NUS National University of Singapore

NTCIR-8 Patent Mining Task

Extracting Technology and Effect Entities in Patents and Research Papers

Jingjing Wang, Han Tong Loh, Wen Feng Lu

Department of Mechanical Engineering, National University of Singapore, Singapore

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

Outline

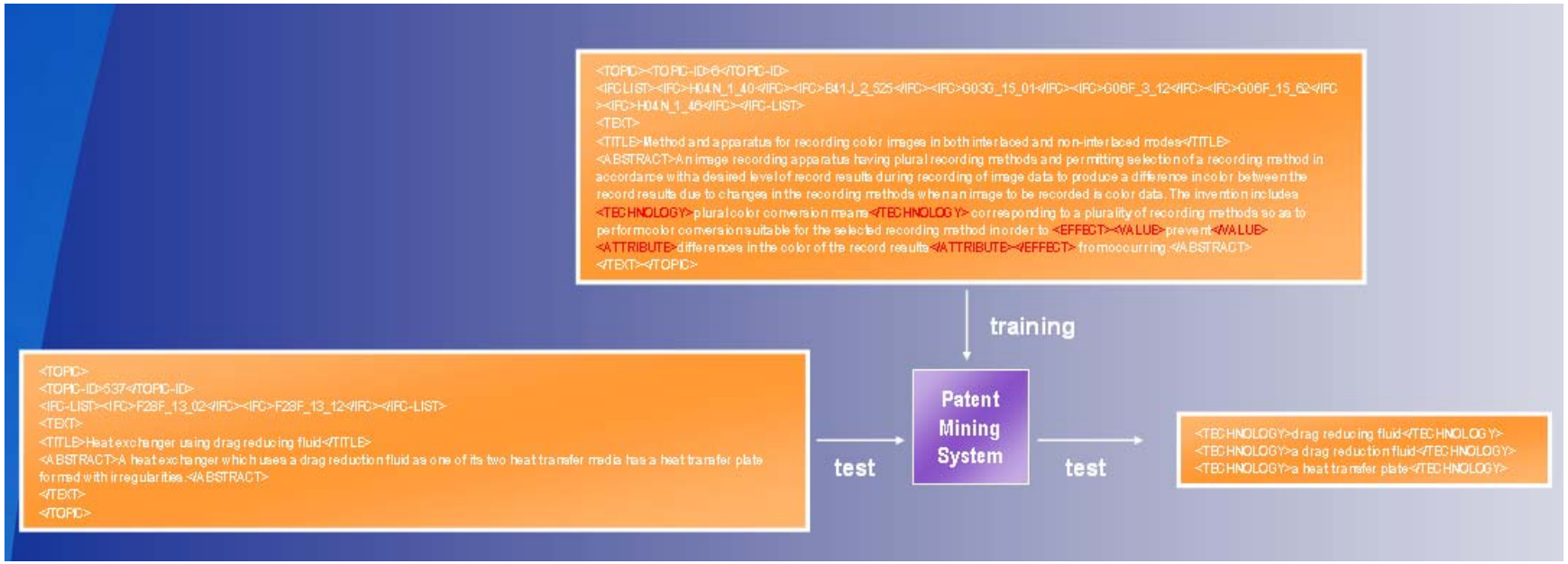
- Introduction
- Our Methods
- Issues Investigated
- Formal Run's Evaluation Results
- Conclusions

Introduction (1/3)

- Technical Trend Map Creation task: to extract expressions of element technologies and their effects from research papers and patents
- Entities: TECHNOLOGY, EFFECT, ATTRIBUTE, and VALUE
- 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
- Evaluation: Recall / Precision / F-measure

Introduction (2/3)

- Diagram



Introduction (3/3)

- Information Extraction / Named Entity Recognition (NER) task
- Challenges
  - No strict definition of all entities; some entities are very long sequences with complex structure e.g. “a control information definition unit for defining control information representing what kind of processing can be performed on mails after reception” is a TECHNOLOGY
  - Effective and efficient features to highlight the entities are unknown
  - Model that can sufficiently utilize all these features is unknown
- Basic Idea
  - Started from an advanced statistical model with many features
  - Slight modification on the original model
  - Further added some patterns and invoked a pattern-based method

Outline

- Introduction
- Our Methods
- Issues Investigated
- Formal Run's Evaluation Results
- Conclusions

Conditional Random Fields (CRFs)

- The probability of a particular label sequence *y* given observation sequence *x* is assigned as a normalized product of potential functions.

$$p(y \mid x, \lambda) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j F_j(y, x) \right)$$

$$F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)$$

- An potential function can be a state function of the label at position *i* and the entire observation sequence

State Feature Functions Adopted

- Seven types of tags (states)
  - BIO (begin, inside, outside) labeling scheme with three kinds of positive tags
  - Positive tags: {“technology-B”, “technology-I”, “value-B”, “value-I”, “attribute-B”, “attribute-I”}
  - Negative tag: “other”

- Observation Sequences
  - n-gram in the original sequence
  - n-gram in the POS-tag sequence
  - current POS-tag with other unigram and its POS-tag

Tag Modification

IF

$$p(Y = \text{"other"} \mid x, \lambda) < t \quad // \; t \text{ is a threshold}$$

THEN

$$p(y \mid x, \lambda) = \max_{Y \neq \text{"other"}} p(Y \mid x, \lambda)$$

$$y := \arg\max_{Y \neq \text{"other"}} p(Y \mid x, \lambda)$$

- A negative tag changes to a positive tag, if the model does not have enough confidence i.e. t = 90%.
- The assigned positive tag has the highest confidence among all positive tags.

Pattern-based Method

- Two problems in above statistical method
  - Fail to involve those indicator tokens that are too far away from current state
  - Treat TECHNOLOGY, ATTRIBUTE, VALUE equally, fail to utilize the relation between ATTRIBUTE and VALUE, lead to ambiguity

- Indicator Words for VALUE
  - An adjective related to polarity opinion, namely good or bad
  - A verb related to making some changes e.g. “improve”, “facilitate”, “adjust”, “reduce” and “prevent”
  - An indicator word list was built according to the semantics (sources: training data and WordNet, which is a thesaurus)

- ATTRIBUTE is usually the nearest noun phrase to the VALUE

Pattern-based Method (Cont.)

- Chunking
  - A POS-based chunker to delimit noun-structure
  - A noun-structure is a boarder concept, compared to noun phrase, since ATTRIBUTE may have a more complex structure.

- Stopword
  - Not all (indicator word, noun-structure) are (VALUE, ATTRIBUTE)
  - Some words in the noun-structure pertaining to the indicator word may be too common
  - A stopword list was built for every indicator word

Learning the Patterns

- Laplacian
  - Built the stopword list by learning from the training data
  - Use Laplacian, in which *c* is the number of correctly matched ATTRIBUTE and *e* is the number of errors, as criterion

$$Laplacian = \frac{e+1}{c+e+1}$$

  - If the Laplacian of a pattern was too big, a stopword was added to reduce the Laplacian, until the Laplacian was acceptable, namely less than 0.5.

- An example of learned pattern
  - Indicator word = “prevented”
  - Direction = before
  - Stopword list: “direction” | “material” | “region” | “size”

System Overview

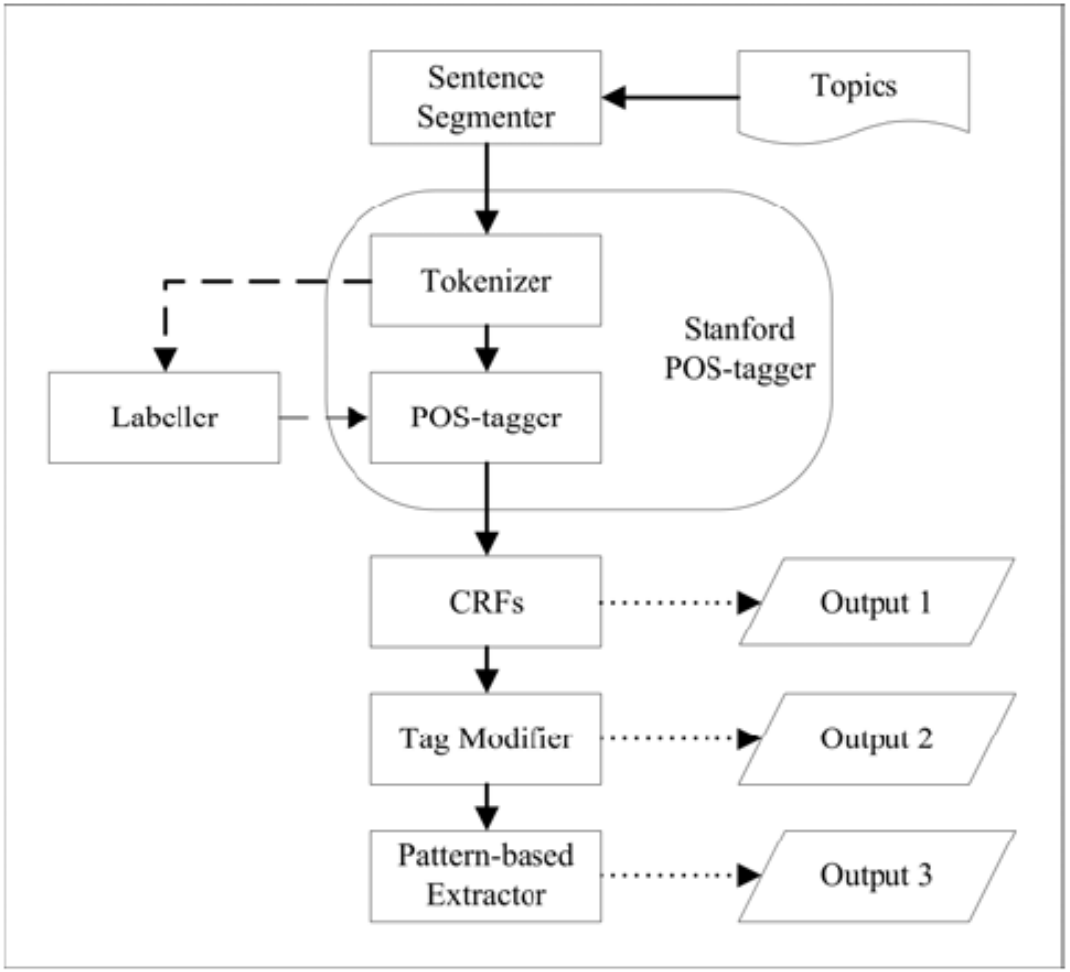


Figure 1. System overview

Outline

- Introduction
- Our Methods
- Issues Investigated
- Formal Run's Evaluation Results
- Conclusions

Issues Investigated (1/3)

- Differences in writing custom between patent and paper
  - Slight difference in representing HTML characters e.g. α is “α” in patent, but “.alpha.” in paper
  - Other differences also exist, but were not considered.
  - We simply separated patent and paper topics initially.
- Hierarchical labeling scheme
  - We tried a hierarchical labeling scheme with two levels, but got bad performance.
- Dependency Parsing
  - It was expected to grasp the relation between VALUE and ATTRIBUTE.
  - It does not offer more information, compared to POS-tagging.

Issues Investigated (2/3)

- Whether the CRFs-based model achieved an acceptable performance?
  - Unfortunately, it did not.

- Whether the tag modification step improved the performance?
  - A big improvement on F-measure was achieved.
  - It may improve the recall, because it forced the CRFs model to output more positive tags, and increase the chance of finding correct entities.
  - It may reduce the precision, because additional output has a high chance of reducing the precision.

Issues Investigated (3/3)

- Whether the manually designed patterns further improved the performance?
  - Both recall and precision pertaining to EFFECT were improved.
  - No influence on TECHNOLOGY
  - The pattern-based method successfully made up for the weakness of the CRFs model, since it utilized the relation between ARRTIBUTE and VALUE, and has no length limit on the sequence.

Outline

- Introduction
- Our Methods
- Issues Investigated
- Formal Run's Evaluation Results
- Conclusions

Formal Run's Evaluation

- Training data
  - 300 patent topics and 300 paper topics

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

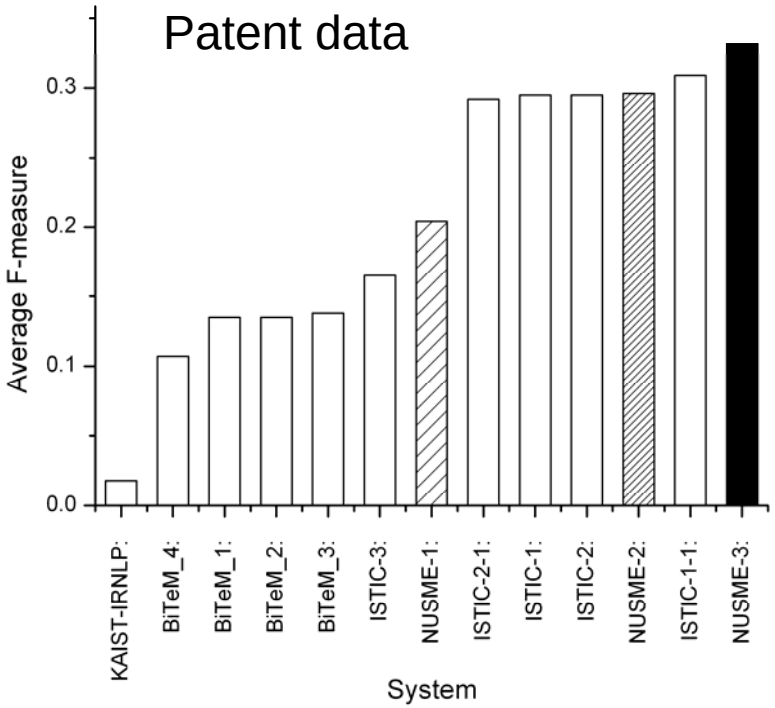
Entity Type	Patent	Paper
Technology entities in title (TT)	39	93
Technology entities in abstract (AT)	847	342
Attribute entities in abstract (AA)	213	204
Value entities in abstract (AV)	198	193

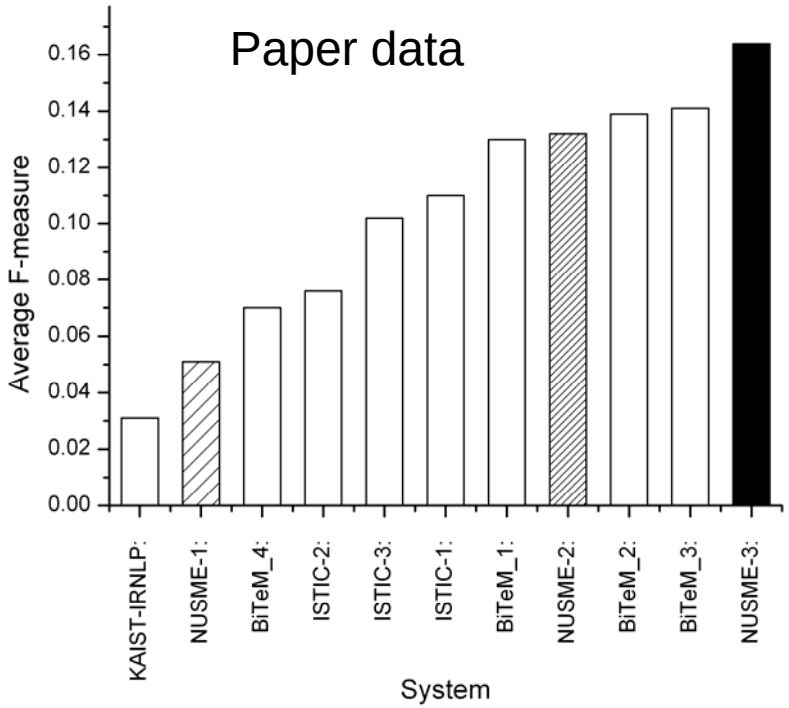
- Test data
  - 200 patent topics and 200 paper topics

Three Submissions

- NUSME-1
  - CRFs-based method
- NUSME-2
  - NUSME-1
  - Tag modification
- NUSME-3
  - NUSME-2
  - Patterns for finding EFFECT
  - Combining results of two methods

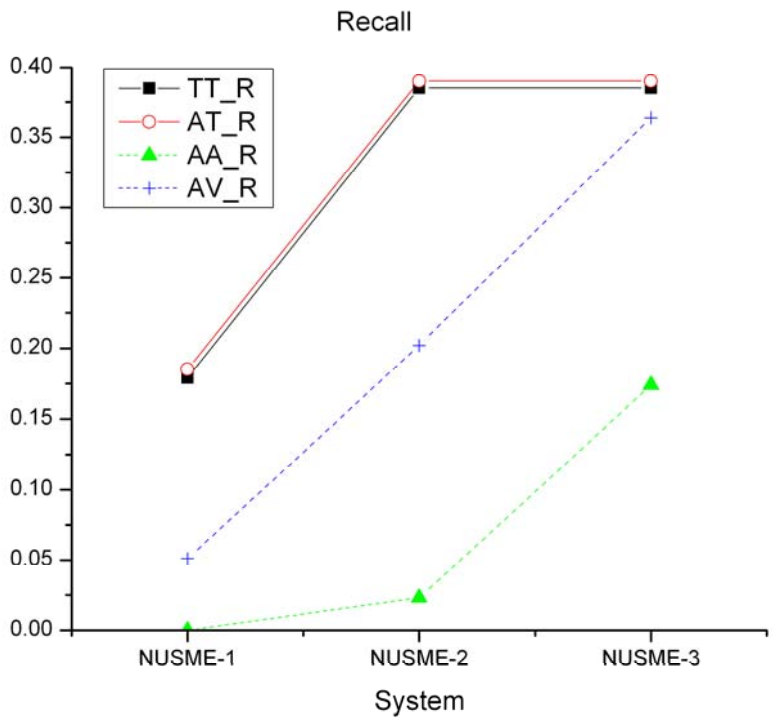
F-measure of all systems





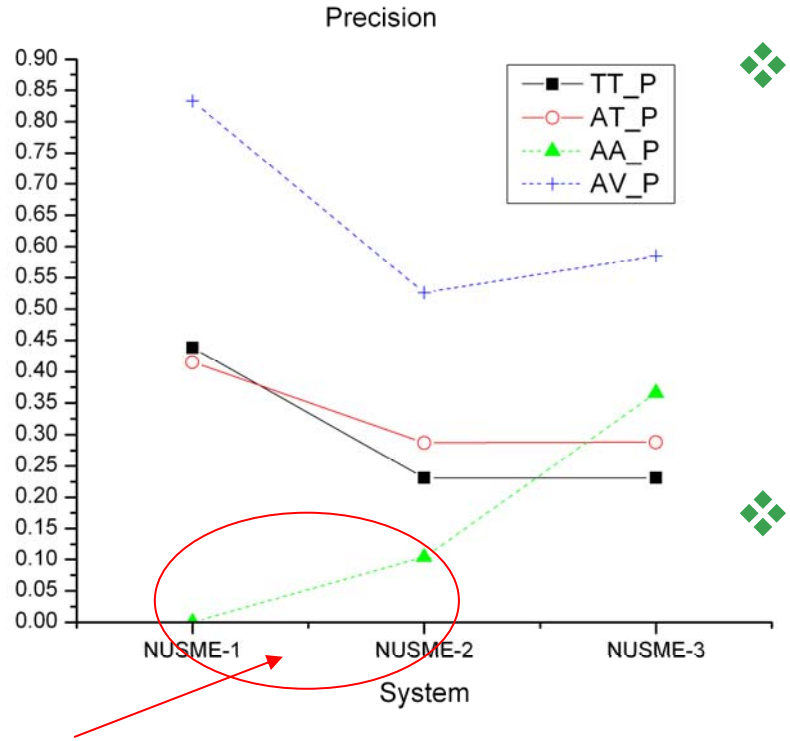
- More efforts obtained better results
- NUSME-2, NUSME-3 achieved relatively good results, a big improvement was achieved by the tag modification step

Recall on Patent data



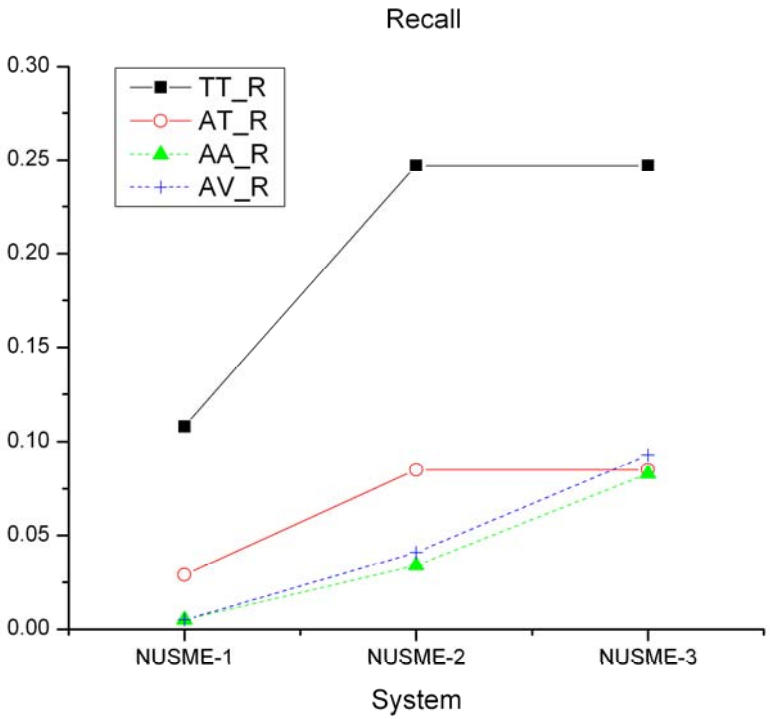
- From NUSME-1 to NUSME-2 (by tag modification step), recall was improved.
- From NUSME-2 to NUSME-3 (by patterns pertaining to EFFECT), recall of AA and AV were improved, recall of TT and AT kept the same.

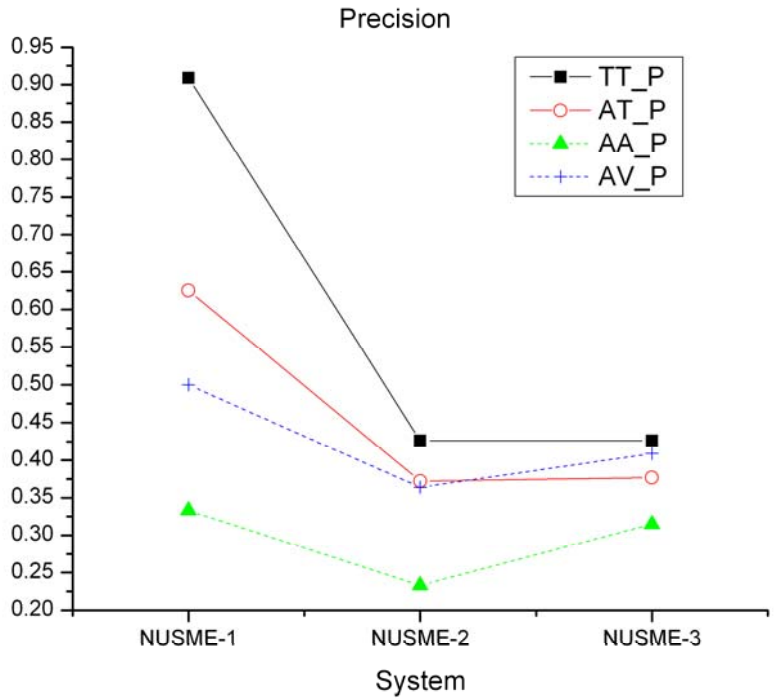
Precision on Patent data



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

Recall & Precision on Paper data





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

Outline

- Introduction
- Our Methods
- Issues Investigated
- Formal Run's Evaluation Results
- Conclusions

Conclusions

- We had tried both statistical method and pattern-based method, and we obtained a relatively good result.
- The tag update rule works.
- In our case, the pattern-based method makes up for the weakness of using statistical method only.
- However, the performance is not good enough.
- Interactive technical trend map creation


NUS
National University
of Singapore
Thank You !