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.

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>Precision</th>
<th>Recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper</td>
<td>Paper</td>
<td>51.97</td>
<td>23.64</td>
<td>32.40</td>
</tr>
<tr>
<td>Paper + Patent</td>
<td>Paper</td>
<td>45.62</td>
<td>18.07</td>
<td>25.75</td>
</tr>
<tr>
<td>Patent</td>
<td>Patent</td>
<td>66.24</td>
<td>40.68</td>
<td>50.35</td>
</tr>
<tr>
<td>Paper + Patent</td>
<td>Patent</td>
<td>63.54</td>
<td>36.65</td>
<td>45.95</td>
</tr>
</tbody>
</table>
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
Our method: Learning method with clustering

Separate training data on the basis of writing styles or contexts

Generalize each problem space defined by each cluster

Learning phase

Annotated Data (Training Data)

Clustering

Modifying cluster

Learning

Rule-CL1

Rule-CL2

Rule-CL3

Selecting Cluster

Applying the rule

Tagged Data

Separate training data on the basis of writing styles or contexts.

Generalize each problem space defined by each cluster.
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,..)
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

\[ \text{AmountBias}(C_i) = \frac{n_i}{n} \]

\[ \text{SimBias}(C_i) = \frac{1}{n} \sum_{j=1}^{n_i} (1 - \text{sim}(x_{cj}, x_{ij}))^2 \]

Small Amount of data

Small problem space
Experiment 1: 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
- EFFECT Tag
  - Pair of ATTRIBUTE and VALUE that are located within two words

<table>
<thead>
<tr>
<th>KIND</th>
<th>TAG</th>
<th>RECALL</th>
<th>PRECISION</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAPER</td>
<td>TECHNOLOGY</td>
<td>9.1</td>
<td>21.9</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>ATTRIBUTE</td>
<td>8.1</td>
<td>15.4</td>
<td>10.6</td>
</tr>
<tr>
<td></td>
<td>VALUE</td>
<td>12.2</td>
<td>26.7</td>
<td>16.8</td>
</tr>
<tr>
<td></td>
<td>EFFECT</td>
<td>2.7</td>
<td>18.2</td>
<td>4.7</td>
</tr>
<tr>
<td>PATENT</td>
<td>TECHNOLOGY</td>
<td>4.7</td>
<td>8.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>ATTRIBUTE</td>
<td>21.9</td>
<td>29.6</td>
<td>25.2</td>
</tr>
<tr>
<td></td>
<td>VALUE</td>
<td>33.8</td>
<td>50.3</td>
<td>40.4</td>
</tr>
<tr>
<td></td>
<td>EFFECT</td>
<td>12.5</td>
<td>33.9</td>
<td>18.2</td>
</tr>
</tbody>
</table>
Experiment 2: 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

<table>
<thead>
<tr>
<th>TRAINING</th>
<th>TEST</th>
<th>PREC</th>
<th>RECALL</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL(CLST=2)</td>
<td>PATENT</td>
<td>63.89%</td>
<td>33.80%</td>
<td>43.75%</td>
</tr>
<tr>
<td>Baseline</td>
<td>PATENT</td>
<td>63.54%</td>
<td>36.65%</td>
<td>45.95%</td>
</tr>
<tr>
<td>CL(CLST=2)</td>
<td>PAPER</td>
<td>52.67%</td>
<td>15.42%</td>
<td>23.74%</td>
</tr>
<tr>
<td>Baseline</td>
<td>PAPER</td>
<td>45.62%</td>
<td>18.07%</td>
<td>25.75%</td>
</tr>
</tbody>
</table>

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)
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
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)

<table>
<thead>
<tr>
<th>CLST ID</th>
<th>PAPER</th>
<th>PATENT</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>5</td>
<td>209</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>173</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>79</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>94</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>209</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>3</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>79</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>94</td>
<td>1</td>
</tr>
</tbody>
</table>

### Context Clustering (Num. of Word)

<table>
<thead>
<tr>
<th>CLST ID</th>
<th>PAPER</th>
<th>PATENT</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>8322</td>
<td>19370</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8137</td>
<td>16683</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1852</td>
<td>3295</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8322</td>
<td>19370</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6285</td>
<td>13388</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1852</td>
<td>3295</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3218</td>
<td>7472</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5104</td>
<td>11898</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6285</td>
<td>13388</td>
</tr>
</tbody>
</table>
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