Task-Oriented Learning of Word Embeddings for Semantic Relation Classification

Kazuma Hashimoto (UT)
Pontus Stenetorp (UCL)
Makoto Miwa (TTI)
Yoshimasa Tsuruoka (UT)

University of Tokyo (UT)
University College London (UCL)
Toyota Technological Institute (TTI)

30/07/2015 CoNLL2015 in Beijing, China
Semantic Relation Classification

- Classifying relations between noun pairs (Hendrickx et al., 2010)
  - Input: Noun pair \( <E_1, E_2> \) in the same sentence
  - Output: Relation label

Financial stress is one of the main causes of divorce.

\[ E_1 \quad \rightarrow \quad \text{Label: Cause-Effect}<E_1, E_2> \quad \rightarrow \quad E_2 \]

The burst has been caused by water hammer pressure.

\[ E_1 \quad \rightarrow \quad \text{Label: Cause-Effect}<E_2, E_1> \quad \rightarrow \quad E_2 \]
Typical Usage of Word Embeddings

- Training based on co-occurrence statistics of words
- Feature engineering for specific tasks (Many papers)

unlabeled corpora

word2vec

Training

Word embeddings

Separated

Relation features
Incorporating the relation classification features into the embedding learning step!
• Predicting **words in between noun pairs**
  
  - Words in between noun pairs are important features to identify the relation labels of the pairs
  - Word embeddings are learned to preserve the information about the word prediction

\[ E_1 \xrightarrow{\text{caused}} \text{the external conflicts are} \xrightarrow{\text{<?>}} \text{by players} \xrightarrow{\text{E}_2} \text{playing tiles to ...} \]
Overview of the Proposed Method

RelEmb

constructing feature vectors

N, W, \tilde{W}, b → S, s

learning embeddings

unlabeled corpora

noun pairs with contexts

labeled corpora

relation classifier

supervised learning
Overview of the Proposed Method

RelEmb

Step 1

unlabeled corpora

noun pairs with contexts

N, W, \tilde{W}, b

construncting feature vectors

S, s

learning embeddings

supervised learning

labeled corpora

relation classifier

30/07/2015 CoNLL2015 in Beijing, China
Context Information with Noun Pairs

- Extracting noun pairs from large corpora
  - With the surrounding context information

Unlabeled corpora

The external conflicts are caused by players playing tiles to ...

<table>
<thead>
<tr>
<th>Noun pair</th>
<th>Between</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>conflicts, players</td>
<td>are caused by</td>
<td>the external</td>
<td>playing tiles to</td>
</tr>
</tbody>
</table>

Typical features for relation classification
Overview of the Proposed Method

RelEmb

Step 2

constructing feature vectors

N, W, \tilde{W}, b

learning embeddings

S, s

supervised learning

relation classifier

unlabeled corpora

noun pairs with contexts

labeled corpora
Predicting Words in the Noun Pairs

Negative sampling (Mikolov et al., 2013)

\[ N(\text{conflicts}) \]
\[ N(\text{players}) \]
\[ W(\text{NULL}) \]
\[ W(\text{are}) \]
\[ W(\text{by}) \]
\[ W(\text{NULL}) \]
\[ \frac{1}{2} (W(\text{the}) + W(\text{external})) \]
\[ \frac{1}{2} (W(\text{playing}) + W(\text{tiles})) \]

\[ f \]

\[ \sigma(\tilde{W}(\text{caused})f + b(\text{caused})) \]
\[ \sigma(\tilde{W}(\text{eaten})f + b(\text{eaten})) \]
\[ \sigma(\tilde{W}(\text{causes})f + b(\text{causes})) \]

logistic regression

the external [conflicts] are caused by [players] playing tiles to ...

context information

target

context information

30/07/2015 CoNLL2015 in Beijing, China
Overview of the Proposed Method

RelEmb

Step 3

constructing feature vectors

N, W, \tilde{W}, b

S, s

relation classifier

unlabeled corpora

noun pairs with contexts

learning embeddings

supervised learning

labeled corpora
Feature Vectors for a Softmax Classifier

- Basic feature set
  - Word embeddings of the noun pairs
  - Averaged \textit{n-gram embeddings} between the pairs
  - Averaged word embeddings outside the pairs

The external conflicts are caused by players playing tiles to ...

\[ [W(\text{NULL}); W(\text{are}); \tilde{W}(\text{caused}); W(\text{by}); W(\text{NULL})] \]

Window size = 2 $\rightarrow$ 5-gram embeddings
Overview of the Proposed Method

**RelEmb**

- **Step 4**
  - Constructing feature vectors
  - Learning embeddings
  - Supervised learning
  - Labeled corpora

- **Unlabeled corpora**
  - Noun pairs with contexts

- **Relation classifier**
Supervised Learning

- Fine-tuning the pre-trained word embeddings

The external conflicts are caused by players playing tiles to...

Feature vector

Softmax classifier

Label prediction

Error backpropagation

$E_1$

$E_2$
Experimental Settings

- Training corpus for the embedding learning step
  - English Wikipedia
    - 1.4 billion pairs of nouns with their contexts
  - 100-dimensional word embeddings
- Task: SemEval 2010 Task 8 (Hendrickx et al., 2010)
  - Semantic relation classification between nominals
  - 19-class classification task
    - Cause-Effect, Effect-Cause, Content-Container, Container-Content, Message-Topic, Topic-Message, …

For more details, please see our paper!
Main Results (1)

- Outperforms competitive baselines
  - W2V-Init: initializing embeddings using word2vec
  - Rand-Init: initializing embeddings randomly
- Task-oriented learning is effective!

<table>
<thead>
<tr>
<th>Features for classifiers</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RelEmb embeddings</td>
<td>82.8</td>
</tr>
<tr>
<td>RelEmb (W2V-Init) embeddings</td>
<td>81.8</td>
</tr>
<tr>
<td>RelEmb (Rand-Init) embeddings</td>
<td>78.2</td>
</tr>
<tr>
<td>SVM bag of words</td>
<td>76.5</td>
</tr>
<tr>
<td>SVM (Rink and Harabagiu, 2010) bag of words, POS, dependency paths, and many other features</td>
<td>82.2</td>
</tr>
</tbody>
</table>
Main Results (2)

• Compares favorably to other more complex neural network models
  – CR-CNN+: using **dataset-specific techniques**
• Convolutional Neural Networks (CNNs) are **much more computationally expensive** than RelEmb

<table>
<thead>
<tr>
<th>Features for classifiers</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RelEmb</td>
<td>82.8</td>
</tr>
<tr>
<td>CR-CNN+ (dos Santos et al., 2015)</td>
<td>84.1</td>
</tr>
<tr>
<td>CR-CNN (dos Santos et al., 2015)</td>
<td>82.7</td>
</tr>
<tr>
<td>FCM (Yu et al., 2014)</td>
<td>80.6</td>
</tr>
</tbody>
</table>
Main Results (3)

- Outperforms other neural network models with **syntactic and semantic** resources
  - Our method
    - is faster and
    - performs better

<table>
<thead>
<tr>
<th>Features for classifiers</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RelEmb+ embeddings, dependency paths, WordNet, named entity</td>
<td>83.5</td>
</tr>
<tr>
<td>FCM+ (Yu et al., 2014) embeddings, dependency paths, named entity</td>
<td>83.0</td>
</tr>
<tr>
<td>CNN (Zeng et al., 2014) embeddings, WordNet</td>
<td>82.7</td>
</tr>
</tbody>
</table>
Salient Patterns for Each Relation Class

- The learned *n-gram embeddings* capture salient patterns for each relation class.

```
<table>
<thead>
<tr>
<th>Cause-Effect(E₁,E₂)</th>
<th>resulted</th>
<th>poverty</th>
<th>caused</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td>caused</td>
<td>stability</td>
<td>caused</td>
<td>the</td>
<td></td>
</tr>
<tr>
<td>generated</td>
<td>coast</td>
<td>resulted</td>
<td>in</td>
<td></td>
</tr>
<tr>
<td>cause</td>
<td>fire</td>
<td>caused</td>
<td>due</td>
<td></td>
</tr>
<tr>
<td>causes</td>
<td>that</td>
<td>resulted</td>
<td>in</td>
<td></td>
</tr>
<tr>
<td>Cause-Effect(E₂,E₁)</td>
<td>after</td>
<td>caused</td>
<td>by</td>
<td>radiation</td>
</tr>
<tr>
<td>from</td>
<td>caused</td>
<td>by</td>
<td>infection</td>
<td></td>
</tr>
<tr>
<td>caused</td>
<td>stomach</td>
<td>caused</td>
<td>by</td>
<td></td>
</tr>
<tr>
<td>triggered</td>
<td>caused</td>
<td>by</td>
<td>genetic</td>
<td></td>
</tr>
<tr>
<td>due</td>
<td>anger</td>
<td>caused</td>
<td>by</td>
<td></td>
</tr>
</tbody>
</table>
```
Salient Patterns for Each Relation Class

- The learned \textit{n-gram embeddings} capture salient patterns for each relation class.

<table>
<thead>
<tr>
<th>Message-Topic($E_1,E_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>discuss</td>
</tr>
<tr>
<td>explaining</td>
</tr>
<tr>
<td>discussing</td>
</tr>
<tr>
<td>relating</td>
</tr>
<tr>
<td>describing</td>
</tr>
<tr>
<td>magazines relating to</td>
</tr>
<tr>
<td>to</td>
</tr>
<tr>
<td>discussion</td>
</tr>
<tr>
<td>concerned about NULL</td>
</tr>
<tr>
<td>interview relates to</td>
</tr>
<tr>
<td>to</td>
</tr>
<tr>
<td>discuss the</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Message-Topic($E_2,E_1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject</td>
</tr>
<tr>
<td>related</td>
</tr>
<tr>
<td>discussed</td>
</tr>
<tr>
<td>documented</td>
</tr>
<tr>
<td>received</td>
</tr>
<tr>
<td>were</td>
</tr>
<tr>
<td>was</td>
</tr>
<tr>
<td>been</td>
</tr>
<tr>
<td>is</td>
</tr>
<tr>
<td>related in</td>
</tr>
<tr>
<td>related in</td>
</tr>
<tr>
<td>discussed in</td>
</tr>
<tr>
<td>is</td>
</tr>
<tr>
<td>related through</td>
</tr>
<tr>
<td>the</td>
</tr>
<tr>
<td>subject of</td>
</tr>
</tbody>
</table>
Task-oriented learning of word embeddings is effective in a semantic relation classification task.

How about
- Other tasks?
- Other domains or datasets?
- Using label data during the embedding learning step?