Multi-level Attention-Based Neural Networks for Distant Supervised Relation Extraction

08/12/2017

Linyi Yang, Tin Lok James Ng, Catherine Mooney, Ruihai Dong
Contents

1 Background
2 Motivation
3 Model
Background of Relation Extraction

Relation Extraction:
Aims to extract relations between two entities from the large-scale unstructured natural language text

Example:

\([\text{Steve Jobs}]_{e1} \text{ is the founder of } [\text{Apple Inc}]_{e2}\).

Relation: Founder
Why Relation Extraction?

- Create new structured knowledge bases, useful for any app
- Augment current knowledge bases
  - Facts: (head, relation, tail) are stored in knowledge base
  - Such as FreeBase or DBPedia.
- Support question answering
- Supervised learning requires labelled corpora
Background of Distant Supervision

Distant Supervision (Mintz et al., 2009):

Automatically generate training data via aligning the NYT news text with Freebase
Problem

- An entity pair only corresponds to one relation

- Counter-Example:

  \((\text{Jobs} \ e_1 , \text{Founder}, \text{Apple Inc} \ e_2)\) and \((\text{Jobs} \ e_1 , \text{CEO}, \text{Apple Inc} \ e_2)\) are both true
Background of Attention

- Attention-based neural networks were first introduced for sequence to sequence learning in machine translation.
- When predicting a target word, it first weighs every location in source sentence and then it calculates a weighted sum.
Background of Attention

- Neural Machine Translation
- (Bahdanau et al., ICLR 2015)
Contents

1. Background
2. Motivation
3. Model
Existing Solutions

- Feature-based logic regression classifier:
  - Multi-instance Multi-label learning allows multiple relations for the same entity pair. (Hoffmann et al., 2011; Surdeanu et al., 2012)
  - Cons: highly rely on NLP toolkits
Neural Networks for Distant Supervised Relation Extraction

- Automatically labelling training data and leverage the availability of big data on the web.

- Input: knowledge base triples + large scale text

- Output: the relation types of the all sentences towards the given entity pair
Existing Solutions

- Convolutional Neural Networks:

Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks (Zeng et al., 2015)
Existing Solutions

- Convolutional Neural Networks:
  - Pros: learning from data; avoid error propagation
  - Cons: lose a large amount of information contained in neglected sentences

Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks (Zeng et al., 2015)
Existing Solutions

- Convolutional Neural Networks + Attention:
  - Pros: dynamically reduce the weights of noisy instances

Neural Relation Extraction with Selective Attention over Instances (Lin et al., 2016)
Motivation

- Dynamically reduce the weights of both noisy words and sentences
- Build a multi-level attention mechanism based on Lin’s work
Multi-level Attention-Based Neural Network
Multi-level Attention-Based Neural Network

- Vector Representation
  - word embedding: capture both semantic and syntactic information of the word
  - position embedding: specify the position information of the word with respect to two target entities (Zeng et al., 2015)


Multi-level Attention-Based Neural Network

- Bidirectional Gated Recurrent Units Networks (Zhou et al., 2016)

\[
\begin{align*}
\tilde{h}_t &= \text{tanh}(W \cdot [r_t \ast h_{t-1}, x_t]) \\
r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\
z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\
h_t &= (1 - z_t) \ast h_{t-1} + z_t \ast \tilde{h}_t \\
h_j(t) &= \overrightarrow{h_j(t)} \oplus \overleftarrow{h_j(t)}
\end{align*}
\]

(Colah, 2015)
Contributions

- Customized Attention Mechanism based on Lin’s work
- Word-level attention: dynamically pay attention to the words in sentences that are more significant for semantic relation information

\[ h_1, h_2, \ldots, h_n \]

Sentence Representation

(Zhou, 2016)
## Dataset

- **Pre-Trained Word Vectors**

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>New York Times Annotated Corpus</td>
</tr>
<tr>
<td>Data type</td>
<td>Pre-Trained Word Vectors</td>
</tr>
<tr>
<td>Source</td>
<td>LDC Data LDC2008T19</td>
</tr>
</tbody>
</table>
Dataset

- Freebase relation instances

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of relations</td>
<td>53 (including “NA”)</td>
</tr>
<tr>
<td>Training data</td>
<td></td>
</tr>
<tr>
<td>sentences</td>
<td>522,611</td>
</tr>
<tr>
<td>entity pairs</td>
<td>281,270</td>
</tr>
<tr>
<td>relational facts</td>
<td>18,252</td>
</tr>
<tr>
<td>Testing data</td>
<td></td>
</tr>
<tr>
<td>sentences</td>
<td>172,448</td>
</tr>
<tr>
<td>entity pairs</td>
<td>96,678</td>
</tr>
<tr>
<td>relational facts</td>
<td>1,950</td>
</tr>
</tbody>
</table>
Results

- Precision and recall curve
Case Study

- Example of the outputs of the sentence-level attention

<table>
<thead>
<tr>
<th>Relation</th>
<th>Place_of_birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td><strong>Ernst Haefliger</strong>, a Swiss tenor who ... roles, died on Saturday in <strong>Davos</strong>, Switzerland, where he maintained a second home</td>
</tr>
<tr>
<td>High</td>
<td><strong>Ernst Haefliger</strong> was born in <strong>Davos</strong> on July 6, 1919, and studied at the Wettinger Seminary ...</td>
</tr>
</tbody>
</table>
Conclusion

- We adopt word-level attention integrated with sentence-level attention to archive better sentence representation.

- We evaluate our model on a widely used dataset to present the effect of multi-level attention mechanism.

- Experimental results show that our model outperforms the state-of-the-art methods.
References

Thanks for listening!

Linyi Yang