Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, Zhi Jin

Sep 21, 2015

Institute of Software, School of EECS, Peking University
Outline

1. Overview of Relation Extraction
2. Linguistic Phenomenon Related to Relation Extraction
3. Deep Neural Networks
   - Multi-Channel Long Short Term Memory Networks
   - Regularization: Dropout
4. Experiments
5. Conclusions
Information Extraction

A trillion gallons of water have been poured into an empty region of outer space.

A trillion gallons of [water]\textsubscript{e1} have been poured into an empty [region]\textsubscript{e2} of outer space.

Entity-Destination ([water]\textsubscript{e1}, [region]\textsubscript{e2})
SemEval 2010 Task 8 - Dataset

- Evaluation Exercises on Semantic Evaluation - ACL SigLex event
- Tasks 8 – Multi-Way Classification of Semantic Relations Between Pairs of Nominals

Training data
- 8000 training samples
  - "The system as described above has its greatest application in an arrayed <e1>configuration</e1> of antenna <e2>elements</e2>.")
  - Component-Whole(e2, e1)

Testing data
- 2717 testing samples
SemEval 2010 Task 8 - Relations

- (1) Cause-Effect
- (2) Instrument-Agency
- (3) Product-Producer
- (4) Content-Container
- (5) Entity-Origin
- (6) Entity-Destination
- (7) Component-Whole
- (8) Member-Collection
- (9) Message-Topic
- (10) Other
Outline

1. Overview of Relation Extraction
2. Linguistic Phenomenon Related to Relation Extraction
3. Deep Neural Networks
   - Multi-Channel Long Short Term Memory Networks
   - Regularization: Dropout
4. Experiments
5. Conclusions
Motivation

- Which type of **sentence structures** can be more appropriate?

- Which type of **linguistic information** can be incorporated?
Dependency Parse Tree

gallons

trillion

have

been

into

region

an

empty

of

space

outer

A

water

prep

det

amod

prep

pobj

pobj

pobj
Shortest Dependency Path (SDP)
Directionality

- Dependency trees are a kind of directed graph
- The entities’ relation distinguishes its directionality

Subpath1: `[water]e_1 \rightarrow \text{of} \rightarrow \text{gallons} \rightarrow \text{poured}`

Subpath2: `[poured] \leftarrow \text{into} \leftarrow [\text{region}]e_2`
Info 1: Word Representations

- **Word Embeddings**
  - Word2vec – map a word to a real-valued vector capturing word’s syntactic and semantic information (Mikolov, NIPS’ 2013)

![word2vec](image)

- **Toy example**
  - Average embeddings + SVM → nearly 79% F1-score
Info 2: POS tags

- Ally each word in the path with its POS tag
- Take a coarse-grained POS category heuristically

```
["NN", "NNS", "NNP", "NNPS"], # noun, proper noun, singular, plural
["IN"], # Preposition or subordinating conjunction
["VBN"], # verb, past participle
["VBD"], # verb, past tense
["VBZ"], # verb, present tense, 3rd person singular
["VBG"], # verb, present participle or gerund
["VBP"], # verb, present tense, not 3rd person singular
["VB"], # verb, base form
["TO"], # to
["JJ", "JJR", "JJS"], # adj
["RB", "RBR", "RBS"], # adv
["CD"], # cardinal number
["DT", "PDT"], # determiner
["PRP"], # personal pronoun
["RP"] # particle
```
A grammatical relation expresses the dependency between a governing word and a dependent word.

Some grammatical relations reflect semantic relation strongly. Like “nsubj”, “dobj”, or “pobj”, etc.

In our experiments, grammatical relations are grouped into 19 classes, mainly based on the category proposed by De Marneffe in LREC’ 2006.
Info 4: Hypernyms

- With the prior knowledge of hypernyms, we know that “water is a kind of substance.”

- This is a hint that the entities, water and region, are more of “Entity-Destination” relations than other relation like “Communication-Topic”, “Cause-Effect”, etc.

- With the help of supersensetagger (Ciaramita, EMNLP’06)
Outline

1. Overview of Relation Extraction
2. Linguistic Phenomenon of Relation Extraction
3. Deep Neural Networks
   - Multi-Channel Long Short Term Memory Networks
   - Regularization: Dropout
4. Experiments
5. Conclusions
The recurrent neural network is suitable for modeling sequential data by nature.

\[ h_t = f(W_{in}x_t + W_{rec}h_{t-1} + b_h) \]

- Gradient vanishing
Long Short Term Memory Networks
Long Short Term Memory Networks

\[ i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \]  \hspace{1cm} (1)

\[ f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \]  \hspace{1cm} (2)

\[ o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \]  \hspace{1cm} (3)

\[ g_t = \tanh(W_g \cdot x_t + U_g \cdot h_{t-1} + b_g) \]  \hspace{1cm} (4)

\[ c_t = i_t \otimes g_t + f_t \otimes c_{t-1} \]  \hspace{1cm} (5)

\[ h_t = o_t \otimes \tanh(c_t) \]  \hspace{1cm} (6)
Framework of SDP-LSTM
Framework of SDP-LSTM
Dropout strategies

- Randomly omitting
Dropout strategies

- Randomly omitting

- Omitting VS Memorizing
  - Dropout different types of neural network layers respectively.

\[
\begin{align*}
    i_t &= \sigma(W_i \cdot D(x_t) + U_i \cdot h_{t-1} + b_i) \\
    f_t &= \sigma(W_f \cdot D(x_t) + U_f \cdot h_{t-1} + b_f) \\
    o_t &= \sigma(W_o \cdot D(x_t) + U_o \cdot h_{t-1} + b_o) \\
    g_t &= \tanh(W_g \cdot D(x_t) + U_g \cdot h_{t-1} + b_g)
\end{align*}
\]
Training Objective

- Penalized cross-entropy error

\[ J = - \sum_{i=1}^{n_c} t_i \log y_i + \lambda \left( \sum_{i=1}^{\omega} \left\| W_i \right\|_F^2 + \sum_{i=1}^{v} \left\| U_i \right\|_F^2 \right) \]
Outline

1. Overview of Relation Extraction
2. Linguistic Phenomenon Related to Relation Extraction
3. Deep Neural Networks
   - Multi-Channel Long Short Term Memory Networks
   - Regularization: Dropout
4. Experiments
5. Conclusions
Effect of Dropout Strategies

(a) Dropout word embeddings
Effect of Dropout Strategies

(a) Dropout word embeddings

(b) Dropout inner cells of memory units
Effect of Dropout Strategies

(a) Dropout word embeddings
(b) Dropout inner cells of memory units
(c) Dropout the penultimate layer
Effects of Different Channels

- Channel effects

<table>
<thead>
<tr>
<th>Channels</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word embeddings</td>
<td>82.35</td>
</tr>
<tr>
<td>+ POS embeddings (only)</td>
<td>82.98</td>
</tr>
<tr>
<td>+ GR embeddings (only)</td>
<td>83.21</td>
</tr>
<tr>
<td>+ WordNet embeddings (only)</td>
<td>83.03</td>
</tr>
<tr>
<td>+ POS + GR + WordNet embeddings</td>
<td>83.70</td>
</tr>
</tbody>
</table>

- Traditional recurrent neural network: 82.8%

- LSTM over one path: 82.2%
## Comparison

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature set</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>POS, WordNet, prefixes and other morphological features, dependency parse, Levin classes, PropBank, FanmeNet, NomLex-Plus, Google n-gram, paraphrases, TextRunner</td>
<td>82.2</td>
</tr>
<tr>
<td>RNN</td>
<td>Word embeddings</td>
<td>74.8</td>
</tr>
<tr>
<td></td>
<td>Word embeddings, POS, NER, WordNet</td>
<td>77.6</td>
</tr>
<tr>
<td>MVRNN</td>
<td>Word embeddings</td>
<td>79.1</td>
</tr>
<tr>
<td></td>
<td>Word embeddings, POS, NER, WordNet</td>
<td>82.4</td>
</tr>
<tr>
<td>CNN</td>
<td>Word embeddings</td>
<td>69.7</td>
</tr>
<tr>
<td></td>
<td>Word embeddings, word position embeddings, WordNet</td>
<td>82.7</td>
</tr>
<tr>
<td>Chain CNN</td>
<td>Word embeddings, POS, NER, WordNet</td>
<td>82.7</td>
</tr>
<tr>
<td>FCM</td>
<td>Word embeddings</td>
<td>80.6</td>
</tr>
<tr>
<td></td>
<td>Word embeddings, dependency parsing, NER</td>
<td>83.0</td>
</tr>
<tr>
<td>CR-CNN</td>
<td>Word embeddings</td>
<td>82.8†</td>
</tr>
<tr>
<td></td>
<td>Word embeddings, position embeddings</td>
<td>82.7</td>
</tr>
<tr>
<td></td>
<td>Word embeddings, position embeddings</td>
<td>84.1†</td>
</tr>
<tr>
<td>SDP-LSTM</td>
<td>Word embeddings</td>
<td>82.4</td>
</tr>
<tr>
<td></td>
<td>Word embeddings, POS embeddings, WordNet embeddings, grammar relation embeddings</td>
<td>83.7</td>
</tr>
</tbody>
</table>
New Mission Impossible

86 %
Outline

1. Overview of Relation Extraction
2. Linguistic Phenomenon Related to Relation Extraction
3. Deep Neural Networks
   - Multi-Channel Long Short Term Memory Networks
   - Regularization: Dropout
4. Experiments
5. Conclusions
Conclusions

- Classifying relation is a challenging task due to the inherent ambiguity of natural languages and the diversity of sentence expression.

- The shortest dependency path can be a valuable resource for relation classification.

- Treating the shortest dependency path as two sub-paths helps to capture the directionality.

- LSTM units are effective in feature detection and propagation along the shortest dependency path.
Thanks!

Q&A