Semi Supervised Preposition-Sense Disambiguation using Multilingual Data

Hila Gonen and Yoav Goldberg
Bar-Ilan University

Data Science Summit Europe, May 2017
Prepositions

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Prepositions

- Prepositions are connective words such as: on, with, at, from, about...
Prepositions

- Prepositions are connective words such as: on, with, at, from, about...

- “The dog jumped over the wall”
- “The book is on the table”
- “The show starts in two hours”
Preposition-sense disambiguation

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Preposition-sense disambiguation

- Prepositions are common, ambiguous and carry different meanings in different contexts.
Preposition-sense disambiguation

- Prepositions are common, ambiguous and carry different meanings in different contexts.

- “You should book a room for 2 nights”
  “For some reason, he is not here yet”
  “I went there to get a present for my mother”
Preposition-sense disambiguation

- Prepositions are common, ambiguous and carry different meanings in different contexts.

- “You should book a room for 2 nights” → DURATION
  “For some reason, he is not here yet”
  “I went there to get a present for my mother”
Preposition-sense disambiguation

- Prepositions are **common**, **ambiguous** and carry **different meanings** in different contexts.

- “You should book a room **for** 2 nights” → **DURATION**
  “**For** some reason, he is not here yet” → **EXPLANATION**
  “I went there to get a present **for** my mother”
Preposition-sense disambiguation

- Prepositions are common, ambiguous and carry different meanings in different contexts.

- “You should book a room **for** 2 nights” → DURATION
  “**For** some reason, he is not here yet” → EXPLANATION
  “I went there to get a present **for** my mother” → BENEFICIARY
Preposition-sense disambiguation

- Prepositions are common, ambiguous and carry different meanings in different contexts.

- “You should book a room for 2 nights” → DURATION
  “For some reason, he is not here yet” → EXPLANATION
  “I went there to get a present for my mother” → BENEFICIARY

- The preposition-sense disambiguation task:
  given a preposition within a sentential context, decide to which category it belongs, or what its role in the sentence is.
A Challenging Task

- Small dataset - 4250 examples (Schneider et al., 2016)

- Can we improve performance by using unannotated data?

- Are translations of prepositions to other languages predictive for this task of sense-disambiguation?
Our first model for preposition classification

- We treat it as a multiclass problem - We feed features into a classifier
- We use an MLP (multi-layer perceptron) as our classifier

\[ y = \arg\max_j MLP_{sense}(\phi(s, i))[j] \]

\( \phi(s, i) \) is the concatenation of 18 contextual features and the preposition’s embedding
Our first model for preposition classification

- We treat it as a multiclass problem - We feed features into a classifier
- We use an MLP (multi-layer perceptron) as our classifier

\[ y = \arg\max_j \text{MLP}_{\text{sense}}(\phi(s, i))[j] \]

\( \phi(s, i) \) is the concatenation of 18 contextual features and the preposition’s embedding
Our first model for preposition classification

- We treat it as a multiclass problem - We feed features into a classifier
- We use an MLP (multi-layer perceptron) as our classifier

\[
y = \arg\max_j MLP_{sense}(\phi(s, i))[j]
\]

\(\phi(s, i)\) is the concatenation of 18 contextual features and the preposition’s embedding
Our first model for preposition classification

- We treat it as a multiclass problem - We feed features into a classifier
- We use an MLP (multi-layer perceptron) as our classifier

\[ y = \arg\max_j MLP_{sense}(\phi(s, i))[j] \]

\( \phi(s, i) \) is the concatenation of 18 contextual features and the preposition's embedding
Our first model for preposition classification

- We treat it as a multiclass problem - We feed features into a classifier.
- We use an MLP (multi-layer perceptron) as our classifier.

\[ y = \arg\max_j MLP_{sense}(\phi(s, i))[j] \]

\( \phi(s, i) \) is the concatenation of 18 contextual features and the preposition’s embedding.

Here we use annotated data only!
Improving the first model

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Improving the first model

- We want to add something to this model to improve it
Improving the first model

- We want to add something to this model to improve it
- Idea: concatenate another vector to the feature vector
Improving the first model

- We want to add something to this model to improve it
- Idea: concatenate another vector to the feature vector
- How can we derive such a vector?
Improving the first model

- We want to add something to this model to improve it
- Idea: concatenate another vector to the feature vector
- How can we derive such a vector?
  - Preferably from unannotated data
Improving the first model

- We want to add something to this model to improve it

- Idea: concatenate another vector to the feature vector

- How can we derive such a vector?
  - Preferably from unannotated data
  - Capture some representation of the context
Improving with Multilingual Data

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Improving with Multilingual Data

- Intuition: we want to derive a representation from unannotated data that is predictive of the preposition-sense.
Improving with Multilingual Data

- Intuition: we want to derive a representation from unannotated data that is predictive of the preposition-sense.
- Ambiguity differs between languages (Dagan et al., 1991):

  “What action will it take to defuse the crisis and tension in the region?” (PLACE)

  “These are only available in English, which is totally unacceptable” (MANNER)
Improving with Multilingual Data

- Intuition: we want to derive a representation from unannotated data that is predictive of the preposition-sense.
- Ambiguity differs between languages (Dagan et al., 1991):
  
  **French: dans**
  “What action will it take to defuse the crisis and tension *in* the region?” (PLACE)

  **French: en**
  “These are only available *in* English, which is totally unacceptable” (MANNER)
Improving with Multilingual Data

- Intuition: we want to derive a representation from unannotated data that is predictive of the preposition-sense.

- Ambiguity differs between languages (Dagan et al., 1991):
  
  - French: *dans*
  
  "What action will it take to defuse the crisis and tension *in* the region?" (PLACE)
  
  - French: *en*
  
  "These are only available *in* English, which is totally unacceptable" (MANNER)

- Hypothesis: a representation that is predictive of the preposition's translation is likely to be predictive also of its sense.
Extracting training data

- Data in 12 languages from Europarl corpus (Koehn, 2005):
  - Bulgarian, Czech, Danish, German, Greek, Spanish, French, Hungarian, Italian, Polish, Romanian and Swedish.
Extracting training data

- Data in 12 languages from Europarl corpus (Koehn, 2005):
  Bulgarian, Czech, Danish, German, Greek, Spanish, French, Hungarian, Italian, Polish, Romanian and Swedish.

The vote will take place tomorrow at 12 p.m.

Le vote aura lieu demain à 12 heures.
Extracting training data

- Data in 12 languages from Europarl corpus (Koehn, 2005):
  Bulgarian, Czech, Danish, German, Greek, Spanish, French, Hungarian, Italian, Polish, Romanian and Swedish.

Training example: (FR, The vote will take place tomorrow at 12 p.m., Le vote aura lieu demain à 12 heures.)

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Learning a context representation

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
Learning a context representation

- Encode the context as a concatenation of two LSTMs:

\[ ctx(s, i) = LSTM_f(w_1:i-1) \circ LSTM_b(w_{n:i+1}) \]
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
  \[ ctx(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1}) \]

- An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.
Learning a context representation

- Encode the context as a concatenation of two LSTMs:

\[
ctx(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1})
\]

- An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
  \[ ctx(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1}) \]

- An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
  
  \[ \text{ctx}(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1}) \]

- An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.

- Context vector is fed into a language specific MLP for predicting the target preposition:
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
  \[
  ctx(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1})
  \]
  - An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.

- Context vector is fed into a language specific MLP for predicting the target preposition:
  \[
  \hat{p} = \text{argmax}_j MLP_L(ctx(s, i))[j]
  \]
Learning a context representation

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
  \[ ctx(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1}) \]

- An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.

- Context vector is fed into a language specific MLP for predicting the target preposition:
  \[ \hat{p} = \text{argmax}_j MLP_L(ctx(s, i))[j] \]
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
  \[ \text{ctx}(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1}) \]

  - An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.

- Context vector is fed into a language specific MLP for predicting the target preposition:
  \[ \hat{p} = \arg\max_j \text{MLP}_L(\text{ctx}(s, i))[j] \]

- Train the context-encoder with all languages together.
- The context-encoder and the word embeddings are shared across languages.
Learning a context representation

- Encode the context as a concatenation of two LSTMs:
  \[ ctx(s, i) = LSTM_f(w_{1:i-1}) \circ LSTM_b(w_{n:i+1}) \]
  - An LSTM is a type of a neural network that allows the representation of arbitrary sized sequences, without limiting the length of the history.

- Context vector is fed into a language specific MLP for predicting the target preposition:
  \[ \hat{p} = \text{argmax}_j MLP_L(ctx(s, i))[j] \]
  - Train the context-encoder with all languages together.
  - The context-encoder and the word embeddings are shared across languages.
Sense classification using the representation

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Sense classification using the representation

- Concatenate the representation obtained from the context encoder to the feature vector.
Sense classification using the representation

- Concatenate the representation obtained from the context encoder to the feature vector.

- Classify prepositions to senses using an MLP network:
Sense classification using the representation

- Concatenate the representation obtained from the context encoder to the feature vector.

- Classify prepositions to senses using an MLP network:

\[ y = \arg\max_j MLP_{\text{sense}}(ctx(s, i) \circ \phi(s, i))[j] \]

- \( ctx(s, i) \) - the output vector of the context-encoder
- \( \phi(s, i) \) - the feature vector
Sense classification using the representation

- Concatenate the representation obtained from the context encoder to the feature vector.

- Classify prepositions to senses using an MLP network:
  \[
  y = \arg\max_j MLP_{sense}(ctx(s, i) \circ \phi(s, i))[j]
  \]

  \(ctx(s, i)\) - the output vector of the context-encoder
  \(\phi(s, i)\) - the feature vector

- The error back-propagates also to the context-encoder and to the word embeddings.
Full Model

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Full Model

6M training examples

French prepositions: dans, en, sur, ..., par
German prepositions: mit, vor, zu, ..., gegen
Spanish prepositions: sobre, con, para, ..., a

Motivation
Method
Results

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Full Model

6M training examples

3397 training examples

French prepositions
German prepositions
Spanish prepositions

dans, en, sur, ..., par
mit, vor, zu, ..., gegen
sobre, con, para, ..., a

MLP_{FR} MLP_{GE} MLP_{SP}

context representation

LSTM_f

he booked a room

LSTM_b

for two nights

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data
Results

- The multilingual representation improves accuracy by 2.86 points:

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>73.34 (71.63-73.97)</td>
</tr>
<tr>
<td>+context</td>
<td>73.76 (71.86-75.38)</td>
</tr>
<tr>
<td>+context(multilingual)</td>
<td><strong>76.20 (74.91-77.26)</strong></td>
</tr>
</tbody>
</table>
Results

- Adding external word embeddings + ensemble further improves the results:

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (Web-reviews corpus)</th>
<th>Accuracy (Semeval 2007 corpus)</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>77.61</td>
<td>79.5</td>
</tr>
<tr>
<td>+context</td>
<td>78.90</td>
<td>81.1</td>
</tr>
<tr>
<td>+context (multilingual)</td>
<td>80.54</td>
<td>81.2</td>
</tr>
<tr>
<td>+both-contexts</td>
<td>79.84</td>
<td>81.7</td>
</tr>
</tbody>
</table>
Conclusion

- Prepositions are tricky, the annotated corpora are small
- Multilingual data can be used as an external signal for preposition disambiguation
- Learning a representation of the context from parallel corpora improves results
Thank you!
Questions?