Construction of Tagged Corpus for Nanodevices Development Papers (2ed report)

Information Extraction by Cascading Named Entity Recognition Tools

Thaer M. Dieb　Masaharu Yoshioka
Graduate School of Information Science and Technology
Hokkaido University, Japan
diebt@kb.ist.hokudai.ac.jp
yoshioka@ist.hokudai.ac.jp

Shinjiro Hara
Center For Integrated Quantum Electronics
Hokkaido University, Japan
hara@rciqe.hokudai.ac.jp

1 Introduction

The nanodevices development process is not well systematized, and it requires both engineering knowledge and craftsmanship skills [1]. In order to support this development process, we have been working on the project called "knowledge transfer environment for nanodevices development". As part of this environment, we want to utilize the information in the nanodevices development papers. We have already proposed to construct a tagged corpus for nanodevices development papers to achieve this goal.

In the first report [2], we have discussed the need for utilizing the information in the papers to enhance the analysis of the experiments results; also we have discussed the design and the reliability of the corpus. We confirmed the ability of constructing such corpus; However, we found that the term boundary identification problem can affect notably the inter annotator agreement ratio.

In this report, we introduce a revised guideline for annotating tags in order to enhance the reliability of the corpus. In addition to that, we also develop an information extraction tool based on this corpus by cascading several named entity recognition tools. This tool will help accelerating the corpus construction process.

2 Tagged corpus for nanodevices development papers

2.1 Background

In order to analyse the results of the experiments of nanodevices development, we built an experiment record management system [3]. In this system using pattern mining techniques, we found that information stored in the system were not enough for detailed analysis. Based on the discussion with field experts, we decided to extract other information from research papers related to these experiments. In order to extract this information, we will construct a tagged corpus.

If we could achieve a well defined corpus, it will be good collaboration schema among researchers in nanodevice, computer science and natural language processing. This will help accelerating the process of dealing with problems in the field of nanoinformatics.

In the first report, we have done a first reliability evaluation experiment. We have defined two metrics for analysis. One is tight agreement that takes into consideration the term boundary issue. The other one is loose agreement that ignores term boundary. We found that K=0.41 in case of tight agreement, and K=0.74 in case of loose agreement. These results were not reliable enough, especially in the tight agreement.

2.2 Second reliability evaluation experiment

2.2.1 Experiment set up

In order to resolve the annotation mismatch issues, we had to revise the corpus construction guideline. The modification includes more examples of desired information, and also new rules to resolve annotation mismatch.

The annotation this time was done using software called XConc Suite [4], an XML-based tool originally developed for annotating biomedical information to construct GENIA corpus [5]. XConc provides assisting functions to the annotators.

The first step was to train the annotators on XConc by using only an abstract. After that a full paper training session was held on the new revised version of the corpus guideline. The second training resulted in some annotation mismatches. Based on the discussion with the annotators, we have slightly modified the guideline (added some example of evaluation parameters). After this training, we asked the two annotators (graduate students of nanode-

vices development different from those of the first experiment)to annotate the same paper [6] independently.

2.2.2 Experiment results

When measuring Kappa statistics coefficient, we found that $K = 0.63$ in case of tight agreement, and $K = 0.77$ in case of loose agreement. 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.

Table 1: Tight agreement between two annotators

	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

Kappa statistics coefficient K=0.63

Table 2: Loose agreement between two annotators

	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

Kappa statistics coefficient K=0.77

2.2.3 Results analysis

We noticed a slight improvement in case of loose agreement, and a considerable one in case of tight agreement. This is because of the enhanced examples of the information to be extracted. However the term boundary issue is still notably affecting the IAA ratio. There still different categories of annotation mismatches. However, we provide the main ones below

- Category mismatch between characteristics of material (SMChar) and final artifact (TArtifact), especially in overlapped terms. That is because characteristics of input material can also define characteristics of final product.
- It seems that recognizing parameters and their values with clear boundaries is a difficult issue. We can confirm that from the numbers in the tables.

2.3 Automatic information extraction

2.3.1 Cascading named entity recognition

Manual annotation of research papers is a time consuming process. In order to speed it up, we are trying to automatically extract information using machine learning techniques. However, because of the overlapped tag structure of the annotation, it is not easy for the machine to learn to set the correct tags information within small window size all at once. It is necessary to separate overlapped tags in the process of training the machine. In this case, we need to break the learning process into cascading levels i.e. cascading named entity recognition [7]. Figure 1 shows an example of overlapped tag structure.

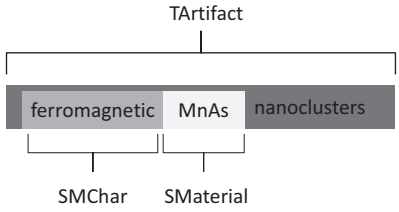

Figure 1: Overlapped tag structure

We separate the tags based on the tag structure into 4 groups, each group will be handled in one level of the training process.

- basic level with most inner tags or non overlapped tags: SMaterial (input material), SMChar (characteristics of material), and MMethod(method of manufacturing).
- overlapping with the first group tags: TArtifact (final product).
- parameter ExP(experiment parameter),EvP(evaluation parameter).
- parameter value: ExPVal,EvPVal

Our approach is described as follows
First of all, we format the paper using POS tagger. We also use assistant chemical entity recognition tool to tag input material. After this, we use cascading statistical learning for each tag group to

train the machine in several levels. Figure 2 shows a model of the cascading named entity recognition.

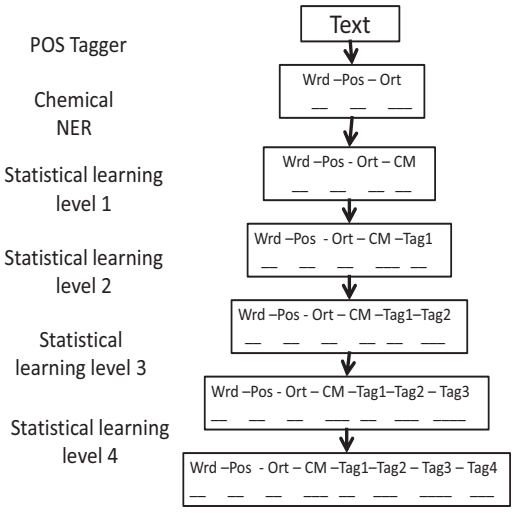

Figure 2: Cascading named entity recognition

Note:Wrd=word, POS=POS tag, Ort=orthogonal feature,CM=chemical compound,Tag1,..Tag4=tag group.

As a POS tagger, we use GPoSTTL [8], enhanced version of Brill tagger. Using this tagger, we are trying to extract the orthogonal feature. Orthogonal feature seems to improve the automatic annotation results[9]. For the automatic annotation, we use YamCha [10], which is open source text chunker oriented towards multiple NLP tasks. Training will be done in the cascading style explained above. In order to support the learning process for the SMaterial tag, we use oscar3-a5 [11], a tool for shallow, chemistry-specific parsing of chemical documents.

2.3.2 Chemical material entity recognition

As explained above, we are using oscar3-a5 tool to support the learning process for SMaterial tag. However, we need to check the performance of this tool to annotate input material. We compare the annotation of oscar3-a5 and the annotation done by field experts on the same paper. This comparison was done only on the SMaterial (or Chemical coMpound in the case of oscar3-a5) tag, and the results were as follows

Table 3: oscar3 chemical entity recognition performance

	standard	
oscar3	107	75
	1	

From the above table, we can see that almost all the SMaterial tags annotated by field experts are being caught by oscar3; However, there are high percentage of oscar3 tags that are not considered as input material tags by field experts. These tags are mainly method name and or group source of some material. As a conclusion, we cant use oscar3 as a solely tool to tag input material, however, it would be of very good assistant in the learning process by YamCha.

2.3.3 Automatic tag annotation

We use YamCha for the automatic annotation in a cascading style mode. However, In order to train the tool, we need to format the input paper in a special format compatible with YamCha. The original format of the paper is XML (the output of XConc Suite). We programmed a tool to transfer our annotated papers with XML format into compatible training format for YamCha. This tool also calls the GPoSTTL tagger to do POS tagging and oscar3 for the chemical entity recognition. Table 4 shows an example of the training data format.

Table 4: Example of the training data

Wrd	POS	Ort	CM	Tag1	Tag2	Tag3	Tag4
Self-assembled	JJ	OtherHyphon	o	o	o	o	o
formation	NN	Lowercase	o	o	o	o	o
of	IN	Lowercase	o	o	o	o	o
ferromagnetic	JJ	Lowercase	o	B-SMChar	B-TArtifact	o	o
MnAs	NNP	TwoCaps	CM	B-SMaterial	I-TArtifact	o	o
nanoclusters	NNS	Lowercase	o	o	I-TArtifact	o	o
on	IN	Lowercase	o	o	o	o	o
GaInAs	NNP	TwoCaps	CM	B-SMaterial	o	o	o

Note:Wrd=word, POS=POS tag, Ort=orthogonal feature,CM=chemical compound,Tag1,..Tag4=tag group.
The training will be done in 4 levels corresponding to each tag group as follows

- training for tag level 1 annotation uses column 1-5
- training for tag level 2 annotation uses column 1-6
- training for tag level 3 annotation uses column 1-7
- training for tag level 4 annotation uses column 1-8

So far, we only have 2 annotated documents. We use one document [12] for the training, and the other one [6] for the test. Table 5 shows an example of the automatically annotated document by YamCha after the training. However, this sample doesnot include the oscar tag CM.

Table 5: Example of automatic annotation output

Wrd	POS	Ort	CM	Tag1	Tag2	Tag3	Tag4
Hexagonal	NNP	InitCap	o	o	o	o	o
ferromagnetic	JJ	Lowercase	o	o	o	o	o
MnAs	NNP	TwoCaps	o	B-SMaterial	I-TArtifact	o	o
nanoclusters	NN	Lowercase	o	o	I-TArtifact	o	o
formation	NN	Lowercase	o	B-SMCharo	o	o	o
on	IN	Lowercase	o	o	o	o	o
GaInAs	NNP	TwoCaps	CM	B-SMaterial	o	o	o

Note:Wrd=word, POS=POS tag, Ort=orthogonal feature,CM=chemical compound,Tag1,..Tag4=tag group.

Until the time of submitting this paper, we didnot have to assess the performance of YamCha. We plan to check the performance using different conditions, like using different POS tagger, different features, or different hierarchy of tag group.

Conclusion

In this paper, we proposed and enhanced guideline to construct an annotated corpus for nanodevices development papers. This guideline increases the reliability of the corpus. However, we are planning for further enhancement in order to increase the reliability ratio of the tight agreement case into a sufficient level. In addition to that, we proposed a cascading style entity recognition to be used in the automatic annotation of the research papers.

In the future, we plan to do several tests on the performances of the automatic annotation tool using different conditions like orthogonal features.

Acknowledgement

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 would like to thank Prof. Fukui, Mr. Yatago and Mr. Sakita for their contribution to make this corpus. Also we would like to thank Prof.Takeuchi (Okayama University), Prof.Kano (Japan Science and Technology Agency) for their discussion and comments about cascading named entity recognition.

References

[1] 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. In *NANOTECHNOLOGY, Vol. 19, pp. 265604-1-8, 2008.*

[2] T.M.Dieb M.Yoshioka. Construction of Tagged Corpus for Nanodevices Development Papers. *NLP Japan 2011.*

[3] M. Yoshioka, K. Tomioka, S. Hara, and T. 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, pp. 869-872, 2010.*

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

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

[6] S. Hara, T. Fukui. Hexagonal ferromagnetic MnAs nanocluster formation on GaInAs/InP 111B layers by metal-organic vapor phase epitaxy. In *APPLIED PHYSICS LETTERS 89, 113111 2006*

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

[8] GPoSTTL *http://gposttl.sourceforge.net/.*

[9] K. Takeuchi and N. Collier, Bio-Medical Entity Extraction using Support Vector Machines, In *Journal of Artificial Intelligence in Medicine, vol. 33, issue 2, pages 125-137, February, 2005.*

[10] YamCha *http://chasen.org/ taku/software/yamcha/.*

[11] OSCAR3 *http://apidoc.ch.cam.ac.uk/oscar3/.*

[12] S. Hara, J. Motohisa, and T. 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).