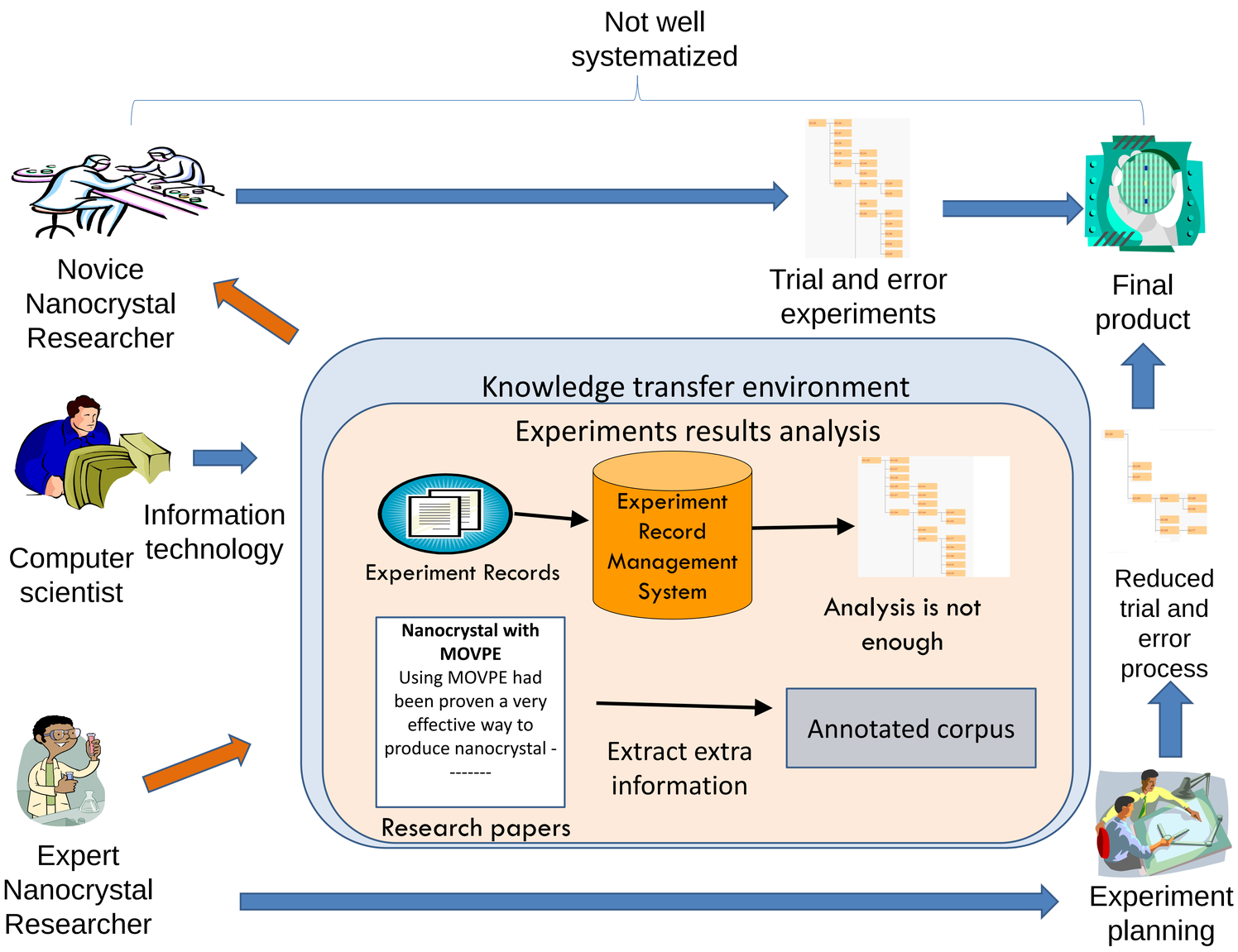
Automatic Information Extraction of Experiments from Nanodevices Development Papers

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

Motivation

Automatic extraction (using machine learning techniques) of useful information from Nanodevices development papers to help analyzing the experimental results of the Nanodevices developing experiments.

Background

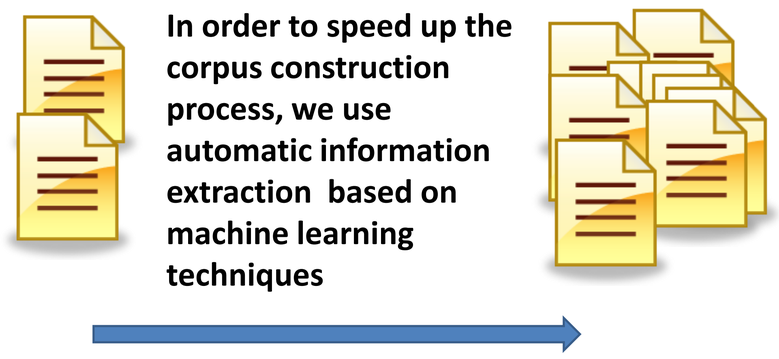


Corpus Example

We fabricated InGaAs nanowires (NWs) in SiO2 mask openings on a GaAs(111)B substrate at growth temperatures of 600–700°C using catalyst-free selective-area metal organic vapor phase epitaxy. At a growth temperature of 600°C, particle-like depositions occurred, but they decreased in number and density when the growth temperature was increased to 650 °C and disappeared above 675 °C. The heights and growth rates of the NWs increased when the growth temperature was increased and the mask opening diameter was decreased from 300 to 50 nm. Photoluminescence (PL) spectra measured for the NWs indicated a blue shift in the peak from 0.95 to 1.3 eV as the growth temperature was increased from 600 to 700 °C, indicating an increase in the Ga composition from 62 to 88% in the InGaAs NWs.

A) Material Information InGaAs
B) Material features (111)B
C) Experimental parameters growth temperatures
D) Value of experimental parameters 600–700°C
E) Evaluation parameters Photoluminescence (PL) spectra
F) Value of the evaluation parameter 0.95 to 1.3 eV
G) Manufacturing method catalyst-free selective-area metal organic vapor phase epitaxy
H) Artifact nanowires

Automatic Information Extraction



Domain Related Issues

Baseline System

Using basic system from bioinformatics field as a reference.
Baseline system learns tags based on word features only.

Word	POS	Orthogonal	Tag category
Inhibition	NN	InitCap	B-NP
of	IN	Lowercase	B-PP
NF-kappaB	NN	OtherHyphon	B-NP
activation	NN	Lowercase	I-NP
reversed	VBD	Lowercase	B-VP
the	DT	Determiner	B-NP

Suggested System

Add a new feature of chemical entity (using chemical entity recognition)
Usage of information of one tag group to assess tags in another tag groups (Cascading).

Word	POS	Orthognal	Chemical Entity	Tag1	Tag2	Tag3	Tag4
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	I-TArtifact	O	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
coupling	NN	Lowercase	O	O	O	I-EvP	O

Level 1 Level 2 Level 3 Level 4

Performance

	Tight /precision		Tight /recall		Loose /precision		Loose /recall	
	Base	Sug.	Base	Sug.	Base	Sug.	Base	Sug.
SMaterial	0.98	0.99	0.88	0.96	0.99	0.99	0.89	0.96
SMChar	0.89	0.89	0.68	0.69	0.89	0.89	0.68	0.69
Mmethod	0.94	0.93	0.79	0.86	0.94	0.95	0.79	0.88
TArtifact	0.87	0.86	0.71	0.73	0.91	0.93	0.75	0.79
ExP	0.91	0.91	0.74	0.81	0.98	0.97	0.80	0.86
EvP	0.76	0.76	0.59	0.60	0.86	0.87	0.66	0.69
ExPVal	0.71	0.72	0.49	0.57	0.90	0.85	0.62	0.67
EvPVal	0.84	0.86	0.57	0.60	0.96	0.97	0.65	0.67
Overall	0.89	0.89	0.71	0.76	0.94	0.94	0.75	0.80

Conclusion

Confirm that chemical named entity recognition has improved the quality of the annotation.
Cascading style may improve the recall.
It is necessary to discuss the effectiveness of the cascading style annotation by using large corpus.