Automatic Annotation of Parameters from Nanodevice Development Research Papers

Thaer M. Dieb
Graduate school of IST, Hokkaido University, Sapporo, Japan
diebt@kb.ist.hokudai.ac.jp

Masaharu Yoshioka
Graduate school of IST, Hokkaido University, Sapporo, Japan
yoshioka@ist.hokudai.ac.jp

Shinjiro Hara
RCIQE, Hokkaido University, Sapporo, Japan
hara@rciqe.hokudai.ac.jp

Marcus C. Newton
Physics & Astronomy, University of Southampton, Southampton, UK
M.C.Newton@soton.ac.uk

Abstract

In utilizing nanodevice development research papers to assist in experimental planning and design, it is useful to identify and annotate characteristic categories of information contained in those papers such as source material, evaluation parameter, etc. In order to support this annotation process, we have been working to construct a nanodevice development corpus and a complementary automatic annotation scheme. Due to the variations of terms, however, recall of the automatic annotation in some information categories was not adequate. In this paper, we propose to use a basic physical quantities list to extract parameter information. We confirmed the efficiency of this method to improve the annotation of parameters. Recall for parameters increases between 4% and 7% depending on the type of parameter and analysis metric.

1 Introduction

"Nanoinformatics" is an emerging interdisciplinary research field in developing a computational framework to support nanoscale research (Karl et al. 2004). 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 (De la Iglesia et al. 2011). In order to support nanodevice development process, we have been working on a project that aims at analyzing the experiment results related to nanodevice development to provide insights for nanodevice novice researchers to help them planning their experiments more effectively (Yoshioka et al. 2010). In this project, we have proposed a framework to annotate useful information from research papers related to nanodevice development (e.g., source material, evaluation parameter, and so on), and use them for analyzing experiment results (Dieb et al. 2011). In order to speed up the annotation process, we have built an automatic annotation framework using machine-learning techniques to annotate research papers (Dieb et al. 2012).

However, due to the variations of terms, this framework may miss to annotate terms that are not in the training data set. Therefore, we have used chemical named entity recognition system to add generalized feature to extract "source material" terms. This generalized information is useful to extract "source material" terms and recall of this category increased. However, there are several other categories whose recall is inadequate.

Proceedings of the 4th International Workshop on Computational Terminology, pages 77–85,
Dublin, Ireland, August 23 2014.

In this paper, we propose to use a physical quantities list for adding generalized feature for extracting parameter terms in two categories ("evaluation parameter" and "experiment parameter"). In those two categories, since "experiment parameter" represents a control parameter for the experimental equipment and "evaluation parameter" represents ones measured by measuring devices, most of the terms are associated with physical quantities and they contains (a) term(s) that represent(s) its characteristics. We use 2 methods for the identification: first one we try to identify parameters (experiment and evaluation) using the new automatic annotated framework. The other one, we identify the parameters (in general) using the automatic annotation framework, and then classify the parameters into experiment and evaluation using SVM (Support Vector Machine) (Li 2005).

Several attempts have been made to use dictionary to enhance machine-learning performance. For example, Usié et al. (2013) is using a dictionary to assist in identifying certain categories of chemical entities in biomedical text. Our method is using the physical quantities list to enhance the identification of parameter information.

This paper has five sections. The first one is introduction. Second one introduces the nanodevice development papers corpus we have developed (Dieb et al, 2011) and the automatic annotation framework we have built (Dieb et al., 2012) in brief. Section 3 discusses parameter identification methods. In section 4, we demonstrate the experiment and discuss the results, and section 5 is a conclusion.

2 Automatic Annotation Framework for Nanodevice Development Papers

2.1 Nanodevice Development Papers Corpus

In order to analyze experiment results related to nanodevice development, we constructed nanodevice development papers corpus that annotates characteristic information from research papers Dieb et al. (2011). It was very critical to decide on what categories of information are necessary and adequate for the analysis of experiment results. Based on discussion with nanodevice development researchers, we were able to build an abstract for a development experiment in order to define the necessary terms for the analysis. We have defined eight categories of information for annotation as below:

- Source Material (SMaterial) e.g., As, InGaAs
- Source Material Characteristic (SMChar) e.g.,(111)B
- Experiment Parameter (ExP) e.g., total pressure
- Value of the Experiment parameter (ExPVal) e.g., 50 nm
- Evaluation Parameter (EvP) e.g., peak energy, FWHMs
- Value of the Evaluation Parameter (EvPVal) e.g., 1.22eV
- Manufacturing Method (MMethod) e.g., SA-MOVPE
- Final Product (TArtifact) e.g., semiconductor

The information in the papers is annotated using XML format. Figure 1 shows an example of an annotated paper.

We demonstrate the successful formation of <TArtifact> <SMChar>ferromagnetic</SMChar> <SMaterial>MnAs</SMaterial> nanoclusters </TArtifact> self-assembled on <SMaterial> GaInAs</SMaterial><SMChar> (1 1 1) B </SMChar> surfaces by <MMethod>metalorganic vapor phase epitaxy </MMethod><MMethod>MOVPE</MMethod>).
The <TArtifact><SMChar>hexagonal</SMChar> <SMaterial>MnAs</SMaterial>nanoclusters</TArtifact> show <EvPVal>strong</EvPVal> <EvP>ferromagnetic coupling</EvP><ExPVal>at room temperature </ExPVal> when the <ExP>external magnetic fields </ExP> are applied <ExPVal>in a direction parallel to the <SMaterial>InP</SMaterial><SMChar>(1 1 1) B</SMChar> wafer planes</ExPVal>.

Fig. 1: Example of annotated paper

Manual annotation of research papers is a time consuming process. Papers were first annotated by graduate students in nanodevice development domain. Several experiments and discussions have been held to improve the reliability of the corpus by resolving mismatches between different annotators. Papers then were corrected manually based on discussion with annotators. So far, we were able to complete five fully annotated papers, which can allow us to start our preliminary experiments concerning extracting and utilizing characteristic information in supporting nanodevice development. These papers are currently from the same research group (e.g., (Hara et al. 2006)). We also started to collaborate with other research groups to enlarge and include different research topics related to nanodevice development. Other information categories can be added to our model as needed along the way of the corpus development. Currently this corpus is still under construction, and not yet available.

2.2 Automatic Annotation Framework

In order to speed up the annotation process that also require domain expert, we have built an automatic annotation framework using the machine learning techniques to annotate the desired categories of information Dieb et al. (2012).

There are two main issues to be discussed in this automatic annotation framework:

(1) Chemical entity recognition:

In literature related to nanodevice development, most of the source material items are chemical compounds. If large training data set is available, the machine might be able to identify source material based on the training data only with no need to additional clue. However, since training data size is still very small, more clues are needed to identify Source Material. Identifying chemical entities can add more clues and will help the machine to recognize Source Material terms.

A new chemical entity recognizer called SERB-CNER (Syntactically Enhanced Rule-Based CNER) is used to enhance the identification of Source Material terms. SERB-CNER is a rule-based chemical entity recognizer that uses regular expressions to identify chemical compounds. In addition to that, SERB-CNER uses syntactic rules to eliminate some mismatches that might occur between chemical entities and general text.

(2) Overlapped term structure:

In nanodevice development domain, terms sometimes can overlap within each other, and not always simple. This might be a common issue in several other domains. Figure 2 shows an example of term overlapping.

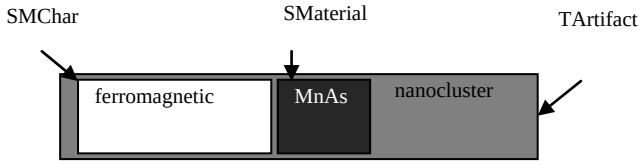

Fig. 2 Example of overlapped term structure

Because of this overlapping between different terms, same chunk of text might have information related to more than one term at the same time. That makes it difficult for the machine to learn to set the correct term information all at once. To tackle this issue, we have separated overlapped term categories into four groups where terms do not overlap within other terms from the same group. Based on these 4 groups, the machine learning process also divided into 4 cascading levels i.e. cascading named entity recognition (Kano et al. 2011).

For the automatic annotation framework, YamCha 0.33 (YamCha) was used, as a machine learning based sequence labelling tools. For the features, we use linguistic features like POS tag and orthogonal feature. POS tag was generated using rb tagger (rb tagger), which is A Simple Ruby Rule-Based Part of Speech Tagger based on the work of Eric Brill. Orthogonal feature was calculated using regular expressions. Chemical entity feature (CM) was calculated using SERB-CNER that identifies chemical entities as we discussed before. Term group's features were estimated in cascading style. In each level of this cascading style, we apply machine-learning technique to estimate the target feature for this level using information estimated from previous levels. For example, in cascading level 1, the machine will try to estimate term group 1 (TG1) using information of {Word, POS, Orth, CM}. In cascading level two, the machine will try to estimate term group 2 (TG2) using information of {Word, POS,

Orth, CM}, and TG1 that the machine estimated in cascading level 1. Figure 3 illustrates the cascading style annotation on an example of input data in IOB format.

Word	POS	Orth	CM	TG1	TG2	TG3	TG4
hexagonal	JJ	Lowercase	O	B-SMChar	B-TArtifact	O	O
MnAs	NNP	TwoCaps	B-CM	B-SMaterial	I-TArtifact	O	O
nanoclusters	NNS	Lowercase	O	O	I-TArtifact	O	O
show	VBP	Lowercase	O	O	O	O	O
strong	JJ	Lowercase	O	O	O	O	B-EvPVal
ferromagnetic	JJ	Lowercase	O	O	O	B-EvP	O

Note: POS=POS tag, Orth=orthogonal feature, CM=chemical compound, TG1, … TG4=term group.

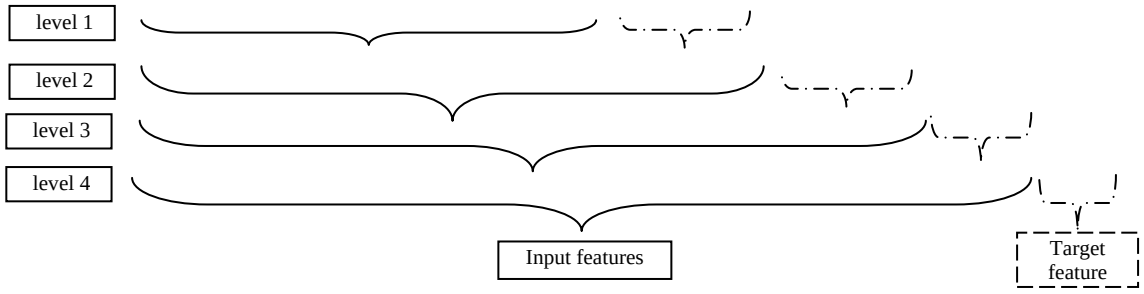

Fig. 3 Cascading style annotation on input data for the automatic annotation framework

We have tested this system using only two fully annotated papers with 10 fold cross validation. We use two metrics for analysis. One is tight agreement, which takes term category and term boundary into consideration and the other one is loose agreement, which checks term categories only. Table 1 shows the result of the automatic annotation framework. As we can see from the table, some categories have low recall, for example, EvP and EvPVal that depends on EvP for assessment.

Table 1: Automatic annotation framework performance

	Tight /precision	Tight /recall	Loose /precision	Loose /recall
SMaterial	0.99	0.96	0.99	0.96
SMChar	0.89	0.69	0.89	0.69
MMethod	0.93	0.86	0.95	0.88
TArtifact	0.86	0.73	0.93	0.79
ExP	0.91	0.81	0.97	0.86
EvP	0.76	0.60	0.87	0.69
ExPVal	0.72	0.57	0.85	0.67
EvPVal	0.86	0.60	0.97	0.67
Overall	0.89	0.76	0.94	0.80

3 Usage of Physical Quantities for Parameter Identification

3.1 Basic Idea

In the framework we proposed (Dieb et. al, 2012), recall and precision of each categories are evaluated, and recall and precision of "evaluation parameter" (EvP) is lower than average. One of the reasons about this problem is a variety of parameter terms used in the paper. Machine learning based sequence labelling tools is good at extracting terms that are also exists in the training data set, but it is necessary to extract more generalized clue to identify terms that do not appear in the training data set.

Identification of chemical named entity is helpful to extract such clue from the corpus, but it is not sufficient.

As Nakagawa and Mori（2003）discussed, most of the compound nouns are constructed from basic noun and identification of terms that contribute to make up compound noun is useful to extract terms.

Since “experiment parameter” (ExP) represents a control parameter for the experimental equipment and “evaluation parameter” (EvP) represents ones measured by measuring devices, most of the terms are associated with physical quantities and they contains (a) term(s) that represent(s) its characteristics. For example, “density of the nanoclusters” and “height of the nanoclusters” contains physical quantity term “density” and “height” respectively. Identification of physical quantities of test data may support to extract new terms. For example, identification of "size" as a physical quantity might support identification of “size of nanoclusters” as a parameter.

There are two types of parameters exist in this nanodevice development papers corpus; Experiment parameter that is used to control the conditions of the experiment like temperature, pressure, gas flow rate, and so on; The other type is evaluation parameter that is used to evaluate the quality of the experiment output like smoothness of the surface, conductivity, and so on. In this paper, a list of physical quantities is used to support extracting terms of those two types of parameters.

3.2 Physical Quantities List

In order to construct a list of physical quantities, we started with a basic list of physical properties of matter, collected by Dr. Anne Marie Helmenstine, Ph.D. in biomedical science. [1] This list includes but not limited to the physical properties of an object. For example, concentration, density, and so on. In addition to this list, we have added several other common parameters that commonly found in nanodevice development research papers. For example, height, conductivity, and so on. Additionally, we have added several keywords that usually in close relation with parameters like ratio, rate, percentage, and so on. We collected these keywords from research papers related to nanodevice development. The compiled list then checked by nanodevice researchers as a basic list for physical quantities.

3.3 Parameter Classification

In the automatic annotation framework, we proposed (Dieb et al, 2012), experiment and evaluation parameters (ExP and EvP) are exclusive categories and identified at the same time. However, in this approach, recall and word boundary identification rate of terms is lower than SMaterial and MMethod. In order to identify good term boundaries to identify parameter terms, it is better to have large variation of compound terms for representing parameter. However, since our corpus size is limited, it is better to merge two parameter categories into one "Parameter" category to enlarge training example for making compound parameter terms. After extracting parameter terms, another machine learning classifier is used for categorize each extracted parameter.

Since two parameter categories (ExP and EvP) represent role of parameter in the paper, it is better to include results of dependency analysis and verbs related to the terms. Therefore, in this paper, we use SVM with following features to identify its category.

1. Information about term
   - A) Surface description of the term
   - B) lemmatized element(s) of the term
   - C) (a) term(s) that is (are) identified as physical quantity by a physical quantities list
2. Dependency structure results
   - A) First verb that given parameter terms directly depends
   - B) First preposition that given parameter terms directly depends
   - C) Lemmas that parameter terms depends.

Following is example of extracted features for a sentence that contains parameter term “temperature”

Figure 4 represents an example of feature extraction. Stanford Core NLP tools are used for generating dependency analysis results. Features that start with “1stVB++”, “1stIN++”, “plemma++”, and “PAR++” represents 2-A,B,C and 1-C respectively. Other elements correspond to 1-A and 1-B.

[1] This list of physical properties of matter is available online at: http://chemistry.about.com/od/matter/a/Physical-Properties-List.htm

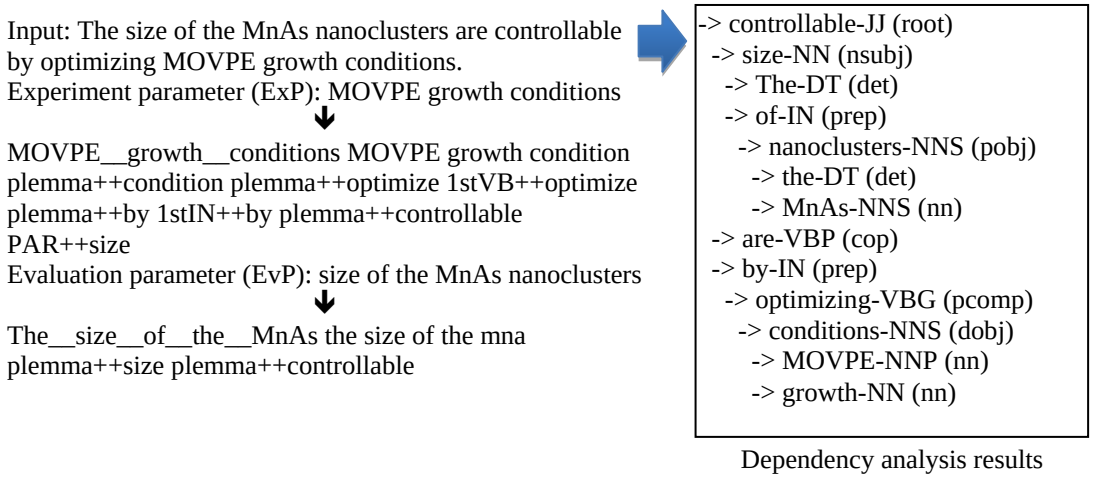

Fig. 4 Example of feature extraction for two parameters from a text

4 Experiments

4.1 Setup

In order to confirm the effectiveness of our proposed framework, we conduct automatic annotation experiments by using nanodevice development papers corpus with following three systems.

- Base line system: we use the automatic annotation framework we previously proposed (Dieb et al. 2012) without physical quantities list for parameter identification.
- Suggested system without parameter classification: we integrated the physical quantities list for parameter identification component in the automatic annotation framework. This component cannot separate experiment (ExP) from evaluation parameter (EvP) because that depends on the context, so the CRF will handle this separation based on training data.
- Suggested system with parameter classification using SVM: after annotating parameters using automatic annotation framework with physical quantities for parameter identification component integrated, SVM will handle the classification of parameters using statistical data into experiment (ExP) and evaluation parameter (EvP).

All systems use CRF++ (CRFpp) implementation of Conditional Random field (CRF) (John et al. 2001) as the machine learning system. Additionally, we have added one more level for the cascading for the SMChar term category, since we found some cases that caused overlap between SMChar and SMaterial that used to be in the same term group. Figure 5 shows an example of the input data in IOB format. Since chemical compound and parameters do not overlap between each other, we use the same feature column for both (CM is chemical entity feature, and PAR is parameter feature based on the physical quantity list).

For experiment 1: the feature CM/PAR has only CM type of values.

For experiment 3: the feature TG4 has Param type of values (general parameter replaces ExP, and EvP without separation. Classification is done independently with SVM)

4.2 Results and Discussion

We have five fully annotated papers. In order to check the performance of these three systems, we use five cross fold validation (training on four papers, and testing on the 5th) for each system. We use tight and loose agreement metrics for analysis (same as explained in section 2.2). Table 2 shows comparative average results for the three experiments. In this experiment, since it is necessary to identify new terms that are only exists in one paper; recall of baseline system is lower than the value in Table 1. Statistical significance test is conducted for the difference between value of baseline system and value of proposed system. "*" represents difference is statistically significant ($P<0.05$) in both side test.

Word	POS	Orth	CM/PAR	TG1	TG2	TG3	TG4	TG5
V/	NP	Other	O	O	O	O	B-ExP	O
Mn	NP	InitCap	B-CM	B-SMaterial	O	O	I-ExP	O
ratio	NN	Lowercase	B-PAR	O	O	O	I-ExP	O
were	VBD	Lowercase	O	O	O	O	O	O
850	CD	DigitNumber	O	O	O	O	O	B-ExPVal
°C	NN	Other	O	O	O	O	O	I-ExPVal

Note: POS=POS tag, Orth=orthogonal feature, CM/PAR=chemical compound/parameter list, TG1, … TG5=term group.

Fig. 5 Example of input data for our suggested system

Table 2: Performance comparison between base line system and suggested one

	Base line system				**Our system without parameter classification**				**Our system with parameter classification using SVM**			
	T_pre	**T_rec**	**L_pre**	**L_rec**	**T_pre**	**T_rec**	**L_pre**	**L_rec**	**T_pre**	**T_rec**	**L_pre**	**L_rec**
SMaterial	0.94	0.92	0.97	0.95	0.94	0.92	0.97	0.95	0.94	0.92	0.97	0.95
MMethod	0.97	0.70	0.98	0.70	0.98	0.72	0.98	0.72	0.98	0.72	0.98	0.72
SMChar	0.87	0.68	0.90	0.70	0.87	0.69	0.90	0.71	0.87	0.69	0.90	0.71
TArtifact	0.88	0.72	0.93	0.76	0.89	0.72	0.93	0.75	0.89	0.72	0.93	0.75
Param	**0.67**	**0.45**	**0.86**	**0.58**	**0.67**	**0.49**	**0.87**	**0.64**	**0.66**	**0.52**	**0.85**	**0.67**
ExP	**0.85**	**0.59**	**0.91**	**0.63**	**0.84**	**0.62**	**0.91**	**0.67**	**0.80**	**0.63**	**0.88**	**0.70**
EvP	**0.50**	**0.32**	**0.78**	**0.51**	**0.50**	**0.35**	**0.78**	**0.55***	**0.50**	**0.38***	**0.76**	**0.57***
ExPVal	0.67	0.42	0.81	0.53	0.67	0.42	0.80	0.52	0.70	0.43	0.82	0.52
EvPVal	0.62	0.34	0.79	0.43	0.63	0.37	0.78	0.46	0.62	0.37	0.78	0.46
overall	0.82	0.63	0.90	0.69	0.82	0.64	0.90	0.71	0.81	0.64	0.89	0.71

Note: T_pre=Tight precision T_rec=Tight recall, L_pre=Loose, L_rec= Loose recall, param= general parameter resulted from merging ExP and EvP

From this result, identification of physical quantities may not affect earlier stage of cascading term extraction (SMaterial, MMethod, SMChar, and TArtifact). For both ExP and EvP, recall is increased when we use physical quantities list, and increase even more when we use SVM classification. Especially for EvP, improvement of recall for loose agreement of both suggested systems are statistically significant at the 5% level. In addition, improvement of recall for tight agreement of using SVM for parameter classification is also statistically significant. These improvements justify usage of physical quantities list is good for improving recall of ExP and EvP. In addition, using both examples of ExP and EvP for identification of compound word boundary is also helpful. Even though precision of ExP and EvP are decreased from baseline system, precision of parameter term (terms that belongs to ExP or EvP) identification is almost same as baseline system.

In order to evaluate detailed behaviour of our proposed system, we check the data whose annotation results are different from baseline system and proposed system. We confirm there are some cases in which new parameters terms that did not exist in training data are extracted. For example, "temperature of (MeCp)2Mn" did not exist at all in the training data, and was not able to be annotated by CRF in test data before adding physical quantities list. After marking "temperature" as parameter, CRF was able to find "temperature of (MeCp)2Mn" by learning compound term construction rule (i.e., PAR of CM may be a parameter term). On the other hand, merging parameters into one category then classifying them seems to be effective, because we can get use of a larger training data. For example, "partial pressure" (where "pressure" is in the physical quantities list) did not exist in the training data. It was not recognized neither by the baseline system nor by the suggested system without parameter classification; However, when we merge the parameters into one category (allowing for larger training data for the identification of parameter), the suggested system with parameter classification was able to identify such entity. Another example, "period of the mask openings" where "period" is in the physical

quantities list. In this example also, only the suggested system with parameter classification were able to identify such entity using physical quantities list.

Regarding the precision, there are several cases whose parameter types are difficult to identify by using sentence information only. For example, "diameter" can be some cases as experiment parameter, and can be evaluation parameter in other cases depending on the context. In "the typical diameter of the initial circular opening" it is ExP, however, in “diameter of the NCs", it is EvP. Even though both texts have similar style, "diameter" can be of different types. In order to improve the quality of parameter type classification, it is better to use other information that cannot be extracted from one sentence (e.g., description of same parameter in other sentences of the same paper, position of first parameter term appearance).

5 Conclusion and Future Development

In this paper, we proposed to use a basic physical quantities list in combination with machine learning technique to improve the extraction of parameter information from research papers related to nanodevice development. We confirmed our proposed system improve recall of parameters between 4% and 7% depending on the type of parameter and analysis metric, and using both example of Experiment parameter and Evaluation parameter for term boundary identification is helpful.

In future studies, and as further manuscripts are annotated, we plan to make use of a larger corpus thus allowing us to evaluate our system on a larger text collection. Additionally, we are planning to develop a new POS tagger for the nanodevice development domain. Brill tagger is developed for general language POS tagging purposes, and changing it with specialized POS tagger might improve the results. It is beneficial to study the impact of each component of the system (POS tagger, the chemical named entity recognizer, or the machine learning system) on the overall performance (Kolluru et al, 2011)

Acknowledgements

This work was partially supported by MEXT/JSPS KAKENHI Grant Number 24240021 and 26540165.

Reference

CRFpp: available online at http://crfpp.googlecode.com/svn/trunk/doc/index.html

GPoSTTL: available online at http://gposttl.sourceforge.net.

Thaer M. Dieb, Masaharu Yoshioka, and Shinjiroh Hara. Automatic Information Extraction of Experiments from Nanodevices Development Papers, IIAIAAI 2012 Proceedings of 2012 IIAI International Conference on Advanced Applied Informatics,pp.42-47,2012

Thaer M. Dieb, Masaharu Yoshioka, and Shinjiroh Hara. Construction of Tagged Corpus for Nanodevices Development Papers, GrC, 2011 Proceedings of the 2011 IEEE International Conference on Granular Computing, pp. 167–170, 2011.

Shinjiroh Hara, and Takahashi Fukui. Hexagonal ferromagnetic MnAs nanocluster formation on GaInAs/InP (111) B layers by metal–organic vapor phase epitaxy. Applied Physics Letters, Vol. 89, 113111, 2006.

Diana de la Iglesia, Stacey Harper, Mark D. Hoover, Fred Klaessig, Phil Lippel, Bettye Maddux, Jeff Morse, André Nel, Krishna Rajan, Rebecca Reznik-Zellen, and Mark Tuominen, (2011, Apr.). Nanoinformatics 2020 Roadmap, National Nanomanufacturing Network. Amherst, MA 01003. [Online]. Available: http://eprints.internano.org/607/1/Roadmap_FINAL041311.pdf, 2011.

Yoshinobu Kano, Makoto Miwa, K Bretonnel Cohen, Lawrence E Hunter, Sophia Ananiadou, and Jun'ichi 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.

BalaKrishna Kolluru, Lezan Hawizy, Peter Murray-Rust, Junichi Tsujii, Sophia Ananiadou. Using Workflows to Explore and Optimise Named Entity Recognition for Chemistry. PLoS ONE, 6(5), 2011. DOI: 10.1371/journal.pone.0020181

John D Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289

Yaoyong Li, Kalina Bontcheva, Hamish Cunningham SVM-based learning system for information extraction. In Proceedings of the First international conference on Deterministic and Statistical Methods in Machine Learning, pp. 319–339, 2005.

Hiroshi Nakagawa and Tatsunori Mori: Automatic term recognition based on statistics of compound nouns and their components, In Terminology, Vol. 9, No. 2, pp. 201-219, 2003

rb tagger: available online at http://rbtagger.rubyforge.org/

Karl Ruping and Woody 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

Anabel Usié, Joaquim Cruz, Jorge Comas, Francesc Solsona, Rui Alves. A tool for the identification of chemical entities (CheNER-BioC). Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2 ,66-69 (2013)

YamCha : available on line at http://chasen.org/ taku/software/yamcha/.

Masaharu Yoshioka, Katsuhiro Tomioka, Shinjiroh Hara, and Takahashi Fukui. Knowledge exploratory project for nanodevice design and manuacturing. In iiWAS '10 Proceedings of the 12th International Conference on Information Integration and Web-based Application & Services, pp. 869-872, 2010.