Learning Orthographic Features in Bi-directional LSTM for Biomedical Named Entity Recognition

Nut Limsopatham and Nigel Collier
“Biomedical NER aims to identify target biomedical entities (e.g. disease, chemical and gene names) from free texts (e.g. medical articles, medical records.”

Characterization of undifferentiated human ES cells and differentiated EBs by antibodies. All monoclonal antibodies were initially selected for their abilities to recognize recombinant proteins in direct ELISAs. A subset were also tested by Western Blot analysis using recombinant proteins and cell lysate to confirm binding to a single epitope. The best clone was later screened for its applications for immunocytochemistry and flow cytometry using various cell lines. MCF-7 cells were used for screening mouse anti-human E-Cadherin and PODXL (podocalyxin-like) antibodies. MG-63 cells were used for screening mouse anti-human GATA1 (GATA binding protein 1) antibody.
Neural NER approaches

- Recent studies have shown that neural network-based models outperform traditional approaches (e.g. CRF and SVM)
  - BLSTM-CNN [1] used **CNN** to model at character level of each word before applying **BLSTM** to model contexts of words
  - BLSTM-CRF [2] used **BLSTM** to model contexts of words and used **CRF** to jointly learn to label named entities from words in a sentence
  - BLSTM-CNN-CRF [3] used **CNN** to model at character level, **BLSTM** to model contexts, and **CRF** to jointly learn to label entities

Effective Bio-NERs are driven by hand-crafted features

- Effective biomedical NER systems (e.g. Gimli [4] and BANNER [5]) are based on CRF or SVM with hand-crafted features, e.g. orthographic features and gazetteers.
We aim to enable neural networks to automatically learn and leverage orthographic features.

Traditionally, orthographic features are:

- Whether a word is in a particular format (e.g. TP53, PR1a, DD-MM-YYYY)
- Whether all characters in a word are lowercase
- Whether all characters in a word are uppercase
- Whether a word starts with an uppercase character followed by lowercase
• Enable BLSTM to learn orthographic features
• Experimental Setup
• Experimental Results
• Conclusions
Orthographic Representation

- Convert each lowercase character, uppercase character, number and punctuation in a word to special characters $c$, $C$, $n$ and $p$, respectively.

<table>
<thead>
<tr>
<th>Input Sentence</th>
<th>Orthographic Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>interaction between CrkII and A-T2</td>
<td>cccccccccc cccccccc CccCC ccc CpCn</td>
</tr>
<tr>
<td>Prognosis of asymptomatic multiple myeloma.</td>
<td>Ccccccccc cc ccccccccccccccc ccccccc cccccgcc cc ccccccccccccccccccc</td>
</tr>
<tr>
<td>activation of 3-hydroxy-3-methylglutaryl</td>
<td>Ccccccccc cc ccccccccccccccc ccccccc cccccggcc cc cc ccccccccccccccccccc</td>
</tr>
<tr>
<td>Modification of dopamine D2 receptor activity</td>
<td>Ccccccccc cc ccccccccccccccc ccccccc cccccggcc cc ccccccccccccccccccc</td>
</tr>
<tr>
<td>G alpha i2 and G alpha i2</td>
<td>C ccccc en ccc C ccccc cc</td>
</tr>
<tr>
<td>TPA induction of FGF-BP gene</td>
<td>CCC ccccccccc cc CCCpCC cccc</td>
</tr>
<tr>
<td>KAP-1 mediated repression in vivo</td>
<td>CCCpn ccccccccc ccccccccc cc ccccc</td>
</tr>
</tbody>
</table>
Our Neural Network Architecture

Interaction
between
CrkII
and
A-T

Word Embeddings
Original Sentence

Character-Based Word Representation

Orthographic Sentence

Ccccccccc
Cccccc
CcCC
ccc
CpCn

Forward LSTM
Backward LSTM

Out
Out
Out
Out
Character-based Word Representation

- Character-based Word Representation
- Max Pooling
- Convolution
- Character Embeddings (Word Matrix)
## Biomedical NER Tasks

<table>
<thead>
<tr>
<th></th>
<th>BC2</th>
<th>BioNLP09</th>
<th>NCBI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target entities</strong></td>
<td>Genes</td>
<td>Bio-molecular events</td>
<td>Diseases</td>
</tr>
<tr>
<td><strong>Type of data</strong></td>
<td>MEDLINE abstracts</td>
<td>MEDLINE abstracts</td>
<td>PubMed articles</td>
</tr>
<tr>
<td><strong>Number of sentences for training</strong></td>
<td>201</td>
<td>1,436</td>
<td>8,662</td>
</tr>
<tr>
<td><strong>Number of sentences for development</strong></td>
<td>488</td>
<td>995</td>
<td>2,872</td>
</tr>
<tr>
<td><strong>Number of sentences for testing</strong></td>
<td>58</td>
<td>2,200</td>
<td>1,036</td>
</tr>
</tbody>
</table>
Experimental Setup

- **Baseline:**
  - FeedForward [6] – A feed forward neural network with the context window size of 5
  - BLSTM – similar to [2] when hand-crafted features are not taken into account
  - CNN-BLSTM (C)
  - CNN-BLSTM [3]

Experimental Results

FeedForward performs worse than BLSTM
Experimental Results

Character-based word representation is more effective than pre-trained word embedding.

- **F1-Score**

  - **BC2**: FeedForward, BLSTM, CNN-BLSTM(C), ORTH-BLSTM
  - **BioNLP09**: CNN-BLSTM(C), ORTH-BLSTM
  - **NCBI**: CNN-BLSTM, ORTH-BLSTM
Experimental Results

Combining character and word levels further improves the performance

- BC2
- BioNLP09
- NCBI

F1-Score

- FeedForward
- BLSTM
- CNN-BLSTM(C)
- CNN-BLSTM
- ORTH-BLSTM
Automatically induced orthographic features improve the performance
Conclusions

• We have investigated a novel approach to enable BLSTM to induce and leverage orthographic features

• We compared our approach against existing effective neural NER approaches on biomedical NER tasks

➢ Our approach was also effective for biomedical NER
Thank you 😊