TCS-TR-B-15-12

TCS Technical Report

NaDev (Nanocrystal Device development) Corpus Annotation Guideline

by

Thaer M. Dieb, Masaharu Yoshioka, Shinjiro Hara

Division of Computer Science

Report Series B

July 14, 2015

Hokkaido University
Graduate School of
Information Science and Technology

Email: diebt@kb.ist.hokudai.ac.jp

Phone: +81-011-706-7161
Fax: +81-011-706-7161

Abstract

In order to support nanocrystal device development, we want to utilize information from related publications. We developed a method for constructing an annotated corpus of publications related to nanocrystal device development to support automatic information extraction. The corpus is named ”NaDev” (Nanocrystal Device Development corpus). In cooperation with domain expert, we defined eight categories of information to be extracted, developed an annotation guideline based on corpus construction experiments, and evaluated the corpus and finalize it with domain expert. NaDev can be used as training data to support an automatic annotation framework to extract experimental information from research papers related to nanocrystal device development. To the best of our knowledge, this is the first attempt to construct a corpus for the development of nanocrystal devices.

NaDev (Nanocrystal Device development) Corpus Annotation Guideline

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

[1] Graduate School of Information Science and Technology, Hokkaido University, Japan

[2] Research Center for Integrated Quantum Electronics, Hokkaido University, Japan

Content

Introduction

Nanocrystal device development is an area of nanoscale research where nanoelectronic devices are developed for future nanoelectronic industry applications using electronic materials. In order to support nanocrystal device development, we want to utilize information from related publications.

We developed a method for constructing an annotated corpus of publications related to nanocrystal device development to support automatic information extraction. The corpus is named ”NaDev” (Nanocrystal Device Development corpus). This is a joint research project between the Research Center for Integrated Quantum Electronics (RCIQE) and the Division of Computer Science at Hokkaido University, Japan. To the best of our knowledge, this is the first attempt to construct a corpus for the development of nanocrystal devices.

In cooperation with domain expert, we defined eight categories of information to be extracted. We developed an annotation guideline in a spiral manner based on feedback from graduate school annotators after each corpus construction experiment. After reaching a reliable Inter-Annotator Agreement (IAA), we evaluated the corpus and finalize it with domain expert.

The corpus currently has 5 fully annotated papers, 392 sentences, and 2,870 annotated terms in eight information categories.

NaDev can be used as training data to support an automatic annotation framework to extract experimental information from research papers related to nanocrystal device development.

If you would like to use the NaDev corpus, please send an e-mail request to (NaDev@kb.ist.hokudai.ac.jp).

Annotation guideline

Version 1.0 (release), July 10, 2015

Motivation

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

Followings are categories annotated in this corpus.

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

Annotation

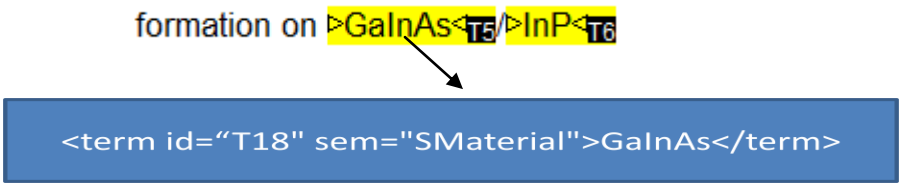
The annotator can use annotation support tool Xconc suite [1] that was originally developed originally to annotate biomedical entities in GENIA project [2]. XCnoc uses XML to represent the annotation.

Below is an example of the annotation

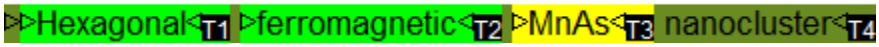
Unannotated text

the growth of GaInAs and InP layers

Annotated text



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



Detailed explanation about each category

- **Material Information (SMaterial)**

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

Example

Hydrogen

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

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

Example

the growth of GaInAs

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

Example

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

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

Example

estimated p[$(MeCp)_2Mn$]

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

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

Example

InP (001) wafers

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

Example

on the III–V compound

group III

➢ **<u>Material Characteristics (MChar)</u>**

Defines electro-chemical characteristics of material.

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

Example

ferromagnetic MnAs nanoclusters

hexagonal NCs

- **Experiment Parameter and Experiment Parameter Value (ExP,ExPVal)**
    - Sometimes the parameter value is not stated concretely, however it is referenced to some value that not stated concretely within the text, in that case we also consider it as parameter value

Example

at room temperature

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

Example

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

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

Example

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

with increasing growth temperatures from 550 to 700 ^{o}C

increasing the MnAs growth time from 3 to 30 min on

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

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

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

Example

top surfaces of the MnAs NCs are atomically flat

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

Example

self-assembled on planar GaInAs surfaces

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

Example

strong ferromagnetic coupling

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

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

➢ **Method (Mmethod)**

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

Example

metal-organic vapor phase epitaxy

MOVPE

buildup fabrication and magnetic domain characterizations

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

Example

The MOVPE growth conditions.

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

Example

SA-MOVPE, LT-MBE

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

Example

ZnO fine particles deposited on silica surfaces

deposited is NOT a method here

➢ **Final product (TArtifact)**

- When a final product is combined with the name of some material, the product name should be annotated as product; however the material name within that product should also be annotated as overlapped term inside the first term.
  Example
  We fabricated InGaAs nanowires (NWs)

- Layers are not considered as final products

Example

GaInAs/InP (1 1 1) B layers,

➢ **General comments**

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

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

Example

of the NWs increased when the growth temperature was

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

Example

above 675 ºC

from 300 to 50 nm

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

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

Example

The heights

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

Example

metal-organic vapor phase epitaxy (MOVPE),

tri-methyl-gallium (TMGa)

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

Example

(TMGa)

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

Example

magnetic domain characterizations of anisotropic-shaped MnAs nanoclusters

The temperature of $(MeCp)_2Mn$

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

The estimated partial pressures for TMIn and TMGa

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

Example

aspect ratio of the initial mask openings

numbers of the MnAs nanoclusters

Appendix

To evaluate the reliability of the annotation guideline, we conducted a corpus construction experiments. In each experiment, two graduate course students were asked to annotate the same paper independently. We used Kappa coefficient to test Inter-Annotator agreement (IAA). Two metrics were used for the analysis, tight agreement, which considers the term boundary and term category, and the other is loose agreement, which ignores the term boundary.

In the first experiment, the kappa coefficient was 41% in case of tight agreement, and 74% in case of loose agreement. These results were not sufficient enough according to Green (1997) [3]. We revised the guideline based on feedback from annotators, and then conducted another experiment. We reached a kappa coefficient of 0.63 for tight agreement and 0.77 for loose agreement.

Even though the corpus construction guideline reach a reliable level in case of loose agreement. It was necessary to evaluate this corpus and finalize it with a domain expert researcher to ensure reliability. We constructed a version of the corpus based on the agreed annotation between the two annotators. Careless mistakes, such as one annotator missing to add an annotation, or typical types of disagreement when annotators misunderstand the guideline, were easily checked in the discussion after each annotation experiment, so it is considered as agreed annotation.

A domain expert checked and evaluated the corpus. To improve annotation consistency, the domain expert suggested few modifications to the guideline. Additionally, the discussion reveals that NaDev has two types of papers, four papers focusing on the synthesis of new nanomaterials [4-7], and one focused on the characterization of nanomaterials [8].

Based on domain expert's revision, we have made a finalized version of the corpus. In order to evaluate annotation reliability of graduate students, we compare this finalized version with the original corpus constructed before the evaluation experiment. Evaluation shows that, if we exclude effect of new guideline modifications made by the domain expert: 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%). It is important to have discussions between annotators after the annotation process. Such discussions can resolve mismatches caused by careless mistakes or misunderstanding of the guideline. Recall is also high (ranging between 91% and 100%). 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.

Acknowledgments

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.

References

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

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

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

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

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

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

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

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