Efficient Dependency-Guided Named Entity Recognition

Zhanming Jie    Aldrian Obaja Muis    Wei Lu

Singapore University of Technology and Design

February 7, 2017

Slides: http://www.statnlp.org/project/depner.html
Table of Contents

Motivation
  Named Entity Recognition
  Dependency
  Relationship between dependency and NER

Related Work

Dependency-Guided NER
  Semi-Markov CRFs
  Dependency-Guided Model
  Time Complexity

Experiments
  Dataset
  Results

Conclusion
Named Entity Recognition (NER)

- **Named Entity Recognition**: important component for many natural language processing task.

- Example:

  Foreign Minister Shlomo Ben-ami gave a talk

<table>
<thead>
<tr>
<th>Foreign</th>
<th>Minister</th>
<th>Shlomo</th>
<th>Ben</th>
<th>-</th>
<th>Ami</th>
<th>gave</th>
<th>a</th>
<th>talk</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>NNP</td>
<td>NNP</td>
<td>NNP</td>
<td>HYPH</td>
<td>NNP</td>
<td>VBD</td>
<td>DT</td>
<td>NN</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
<td>B-PER</td>
<td>I-PER</td>
<td>I-PER</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
- **Dependency Tree**: focus on the relationship between words in a sentence.

- Example:
Relationship between dependency and NER

Foreign Minister Shlomo Ben - Ami gave a talk

The House of Representatives votes on the measure
Table of Contents

Motivation
  Named Entity Recognition
  Dependency
  Relationship between dependency and NER

Related Work

Dependency-Guided NER
  Semi-Markov CRFs
  Dependency-Guided Model
  Time Complexity

Experiments
  Dataset
  Results

Conclusion
Related Work

- Dependency information as features for NER. (Cucchiarelli and Velardi 2001; Sasano and Kurohashi 2008; Ling and Weld 2010)

- Skip-chain CRFs model (Liu, Huang, and Zhu 2010): loopy graphical model.

Our model is more efficient than the semi-Markov CRFs model and performs competitive performance.
Semi-Markov CRFs

- $x$: input sentence
- $y$: output sequence (e.g., a named entity label sequence in our case)

$$ p(y|x) = \frac{\exp(w^T f(x, y))}{Z(x)} $$

$$ Z(x) = \sum_{i=1}^{n} \sum_{l=1}^{L} \sum_{y' \in T} \sum_{y \in T} \exp(w^T f(x, y', y, i - l, i)) $$

- $f(x, y)$: feature vector
- $Z(x)$: partition function
Semi-Markov CRFs

- **Orange**: PERSON entity
- **Red**: MISC entity
- **Blue Path**: the gold path for the input sentence.

**Figure**: Illustrations of possible combinations of entities for the conventional semi-CRFs model and the example sentence

Find the gold path among all the possible edges.
Definition (Valid Span)

- a single word or a word sequence
- covered by a chain of (undirected) arcs where no arc is covered by another.

This leads to the following new partition function:

\[
Z(x) = \sum_{(i,j) \in S_L(x)} \sum_{y' \in T} \sum_{y \in T} \exp(w^T f(x, y', y, i, j)) \tag{1}
\]

\(S_L(x)\) refers to its subset that contains only those valid spans whose lengths are no longer than \(L\).
Dependency-Guided Model (DGM)

Figure: Illustrations of possible combinations of entities for our DGM model, as well as the example sentence with its dependency structure.
Time Complexity

- **Best case:** $O(nT^2)$
- **Worst case:** $O(nLT^2)$

**Figure:** The best-case and worst-case scenarios of DGM.
The average number of valid spans is:

\[ n \left(1 + \frac{1}{n}\right)^{n-1} \leq n \cdot e \]

This shows that the average-case time complexity of our model is \( O(n|T|^2) \).
Besides DGM model, another variant where we restrict

- the chain (of arcs) to be of length 1 (i.e., single arc) only.

Time complexity is always $O(n|T|^2)$:

- less running time
- produces promising results though less accurate than DGM.
Table of Contents

Motivation
   Named Entity Recognition
   Dependency
   Relationship between dependency and NER

Related Work

Dependency-Guided NER
   Semi-Markov CRFs
   Dependency-Guided Model
   Time Complexity

Experiments
   Dataset
   Results

Conclusion
Dataset

- Broadcast News section from OntoNotes 5.0 (Finkel and Manning 2009).
  - 7 subsections: ABC, CNN, MNB, NBC, P25, PRI and VOA.

<table>
<thead>
<tr>
<th></th>
<th># Sent.</th>
<th># Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ALL</td>
</tr>
<tr>
<td>Train</td>
<td>9,996</td>
<td>18,855</td>
</tr>
<tr>
<td>Test</td>
<td>3,339</td>
<td>5,742</td>
</tr>
</tbody>
</table>

Table: Dataset statistics.
## Results

<table>
<thead>
<tr>
<th>Dependency</th>
<th>Model</th>
<th>ABC</th>
<th>CNN</th>
<th>MNB</th>
<th>NBC</th>
<th>P2.5</th>
<th>PRI</th>
<th>VOA</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Given</strong></td>
<td>Linear-CRFs</td>
<td>70.2</td>
<td>75.9</td>
<td><strong>75.7</strong></td>
<td>65.9</td>
<td>70.8</td>
<td>83.2</td>
<td>84.6</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td>Semi-CRFs</td>
<td><strong>71.9</strong></td>
<td><strong>78.2</strong></td>
<td>74.7</td>
<td><strong>69.4</strong></td>
<td>73.5</td>
<td><strong>85.1</strong></td>
<td>85.4</td>
<td>79.6</td>
</tr>
<tr>
<td></td>
<td>DGM-S</td>
<td>71.4</td>
<td>77.0</td>
<td>73.4</td>
<td>68.4</td>
<td>72.8</td>
<td><strong>85.1</strong></td>
<td>85.2</td>
<td>79.0</td>
</tr>
<tr>
<td></td>
<td>DGM</td>
<td><strong>72.3</strong></td>
<td><strong>78.6</strong></td>
<td><strong>76.3</strong></td>
<td><strong>69.7</strong></td>
<td><strong>75.5</strong></td>
<td><strong>85.5</strong></td>
<td><strong>86.8</strong></td>
<td><strong>80.5</strong></td>
</tr>
<tr>
<td><strong>Predicted</strong></td>
<td>Linear-CRFs</td>
<td>68.4</td>
<td>75.4</td>
<td>74.4</td>
<td>66.3</td>
<td>70.8</td>
<td>83.3</td>
<td>83.7</td>
<td>77.3</td>
</tr>
<tr>
<td></td>
<td>Semi-CRFs</td>
<td><strong>71.6</strong></td>
<td><strong>78.0</strong></td>
<td>73.5</td>
<td><strong>71.5</strong></td>
<td><strong>73.7</strong></td>
<td><strong>84.6</strong></td>
<td><strong>85.3</strong></td>
<td><strong>79.5</strong></td>
</tr>
<tr>
<td></td>
<td>DGM-S</td>
<td>70.6</td>
<td>76.4</td>
<td>73.4</td>
<td>68.7</td>
<td>71.3</td>
<td><strong>83.9</strong></td>
<td>84.4</td>
<td>78.2</td>
</tr>
<tr>
<td></td>
<td>DGM</td>
<td><strong>71.9</strong></td>
<td><strong>77.6</strong></td>
<td><strong>75.4</strong></td>
<td><strong>71.4</strong></td>
<td><strong>73.9</strong></td>
<td><strong>84.2</strong></td>
<td><strong>85.1</strong></td>
<td><strong>79.4</strong></td>
</tr>
</tbody>
</table>

**Table:** NER results for all models, when given and predicted dependency trees are used and dependency features are used. Best values and the values which are not significantly different in 95% confidence interval are put in bold.
### Results

<table>
<thead>
<tr>
<th>Dependency</th>
<th>Model</th>
<th>ABC</th>
<th>CNN</th>
<th>MNB</th>
<th>NBC</th>
<th>P2.5</th>
<th>PRI</th>
<th>VOA</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Given</strong></td>
<td>Linear-CRFs</td>
<td>66.5</td>
<td>74.1</td>
<td>74.9</td>
<td>65.4</td>
<td>70.8</td>
<td>82.9</td>
<td>82.3</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>Semi-CRFs</td>
<td><strong>72.3</strong></td>
<td>76.6</td>
<td><strong>75.0</strong></td>
<td>69.3</td>
<td>73.7</td>
<td>84.1</td>
<td>83.3</td>
<td><strong>78.5</strong></td>
</tr>
<tr>
<td></td>
<td>DGM-S</td>
<td>69.4</td>
<td>76.1</td>
<td>73.4</td>
<td>68.0</td>
<td>72.5</td>
<td>85.2</td>
<td>85.1</td>
<td>78.6</td>
</tr>
<tr>
<td></td>
<td>DGM</td>
<td><strong>72.7</strong></td>
<td><strong>77.2</strong></td>
<td><strong>75.8</strong></td>
<td><strong>68.5</strong></td>
<td><strong>76.8</strong></td>
<td><strong>86.2</strong></td>
<td><strong>85.5</strong></td>
<td><strong>79.9</strong></td>
</tr>
<tr>
<td><strong>Predicted</strong></td>
<td>Linear-CRFs</td>
<td>66.5</td>
<td>74.1</td>
<td>74.9</td>
<td>65.4</td>
<td>70.8</td>
<td>82.9</td>
<td>82.3</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>Semi-CRFs</td>
<td><strong>72.3</strong></td>
<td><strong>76.6</strong></td>
<td><strong>75.0</strong></td>
<td><strong>69.3</strong></td>
<td><strong>73.7</strong></td>
<td><strong>84.1</strong></td>
<td><strong>83.3</strong></td>
<td><strong>78.5</strong></td>
</tr>
<tr>
<td></td>
<td>DGM-S</td>
<td>69.1</td>
<td>75.6</td>
<td>73.8</td>
<td>67.2</td>
<td>72.0</td>
<td>84.5</td>
<td><strong>84.2</strong></td>
<td>78.0</td>
</tr>
<tr>
<td></td>
<td>DGM</td>
<td>71.3</td>
<td><strong>76.2</strong></td>
<td><strong>75.9</strong></td>
<td><strong>68.8</strong></td>
<td><strong>74.6</strong></td>
<td><strong>85.1</strong></td>
<td><strong>84.3</strong></td>
<td><strong>78.8</strong></td>
</tr>
</tbody>
</table>

**Table**: NER results for all models, when given and predicted dependency trees are used but dependency features are not used. Best values and the values which are not significantly different in 95% confidence interval are put in bold.
Figure: Training time per iteration of all the models.
Conclusion

- DGM explicitly exploit global structured information conveyed by dependency trees.
- Experiments show that our model performs competitively with the semi-Markov CRFs model.
- Future investigation on the structural relations between dependency trees and named entities.

Our code and system available for download at http://statnlp.org/research/ie/. 