Text Classification & Summarization
(Using Natural Language Processing and Machine Learning Techniques)

Ko, Youngjoong

July 5, 2017

Dept. of Computer Engineering,
Dong-A University
1. Warming Up

2. Text Classification

3. Text Summarization

4. Combination of TC and TS

5. Latest Trend of TC and TS

6. References
The process of the pattern classification system

**Pattern classification (Duda & Hart)**

- **Input**: sensing
  - segmentation
  - feature extraction
  - classification
  - post-processing

- **Decision**: classify

**Design cycle of the pattern classification system**

- **Start**: collect data
  - choose features
  - choose model
  - train classifier
  - evaluate classifier

- **End**
Warming up!!

- Machine Learning
- Natural Language Processing
- Information Retrieval

Text Classification & Summarization
Text Classification

Introduction

- Classify documents into one (or several) of a set of *pre-defined categories (topics of interest)*

- Prominent status in the information system field
  - Explosion of electronic texts from article, WWW, e-mail, digital library, CRM, biomedical text etc.

- The machine learning paradigm
  - Supervised learning: Find decision rule from an example set of labeled documents
From: xxx@sciences.sdsu.edu
Newsgroups: comp.graphics
Subjects: Need specs on Apple QT

I need to get the specs, or at least a very verbose interpretation of the specs. For QuickTime Technical articles from magazines and references to books would be nice, too.

I also need the specs in a format usable on a Unix or MS-DOS system. I can’t do much with the QuickTime stuff they have on …
Text Classification

Vector Space Model

- Multi-dimensional vector space
  - A document is represented as a vector in vector space

- Each dimension
  - Term or concept

![Diagram of vector space model with documents and query representation]
Text Classification

Term Weighting Scheme

- **TFIDF term weight**
  
  \[
  \text{tfidf}(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#(t_k)}
  \]

- **Cosine Normalization**
  - The weights resulting from tfidf, so as to account for document length
  
  \[
  w_{kj} = \frac{\text{tfidf}(t_k, d_j)}{\sqrt{\sum_{s=1}^{r} (\text{tfidf}(t_s, d_j))^2}}
  \]
Text Classification

Semi-supervised Learning Based Text Classification

- **Difficulties of supervised learning in TC**
  - Require large, often prohibitive, number of labeled training data

- **Semi-supervised learning in TC**
  - Automatically constructs labeled training data from unlabeled documents and the title word of each category

- **How can we automatically generate labeled training documents (machine-labeled data) from only title words?**
  - Bootstrapping Framework

- **How can we handle incorrectly labeled documents in the machine-labeled data?**
  - TCFP Classifier

Semi-supervised Learning Based Text Classification

Pre-defined category? If so, I can know title words!

Bootstrapping! Robust classifier from noisy data!

Automobile

Keywords
- car, gear, transmission, sedan

Context-Cluster
- car, parking, gas station, ...
- road, highway, driver, ...

Learning
Generative Model (NB)

Machine-Labeled Data
- Auto: gear, driver, clutch, ...

Learning
TCFP Classifier
Text Classification

Semi-supervised Learning Based Text Classification

- Measuring similarity based on word & context similarity
  - Two similarity matrices

- Naïve Bayes with Minor Modification
  - Kullback-Leibler Divergence
  
  \[
  P(c_j | d, \hat{\theta}) = \frac{P(c_j | \hat{\theta})P(d | c_j, \hat{\theta})}{P(d | \hat{\theta})} = P(c_j | \hat{\theta}) \prod_{i=1}^{|d|} P(w_i | c_j, \hat{\theta})^{N(w_i, d_i)} \\
  \approx \alpha \frac{\log P(c_j | \hat{\theta})}{n} + \sum_{i=1}^{|d|} P(w_i | d_i, \hat{\theta}) \log \left( \frac{P(w_i | c_j, \hat{\theta})}{P(w_i | d_i, \hat{\theta})} \right)
  \]

  - Laplace Parameter Estimation

  \[
  \hat{\theta}_{w | c_j} = P(w_i | c_j, \hat{\theta}) = \frac{1 + N(w_i, G_{c_j})}{|V| + \sum_{i=1}^{|V|} N(w_i, G_{c_j})} \\
  \hat{\theta}_{c_j} = P(c_j | \hat{\theta}) = \frac{1 + G_{c_j}}{|C| + \sum_{c_i} G_{c_i}}
  \]
Text Classification

Semi-supervised Learning Based Text Classification

An example of feature projections in Text Categorization

\[ d=(f_1, f_2), \quad c_1=\{d_1, d_2, d_3\}, \quad c_2=\{d_4, d_5, d_6\} \]

\[
\begin{align*}
&d_1=(1.0, 0.0), \\
&d_2=(0.96, 0.25), \\
&d_3=(0.91, 0.4), \\
&d_4=(0.4, 0.91), \\
&d_5=(0.25, 0.96), \\
&d_6=(0.1, 0.99)
\end{align*}
\]

\[ w: \text{weight}, \quad d: \text{document}, \quad c: \text{category} \]

\[
\begin{array}{ccc}
w, & d, & c \\
1.0, & d_1, & c_1 \\
0.96, & d_2, & c_1 \\
0.91, & d_3, & c_1 \\
0.4, & d_4, & c_2 \\
0.25, & d_5, & c_2 \\
0.1, & d_6, & c_2 \\
\end{array}
\]

Feature representation in feature projections

Document representation in a conventional vector space

feature projections on feature \( f_1 \)

feature projections on feature \( f_2 \)
Semi-supervised Learning Based Text Classification

Text Classification

Final Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Newsgroups</td>
<td>79.36</td>
<td>83.46</td>
<td>83</td>
<td>79.95</td>
<td>82.49</td>
<td>86.19</td>
<td>91.72</td>
</tr>
<tr>
<td>WebKB</td>
<td>73.63</td>
<td>73.22</td>
<td>75.28</td>
<td>68.04</td>
<td>73.74</td>
<td>75.47</td>
<td>85.29</td>
</tr>
<tr>
<td>Reuters</td>
<td>88.62</td>
<td>88.23</td>
<td>86.26</td>
<td>85.65</td>
<td>87.41</td>
<td>89.09</td>
<td>91.64</td>
</tr>
</tbody>
</table>
Text Classification Using Revised EM Algorithm

- Problem of the one-against-the-rest method in TC
  - Negative examples in the one-against-the-rest method have noisy examples

- Solutions
  - Automatically removing noisy examples by the sliding window technique and the revised EM (Expectation Maximization) algorithm
    - How can we find a boundary area containing many noisy documents?
      - Sliding window technique
    - How can we deal with noisy documents found from the boundary?
      - Revised EM algorithm

[IPM 2007, AIRS 2004]
The multi-class setting with four categories changed into the binary setting using the One-Against-the-Rest method.
Text Classification

Text Classification Using Revised EM Algorithm

Finding max/min threshold using Sliding Window and Entropy

Separating into three classes

Positive

Unlabeled

Negative
Text Classification

Text Classification Using Revised EM Algorithm

### Final Results

<table>
<thead>
<tr>
<th>Data Set</th>
<th>k-NN (origin)</th>
<th>k-NN (proposed)</th>
<th>NB (origin)</th>
<th>NB (proposed)</th>
<th>Rocchio (origin)</th>
<th>Rocchio (proposed)</th>
<th>SVM (origin)</th>
<th>SVM (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Newsgroups</strong></td>
<td>86.07</td>
<td>87.96 (+2.19)</td>
<td>83.17</td>
<td>84.86 (+2.03)</td>
<td>82.84</td>
<td>84.48 (+1.98)</td>
<td>88.34</td>
<td>89.08 (+0.84)</td>
</tr>
<tr>
<td>(micro-avg.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>84.58</td>
<td>87.03 (+2.89)</td>
<td>82.87</td>
<td>84.55 (+2.03)</td>
<td>81.5</td>
<td>83.57 (+2.54)</td>
<td>87.73</td>
<td>89.08 (+1.53)</td>
</tr>
<tr>
<td><strong>WebKB</strong></td>
<td>84.97</td>
<td>86.74 (+2.08)</td>
<td>85.67</td>
<td>87.21 (+1.8)</td>
<td>86.52</td>
<td>88.26 (+2.01)</td>
<td>92.12</td>
<td>92.64 (+0.56)</td>
</tr>
<tr>
<td>(micro-avg.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>82.13</td>
<td>85.55 (+4.16)</td>
<td>83.58</td>
<td>86.53 (+3.55)</td>
<td>83.71</td>
<td>87.03 (+3.96)</td>
<td>91.52</td>
<td>92.17 (+0.71)</td>
</tr>
<tr>
<td><strong>WebKB</strong></td>
<td>91.47</td>
<td>94.27 (+3.06)</td>
<td>90.80</td>
<td>93.86 (+3.37)</td>
<td>89.24</td>
<td>91.80 (+2.86)</td>
<td>94.66</td>
<td>95.52 (+0.91)</td>
</tr>
<tr>
<td>(macro-avg.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>82.66</td>
<td>85.43 (+3.34)</td>
<td>81.26</td>
<td>86.38 (+6.31)</td>
<td>77.56</td>
<td>83.55 (+7.71)</td>
<td>89.86</td>
<td>90.72 (+0.96)</td>
</tr>
</tbody>
</table>
Text Classification

Improving Indexing Technique in Text Classification

- The conventional indexing technique in TC
  - Vector space model in TFIDF

- Improvement point
  - Each sentence has different importance for identifying the content of document
    - Text summarization techniques
      - Measuring similarity between the title and each sentence
    - Revised term weight
      - Modified by the sentence importance value

[IPM 2004, Coling 2004]
Text Classification

Improving Indexing Technique in Text Classification

Training Module

- Training data
- Preprocessing
- Feature Selection ($\chi^2$ statistics)
- Calculating the importance of sentence
- Indexing

Text Classification Module

- Input document
- Preprocessing
- Calculating the importance of sentence
  - By using the title
  - By using the importance of terms (centroid)
- Indexing
- Text Classifier (Naïve Bayes, kNN, Rocchio, SVM)
- Assigning Category

19
## Text Classification

### Improving Indexing Technique in Text Classification

- **Results in English Newsgroup data set**

<table>
<thead>
<tr>
<th></th>
<th>Naïve Bayes</th>
<th></th>
<th>Rocchio</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basis</td>
<td>Proposed</td>
<td>Basis</td>
<td>Proposed</td>
</tr>
<tr>
<td></td>
<td>system</td>
<td>system</td>
<td>system</td>
<td>system</td>
</tr>
<tr>
<td>macro-avg $F_1$</td>
<td>83.2</td>
<td>84.4</td>
<td>79.8</td>
<td>80.5</td>
</tr>
<tr>
<td>micro-avg $F_1$</td>
<td>82.9</td>
<td>84.3</td>
<td>79.4</td>
<td>80.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$k$-NN</th>
<th></th>
<th>SVM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basis</td>
<td>Proposed</td>
<td>Basis</td>
<td>Proposed</td>
</tr>
<tr>
<td></td>
<td>system</td>
<td>system</td>
<td>system</td>
<td>system</td>
</tr>
<tr>
<td>macro-avg $F_1$</td>
<td>81.3</td>
<td>82.7</td>
<td>85.8</td>
<td>86.4</td>
</tr>
<tr>
<td>micro-avg $F_1$</td>
<td>81.1</td>
<td>82.5</td>
<td>85.8</td>
<td>86.3</td>
</tr>
</tbody>
</table>
## Text Classification

### Improving Indexing Technique in Text Classification

- **Results in Korean Newsgroup data set**

<table>
<thead>
<tr>
<th></th>
<th>Naïve Bayes</th>
<th>Rocchio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basis system</td>
<td>Proposed system</td>
</tr>
<tr>
<td>macro-avg $F_1$</td>
<td>78.4</td>
<td>80.8</td>
</tr>
<tr>
<td>micro-avg $F_1$</td>
<td>79.1</td>
<td>81.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>k-NN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basis system</td>
<td>Proposed system</td>
</tr>
<tr>
<td>macro-avg $F_1$</td>
<td>78.6</td>
<td>80.6</td>
</tr>
<tr>
<td>micro-avg $F_1$</td>
<td>79.9</td>
<td>81.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Basis system</th>
<th>Proposed system</th>
</tr>
</thead>
<tbody>
<tr>
<td>macro-avg $F_1$</td>
<td>77.8</td>
<td>79.2</td>
</tr>
<tr>
<td>micro-avg $F_1$</td>
<td>78.7</td>
<td>80.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Basis system</th>
<th>Proposed system</th>
</tr>
</thead>
<tbody>
<tr>
<td>macro-avg $F_1$</td>
<td>84.8</td>
<td>85.5</td>
</tr>
<tr>
<td>micro-avg $F_1$</td>
<td>86.0</td>
<td>86.5</td>
</tr>
</tbody>
</table>
Can we develop a novel term-weighting scheme for specialized text classification better than those used in information retrieval?

Class Information

- Supervised-learning based text classification has the training data labeled as either positive or negative for each class

  - Novel term weighting scheme
    - Substitution for idf

Different distributions of positive and negative classes in a whole collection
Text Classification

Term Weighting Scheme Using Class Information

- New Term Weighting Scheme Using Term Relevance Ratio (TRR)

$$\log \text{tf.TRR} = (\log tf_{t_i} + 1) \cdot \log \frac{P(t_i|cl)}{P(t_i|\bar{cl})}$$

- TRR Estimation
  - Maximum Likelihood Estimation (MLE)
    
    $$P(t_i|cl) = \frac{\sum_{j=1}^{|\mathcal{T}_c|} tf_{t_i,j}}{\sum_{k=1}^{|\mathcal{T}_c|} \sum_{j=1}^{|\mathcal{T}_c|} tf_{t_k,j}}$$
    $$P(t_i|\bar{cl}) = \frac{\sum_{j=1}^{|\mathcal{T}_c|} tf_{t_i,j}}{\sum_{k=1}^{|\mathcal{T}_c|} \sum_{j=1}^{|\mathcal{T}_c|} tf_{t_k,j}}$$

  - Reformation of the MLE

    $$P(t_i|cl) = \sum_{k=1}^{|\mathcal{T}_c|} P(t_i|d_k) \cdot P(d_k|cl)$$
    $$P(t_i|\bar{cl}) = \sum_{k=1}^{|\mathcal{T}_c|} P(t_i|d_k) \cdot P(d_k|\bar{cl})$$
## Term Weighting Scheme Using Class Information

### Final Results

<table>
<thead>
<tr>
<th></th>
<th>tf.idf</th>
<th>log tf.idf</th>
<th>log tf.chi</th>
<th>tf.rf</th>
<th>Delta tf.idf</th>
<th>log tf.TRR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reuters (BEP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kNN</td>
<td>85.66</td>
<td>86.95</td>
<td>83.52</td>
<td>85.52</td>
<td>86.94</td>
<td>90.48 (+4.06)</td>
</tr>
<tr>
<td>micro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>92.46</td>
<td>93.29</td>
<td>89.99</td>
<td>92.68</td>
<td>92.46</td>
<td>94.90 (+1.72)</td>
</tr>
<tr>
<td>macro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>89.89</td>
<td>90.51</td>
<td>89.89</td>
<td>90.12</td>
<td>88.13</td>
<td>91.68 (+1.29)</td>
</tr>
<tr>
<td>micro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NG (F₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kNN</td>
<td>81.13</td>
<td>85.13</td>
<td>71.53</td>
<td>82.96</td>
<td>86.22</td>
<td>87.78 (+1.81)</td>
</tr>
<tr>
<td>macro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>micro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>81.20</td>
<td>85.17</td>
<td>71.55</td>
<td>82.79</td>
<td>86.01</td>
<td>87.75 (+2.02)</td>
</tr>
<tr>
<td>macro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNG (F₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kNN</td>
<td>86.78</td>
<td>87.46</td>
<td>80.54</td>
<td>87.13</td>
<td>86.35</td>
<td>88.09 (+0.72)</td>
</tr>
<tr>
<td>macro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>micro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>87.07</td>
<td>87.74</td>
<td>81.41</td>
<td>87.37</td>
<td>86.64</td>
<td>88.43 (+0.78)</td>
</tr>
<tr>
<td>macro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>77.92</td>
<td>79.34</td>
<td>63.34</td>
<td>69.9</td>
<td>79.73</td>
<td>83.24 (+4.40)</td>
</tr>
<tr>
<td>micro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>79.69</td>
<td>80.66</td>
<td>66.77</td>
<td>74.54</td>
<td>81.69</td>
<td>84.36 (+3.27)</td>
</tr>
<tr>
<td>macro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>83.64</td>
<td>84.14</td>
<td>74.84</td>
<td>81.61</td>
<td>80.41</td>
<td>84.67 (+0.62)</td>
</tr>
<tr>
<td>micro-averaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>85.12</td>
<td>85.47</td>
<td>76.84</td>
<td>83.34</td>
<td>81.83</td>
<td>85.70 (+0.27)</td>
</tr>
</tbody>
</table>
**Text Classification**

**Selected Published Papers**

[International Journal Papers]


Text Classification

Selected Published Papers

[International Conference Papers]


Reducing the size of a document while preserving its content
- To extract content and present the most important content to a user in a condensed form

Text summarization (TS) system
- Identify the most salient information in a document
- The most widespread summarization strategy: sentence extraction
Text Summarization

Fundamental Processes

- **Summary Generation**
  - **post-processing**
  - **salient sentences extraction**
  - **keyword extraction**
  - **POS tagging**
  - **sentence segmentation**
  - **text extraction**

Input (html/xml documents)

Text Summarization

- **Generic Summarization**
  - Based on the content of a given text, ATS systems often produce generic summaries that highlight the most salient points of a given text.

- **Query-based Summarization**
  - Most summaries of Web search engines such as Google are based on the query terms from the user’s search: Snippet
Text Summarization

Fundamental Processes

- **Linguistic approaches**
  - High performance
  - Require high quality linguistic analysis tools (parser etc.) and linguistic resources (WordNet etc.)

- **Statistical approaches**
  - Easy to understand and implement, low cost
  - Generally low performance
Text Summarization

General Statistical Methods

- **Title Method**
  - How many words are commonly used between the sentence and title
  - Boolean weighted vector space model

\[
Score(S_i) = sim(S_i, Q) \quad sim(S_i, Q) = \sum_{k=1}^{n} w_{ik} w_{jk}
\]

- **Location Method**
  - Leading several sentences of an article are important and a good summary

\[
Score(S_i) = 1 - \frac{i - 1}{N}
\]
Text Summarization

General Statistical Methods

- **Aggregation Similarity Method**
  - The sum of similarity with other all sentence vectors

  \[ \text{sim}(S_i, S_j) = \sum_{k=1}^{n} w_{ik} w_{jk} \]
  \[ \text{asim}(S_i) = \sum_{j=1, j \neq i}^{n} \text{sim}(S_i, S_j) \]

- **Frequency Method**
  - The sum of *tf-idf* term weights of words in each sentence

  \[ \text{Score}(S_i) = \sum_{k=1}^{n} (tf_i \times \log \frac{N}{df_i}) \]
TF-based query method

- For no-title cases
- High frequent words as keywords instead of title

\[
\text{sim}(S_i, Tfq) = \sum_{k=1}^{n} w_{ik} w_{Tfqk}
\]
# Text Summarization

## The Performance of Generic Statistical Methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>30%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>48.8</td>
<td>43.5</td>
</tr>
<tr>
<td>Location</td>
<td>49.4</td>
<td>46.6</td>
</tr>
<tr>
<td>TF based query</td>
<td>45.6</td>
<td>46.5</td>
</tr>
<tr>
<td>Aggregation Similarity</td>
<td>40.6</td>
<td>23.9</td>
</tr>
<tr>
<td>Frequency</td>
<td>35.2</td>
<td>13.0</td>
</tr>
</tbody>
</table>
Topic keyword identification using lexical clustering

- Automatic detection of topic words is very useful in Text Summarization

The proposed method

- Keyword identification for text summarization
  - Using co-occurrence statistics
    - Context vector space
  - Using context vector space and k-means algorithm
    - Lexical clustering

[IEICE 2003]
Text Summarization

Two-step Text Summarization

- High performance summarization method to efficiently combine statistical approaches

**Statistical Based Text Summarization Method**

- Feature Sparseness Problem
- Bi-gram Pseudo Sentence
- Low Performance
- Two-step Combinational Method
- No-title Documents
- TF Based query Method

[ PRL 2008, AIRS 2004 ]
First Step: Removing Noisy Sentences

- **The goal of the first step**
  - Not extract salient sentences but reduce noisy sentences

- **Bi-gram pseudo sentences**
  - Solve the feature sparseness problem
    - Can get few feature information from only single sentence
  - A new meaning unit
    - Two adjacent sentences

- **The linear combination method in the first step**
  - Title and Location methods
  - Remove about 50% noisy bi-gram pseudo sentences

\[
Score(S_i) = sim(S_i, Q) + (1 - \frac{i - 1}{N})
\]
The goal of the second step

- Generate summary from extracting the salient original single sentences

Separate the remaining bi-gram pseudo sentences into original single sentences

Aggregation Similarity Method

- Since noisy sentences are eliminated in the first step, we can get important score with good quality.

\[
Score(S_i) = sim(S_i, Q) + (1 - \frac{i - 1}{N}) + w \cdot asim(S_i)
\]
Text Summarization

Evaluation Results

- Comparing with other summarization methods with title
  - Title, Location, DOCUSUM

<table>
<thead>
<tr>
<th></th>
<th>DOCUSUM</th>
<th>Two-step</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>52.2</td>
<td>53.4</td>
<td>+1.2</td>
</tr>
<tr>
<td>30%</td>
<td>50.3</td>
<td>55.3</td>
<td>+5.0</td>
</tr>
</tbody>
</table>
Using Relevance Feedback Technique

- Each sentence is segmented.
- Relevant and non-relevant sentences are separated by whether or not each sentence includes a query term.
- The relevance weight of candidate terms from relevant sentences is estimated by the statistical weighting function and the initial query is expanded by using the candidate terms with the high relevance weight (TSV).
- The important score of each sentence is estimated by using TSV and the location information of expanded query terms.
- Finally, a snippet is generated by sentences with a high important score.

[ IPL 2008, SIGIR 2007 ]
Text Summarization

Query-based Summarization

- Term Selection Value (TSV)

\[
    w_t = TSV_t = \log \frac{p(1 - q)}{q(1 - p)} = \log \frac{(r + 0.5)(S - s + 0.5)}{(R - r + 0.5)(s + 0.5)}
\]

- The Importance Score Estimation using TSVs

\[
    \text{Score}(S_i) = \alpha \left( \frac{RW\text{score}(S)}{RW\text{scoreMax}} \right) + (1 - \alpha) \left( 1 - \frac{i - 1}{N} \right)
\]
## Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>Title Method</th>
<th>Proposed Method</th>
<th>Search Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Naver Data Set</strong></td>
<td>56.6%</td>
<td>67.1%</td>
<td>20.4%</td>
</tr>
<tr>
<td><strong>Google Data Set</strong></td>
<td>57.4%</td>
<td>68.7%</td>
<td>59.5%</td>
</tr>
</tbody>
</table>
Selected Published Papers

[International Journal Papers]


[International Conference Papers]

Combination of TC and TS

Interactive Framework of the Summarization and Categorization

Enhancing Summarization

- Importance score of sentence using Category-based Language Model

\[
\alpha \times \sigma(\text{Score}_d(S_i)) + \hspace{1cm} \text{Intra-document Information}
\]

\[
\text{Score}(S_i) = (1.0 - \alpha) \times \beta \times \sigma(\text{Score}_c(S_i)) + \hspace{1cm} \text{Intra-Category Information}
\]

\[
(1.0 - \alpha) \times (1.0 - \beta) \times \sigma(\text{Score}_\text{COL}(S_i)) \hspace{2cm} \text{Global Information}
\]

\[
\text{Size of document} \hspace{2cm} \text{Size of category}
\]

\[
\alpha = \frac{|d|}{|d| + \mu_1}, \quad \beta = \frac{|c|}{|c| + \mu_2}
\]

\[
\sigma(x) = \frac{1.0}{1.0 + \exp(-x)}
\]

Size of document average size of category average size of document average size of category

\[
\text{Intra-document Information} \hspace{3cm} \text{Intra-Category Information} \hspace{3cm} \text{Global Information}
\]
Combination of TC and TS

Enhancing Summarization

- **Applying Probabilistic Latent Class Model**
  - Similar to Soft Clustering or Topic Modeling
  - Deal with the error propagation problem from the classifier

\[
\alpha \times \sigma(Score_d(S_i)) + \\
(1.0 - \alpha) \beta \times \sum_c P(c|d) \times \sigma(Score_c(S_i)) + \\
(1.0 - \alpha)(1.0 - \beta) \times \sigma(Score_{COL}(S_i))
\]

, where \( \sum_c P(c|d) = 1 \)
Combination of TC and TS

Enhancing Classification

Background

- Previous classification methods use only frequencies of each own term
- The term-frequency does not reflect the importance of sentence

Proposition

- Apply the term relevance to the document representation of classification

\[ w'_j = w_j \times (0.5 + \sigma(TR_s(t_j))) \]

\[ \sigma(x) = \frac{1.0}{1.0 + \exp(-x)} \]

\[ TR_s(t) = \log\left\{ \frac{(r_s + 0.5)(S_s - s_s + 0.5)}{(R_s - r_s + 0.5)(s_s + 0.5)} \right\} \]

, where \( R_s + S_s = |d| \)

- \( w_j \): the typical feature weight of \( j \)-th term \( t_j \)
- \( w'_j \): the proposed feature weight of \( t_j \)
- \( r_s \): the # of summary sentences that include \( t_j \) in document \( d \)
- \( s_s \): the # of non-summary sentences that include \( t_j \) in \( d \)
- \( R_s \): the # of relevant sentences in \( d \)
- \( S_s \): the # of irrelevant sentences in \( d \)
## Summary of Evaluations

### Evaluation Results

#### In Document Summarization

<table>
<thead>
<tr>
<th>Summarization Methods</th>
<th>KORDIC</th>
<th>AbleNews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Category Information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed system</td>
<td>0.567</td>
<td>0.612</td>
</tr>
<tr>
<td>Previous best system (Contextual Info)</td>
<td>0.553</td>
<td>0.564</td>
</tr>
<tr>
<td>With Category Information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed system</td>
<td>0.614</td>
<td>0.664</td>
</tr>
<tr>
<td>Previous best system (Contextual Info)</td>
<td>0.588</td>
<td>0.614</td>
</tr>
</tbody>
</table>

#### In Document Classification

<table>
<thead>
<tr>
<th>Classification Methods</th>
<th>KORDIC</th>
<th>AbleNews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed system</td>
<td>0.784</td>
<td>0.890</td>
</tr>
<tr>
<td>Previous best system (SVM with TF-IDF)</td>
<td>0.760</td>
<td>0.852</td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed system</td>
<td>0.545</td>
<td>0.598</td>
</tr>
<tr>
<td>Previous best system (LDA with TF)</td>
<td>0.518</td>
<td>0.549</td>
</tr>
</tbody>
</table>
Deep eXtreme Multi-label Learning (XML)

- **XML focuses on tackling the problem of extremely high input dimensions for both input feature dimension and label dimension.**
  - Allows for the co-existence of more than one label for a single data sample
  - One-against-the-rest classifiers are not feasible since it will be almost computationally intractable to train a massive number of one million classifiers.

- **Tree based method and embedding based method**

- **Embedding based method**: Projecting the high dimensional label vectors onto a low dimension linear subspace

---

Latest Trend of TC and TS

Deep eXtreme Multi-label Learning (XML)
A neural model, CRSum, to take a sentence’s contextual relations with its surrounding sentences into consideration for extractive summarization

Contextual relations with a two-level attention mechanism in CRSum

CNN (Convolutional Neural Network), RNN (Rucurrent Neural Network) and Attention Mechanism

Latest Trend of TC and TS

Extractive Summarization Using a Neural Attention Model
The Latest Trend of TC and TS

Extractive Summarization Using a Neural Attention Model
Thank you for your attention!

고 영 중 (Ko, Youngjoong)
web.donga.ac.kr/yjko