Named Entity Recognition on Transcription using cascaded classifiers

Firoj Alam\textsuperscript{1,2}

\textsuperscript{1}SIS Lab, Department of Information Engineering and Computer Science, University of Trento

\textsuperscript{2}HLT Unit, FBK-irst
Overview

- Named Entity Recognition (NER)
- Experiment
  - Written vs spoken documents (e.g. transcription)
  - System architecture
  - Case restoration model
- Results
- Conclusion and future study
Named Entity Recognition system

NER is the subtask of Information Extraction (IE) aiming to detect and classify entities in texts into predefined categories such as person, location, organization, time expressions and so on.
Written vs Spoken documents

- **Written documents**: Text appears as standard written form e.g. newspaper articles.

- **Spoken documents**: Speech (e.g. broadcast news) are transcribed using Automatic Speech Recognition (ASR) system.

- Three factors of recognizing NEs in spoken documents:
  - Case information is missing
  - Punctuation marks is missing
  - ASR errors

Firoj, EVALITA-2011
Written vs Spoken documents

- **Examples of written text:**

  Dal 2000 ad oggi sono stati così sottratti alle casse dello Stato ben 14 milioni di euro.

- **Examples of spoken text:**

  **Automatic Transcription:**
  dieta dimagrante parla ventidue ridotti da venticinque a quattordici membri del Cda ha cambiato lo statuto l' altoatesino Prada Acer verso la presidenza Duiella probabile amministratore delegato

  **Manual transcription:**
  per la A venticinque a quattordici membri del CDA cambiato lo Statuto l' altoatesino Pardatscher verso la Presidenza Duiella probabile amministratore delegato
Written vs Spoken documents

- Final classification has been done using Yamcha

- The word error rate (WER) of the ASR is **16.39%**, unit accuracy is **83.61%** and percent correct is **87.48%**

Firoj, EVALITA-2011
System Architecture

- Approach is similar to Typhoon developed by HLT unit at FBK.
- Second classifier is based on CRF instead of HMM
Second Classifier

- Using unlabeled datasets as additional features
  1. *First classifier* (CRF++) is trained on annotated data (training set)
  2. Annotate unlabeled data by *first classifier*
  3. *Second classifier* is trained on datasets that is produced by first classifier in step 2 and it classifies training and test sets to integrate additional features.
  4. Finally, retrain the *first classifier* on the training set produced in step 3 and classify the test data
Case and Punctuation Restoration

- L’adige corpus
- Classifier is based on CRF
- Performance of this model is 96.49

<table>
<thead>
<tr>
<th></th>
<th>Without case restoration</th>
<th>With case restoration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>63.84</td>
<td>66.53</td>
</tr>
<tr>
<td>Recall</td>
<td>58.91</td>
<td>61.27</td>
</tr>
<tr>
<td>F1</td>
<td>58.15</td>
<td>62.06</td>
</tr>
</tbody>
</table>

---

Firoj, EVALITA-2011
Official results on closed task

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>61.76%</td>
<td>60.23%</td>
<td>60.98</td>
</tr>
<tr>
<td>GPE</td>
<td>81.79%</td>
<td>78.52%</td>
<td>80.12</td>
</tr>
<tr>
<td>LOC</td>
<td>65.22%</td>
<td>47.87%</td>
<td>55.21</td>
</tr>
<tr>
<td>ORG</td>
<td>50.21%</td>
<td>43.85%</td>
<td>46.82</td>
</tr>
<tr>
<td>PER</td>
<td>47.28%</td>
<td>55.26%</td>
<td>50.96</td>
</tr>
</tbody>
</table>
## Official results on Open task

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>65.55%</td>
<td>61.69%</td>
<td>63.56</td>
</tr>
<tr>
<td>GPE</td>
<td>80.33%</td>
<td>80.44%</td>
<td>80.38</td>
</tr>
<tr>
<td>LOC</td>
<td>76.36%</td>
<td>44.68%</td>
<td>56.38</td>
</tr>
<tr>
<td>ORG</td>
<td>60.51%</td>
<td>47.52%</td>
<td>53.24</td>
</tr>
<tr>
<td>PER</td>
<td>48.92%</td>
<td>54.39%</td>
<td>51.51</td>
</tr>
</tbody>
</table>
Official results of manually transcribed test set

<table>
<thead>
<tr>
<th>Task</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed task</td>
<td>79.33%</td>
<td>79.80%</td>
<td>79.57</td>
</tr>
<tr>
<td>Open task</td>
<td>82.82%</td>
<td>81.27%</td>
<td>82.04</td>
</tr>
</tbody>
</table>
Conclusion and future study

- Case and punctuation model improve the performance of the system
- Exploiting unlabeled datasets helps to improve the performance
- Future Study:
  - Using unlabeled transcribed data
  - Adapting relevant sentences from unlabeled data
  - This system is going to include into typhoon which is available as a part of Textpro (http://textpro.fbk.eu/).
Thank you

???