pioNER: Datasets and Baseline for Armenian Named Entity Recognition

speaker: Tsolak Ghukasyan¹
tsggukasyan@ispras.ru

Garnik Davtyan²
garnik.davtyan@ispras.ru

Karen Avetisyan³
karavet@ispras.ru

Ivan Andrianov⁴
ivan.andrianov@ispras.ru

Ivannikov Laboratory for System Programming at Russian-Armenian University¹,²,³, Yerevan
Ivannikov Institute for System Programming of the Russian Academy of Sciences⁴, Moscow
Motivation

- Create named entity corpora for the Armenian language
- Establish baseline results for named entity recognition
Main contributions

- Silver-standard named entity corpus
- Gold-standard named entity corpus
- GloVe word embeddings
- Baseline results
Plan

• Datasets
  • Automatic train set generation
  • Annotation of test set

• Word embeddings training

• Evaluation of NER models
Train set generation

- Steps: Extract article fragments with outgoing links

Moscow

From Wikipedia, the free encyclopedia

Moscow is the capital and most populous city of Russia, with 13.2 million residents within the city limits and 17 million within the urban area.
Train set generation


- Steps: Select outgoing links in the article fragment
Train set generation

- Steps: *Retrieve the links’ target articles*
Train set generation

- Steps: Extract articles’ “instance of”, “subclass of” categories from their Wikidata element
Train set generation

- Steps: Classify articles into named entity types using “instance of”, “subclass of” values

```plaintext
Moscow
- Location
  - From Wikipedia, the free encyclopedia
  - Moscow is the capital and most populous city of Russia, with 13.2 million residents within the city limits and 17 million within the urban area.

City limits
- Not a named entity
  - The terms city limit and city boundary

Capital city
- Not a named entity
  - A capital city (or simply capital) is the

Urban area
- Not a named entity
  - An urban area or urban agglomeration

Russia
- Location
  - From Wikipedia, the free encyclopedia

List of cities and towns in Russia by population
- Not a named entity
  - This is a list of cities and towns in Russia with a population over 50,000
```

Wikidata item
- instance of: city
  - subclass of: human settlement

Wikidata item
- instance of: city
  - subclass of: human settlement

Wikidata item
- instance of: city
  - subclass of: human settlement

Wikidata item
- instance of: city
  - subclass of: human settlement

Wikidata item
- instance of: city
  - subclass of: human settlement
Train set generation

- Steps: *Label links’ text according to their target article’s type*
Test set annotation

- Over 250 news articles (sports, politics, art etc)
- Location, organization, person entities
- Annotation tool: BRAT NLP
- BBN Technologies guidelines for TREC 2002 question answering track

Distribution of news articles’ topics in the test set

Geographical distribution of news articles’ content in the test set
Test set statistics

- Comparable in size with benchmark datasets for other languages

Table 1. Comparison of Armenian, English, German, Spanish and Russian test sets: sentence, token, and named entity counts

<table>
<thead>
<tr>
<th>Test set</th>
<th>Tokens</th>
<th>LOC</th>
<th>ORG</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenian</td>
<td>53453</td>
<td>1306</td>
<td>1337</td>
<td>1274</td>
</tr>
<tr>
<td>English CoNLL03</td>
<td>46435</td>
<td>1668</td>
<td>1661</td>
<td>1617</td>
</tr>
<tr>
<td>German CoNLL03</td>
<td>51943</td>
<td>1035</td>
<td>773</td>
<td>1195</td>
</tr>
<tr>
<td>Spanish CoNLL02</td>
<td>51533</td>
<td>1084</td>
<td>1400</td>
<td>735</td>
</tr>
<tr>
<td>Russian factRuEval-2016</td>
<td>59382</td>
<td>1239</td>
<td>1595</td>
<td>1353</td>
</tr>
</tbody>
</table>
Word Embeddings

• Data
  • 79 million tokens
  • Source: Armenian Wikipedia, The Armenian Soviet Encyclopedia, a subcorpus of Eastern Armenian National Corpus, news articles, blog posts

• Model
  • GloVe
  • Dimensions: 50, 100, 200, 300
  • Vocabulary size: 400,000
Baselines

• Stanford NER
• SpaCy 2.0 NER
• biLSTM+CRF over GloVe+charBiLSTM features
Evaluation

- Generated data randomly split into 80% as train, 20% as development set
- Manually annotated data as test set
- IOB tagging scheme

Table 2. Evaluation results for named entity recognition algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Stanford NER</td>
<td>76.86</td>
<td>70.62</td>
</tr>
<tr>
<td>spaCy 2.0</td>
<td>68.19</td>
<td>71.86</td>
</tr>
<tr>
<td>Char-biLSTM+biLSTM+CRF</td>
<td>77.21</td>
<td>74.81</td>
</tr>
</tbody>
</table>
Error analysis

On development set:

- Errors caused by incompleteness of generated annotations
- Low precision for organization entities
- Error rate higher on descriptor words

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>B-PER</th>
<th>B-ORG</th>
<th>B-LOC</th>
<th>I-ORG</th>
<th>I-PER</th>
<th>I-LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>26707</td>
<td>100</td>
<td>57</td>
<td>249</td>
<td>150</td>
<td>78</td>
<td>129</td>
</tr>
<tr>
<td>B-PER</td>
<td>107</td>
<td>712</td>
<td>6</td>
<td>32</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>B-ORG</td>
<td>93</td>
<td>6</td>
<td>259</td>
<td>58</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B-LOC</td>
<td>226</td>
<td>25</td>
<td>32</td>
<td>1535</td>
<td>5</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>I-ORG</td>
<td>67</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>289</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>I-PER</td>
<td>46</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>660</td>
<td>8</td>
</tr>
<tr>
<td>I-LOC</td>
<td>145</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>45</td>
<td>11</td>
<td>597</td>
</tr>
</tbody>
</table>

Table 3. Confusion matrix of charBiLSTM+biLSTM+CRF on the development set
Future work

• Improved approaches for dataset generation (e.g. WiNER)
• Richer annotation
  • more entity types (e.g. Event)
  • more fine-grained types (e.g. City, Country, Region instead of Location)
Thank you!