pp. 525–528, 2004.

[2] National Nanomanufacturing Network. *Workshop on Nanoinformatics Strategies.* 2007. http://128.119.56.118/ nnn01/Workshop.html.

[3] Fernando Martìn-Sanchez, Victoria Lòpez-Alonso, Isabel Hermosilla-Gimeno, and Guillermo Lopez-Campos. *A Primer in Knowledge Management for Nanoinformatics in Medicine*. Springer-Verlag GmbH, 2008. LNCS 5178.

[4] Masaharu Yoshioka, Katsuhiro Tomioka, Shinjiroh Hara, and Takashi Fukui. Knowledge exploratory project for nanodevice design and manufacturing. In *iiWAS '10 Proceedings of the 12th International Conference on Information Integration and Web-based Application & Services*, 2010.

[5] T. Fukui, S. Ando, Y. Tokura, and T. Toriyama. GaAs tetrahedral quantum dot structures fabricated using selective area metalorganic chemical vapor-deposition. *APPLIED PHYSICS LETTERS*, Vol. 58, pp. 2018–2020, 1991.

[6] J. Noborisaka, J. Motohisa, S. Hara, and T. Fukui. Fabrication and characterization of freestanding GaAs/AlGaAs core-shell nanowires and AlGaAs nanotubes by using selective-area metalorganic vapor phase epitaxy. *APPLIED PHYSICS LETTERS*, Vol. 87, , 2005.

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

[8] Masatoshi Yoshimura, Katsuhiro Tomioka, Kenji Hiruma, Shinjiro Hara, Junichi Motohisa, and Takashi Fukui. Growth and characterization of InGaAs nanowires formed on GaAs(111)B by selective-area metal organic vapor phase epitaxy. *Japanese Journal of Applied Physics*, Vol. 49, No. 4, pp. 04DH08–1–5, 2010.

[9] Shinjiro Hara, Junichi Motohisa, and Takashi Fukui. Self-assembled formation of ferromagnetic MnAs nanoclusters on GaInAs/InP (1 1 1) B layers by metal-organic vapor phase epitaxy. *Journal of Crystal Growth*, Vol 298, January 2007, Pages 612-615 Thirteenth International Conference on Metal Organic Vapor Phase Epitaxy (ICMOVPE XIII).

[10] Xconc Suite *http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=XConc+Suite.*

[11] GENIA Project *http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi.*