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

Context of the study

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 ?

Data

Flickr30k Entities (F30kE) (Plummer et al., 2017)

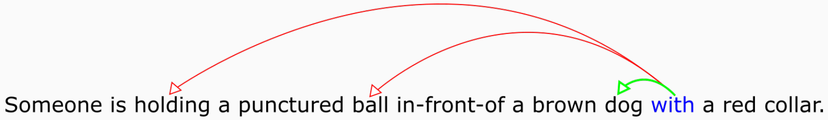
- A 30k images corpus with 5 captions per image

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 starring 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



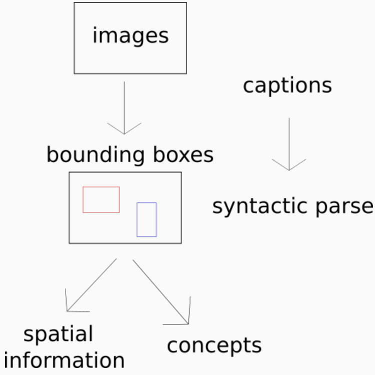
Data: the PP-corpus

- 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 Peen TreeBank
- 75% of the PP-attachments are well predicted by the parser
    - Presence of a true ambiguity in attachments
    - Noticeable deferences 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



Error Prediction classifier

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.
NN PP NN

Error prediction classifier

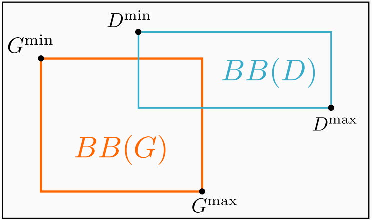
Conceptual features

- 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 8^{th} class is used for word without bounding box

Error prediction classifier

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

Error Prediction classifier

Adaboost based classifier

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

Features	Accuracy
Baseline	0.75
Textual	0.88
Conceptual	0.83
Visual	0.77
T + C	0.90
T + C + V	0.89

- 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:

$$
\begin{array}{lll}
0 & X \rightarrow p & \Rightarrow G_p = \{X\} \\
1 & N \leftarrow V \rightarrow p & \Rightarrow G_p = G_p \cup \{N\} \\
2 & N \leftarrow P \leftarrow V \rightarrow p & \Rightarrow G_p = G_p \cup \{N\} \\
3 & N' \leftarrow N \rightarrow p & \Rightarrow G_p = G_p \cup \{N'\} \\
4 & N' \leftarrow P \leftarrow N \rightarrow p & \Rightarrow G_p = G_p \cup \{N'\} \\
5 & N' \rightarrow X \rightarrow N \rightarrow p & \Rightarrow G_p = G_p \cup \{N'\} \\
6 & N \rightarrow N \rightarrow p & \Rightarrow G_p = G_p \cup \{N\} \\
7 & V \rightarrow N \rightarrow p & \Rightarrow G_p = G_p \cup \{V\}
\end{array}
$$

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

Features	Accuracy
Baseline	0.75
T	0.85
C	0.82
V	0.77
T + C	0.86
T + V	0.86
C + V	0.82
T + C + V	0.86

- 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

Prep	Occ	BL	T	C	V	TCV
with	310	0.65	0.78	0.75	0.66	0.79
during	41	0.71	0.76	0.73	0.71	0.76
around	59	0.73	0.81	0.73	0.71	0.83
behind	35	0.74	0.86	0.83	0.77	0.89

Conclusion

- 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

Thank you

`Manual annotation of PP-corpus available on request`