Correcting prepositional phrase attachments using multimodal corpora

Sebastien Delecraz and Alexis Nasr and Frederic Bechet and Benoit Favre
Pisa, September 21, 2017

Aix-Marseille Univ, LIF, CNRS
Multimodal processing for semantic and syntactic disambiguation

- PP-Attachment is one of the main sources of errors for syntactic parsers
  - Syntactic and semantic ambiguities
  - *Example: John look at a man with a telescope.*

- Multimodal corpora with both images and text are widely available
  - images with captions
  - videos with captions and speech

- Can image and text features be combined for solving PP-ambiguities in a multimodal corpus?
Flickr30k Entities (F30kE) (Plummer et al., 2017)

- A 30k images corpus with 5 captions per image

1. someone is holding out a punctured ball in front of a brown dog with a red collar.
2. A man holding out a deflated soccer ball to a gray dog.
3. The owner tries to hand a deflated ball to his dog.
4. Large gray dog being handed a white soccer ball.
5. A brown dog staring at a soccer ball.

Someone (people) is holding out a punctured ball (other) in front of a brown dog (animals) with a red collar (clothing).
Data: our annotation

- POS tagging of the captions
- Captions containing ambiguous PP-attachment have been identified using two simple regular expressions:
  - X* Noun X* Noun X* p X*
  - X* Verb X* Noun X* p X*
- Adding manual annotation for PP-attachment for ambiguous captions

Someone is holding a punctured ball in-front-of a brown dog with a red collar.
Data: the PP-corpus

- PP-corpus consists in 29068 PP-attachment manually annotated over 22800 captions
  - With the full Flickr30k entities annotation
  - And a parse of all caption produced by a transition based parser trained on the Penn TreeBank
- 75% of the PP-attachments are well predicted by the parser
  - Presence of a true ambiguity in attachments
  - Noticeable differences between data used to train the parser and the captions
Error Prediction classifier

Goal: predict if a PP-attachment produced by the parser is correct

- Multimodal Features
  - From captions: part of syntactic parse
  - From images: spatial information, conceptual information

```
images

captions

bounding boxes

syntactic parse

spatial information

concepts```

Textual features

- the preposition, its governor, its dependent:
  - lemma
  - part-of-speech
  - syntactic dependency
  - distance between words

Someone is holding a punctured ball in-front-of a brown dog with a red collar.
Conceptual features

- 7 concepts are used: animals, body parts, instruments, vehicles, people, scene, other
- Extracted from the reference of the Flickr30k entities corpus
- Only governor and dependent concept are used as input of the classifier.
- An 8th class is used for word without bounding box
Visual features

- Limited to geometric features: information from pixel are not used

- Relative position of the dependent bounding box compared to the governor box

- Areas ratio

- For words without bounding boxes zero values are given
Adaboost based classifier

- Train: 23254 PP-attachments
- Dev: 2907 PP-attachments
- Test: 2907 PP-attachments

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.75</td>
</tr>
<tr>
<td>Textual</td>
<td>0.88</td>
</tr>
<tr>
<td>Conceptual</td>
<td>0.83</td>
</tr>
<tr>
<td>Visual</td>
<td>0.77</td>
</tr>
<tr>
<td>T + C</td>
<td>0.90</td>
</tr>
<tr>
<td>T + C + V</td>
<td>0.89</td>
</tr>
</tbody>
</table>

- Baseline corresponds to select the majority class
- Text gives the best results
- When used alone, visual features increase accuracy
Correction Strategy

When the classifier considers a PP-attachment not correct a set $G_p$ of candidate governors is identified using the simple following rules:

0 \hspace{1cm} X \rightarrow p \quad \Rightarrow \quad G_p = \{X\}

1 \hspace{1cm} N \leftarrow V \rightarrow p \quad \Rightarrow \quad G_p = G_p \cup \{N\}

2 \hspace{1cm} N \leftarrow P \leftarrow V \rightarrow p \quad \Rightarrow \quad G_p = G_p \cup \{N\}

3 \hspace{1cm} N' \leftarrow N \rightarrow p \quad \Rightarrow \quad G_p = G_p \cup \{N'\}

4 \hspace{1cm} N' \leftarrow P \leftarrow N \rightarrow p \quad \Rightarrow \quad G_p = G_p \cup \{N'\}

5 \hspace{1cm} N' \rightarrow X \rightarrow N \rightarrow p \quad \Rightarrow \quad G_p = G_p \cup \{N'\}

6 \hspace{1cm} N \rightarrow N \rightarrow p \quad \Rightarrow \quad G_p = G_p \cup \{N\}

7 \hspace{1cm} V \rightarrow N \rightarrow p \quad \Rightarrow \quad G_p = G_p \cup \{V\}

With a use of 1.5 rules on average, $G_p$ contains the correct governor in 92.28% of the cases.
Correction Strategy

- Focus only on PP-attachment considered as erroneous by the error prediction classifier
- Compute the set $G_p$ and the output scores of the classifier for each candidate governor
- The governor with the best score for the CORRECT class is selected

More efficient than parse reranking
Experiments

PP-attachment accuracy on the test set after using the correction strategy

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.75</td>
</tr>
<tr>
<td>T</td>
<td>0.85</td>
</tr>
<tr>
<td>C</td>
<td>0.82</td>
</tr>
<tr>
<td>V</td>
<td>0.77</td>
</tr>
<tr>
<td>T + C</td>
<td>0.86</td>
</tr>
<tr>
<td>T + V</td>
<td>0.86</td>
</tr>
<tr>
<td>C + V</td>
<td>0.82</td>
</tr>
<tr>
<td>T + C + V</td>
<td>0.86</td>
</tr>
</tbody>
</table>

- Textual features, which are the most specific, are the most relevant feature set
- Conceptual features results are close to textual features
- Visual features improve accuracy when used alone without impacted when mixed with other features
Experiments

PP-attachment accuracy on the test set for some prepositions

<table>
<thead>
<tr>
<th>Prep</th>
<th>Occ</th>
<th>BL</th>
<th>T</th>
<th>C</th>
<th>V</th>
<th>TCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>with</td>
<td>310</td>
<td>0.65</td>
<td>0.78</td>
<td>0.75</td>
<td><strong>0.66</strong></td>
<td>0.79</td>
</tr>
<tr>
<td>during</td>
<td>41</td>
<td>0.71</td>
<td>0.76</td>
<td>0.73</td>
<td><strong>0.71</strong></td>
<td>0.76</td>
</tr>
<tr>
<td>around</td>
<td>59</td>
<td>0.73</td>
<td>0.81</td>
<td>0.73</td>
<td><strong>0.71</strong></td>
<td>0.83</td>
</tr>
<tr>
<td>behind</td>
<td>35</td>
<td>0.74</td>
<td>0.86</td>
<td>0.83</td>
<td><strong>0.77</strong></td>
<td>0.89</td>
</tr>
</tbody>
</table>
• Correction strategy which use multimodal features with good performance on PP-attachment
• As expected the most relevant features set is the textual one
• Visual features, limited to spatial information in our case, can improve the accuracy of the PP-attachment
• Work in progress to use pixel information from images to increase visual features impact
Conclusion

Thank you

Manual annotation of PP-corpus available on request