A cognitively plausible model for anaphora resolution
Olga Seminck and Pascal Amsili
Université Paris Diderot Paris 7

Abstract
Anaphora resolution is a complex problem as it deals with syntax, semantics and discourse. The subject is well studied in the field of psycholinguistics, where multiple preferences were discovered, and in the field of computational linguistics, where many systems have been developed to perform anaphora resolution in documents. Nevertheless, the work done in the two fields remains disconnected.

We investigate how we can bridge the gap by exploiting the options for making a cognitively plausible model for anaphora resolution. Such a model can be beneficial for both fields as it can inspire computational linguistics with findings about how humans process anaphora, and help the psycholinguistic community developing large coverage, incremental models simulating the human processing of anaphora.

The project

Entropy over the Bell Tree as a cost metric

Entropy reflects the uncertainty of a random variable
Entropy is maximal when all possible outcomes have equal probabilities.

\[ H(X) = - \sum_{i \in X} p(X = i) \cdot \log_2(p(X = i)) \]

The space of coreference resolution can be presented by the Bell Tree [1].
The paths of the tree from the root to the leaves can receive probability
All the paths form a probability distribution \( \rightarrow \) calculate entropy

We formulate the cost of an anaphor as the ratio between the entropy over the probability distribution the paths of the Bell Tree and its maximal entropy.

Emma sees a cow in the meadow.

1. \( \{\text{Emma}\} \) 1.0
2. \( \{\text{Emma}, \text{a cow}\} \)
3. \( \{\text{Emma}, \text{cow, the meadow}\} \)

Calculate Entropy

Cost

Model's Prediction

▶ Every referential expression of a text can receive a difficulty score.
▶ Further research can indicate if the cost metric is valuable of all types of coreference, or only certain types of anaphora resolution.

Validation of the Model

▶ We want to validate our model on reading times on corpus.
▶ Dundee Eye-tracking Corpus [2]
▶ French Treebank Corpus [3]
▶ Challenges:
Anaphora are often very short (pronouns) and are often not fixated.
We need to control for low level factors: word frequency, word length, position on the line...

\[ \Rightarrow \text{We have to build a mixed effect model} \]

Conclusion
Our model of processing cost seems reasonable, but is not yet validated on corpus. A next step in research is to take into account more factors, especially factors coming from syntax. Furthermore, we need to reflect on a way of measuring reading time for anaphora. Once we developed a method for measuring reading time of anaphora, we will try to validate our model with a mixed effect model on an eye-tracking corpus.

References

Acknowledgments
This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the ‘Investissements d’Avenir’ program (reference: ANR-10-LABX-0083).
This work is supported by the PhD program Frontières du Vivant of the Center for Research and Interdisciplinarity.