Experimental Information Extraction from Nanocrystal Device Development Research Papers.

Thaer M. Dieb
National Institute for Material Science
Graduate School of Frontier Sciences, The University of Tokyo

Contents

- Background and Motivation
- NaDev corpus construction

  Annotated corpus for nanocrystal device research papers
- NaDevEx framework development

  Automatic information extraction framework for nanocrystal device research papers using machine learning
- Chemical named entity recognition using ensemble-learning
- Utilization of extracted information
- Conclusion and future work
- Publication list

Contents

- Background and Motivation
- NaDev corpus construction

  Annotated corpus for nanocrystal device research papers
- NaDevEx framework development

  Automatic information extraction framework for nanocrystal device research papers using machine learning
- Chemical named entity recognition using ensemble-learning
- Utilization of extracted information
- Conclusion and future work
- Publication list

Nanocrystal development

Nanocrystal Researcher

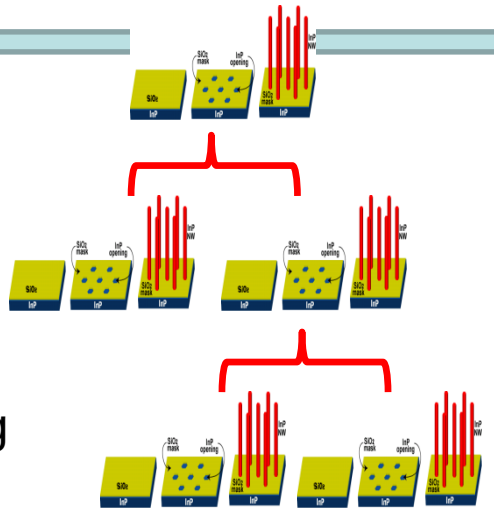
First Principle

+

Tacit knowledge about Manufacturing process

⇩

Depends on experience and not well systematized



Trial and error experiments

Final product (solar cell, …)

Approach

?

Solar cell

Accelerate device development process

Find papers that discuss similar experiment settings

Research papers

Final product: Transistor
Material: GaInAs, AlAs, …
Parameters: temperature,…

Final product: Solar cell
Material: MnAs, GaInAs, …
Parameters: Pressure, …

Final product: Nanowire
Material: MnAs, AlAs, …
Parameters: Gas flow rate, …

The development of solar cell using SAMOVPE…
Introduction
……
Parameter settings
.....
……
Evaluation criteria
……
Conclusion
……..

Objectives

- Provide a framework to extract experimental information in nanocrystal device research papers based on an annotated corpus approach.
  - Annotated corpus construction (NaDev)
  - Automatic annotation framework (NaDevEx) based on machine learning
- This framework is an application of knowledge engineering related to this domain.

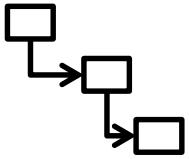
Utilization of research papers' information using text mining

Bioinformatics

Genes, protein,…

Research papers

Pathway



Chemical information

MnAs, …, AlAs, ….

Research papers

MnAs, AlAs

Topic1

MnAs, …, AlAs,

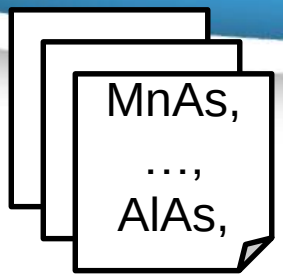
MnAs 100
AlAs 70
…

Topic2

SiO_2, …, Cu, ….

SiO_2 100
Cu 60

Machine learning based using annotated corpus approach



Corpus construction

Machine learning based on a sequence labeling task

Sentence: Germanic acid is inorganic. Mucic acid is a nitric acid.

Token	Lemma	POS	chemical	target
Germanic	Germanic	Noun	O	B-CM
acid	acid	Noun	Chem	I-CM
is	be	Verb	O	O
inorganic	inorganic	Noun	O	O
Mucic	Music	Noun	O	B-CM
acid	acid	Noun	Chem	I-CM
…..	…..	…..	…..	…..

It is crucial to construct a well-defined corpus

Corpus construction related to nanocrystal device development

	Chemical named entity recognition		Nanoinformatics		
Corpus	SCAI	CHEMDNER	nanotoxicology	nanomedicine	NaDev
Materials	○	○	○	○	○
Material characteristics	-	-	○	○	○
Target Product	-	-	○	○	○
Parameters and values	-	-	-	-	○
Manufacturing Method	-	-	-	-	○

SCAI: Kolarik et al. 2008, Klinger et al. 2008
CHEMDNER: Krallinger et al. 2015
Nanotoxicology: Garcia-Remesal et. Al. Biomed. Res. Int.
Nanomedicine: Gaheen et. Al. Comput. Sci. Disc 2013

Contents

- Background and Motivation
- NaDev corpus construction

  Annotated corpus for nanocrystal device research papers
- NaDevEx framework development

  Automatic information extraction framework for nanocrystal device research papers using machine learning
- Chemical named entity recognition using ensemble-learning
- Utilization of extracted information
- Conclusion and future work
- Publication list

Objectives

- Tag set design
  - Extracting useful information for domain researchers.
- Annotation guideline
  - Support consistent annotation by human annotators.

Dieb, T.M., Yoshioka, M., Hara,S.: Construction of tagged corpus for Nanodevices development papers. 2011 IEEE International Conference on Granular Computing (GrC),

NaDev corpus construction

Corpus construction approach

Tag set design
(Information to be extracted)

Annotation guideline design
(detailed explanation with examples on annotation for novices)

Annotation reliability experiment
(check if guideline examples are enough for reliable annotation)

Corpus evaluation with domain expert

NaDev corpus

Collaboration with domain expert

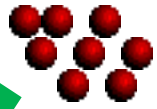
Collaboration with domain graduate students

Collaboration with domain expert

NaDev corpus construction

Preliminary tag set design

MnAs Source material

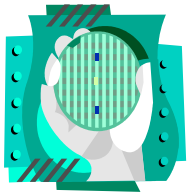


Hexagonal (Source material characteristics)

MOVPE (Manufacturing method)

Growth temperature (Experiment parameters) → 30 C (Values)

Nanowires (Product)

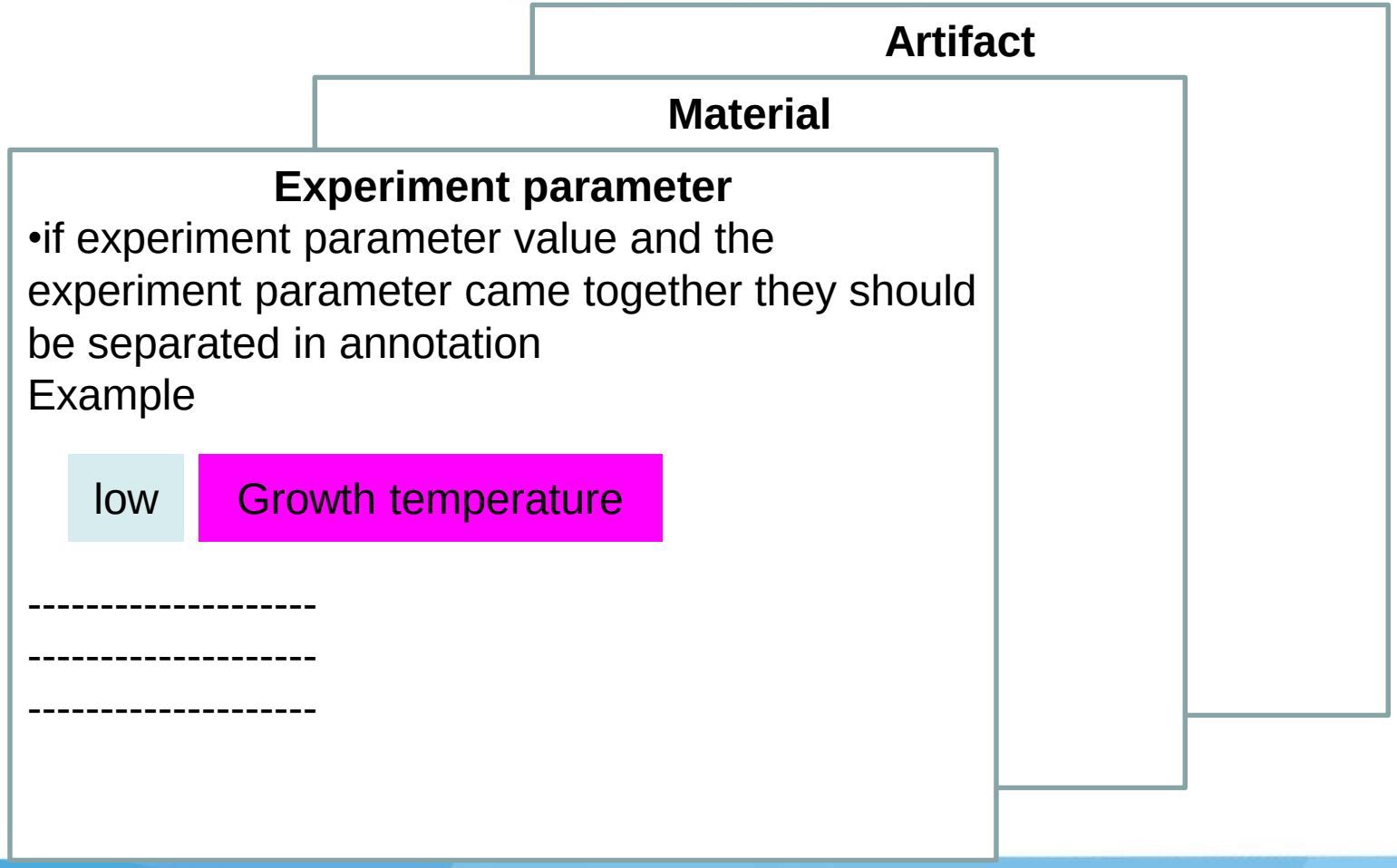


Nanowires conductivity (Evaluation parameters) → 2 µS/cm (Values)

NaDev corpus construction

Construction guideline

In collaboration with domain expert, we developed an annotation guidelines



Reliability measures

Tight and loose agreement

First annotator

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.

Second annotator

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.

Tight Agreement

Loose Agreement

Inter annotator agreement

- Kappa coefficient
  - Tight agreement
    - K=0.63 good agreement but still not sufficient.
  - Loose agreement
    - K=0.77 reliable*

*Interpretation of Kappa statistics coefficient Green, Annette M. (1997). Kappa statistics for multiple raters using categorical classifications. In Proceedings of the Twenty-Second Annual SAS Users Group International Conference (online), San Diego, CA

Domain expert evaluation

- Data separation
  - Agreed part of 2 annotators
  - Disagreement part of 2 annotators
- Domain expert tasks (Prof. Hara of RCIQE)
  - appropriateness of the agreed annotations
  - Choose the appropriate annotation for each disagreed-annotation case
  - Annotate any terms that had not been annotated

NaDev corpus construction

Data setup

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

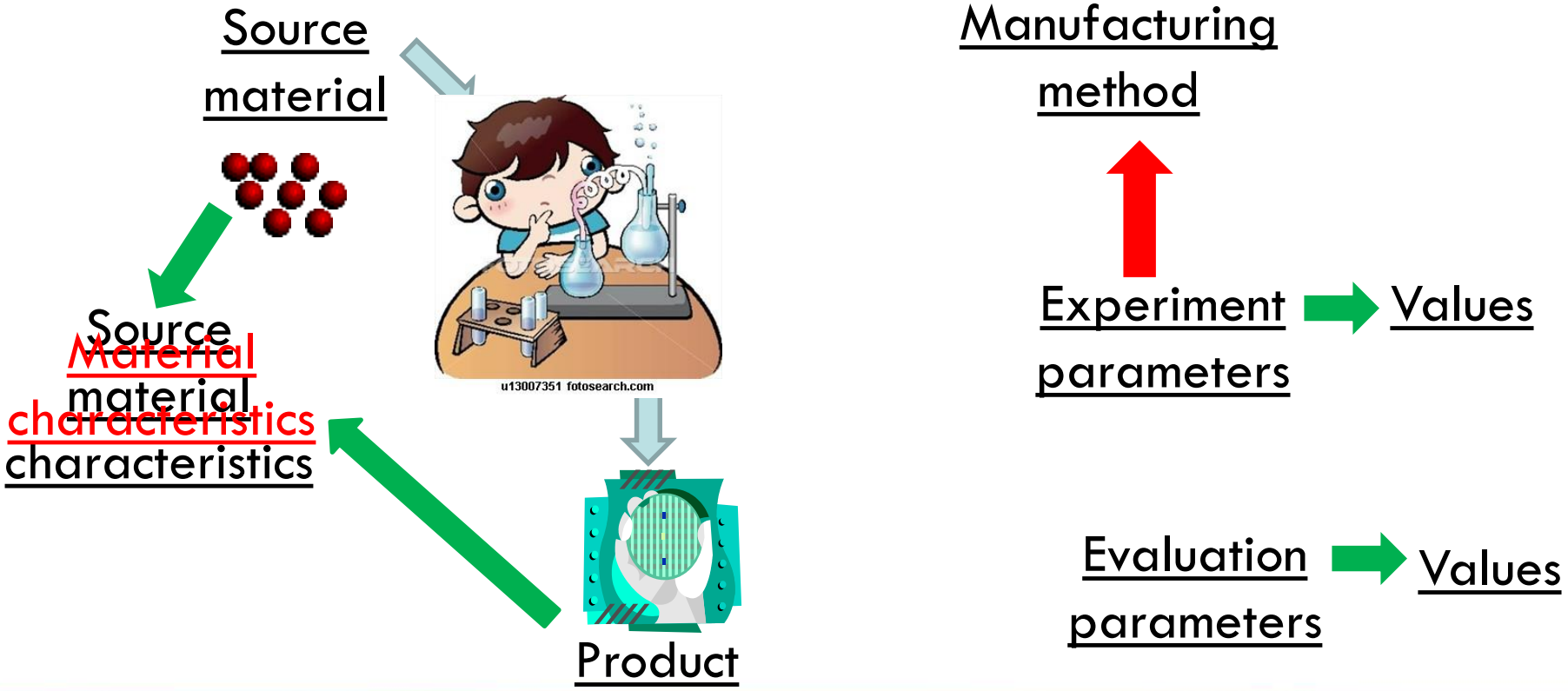
ferromagnetic behavior

Corpus papers types

- The corpus has 5 papers of 2 types based on writing style:
  - Synthesis papers (1-4): focus on the synthesis of new materials.
  - Characterization papers (5): focuses on the analysis and characterization of materials.

Tag set modification

To improve annotation consistency, domain expert suggested few guideline modifications and a revised tag set.



Annotation quality analysis

- Final version of the corpus constructed with all modifications by domain expert.
- We compared this corpus with the original corpus constructed for the evaluation experiment.

Expert Vs. novice

Information category	average	
	Precision	Recall
Source Material	0.97(0.97)	0.79(0.99)
Source Material Characteristics	0.93(0.96)	0.84(0.93)
Manufacturing Method	1.00(1.00)	0.91(0.91)
Target Artifact	0.99(0.99)	0.90(0.90)
Experiment Parameter	1.00(1.00)	0.91(0.91)
Evaluation Parameter	0.98(0.980	0.91(0.91)
Experiment Parameter Value	0.99(0.99)	0.97(0.97)
Evaluation Parameter Value	1.00(1.00)	0.86(0.86)
total	0.98(0.98)	0.86(0.92)

Number in parenthesis are excluding guideline modification effect

NaDev Corpus release

- 5 full annotated papers
- 392 sentences,
- 2870 terms
- 8 information categories.

NaDev corpus construction

Corpus example

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/Mn ratio of the supplied source gases.

==========

Source Material (SMaterial): SiO2
Material Characteristic feature (MChar): (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

Contents

- Background and Motivation
- NaDev corpus construction

  Annotated corpus for nanocrystal device research papers
- NaDevEx framework development

  Automatic information extraction framework for nanocrystal device research papers using machine learning
- Chemical named entity recognition using ensemble-learning
- Utilization of extracted information
- Conclusion and future work
- Publication list

Objectives

- Construction of machine learning based extraction system that uses NaDev corpus.
  - Adaptation of similar technique in bioinformatics.
    - Analysis of tag structure
    - Utilization of external resources.
  - Evaluation of the system based on the corpus.

Dieb, T.M., Yoshioka, M., Hara,S., Newton,M.: Framework for automatic information extraction from research papers on nanocrystal devices. Beilstein Journal of Nanotechnology, 6, 1872–1882.
Dieb, T.M., Yoshioka, M.:Extraction of Chemical and Drug Named Entities by Ensemble Learning Using Chemical NER Tools Based on Different Extraction Guidelines. Transactions on Machine Learning and Data Mining, 8, 2 pp. 61-76.

Entity dependency

- MnAs + “nanoclusters” → MnAs nanoclusters
- <SMaterial> nanocluster → <Tartifact>
- Basic features such as POS, Orthogonal are too general to identify tags in some cases.
- Defining common characteristics of certain entities can help identifying new similar entities.
- Most of the chemicals are source materials

Training	MnAs nanocluster	POS: noun Orth: InitalCap CH
Testing	AlAs nanocluster	POS: noun Orth: InitalCap CH

Physical quantity list

- Most parameters contain head nouns that are physical quantities, temperature, pressure..
- Identifying physical quantities can support identifying parameters.
- A physical quantity list was prepared.

Training Temperature of (MeCp)2Mn — POS: noun, Orth: InitalCap, PAR

Testing Pressure of AsH3 — POS: noun, Orth: InitalCap, PAR

Tag groups

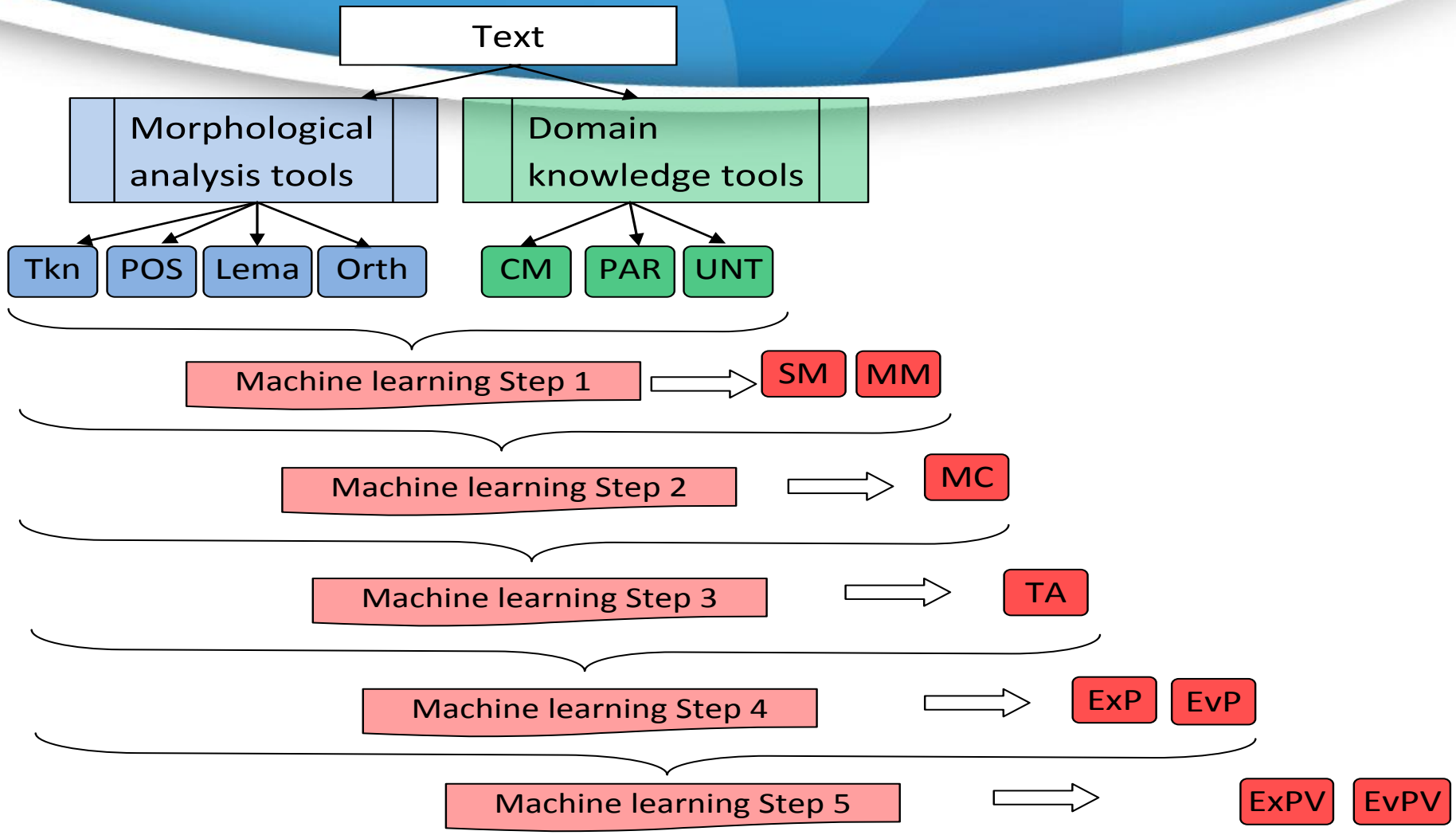
- Tag group 1: Source material, and Manufacturing method SMaterial, and MMethod.
- Tag group 2: Material characteristics MChar
- Tag group 3: Final product TArtifact.
- Tag group 4: parameters ExP, EvP.
- Tag group 5: parameters value: ExPVal, EvPVal

Cascading style annotation

- Using information of one tag to assess another tag.
- Separate the annotation into 5 step-by-step levels based on overlapping structure between entities.

NaDevEx framework development

NaDevEx outline



Legend: Tkn: token, POS: part of speech, Lema: lemmatization, Orth: orthogonal, CM: chemical named entity, PAR: physical quantity matching, UNT: measurement unit list matching, SM: SMaterial, MM: MMethod, MC: MChar, TA: TArtifact, ExP: ExP, EvP: EvP, ExPV: ExPVal, and EvPV: EvPVal

NaDevEx evaluation

- Comparing with human annotators
  - 5 fold cross validation (training on 4 papers and testing on the 5th)
  - Evaluation metrics
    - Precision: fraction of identified entities that are correct.
    - Recall: fraction of correct entities which are identified.

NaDevEx framework development

NaDevEx Vs. human annotators

	Human		NaDevEx	
	Precision	**Recall**	**Precision**	**Recall**
SMaterial	0.97(0.97)	0.79(0.99)	0.95	0.94
MMethod	1.00(1.00)	0.91(0.91)	0.97	0.73
MChar	0.93(0.96)	0.84(0.93)	0.94	0.67
TArtifact	0.99(0.99)	0.90(0.90)	0.88	0.73
ExP	1.00(1.00)	0.91(0.91)	0.93	0.68
EvP	0.98(0.980	0.91(0.91)	0.78	0.55
ExPVal	0.99(0.99)	0.97(0.97)	0.80	0.53
EvPVal	1.00(1.00)	0.86(0.86)	0.75	0.39
Total	0.98(0.98)	0.86(0.92)	0.89	0.69

Number in parenthesis are excluding guideline modification effect
Underlining indicates that a difference is statistically insignificant at the 5% level (P >= 0.05).

Results analysis

- NaDevEx is good for precision, not as much for recall.
- Performance with information category with rich domain knowledge (SMaterial), is almost comparable to human annotators.

Contents

- Background and Motivation
- NaDev corpus construction

  Annotated corpus for nanocrystal device research papers
- NaDevEx framework development

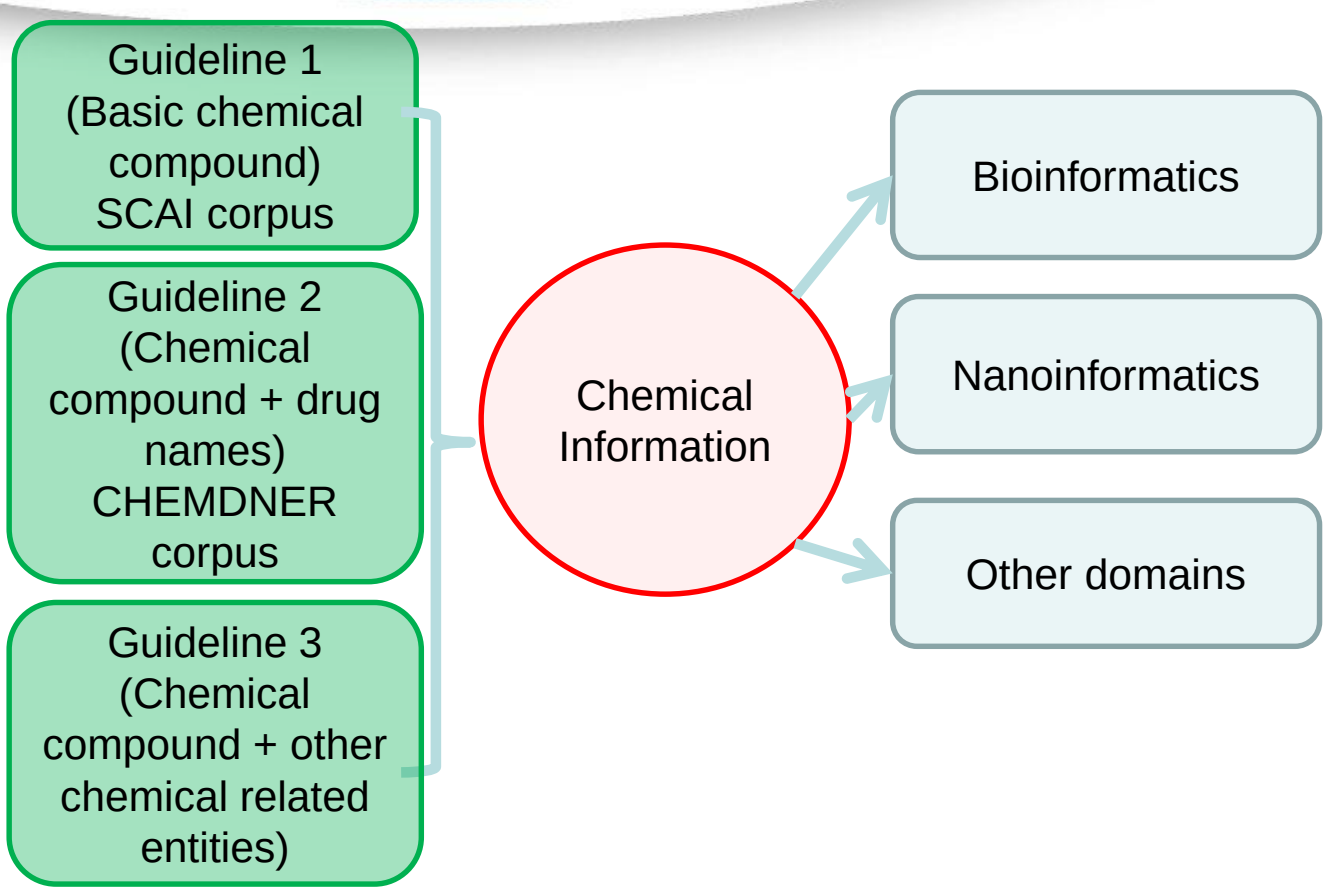
  Automatic information extraction framework for nanocrystal device research papers using machine learning
- Chemical named entity recognition using ensemble-learning
- Utilization of extracted information
- Conclusion and future work
- Publication list

Material information as chemicals

- Considerable amount of chemical information exist in nanocrystal device related papers.
- To improve chemical entity recognition, we developed a chemical entity recognizer using ensemble learning.

Chemical entity recognition

Different annotation guideline



Different annotation guideline

- Annotation guidelines are not exactly the same
  - Different definition of what is chemical entity.
    - "Fatty acid" is chemical entity in CHEMDNER corpus but not in SCAI corpus.
  - Different entities of interest
    - drug names are added in CHEMDNER corpus.
    - Modifiers are added in SCAI corpus
      - Derivatives
  - Boundary mismatch
    - Carbamate **and** N- sulfocarbamoyl toxins CHEMDNER.
    - Sulfoxides and sulfones SCAI

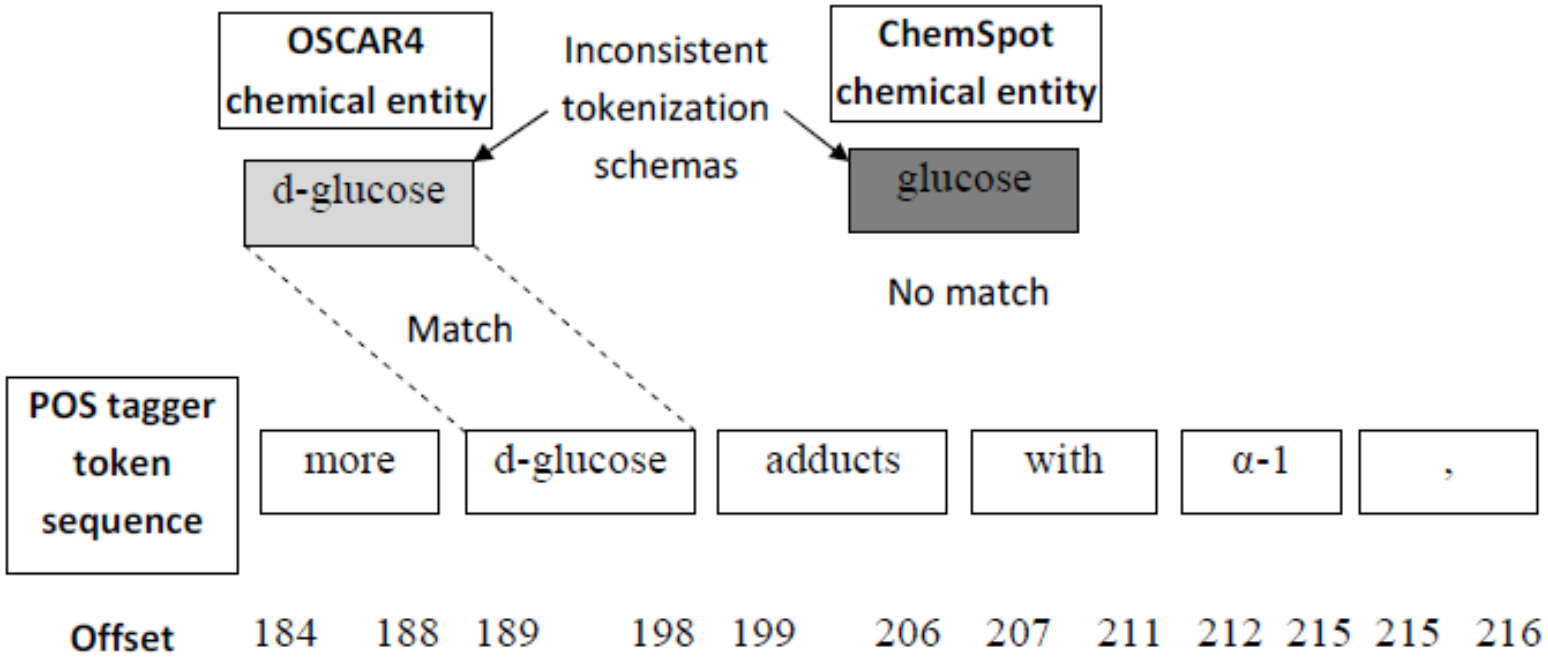
Chemical entity recognition

Tools

- Different CNER tools with different charactaristics
  - SERB-CNER
    - **C**hemical **N**amed **E**ntity **R**ecognizer
    - Rule-based using regular expression
    - Additional syntactic rules to enhance precision
    - Used for extracting chemical compounds in nanodevice development publications.
  - Oscar 3 and 4: Rule based with dictionary
    - https://bitbucket.org/wwmm/oscar4/wiki/Home
  - ChemSpot: Hybrid approach of machine learning and dictionary
    - https://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/chemspot/chemspot/

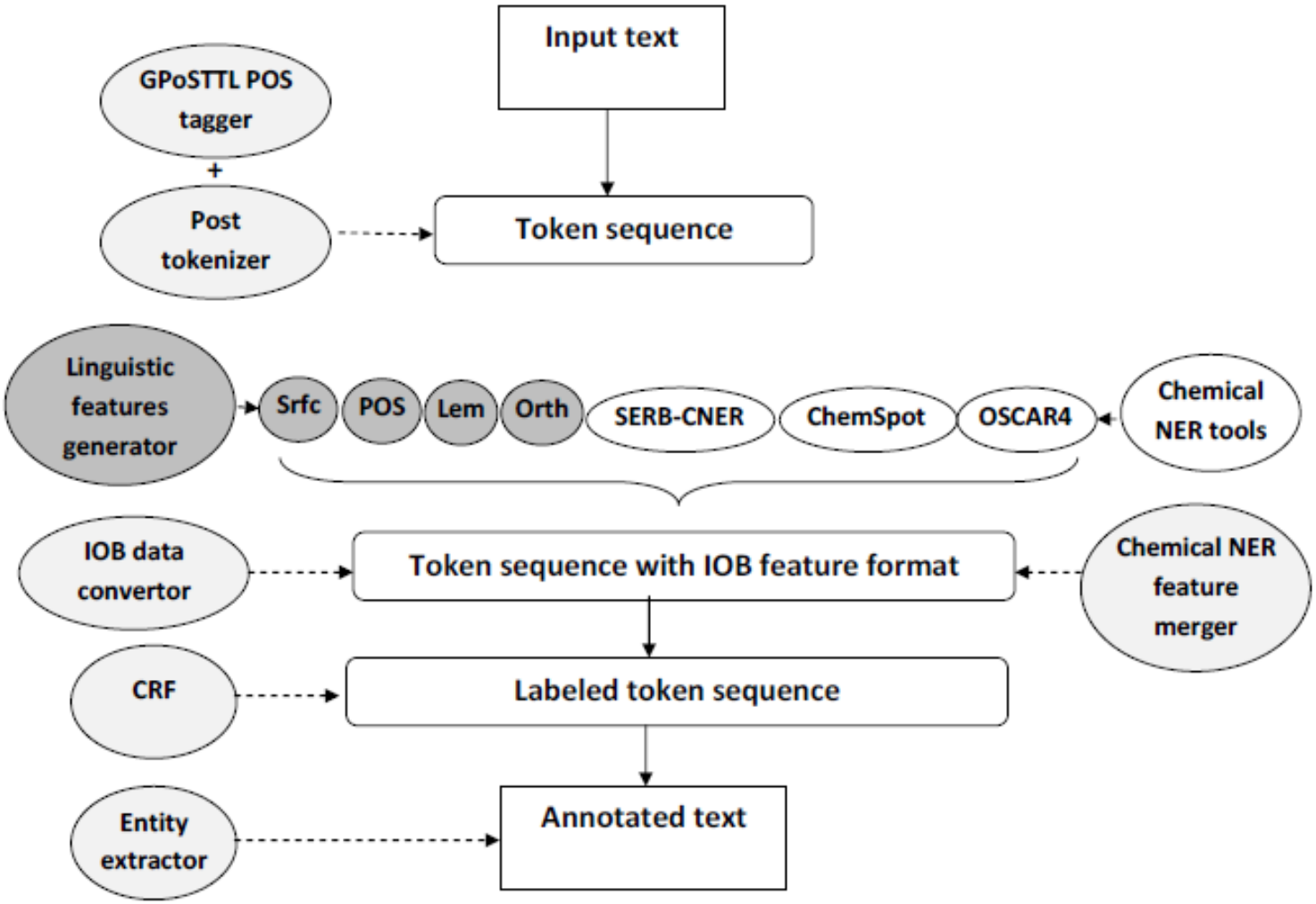
Chemical entity recognition

Tokenization Mechanism



Chemical entity recognition

System activity diagram



Chemical entity recognition

Ensemble Approach Compared to Simple Domain Adaptation

Average system performance on the BioCreative IV, CHEMDNER corpus

	Macro-average			Micro-average		
	Precision	Recall	F-score	Precision	Recall	F-score
SERB-CNER+CRF	85.31	69.52	74.23	89.24	68.15	77.28
ChemSpot+CRF	85.26	76.77	78.84	88.10	76.21	81.72
OSCAR4+CRF	86.00	76.41	78.88	88.65	74.67	81.06
Ensemble	78.72	70.83	72.72	82.26	70.86	76.13
Ensemble/p.tok	86.62 $*	79.46 $*#	81.13 $*#	88.76 *	78.60 $*#	83.37 $*#

CRF: Conditional Random Field. Ensemble = (SERB-CNER+ChemSpot+OSCAR4+CRF) without post-tokenization. Ensemble/p.tok = (SERB-CNER+ChemSpot+OSCAR4+CRF) with post-tokenization. Underlining indicates significant values for the ensemble system compared with the performance before post-tokenization. A dollar sign ($) indicates a significant value compared with SERB-CNER combined with CRF. An asterisk (*) indicates a significant value compared with ChemSpot combined with CRF. A hash (#) indicates a significant value compared with OSCAR4 combined with CRF. All
significant measures were at the 0.05 level (P < 0:05).

Discussion

- Ensemble approach is generally promising,
  - each chemical NER tool can contribute some unique new findings, thereby leveraging the performance.
- Text-tokenization method considerably affects the performance of the system.

Contents

- Background and Motivation
- NaDev corpus construction

  Annotated corpus for nanocrystal device research papers
- NaDevEx framework development

  Automatic information extraction framework for nanocrystal device research papers using machine learning
- Chemical named entity recognition using ensemble-learning
- Utilization of extracted information
- Conclusion and future work
- Publication list

Utilization of extracted information

- We propose several methods to utilize the extracted information
  - Paper retrieval system
  - Finding similar papers through paper clustering

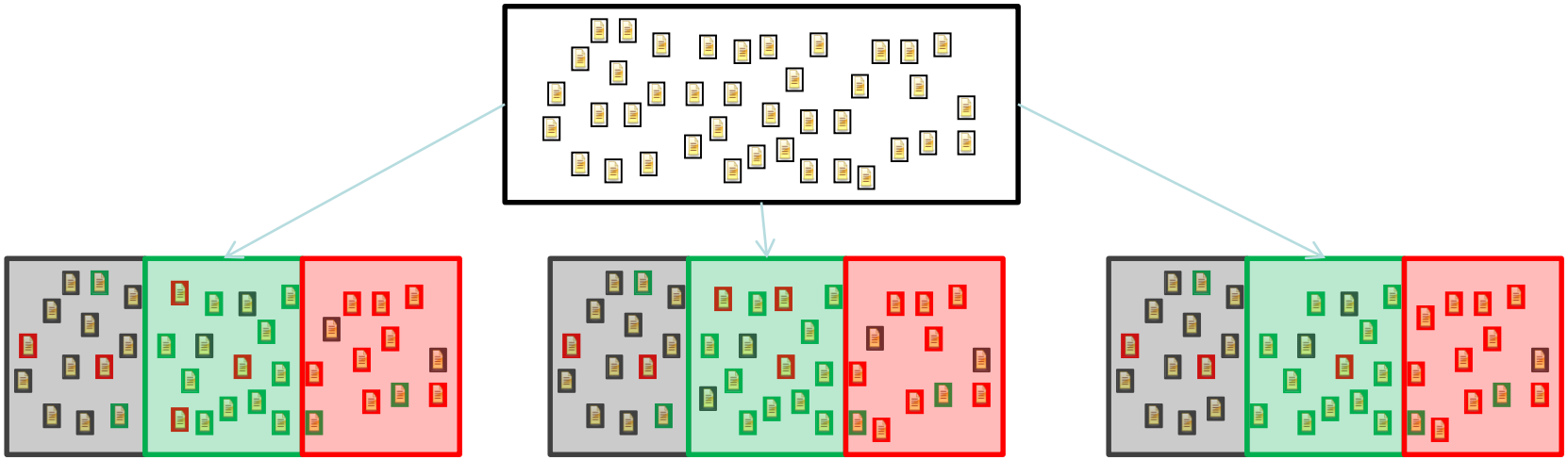
Paper retrieval system

- User could find papers that involve certain material in developing certain final product.
  - MnAs for solar cell
- Helpful in finding recent analysis in research papers.
- Support the data-collection process.

Utilization of extracted information

Paper clustering

- Find similarity between research papers based on different similarity metrics.



Contents

- Background and Motivation
- NaDev corpus construction

  Annotated corpus for nanocrystal device research papers
- NaDevEx framework development

  Automatic information extraction framework for nanocrystal device research papers using machine learning
- Chemical named entity recognition using ensemble-learning
- Utilization of extracted information
- Conclusion and future work
- Publication list

Conclusion

- We developed a framework to support experimental information extraction from nanocrystal device development papers.
  - An annotated corpus was developed in collaboration with a domain expert.
  - NaDev construction guideline was released, NaDev can be distributed upon request.
  - An automatic information extraction system was build using that corpus.
    - This system uses cascading style machine learning and NLP techniques
    - This system is almost not defeated by human annotators for domain knowledge rich feature

Next step

- Developing application to utilize the extracted information.
  - Accelerate the development of new devices
  - Highly depend on quality of information extraction.
    - Improve the quality of NaDevEx
      - Increase the size of the corpus.
      - Construct resources for representing domain knowledge.
        - » Physical quantity list is not enough to identify parameters.

Conclusion and future work

Corpus size increment

- Based on experience from other domains
  - Corpus start with smaller size
  - Attract researchers to expand this research activity
- NaDev uses full text of research papers instead of abstract
  - Abstracts usually do not contain detailed explanation about experiments' parameters in relation with output evaluation.
  - Abstract can offer wider variety of experimental information.
  - Extend the corpus using large number of abstracts.

Future vision

- Support data collection process in nanoinformatics domain to accelerate development of new devices.
  - NaDevEx can be used to find research papers that contain recent analysis results on nanocrystal devices in a precision oriented manner.

有難う御座いました

Thank you for listening