Empowered by Innovation NEC

An information extraction method for multiple data sources

Hironori Mizuguchi and Dai Kusui

Information and Media Processing Laboratories, NEC Corporation

Background

- Information extraction from various documents
  - Internet web pages consist of news, blogs, product information,,,
  - Intranet enterprise documents consist of presentations, papers, patents,,,
- Different writing style and different in these information
- Information extraction methods in previous study cannot achieve good accuracy
  - Because of different contexts even if same annotations.

Training	Test	Precision	Recall	F
Paper	Paper	51.97	23.64	32.40
Paper + Patent	Paper	45.62	18.07	25.75
Patent	Patent	66.24	40.68	50.35
Paper + Patent	Patent	63.54	36.65	45.95

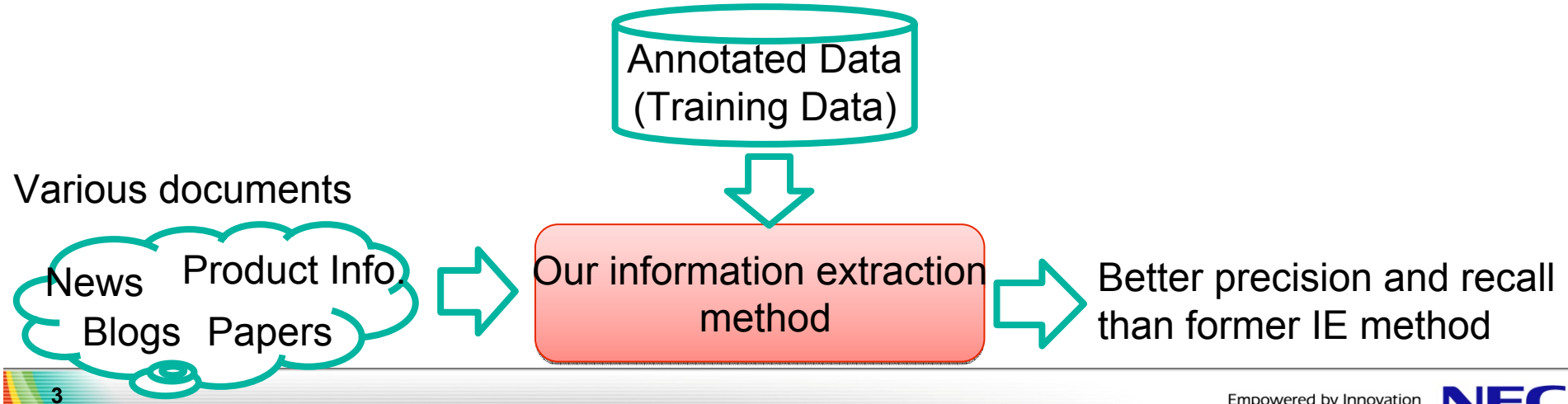
⇩ down

⇩ down

Empowered by Innovation NEC

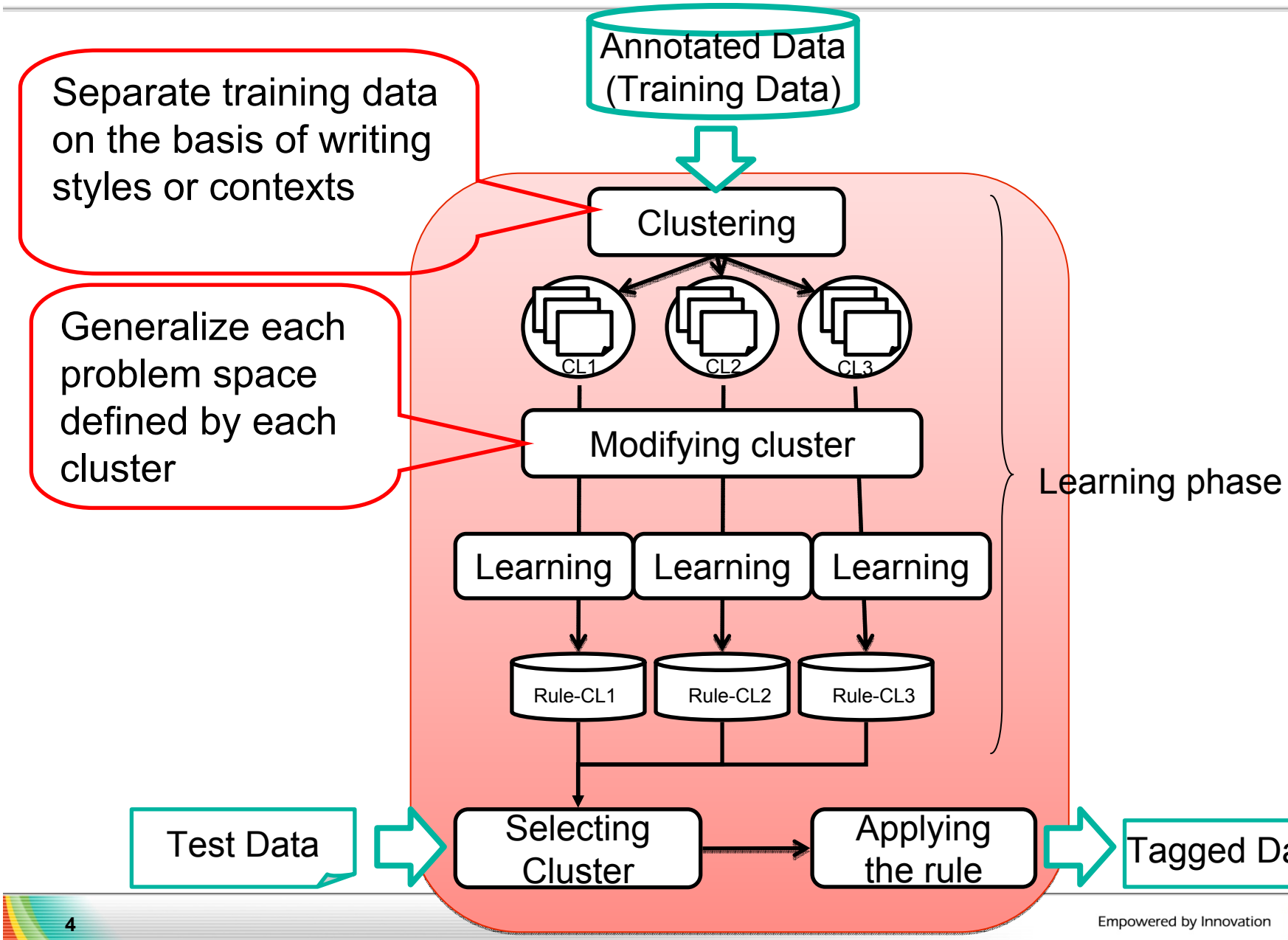
Purpose

- Develop information extraction method from various documents with different writing styles
- Keep accuracy even if we don't care about kinds of data
  - Input: Various documents
    - Ex: Papers, Patents,,, (we don't need to care about kinds of each data)
  - Use: Various annotated training data
    - Ex: TECHNOLOGY tag, ATTRIBUTE tag,,,
  - Output: Extracted information
    - Better precision and recall than previous method by focusing on difference about writing styles and contexts



Empowered by Innovation NEC

Our method: Learning method with clustering



Empowered by Innovation NEC

Clustering

- Separate training data on the basis of writing styles or different contexts
- Clustering target
  - **Document**
    - Document clusters according to document styles
  - **Context** (information of each word and its neighbor words)
    - Context clusters according to contexts
- Clustering method
  - **K-means with cosine similarity**
  - CoClustering (pLSA, NMF,,)

Empowered by Innovation NEC

Modifying cluster

Generalize each problem space defined by each cluster

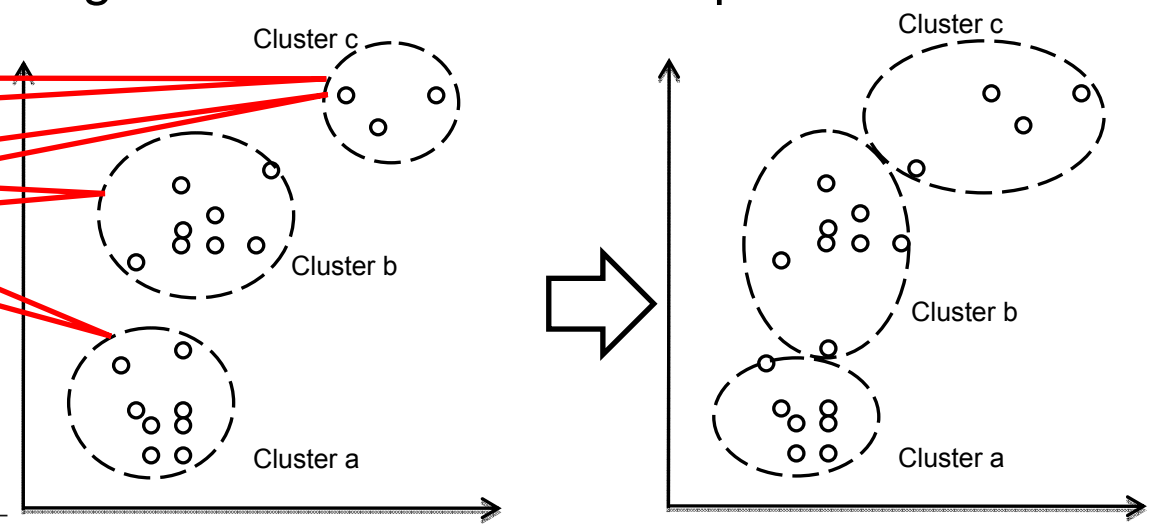
- Two biases
  - Different amount of data in each cluster
  - Small Problem space
- Modify clusters to limit the effects of these biases, modify clusters
  - Calculate two bias score of each cluster (AmountBias, SimBias)
  - Move data from large dense cluster to small sparse cluster

Small Amount of data

Small problem space

$$AmountBias(C_i) = \frac{n_i}{\bar{n}}$$

$$SimBias(C_i) = \frac{\frac{1}{n}\sum_{j=1}^{n_i}(1 - sim(x_{ci}, x_{ij}))^2}{\frac{1}{N}\sum_{k=1}^{k}\sum_{j=1}^{n_i}(1 - sim(x_{ck}, x_{kj}))^2}$$



Empowered by Innovation NEC

Experiment1:NTCIR8 formal run

- Extract 4 tags from papers and patents
  - TECHNOLOGY, ATTRIBUTES, VALUE, EFFECT
- Settings
  - Clustering training data
    - Do not experiment (because of time limitations)
  - Modifying clusters
    - Do not experiment (because of time limitations)
  - Learning
    - SVM classifying each word into 4 class
      - TECHNOLOGY, ATTRIBUTE, VALUE, OTHER
  - EFECT Tag
    - Pair of ATTRIBUTE and VALUE that are located within two words

KIND	TAG	RECALL	PRECISION	F
PAPER	TECHNOLOGY	9.1	21.9	12.9
	ATTRIBUTE	8.1	15.4	10.6
	VALUE	12.2	26.7	16.8
	EFFECT	2.7	18.2	4.7
PATENT	TECHNOLOGY	4.7	8.0	6.0
	ATTRIBUTE	21.9	29.6	25.2
	VALUE	33.8	50.3	40.4
	EFFECT	12.5	33.9	18.2

Empowered by Innovation NEC

Experiment2: Apply our method to NTCIR8

- Apply our method to NTCIR8 Technical Trend Map Creation subtask data
- Extract 3 tags from papers and patents
  - TECHNOLOGY, ATTRIBUTES, VALUE

TRAINING	TEST	PREC	RECALL	F
CL(CLST=2)	PATENT	63.89%	33.80%	43.75%
Baseline	PATENT	63.54%	36.65%	45.95%
CL(CLST=2)	PAPER	52.67%	15.42%	23.74%
Baseline	PAPER	45.62%	18.07%	25.75%

- Settings
  - Clustering training data
    - Document data clustering
    - K-means with cosine similarity (K=2)
  - Modifying clusters
    - Do not experiment (because of time limitations)
  - Learning
    - SVM classifying each word into 4 classes (TECHNOLOGY, ATTRIBUTE, VALUE, OTHER)

Empowered by Innovation NEC

The number of cluster and the effect

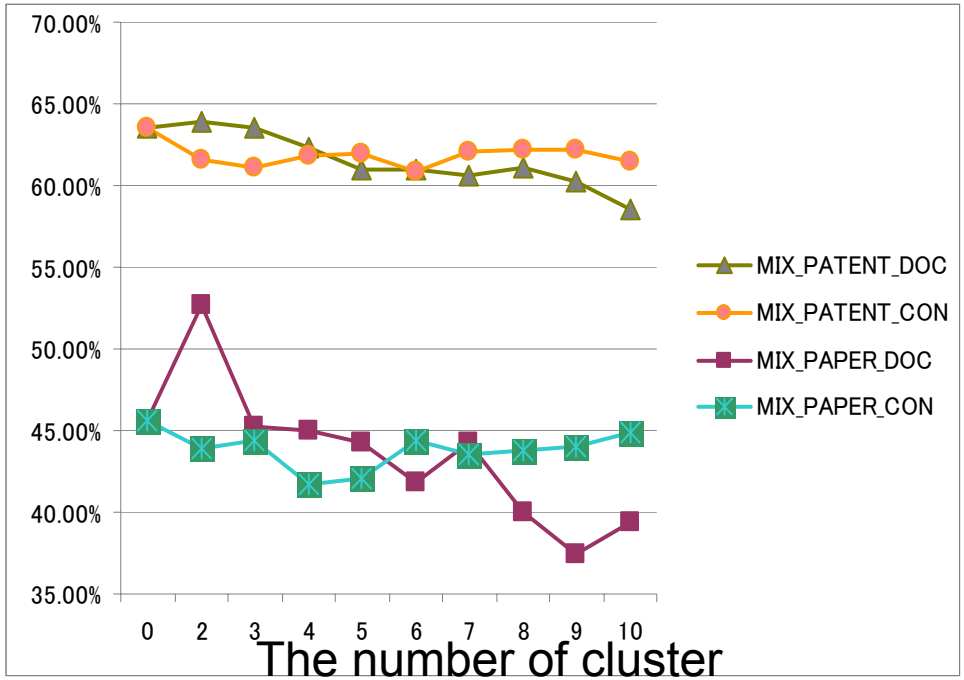
Precision

- Document Clustering (DOC_PATENT, DOC_PAPER):
  - Performance of two cluster is best, and better than baseline.
- Context Clustering (CON_PATENT,CON_PAPER):
  - Performance of no cluster is better than baseline.

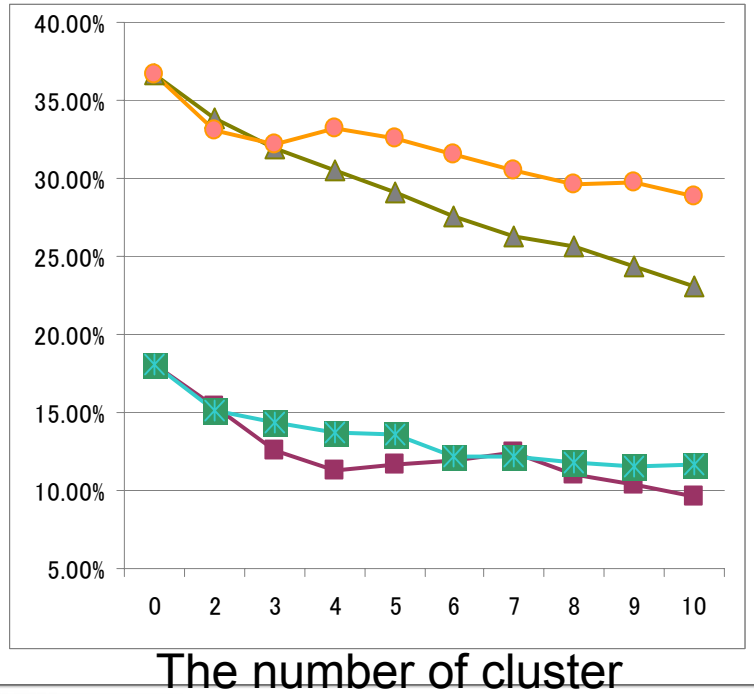
Recall

- Decreased recall because of not modified clusters

Precision



Recall



Empowered by Innovation NEC

Amount of data in each cluster

- Document Clustering can divide documents into each kind
  →good effect for precision
- Context Clustering can not divide documents into each kind

Document clustering (Num. of Doc)

CLST	ID	PAPER	PATENT	SUM
2	0	5	**209**	214
	1	**173**	2	175
3	0	**79**	1	80
	1	**94**	1	95
	2	5	**209**	214
4	0	3	**124**	127
	1	2	**85**	87
	2	**79**	1	80
	3	**94**	1	95

Context Clustering (Num. of Word)

CLST	ID	PAPER	PATENT	SUM
2	0	8322	19370	27692
	1	8137	16683	24820
3	0	1852	3295	5147
	1	8322	19370	27692
	2	6285	13388	19673
4	0	1852	3295	5147
	1	3218	7472	10690
	2	5104	11898	17002
	3	6285	13388	19673

Empowered by Innovation NEC

Discussions

Our method outperforms prior method when the number of cluster was applicable

- Divide documents into the kinds of data
- Applicable cluster size can be testing with training data.
- If we use various kinds of data like internet web pages and intranet enterprise documents, our method can increase the accuracy of information extraction.

Clustering method and cluster size of context clustering should be modified

- The number of feature vector dimensions is huge (760,000)
  → Clustering methods that can compress huge dimensions, such as a pLSA, is better than K-means
- Kinds of contents is more various than documents
  → Cluster size should be bigger than 10
  In simplified experiment, cluster size : 1000, precisions for patent and paper are 76.68 and 53.85, respectively.

Conclusions

- Developed an information extraction method from various documents
- Applied it to NTCIR8 Technical Trend Map Creation subtask
- Our method improved precision by 7%, compared to previous method, with applicable cluster size.
- Future work
  - Experiment with the modification of clusters and evaluate recall effects.
  - Use another clustering method like pLSA in context clustering

Empowered by Innovation NEC