NTCIR-8 Patent Mining Task:
Extracting Technology and Effect Entities in Patents and Research Papers

NUS National University of Singapore

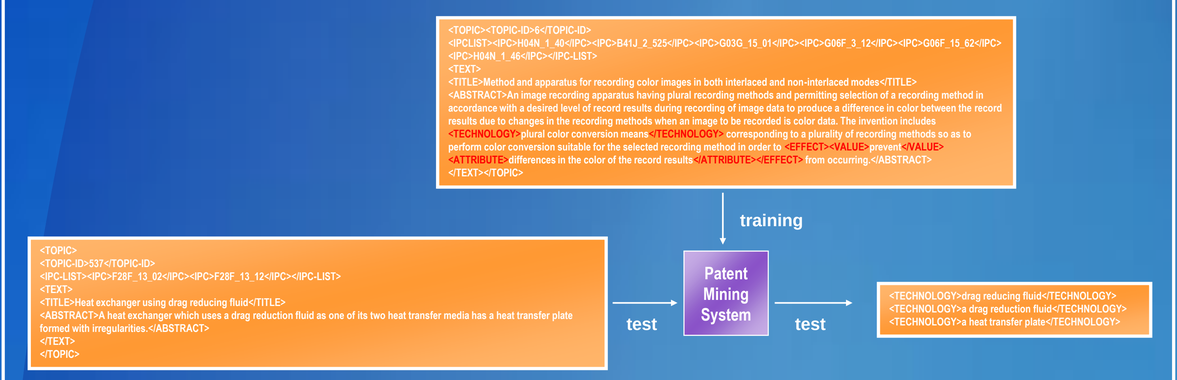
Department of Mechanical Engineering

Jingjing Wang, Han Tong Loh, Wen Feng Lu

{wang_jingjing, mpelht, mpelwf} @ nus.edu.sg

Introduction

The purpose of the Technical Trend Map Creation task is to extract expressions of element technologies and their effects from research papers and patents. Entities involved are TECHNOLOGY, EFFECT, ATTRIBUTE, and VALUE. The organizer offered tagged topics for training and untagged topics for test, in which raw text of each topic is the title and the abstract of a patent or a paper. The evaluation is based on Recall / Precision / F-measure.



We considered it as a Information Extraction / Named Entity Recognition (NER) task, and we started from an advanced statistical model with many features, then we did slight modification on the original model in order to improve the performance, and we further added some patterns and invoked a pattern-based method.

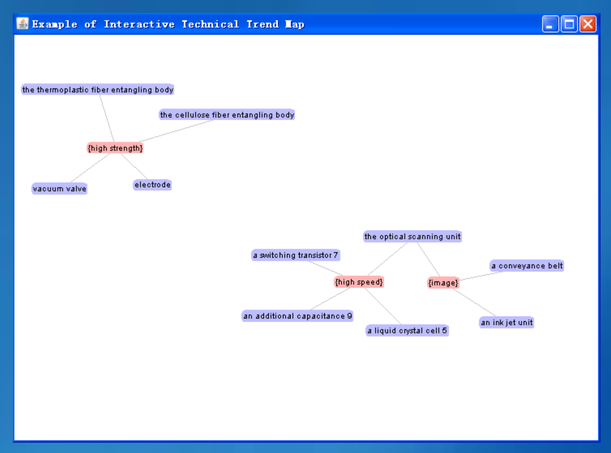
Achievements & Future Works

Achievements:

1. We had tried both statistical method and pattern-based method and we obtained a relatively good result.
2. The tag update rule works. A big improvement was achieved.
3. In our case, the pattern-based method makes up for the weakness of using statistical method only.

Future works:

1. Improvement on current performance, which is not good enough
2. Interactive technical trend map creation



Framework & Formal Run's Evaluation

The architecture of the system, which includes three output modules, is shown in Figure 1. It contains the following six parts:

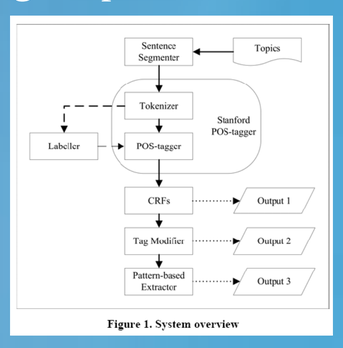


Figure 1. System overview

1. Sentence Segmenter
2. Tokenizer and POS-tagger
3. Labeller
4. CRFs
   - Conditional Random Fields model
5. Tag Modifier
   - A negative tag changes to a positive tag, if the model does not have enough confidence.
   - The assigned positive tag has the highest confidence among all positive tags.
6. Pattern-based Extractor
   - Use indicator words for VALUE and the fact that ATTRIBUTE is usually the nearest noun phrase to the VALUE
   - Include chunking, stopword, and Laplacian

The formal run's evaluation consists of 300 patent topics and 300 paper topics for training plus 200 patent topics and 200 paper topics for test. The distributions of entities in training and test data are as follows:

Distribution of the entities in training data

Entity Type	Patent	Paper
Technology entities in title (TT)	73	92
Technology entities in abstract (AT)	1277	294
Attribute entities in abstract (AA)	223	238
Value entities in abstract (AV)	195	226

Distribution of the desired entities

Entity Type	Patent	Paper
Technology entities in title (TT)	39	93
Technology entities in abstract (AT)	847	342
Attribute entities in abstract (AA)	213	204
Value entities in abstract (AV)	198	193

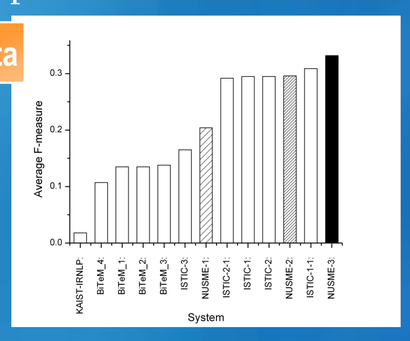
Three submissions

- NUSME-1: CRFs-based method
- NUSME-2: NUSME-1 + Tag modification
- NUSME-3: NUSME-2 + patterns for finding EFFECT

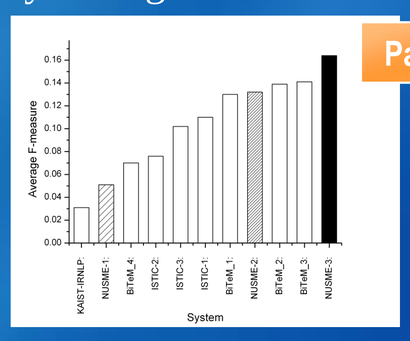
Evaluation Results

1. F-measure of All Systems
   - More effort obtained better results
   - NUSME-2, NUSME-3 achieved relatively good results, a big improvement was achieved by the tag modification step

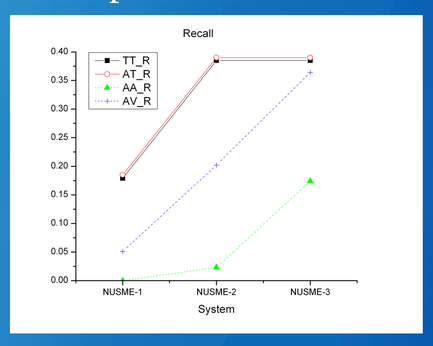
Patent Data

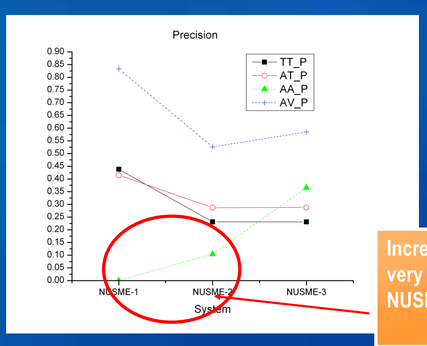


Paper Data



2. Recall & Precision on Patent Data
   - From NUSME-1 to NUSME-2 (by tag modification step), recall was improved.
   - From NUSME-2 to NUSME-3 (by patterns pertaining to EFFECT), recall of AA and AV were improved, recall of TT and AT kept the same.





Increase of AA is due to the very bad precision of AA in NUSME-1 i.e. zero precision.

- From NUSME-1 to NUSME-2 (by tag modification step), precision was reduced.
- From NUSME-2 to NUSME-3 (by patterns pertaining to EFFECT), precision of AA and AV were improved, precision of TT and AT kept the same.

3. The results of paper data were similar to that of patent data.