Document Classification through Image-Based Character Embedding and Wildcard Training

Daiki Shimada, Ryunosuke Kotani, Hitoshi Iyatomi

Applied Informatics, Graduate School of Science and Engineering, Hosei University, Japan
Introduction

- Difficulty of processing Japanese / Chinese text
  - No word boundary
    - Word segmentation preprocess
    - Hard to segment words include coinages and slang words
  - Large number of different characters
    - More than 2,000 different characters for daily use (Japanese)

Example: メロスは激怒した。 Melos was enraged.
Introduction

- Character-level approaches to Japanese / Chinese text
  - Character-level N-gram feature
  - **Character-level Convolutional Neural Networks (CLCNN)** [Zhang et al. 2015]
    - State-of-the-art in English document classification
    - *Vectorization of character (e.g. one-hot vector, lookup table)*
    - *Data augmentation by using paraphrase*

Introduction

- Character-level approaches to Japanese / Chinese text
  - Character-level N-gram feature
  - Character-level Convolutional Neural Networks (CLCNN) [Zhang et al. 2015]
    - State-of-the-art in English document classification
    - Vectorization of character (e.g. one-hot vector, lookup table)
    - Data augmentation by using paraphrase

These strategies is NOT appropriate for Japanese and Chinese.

Two New Document Analysis Techniques for CLCNN

i. Image-based Character Embedding

Water, ice

Salmon, shark

ii. Data augmentation without word segmentation, "wildcard training"

メロスは激怒した。 → メロス＊激＊した。
Key Concept – (i) Image-based Character Embedding

- Focus on Ideographic of Japanese / Chinese characters
  - Most of them imply their meanings.
  - Similar character shapes have similar meanings to each other.

水 氷 icht 「 [ ] <
water  ice

鮫 鮫 ( { [
_salmon  shark  brackets

Our model handles characters through their “images.”
Key Concept – (ii) Data Augmentation without Word Segmentation

- Introducing “wildcard” character — Wildcard Training
  - *Wildcard is defined as a zero-vector in the embedded space.*
  - *It replaces some input characters randomly (like dropout [Hinton et al. 2012]).*

**Input text**

<table>
<thead>
<tr>
<th>Input text</th>
<th>Augmented texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>メロスは激怒した。</td>
<td>メロス＊激＊した。</td>
</tr>
<tr>
<td><em>メロスは激怒した。</em></td>
<td>*ロ＊は激＊した。</td>
</tr>
<tr>
<td><em>メロスは怒し＊。</em></td>
<td></td>
</tr>
</tbody>
</table>

Wildcard training

The Proposed Method

a. Image-based Character Embedding (CAE)

b. Character-level Classifier with Wildcard Training (CLCNN)

Input text

Converting to image

Convolutional Autoencoder (CAE)  Character-level Convolutional Neural Networks (CLCNN)

Output activation
The Proposed Method

a. Image-based Character Embedding (CAE)

b. Character-level Classifier with Wildcard Training (CLCNN)
a. Image-based Character Embedding

- Convolutional Autoencoder (CAE) [Masci et al. 2011] is composed of Encoder and Decoder have conv. and pooling layers.

- CAE is trained by reconstruction loss beforehand.

- Our CAE encodes 6,631 character images into 64-dimensional space.

The Proposed Method

a. Image-based Character Embedding (CAE)

b. Character-level Classifier with Wildcard Training (CLCNN)
b. Character-level Convolutional Neural Networks (CLCNN)

- CLCNN performs hierarchical feature extraction and classification.
- It takes image-based embedded characters as input.
- It’s trained with **wildcard training (WT)**, dropping some characters randomly.
  - Wildcard training augments the combinations of characters.
Experiments and Results

(1) Author Estimation of Japanese Novels (10 classes)
- 104 novels written by 10 authors (almost 10 each)
- Training Dataset: 81 novels (2,010,000 characters)

(2) Publisher Estimation from Japanese Newspaper Articles (4 classes)
- 22,440 articles from four major newspapers (5,610 each) from economics, politics, international sections
- Training Dataset: 17,952 articles (55,420,000 characters)

Comparative approaches
- Character-level N-gram + TF-IDF + Logistics Regression (LR)
- Word segmentation + TF-IDF + LR
- Latent Semantic Indexing (LSI) / Latent Dirichlet Allocation (LDA) + LR
## Experiments and Results

(1) **Author Estimation of Japanese Novels**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(proposed) CAE + CLCNN + WT</td>
<td>69.57</td>
</tr>
<tr>
<td>(proposed) CAE + CLCNN w/o WT</td>
<td>52.17</td>
</tr>
<tr>
<td>(proposed) Lookup Table + CLCNN + WT</td>
<td>69.57</td>
</tr>
<tr>
<td>Lookup Table + CLCNN w/o WT</td>
<td>65.22</td>
</tr>
<tr>
<td>Character-level 3-gram* + TF-IDF</td>
<td>56.52</td>
</tr>
<tr>
<td>Word segmentation* + TF-IDF</td>
<td>47.83</td>
</tr>
<tr>
<td><strong>LSI (# topics = 60)</strong></td>
<td><strong>73.90</strong></td>
</tr>
<tr>
<td>LDA (# topics = 30)</td>
<td>52.10</td>
</tr>
</tbody>
</table>

* 3-gram and Word segmentation use top-50,000 most frequently tokens.

- In spite of no preprocessing, our method shows the second-best.
- Wildcard training (WT) raises the performance of CLCNN.
  - Wildcard training is effective for eliminating overfitting in the classifier
## Experiments and Results

(2) Publisher Estimation from Japanese Newspaper Articles

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(proposed) CAE + CLCNN + WT</td>
<td>86.72</td>
</tr>
<tr>
<td>(proposed) CAE + CLCNN w/o WT</td>
<td>80.95</td>
</tr>
<tr>
<td>(proposed) Lookup Table + CLCNN + WT</td>
<td>79.66</td>
</tr>
<tr>
<td>Lookup Table + CLCNN w/o WT</td>
<td>73.13</td>
</tr>
<tr>
<td>Character-level 3-gram* + TF-IDF</td>
<td>84.27</td>
</tr>
<tr>
<td>Word segmentation** + TF-IDF</td>
<td>67.22</td>
</tr>
<tr>
<td>LSI (# topics = 2,000)</td>
<td>84.00</td>
</tr>
<tr>
<td>LDA (# topics = 70)</td>
<td>56.10</td>
</tr>
</tbody>
</table>

* 3-gram approach uses top-30,000 most frequently tokens.
** Word segmentation approach uses all of morphemes in training data.

- Our methods shows the best score in this task.
- Other character-level methods also shows higher score.
  - Newspaper text is hard to segment words because of many coinages.
Experiments and Results

2-D Mapping of Embedded Character Vectors by t-SNE

- Some characters form clusters.
- Similar shape characters have similar vector representation.
Conclusion and Future works

- A new document analysis method for Japanese
  - Tackling much larger number of characters with “Image-based embedding”
  - Data augmentation without word segmentation
- Towards applying to different languages / NLP tasks
  - Chinese, Korean etc.
  - Tasks that need normalization process (e.g. Entity-linking)
Appendix | Loss Curve of CLCNN Training (Author Estimation)
Appendix | Loss Curve of CLCNN Training (Publisher Estimation)