Topic Models for Word Sense Disambiguation and Token-based Idiom Detection

Linlin Li, Benjamin Roth and Caroline Sporleder

Cluster of Excellence, MMCI
Saarland University, Germany

ACL 2010
What is Sense Disambiguation?

Words
What is Sense Disambiguation?

Words

bank?
What is Sense Disambiguation?

Words

bank?
What is Sense Disambiguation?

Words

bank?
What is Sense Disambiguation?

Words

bank?
What is Sense Disambiguation?

**Phrases**
What is Sense Disambiguation?

Phrases

spill the beans?
What is Sense Disambiguation?

**Phrases**

spill the beans?
What is Sense Disambiguation?

Phrases

"You can't retire. You know too much. You might talk."
What is Sense Disambiguation?

**Phrases**

- spill the beans?

"You can't retire. You know too much. You might talk."
Introduction

The Sense Disambiguation Model

Experimental Setup

Experiments

Conclusion

Overview

context\(c\)

Target?

SDM
Overview

- context($c$)
- Target?
- SDM

The Sense Disambiguation Model

Experiments

Conclusion
Overview

c\text{context}(c)\rightarrow\text{Target?}\quad p(s|c)\rightarrow\text{sense paraphrase}_i\rightarrow\text{SDM}
A Topic Model

PLSA (Hofmann, 1999)

\[ p(w|d) = \sum_z p(z|d) p(w|z) \]

- A generative model, decompose the conditional probability \( p(w|d) \) into a \( \text{word-topic} \) distribution \( p(w|z) \) and a \( \text{topic-document} \) distribution \( p(z|d) \)
- Each semantic topic \( z \) is represented as a distribution over words \( p(w|z) \)
- Each document \( d \) is represented as a distribution over semantic topics \( p(z|d) \)

Bayesian version, LDA (Blei et al., 2003)
Gibbs Sampling (Griffiths and Steyvers, 2004)
The Sense Disambiguation Model

**Latent Topics for Sense Disambiguation**

**Basic Idea**

- Find the sense which maximizes the conditional probability of senses given a context

\[ s = \arg \max_{s_i} p(s_i | c) \]

- This conditional probability is decomposed by incorporating a hidden variable \( z \)
The Sense Disambiguation Model

Latent Topics for Sense Disambiguation

**Basic Idea**

- Find the sense which maximizes the conditional probability of senses given a context
  \[ s = \arg \max_{s_i} p(s_i|c) \]

- This conditional probability is decomposed by incorporating a hidden variable \( z \)

More about the sense disambiguation model...

- A sense \( (s_i) \) is represented as a sense paraphrase that captures (some aspect of) the meaning of the sense.
The Sense Disambiguation Model

Latent Topics for Sense Disambiguation

Basic Idea

- Find the sense which maximizes the conditional probability of senses given a context
  \[ s = \text{arg max}_{s_i} p(s_i|c) \]

- This conditional probability is decomposed by incorporating a hidden variable \( z \)

More about the sense disambiguation model...

- A sense \( (s_i) \) is represented as a sense paraphrase that captures (some aspect of) the meaning of the sense.

- These paraphrases can be taken from existing resource such as WordNet (WSD tasks) or supplied by users (idiom task)
The Sense Disambiguation Model

Latent Topics for Sense Disambiguation

Basic Idea

- Find the sense which maximizes the conditional probability of senses given a context

\[ s = \arg \max_{s_i} p(s_i | c) \]

- This conditional probability is decomposed by incorporating a hidden variable \( z \)

More about the sense disambiguation model...

- A sense \((s_i)\) is represented as a sense paraphrase that captures (some aspect of) the meaning of the sense.

- These paraphrases can be taken from existing resource such as WordNet (WSD tasks) or supplied by users (idiom task)

- We proposed three models of how to incorporate the topic hidden variable
The Sense Disambiguation Model

**Model I**

Contexts and senses paraphrases are both treated as documents

\[ s = \arg \max_{ds_i} p(ds_i|dc) \]
The Sense Disambiguation Model

Model I

Contexts and senses paraphrases are both treated as documents

\[ s = \arg \max_{d_{si}} p(d_{si} \mid dc) \]

- Assume \( ds \) is conditionally independent of \( dc \), given \( z \)

\[ p(ds \mid dc) = \sum_{z} p(z \mid dc)p(ds \mid z) \]
The Sense Disambiguation Model

Model I

Contexts and senses paraphrases are both treated as documents

\[ s = \arg \max_{ds_i} p(ds_i|dc) \]

- Assume \( ds \) is conditionally independent of \( dc \), given \( z \)

\[ p(ds|dc) = \sum_z p(z|dc)p(ds|z) \]

- No direct estimation of \( p(ds|z) \)

\[ p(ds|dc) = p(ds) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)} \]
Use prior sense information $p(s)$ to approximate $p(ds)$

$$p(ds|dc) \approx p(s) \sum_z \frac{p(z|dc)p(z|ds)}{p(z)}$$

- The sense distribution in real corpus is often highly skewed (McCarthy, 2009)
- $p(s)$ can be taken from existing resource (e.g., sense frequency given in WordNet)
- Assume topic distribution is uniform

$$p(ds|dc) \propto p(s) \sum_z p(z|dc)p(z|ds)$$
The Sense Disambiguation Model

Inference

- The test set and sense paraphrase set are relatively small.
- Estimate topics from a very large corpus (a Wikipedia dump), with broad thematic diversity and vocabulary coverage.
- Represent sense paraphrase documents and context documents by topics $p(z|dc)$, $p(z|ds)$. 
The Sense Disambiguation Model

**Model II**

*In case no prior sense information is available*

\[
p(ds|dc) \propto p(s) \sum_z p(z|dc)p(z|ds)
\]

*Vector-space model on inferred topic frequency statistics \(v(z|d)\)*

*Maximizing the cosine value of two document vectors \(\cos(ds, dc)\)*

\[
\arg \max_{ds_i} \cos(v(z|dc), v(z|ds_i))
\]
Sometimes, a sense paraphrase is characterized only by one typical, strongly connected word

- Consider sense paraphrase $ds$ as a collection of conditionally independent words, given context documents

$$p(ds|dc) = \prod_{w_i \in ds} p(w_i|dc)$$

- Take the maximum instead of the product
  - "rock the boat" $\rightarrow \{"break the norm", "cause trouble"\}$
  - $p("break the norm, cause trouble"|dc)$, very strong requirement
  - $p("norm"|dc) \text{ OR } p("trouble"|dc) \Rightarrow$ idiomatic sense

Model III:

$$\arg \max_{qs_j} \{ \max_{w_i \in qs_j} \sum_z p(w_i|z)p(z|dc) \}$$
Data

Coarse-grained WSD

- SemEval-2007 Task-07 benchmark dataset (Navigli et al., 2009)
- Sense categories were obtained by clustering senses from WordNet 2.1 sense inventory (Navigli, 2006)

Fine-grained WSD

- SemEval-2007 Task-17 dataset (Pradhan et al., 2009)
- The sense inventory is from WordNet 2.1

Idiom Sense Disambiguation

- The idiom dataset (Sporleder and Li, 2009)
- 3964 instances of 17 potential English idiomatic expressions, manually annotated as literal or idiomatic
Sense Paraphrases

WSD Tasks
- The word forms, glosses and example sentences of
  - the sense synset
  - the reference synsets (excluding hypernym)

Idiom Task
- Paraphrases the nonliteral meaning from several online idiom dictionaries
  - e.g., rock the boat → {"break the norm", "cause trouble"}
- For the literal sense, we use 2-3 manually selected words
  - e.g., break the ice → {"ice", "water", "snow"}
## Coarse-grained WSD: Results

<table>
<thead>
<tr>
<th>System</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPV-WSD</td>
<td>79.33</td>
<td>72.76</td>
<td>84.53</td>
<td>81.52</td>
<td>78.63*</td>
</tr>
<tr>
<td>TKB-UO</td>
<td>70.76</td>
<td>62.61</td>
<td>78.73</td>
<td>74.04</td>
<td>70.21’</td>
</tr>
<tr>
<td>MII–ref</td>
<td>78.16</td>
<td>70.39</td>
<td>79.56</td>
<td>81.25</td>
<td>76.64</td>
</tr>
<tr>
<td>MII+ref</td>
<td>80.05</td>
<td>70.73</td>
<td>82.04</td>
<td>82.21</td>
<td>78.14’</td>
</tr>
<tr>
<td>MI+ref</td>
<td>79.96</td>
<td>75.47</td>
<td>83.98</td>
<td>86.06</td>
<td>79.99*</td>
</tr>
<tr>
<td>BL_{mfs}</td>
<td>77.44</td>
<td>75.30</td>
<td>84.25</td>
<td>87.50</td>
<td>78.99*</td>
</tr>
</tbody>
</table>

- MII (without annotated data, without sense prior) outperforms the best system within the same type (TKB-UO)
- MI (without annotated data, with sense prior) outperforms the best system within the same type (UPV-WSD)
- MI also outperforms the most frequent sense baseline
- Including selected reference synsets in the sense paraphrases increases the performance
### Coarse-grained WSD: Results

<table>
<thead>
<tr>
<th>System</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPV-WSD</td>
<td>79.33</td>
<td>72.76</td>
<td>84.53</td>
<td>81.52</td>
<td>78.63*</td>
</tr>
<tr>
<td>TKB-UO</td>
<td>70.76</td>
<td>62.61</td>
<td>78.73</td>
<td>74.04</td>
<td>70.21’</td>
</tr>
<tr>
<td>MII–ref</td>
<td>78.16</td>
<td>70.39</td>
<td>79.56</td>
<td>81.25</td>
<td>76.64</td>
</tr>
<tr>
<td>MII+ref</td>
<td>80.05</td>
<td>70.73</td>
<td>82.04</td>
<td>82.21</td>
<td>78.14’</td>
</tr>
<tr>
<td>MI+ref</td>
<td>79.96</td>
<td>75.47</td>
<td>83.98</td>
<td>86.06</td>
<td>79.99*</td>
</tr>
<tr>
<td>BL&lt;sub&gt;mfs&lt;/sub&gt;</td>
<td>77.44</td>
<td>75.30</td>
<td>84.25</td>
<td>87.50</td>
<td>78.99*</td>
</tr>
</tbody>
</table>

- MI (without annotated data, without sense prior) outperforms the best system within the same type (TKB-UO)
- MI (without annotated data, with sense prior) outperforms the best system within the same type (UPV-WSD)
- MI also outperforms the most frequent sense baseline
- Including selected reference synsets in the sense paraphrases increases the performance
### Coarse-grained WSD: Results

<table>
<thead>
<tr>
<th>System</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPV-WSD</td>
<td>79.33</td>
<td>72.76</td>
<td>84.53</td>
<td>81.52</td>
<td>78.63*</td>
</tr>
<tr>
<td>TKB-UO</td>
<td>70.76</td>
<td>62.61</td>
<td>78.73</td>
<td>74.04</td>
<td>70.21'</td>
</tr>
<tr>
<td>MII–ref</td>
<td>78.16</td>
<td>70.39</td>
<td>79.56</td>
<td>81.25</td>
<td>76.64</td>
</tr>
<tr>
<td>MII+ref</td>
<td>80.05</td>
<td>70.73</td>
<td>82.04</td>
<td>82.21</td>
<td>78.14'</td>
</tr>
<tr>
<td>MI+ref</td>
<td>79.96</td>
<td>75.47</td>
<td>83.98</td>
<td>86.06</td>
<td>79.99*</td>
</tr>
<tr>
<td>BL&lt;sub&gt;mfs&lt;/sub&gt;</td>
<td>77.44</td>
<td>75.30</td>
<td>84.25</td>
<td>87.50</td>
<td>78.99*</td>
</tr>
</tbody>
</table>

- MII (without annotated data, without sense prior) outperforms the best system within the same type (TKB-UO)
- MI (without annotated data, with sense prior) outperforms the best system within the same type (UPV-WSD)
- MI also outperforms the most frequent sense baseline
- Including selected reference synsets in the sense paraphrases increases the performance
## Coarse-grained WSD: Results

<table>
<thead>
<tr>
<th>System</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPV-WSD</td>
<td>79.33</td>
<td>72.76</td>
<td>84.53</td>
<td>81.52</td>
<td>78.63*</td>
</tr>
<tr>
<td>TKB-UO</td>
<td>70.76</td>
<td>62.61</td>
<td>78.73</td>
<td>74.04</td>
<td>70.21'</td>
</tr>
<tr>
<td>MII–ref</td>
<td>78.16</td>
<td>70.39</td>
<td>79.56</td>
<td>81.25</td>
<td>76.64</td>
</tr>
<tr>
<td>MII+ref</td>
<td>80.05</td>
<td>70.73</td>
<td>82.04</td>
<td>82.21</td>
<td>78.14'</td>
</tr>
<tr>
<td>MI+ref</td>
<td>79.96</td>
<td>75.47</td>
<td>83.98</td>
<td>86.06</td>
<td>79.99*</td>
</tr>
<tr>
<td>(BL_{mfs})</td>
<td>77.44</td>
<td>75.30</td>
<td>84.25</td>
<td>87.50</td>
<td>78.99*</td>
</tr>
</tbody>
</table>

- MII (without annotated data, without sense prior) outperforms the best system within the same type (TKB-UO)
- MI (without annotated data, with sense prior) outperforms the best system within the same type (UPV-WSD)
- MI also outperforms the most frequent sense baseline
- Including selected reference synsets in the sense paraphrases increases the performance
Fine-grained WSD: Results

<table>
<thead>
<tr>
<th>System</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACAI</td>
<td>52.7 ±4.5</td>
</tr>
<tr>
<td>BL_{mfs}</td>
<td>55.91±4.5</td>
</tr>
<tr>
<td>MI+ref</td>
<td><strong>56.99±4.5</strong></td>
</tr>
</tbody>
</table>

- Model I performs better than the best unsupervised system RACAI (Ion and Tufis, 2007)
- Model I also performs better than the most frequent sense baseline (BL_{mfs})
### Idiom Sense Disambiguation: Results

<table>
<thead>
<tr>
<th>System</th>
<th>Prec$_i$</th>
<th>Rec$_i$</th>
<th>F$_i$</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base$_{maj}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>78.25</td>
</tr>
<tr>
<td>co-graph</td>
<td>50.04</td>
<td>69.72</td>
<td>58.26</td>
<td>78.38</td>
</tr>
<tr>
<td>boot.</td>
<td>71.86</td>
<td>66.36</td>
<td>69.00</td>
<td>87.03</td>
</tr>
<tr>
<td>Model III</td>
<td><strong>67.05</strong></td>
<td><strong>81.07</strong></td>
<td><strong>73.40</strong></td>
<td><strong>87.24</strong></td>
</tr>
</tbody>
</table>

- The system significantly outperforms the majority baseline.
- The system also significantly outperforms one of the state-of-the-art systems, cohesion-graph based approach (Sporleder and Li, 2009).
- It also quantitatively outperforms the bootstrapping system (Li and Sporleder, 2009).
We propose three models for sense disambiguation tasks by incorporating a hidden variable which is estimated from a Wikipedia dump:

- Model I directly optimizes the conditional probability of a sense paraphrase
- Model II is a vector space model on topic frequencies
- Model III maximizes the conditional probability of a particular word in the paraphrase

The proposed models outperform comparable state-of-the-art systems.

The model can be potentially used for other application tasks when class paraphrases are available.