Supersense Tagging for Arabic: The MT-in-the-Middle Attack

Nathan Schneider
Behrang Mohit
Chris Dyer
Kemal Oflazer
Noah A. Smith
Supersense Tagging
Baselines
MT-in-the-Middle
Analysis
Outlook
Supersense Tagging

- A coarse form of **word sense disambiguation** (partitioning of WordNet synsets)
- Generalizes NER beyond proper names; **26 noun categories** (Ciaramita & Johnson 2003)
- Categories broadly applicable across domains
- Scheme suitable for direct annotation (Schneider et al. 2012)
Supersense Tagging

- **English** resources
  - WordNet (Fellbaum 1998)
  - Tagger trained on English SemCor (Ciaramita & Altun 2006) 77% F₁ in-domain

- **Arabic** resources
  - Arabic WordNet (El Kateb et al. 2006)
  - Named entities in OntoNotes (Hovy et al. 2006)
  - Supersense-tagged Wikipedia corpus (Schneider et al. 2012)
    - 65k words—1/6 the size of SemCor
Baselines

• Heuristic matching of Arabic WordNet entries + OntoNotes NEs
  ▸ only covers 33% of nouns in our corpus

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann-A</td>
<td>32</td>
<td>16</td>
<td>21.6</td>
</tr>
<tr>
<td>Ann-B</td>
<td>29</td>
<td>15</td>
<td>19.4</td>
</tr>
</tbody>
</table>

• Unsupervised sequence model
  ▸ feature-rich (Berg-Kirkpatrick et al. 2010)

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann-A</td>
<td>20</td>
<td>16</td>
<td>17.5</td>
</tr>
<tr>
<td>Ann-B</td>
<td>14</td>
<td>10</td>
<td>11.6</td>
</tr>
</tbody>
</table>

[evaluating on Arabic Wikipedia test set—18 articles, 40k words]
MT-in-the-Middle (cf. Zitouni & Florian 2008; Rahman & Ng 2012)

تنتكون الذرة من سحابة من الشحنات السالبة (الإلكترونات) تحوم حول نواة موجبة الشحنة صغيرة جدا في الوسط.
The corn is composed of negative shipments (electronics) cloud hovering over the nucleus of a very small positive body shipment in the center.
The corn is composed of negative shipments (electronics) cloud hovering over the nucleus of a very small positive body.
The corn is composed of negative shipments (electronics).

Cloud hovering over the nucleus of a very small positive body.

Shipment in the center.
MT-in-the-Middle

- Heuristic lexicon matching:

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann-A</td>
<td>32</td>
<td>16</td>
<td>21.6</td>
</tr>
<tr>
<td>Ann-B</td>
<td>29</td>
<td>15</td>
<td>19.4</td>
</tr>
</tbody>
</table>

- MT-in-the-Middle:

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann-A</td>
<td>37</td>
<td>31</td>
<td>33.8</td>
</tr>
<tr>
<td>Ann-B</td>
<td>38</td>
<td>32</td>
<td>34.6</td>
</tr>
</tbody>
</table>
### MT-in-the-Middle

- **MT-in-the-Middle:**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann-A</td>
<td>37</td>
<td>31</td>
<td>33.8</td>
</tr>
<tr>
<td>Ann-B</td>
<td>38</td>
<td>32</td>
<td>34.6</td>
</tr>
</tbody>
</table>

- **Hybrid:**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann-A</td>
<td>35</td>
<td>36</td>
<td>35.5</td>
</tr>
<tr>
<td>Ann-B</td>
<td>36</td>
<td>36</td>
<td>36.0</td>
</tr>
</tbody>
</table>
Analysis

- Pipeline has many places for noise: MT, English supersense tagging, and projection
- We focus on the impact of translation
Analysis

• Compare **cdec** vs. an off-the-shelf Arabic-English system from **QCRI**

• Translation quality:

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>METEOR</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>QCRI</td>
<td>32.86</td>
<td>32.10</td>
<td>0.46</td>
</tr>
<tr>
<td>cdec</td>
<td>28.84</td>
<td>31.38</td>
<td>0.49</td>
</tr>
</tbody>
</table>

• ...but for MTiTM supersense tagging, **cdec** is consistently better (by 2–4 points). Why?
Analysis

- Observation: overall MT scores do not necessarily measure \textit{preservation of coarse lexical semantics}
  
  - We really care about (rough) semantic adequacy for noun phrases
  
  - We elicited \textit{lexical translation acceptability} judgments for a sample of sentences  
    (cf. Carpuat 2013: SSSST)
Analysis

• Lexical acceptability rates: 91.9% for QCRI, 90.0% for cdec

• Example errors
  ▸ corn, maize for atom
  ▸ shipments for charges
  ▸ electronics for electrons
  ▸ transliteration: IMAX for EMACS, genoa lynx for GNU Linux
Analysis

- So lexical translation is mostly OK, and QCRI does slightly better at it
- cdec’s strength: providing better input to projection
  - It produces *word* alignments, whereas QCRI gives *phrase* alignments
Outlook

• Supersense tagging can be accomplished (noisily) for a language so long as it can be automatically translated to English

• Further gains should come from:
  ‣ better MT—lexical translations and word alignments
  ‣ better English supersense tagging
  ‣ better lexicon & corpus resources
Thanks

• Francisco Guzman & Preslav Nakov @ QCRI
• Wajdi Zaghouani
• Waleed Ammar
• QNRF
• All of you for listening!