We participated in the Subtask1-CR-JA & Subtask1-RR-JA, which were NER tasks with limited labeled data of Japanese medical documents. We trained 2 models respectively with augmented training data to extract named entities from the provided Japanese case reports and radiographic reports. From the aspect of Entity-F1 of all entities, our models ranked 2nd in Subtask1-CR-JA and 3rd in Subtask1-RR-JA.

Overview

Challenges and Related Works

Challenges:
- Deep Learning applications need a huge amount of data, and Japanese medical documents are relatively difficult to acquire and annotate.
- The inconsistency in a small dataset may affect models’ output. [1]

Approaches:
- To assure the annotation quality
- To reduce the necessary data volume

Existing approaches for dealing with small dataset in NER:
- Semi-Supervised Learning
- Transfer Learning
- Active Learning

Our Approach

Annotation Inconsistency Detection
- Automatically detecting entities labeled with different tags in KWIC format
- Manually correcting inappropriate tags

Data Augmentation by Synonym Replacement [2]
- Using a binomial distribution to determine whether each token should be replaced.
- Using synonyms from WordNet to replace a token.

Results

Official Results of Entity level

<table>
<thead>
<tr>
<th>Subtask</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR-JA</td>
<td>62.26</td>
<td>61.53</td>
<td>61.89</td>
<td>2</td>
</tr>
<tr>
<td>Others’ Best</td>
<td>61.96</td>
<td>68.91</td>
<td>65.25</td>
<td></td>
</tr>
<tr>
<td>RR-JA</td>
<td>86.96</td>
<td>87.09</td>
<td>87.03</td>
<td>3</td>
</tr>
<tr>
<td>Others’ Best</td>
<td>89.07</td>
<td>89.45</td>
<td>89.26</td>
<td></td>
</tr>
</tbody>
</table>

References