An Incremental Model for the Coreference Resolution Task of BioNLP 2011

University of Zurich[UZH]

Don Tuggener, Fabio Rinaldi, Manfred Klenner
Gerold Schneider, Simon Clematide

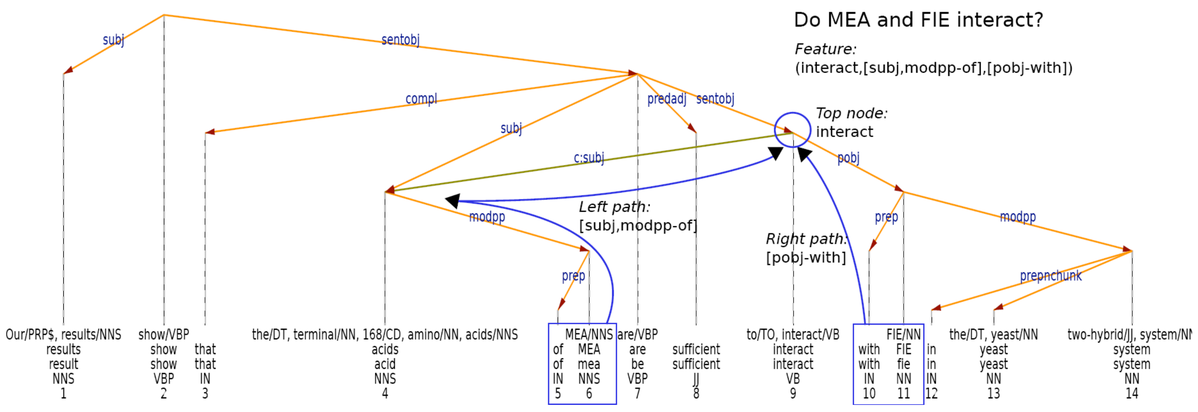
Coreference Resolution Task of BioNLP2011

The Coreference Resolution task of BioNLP focused on finding anaphoric references to proteins and genes. Only antecedent-anaphora pairs are considered in evaluation and not full coreference sets. Although it might not seem to be necessary to generate full coreference sets, anaphora resolution still benefits from their establishment. Our incremental approach naturally enforces transitivity constraints and thereby reduces the number of potential antecedent candidates. The system achieved good results in the BioNLP 2011 shared task:

Team	R	P	F1
A	22.18	73.26	34.05
Our model	21.48	55.45	30.96
B	19.37	63.22	29.65
C	14.44	67.21	23.77
D	3.17	3.47	3.31
E	0.70	0.25	0.37

Preprocessing

Our system is organized around a pipeline of NLP tools, which perform tasks such as sentence splitting, tokenization, part-of-speech tagging, lemmatization, term extraction, recognition of noun and verb groups, dependency parsing. All the information derived by the various stages of linguistic processing can be leveraged upon for text mining purposes. When preprocessing finishes, all sentences are tokenized and their borders are detected; each token is lemmatized; tokens which belong to terms are grouped; each chunk has a type (NP or VP) and a head token; each sentence is described as a syntactic dependency structure.



Our General Coreference Resolution Architecture

Our incremental entity-mention model adresses the main problems of the commonly known mention-pair model (f.e. [1]). The main features are:

- Limited number of antecedent candidates
- Incremental one pass resolution: No clustering, no additional enforcement of transitivity constraints needed
- No underspecified antecedent candidates

Furthermore, the system is light-weighted and can easily be adapted to other systems such as Ontogene[3]:

- No Machine Learning
- Simple salience based antecedent selection
- Very fast

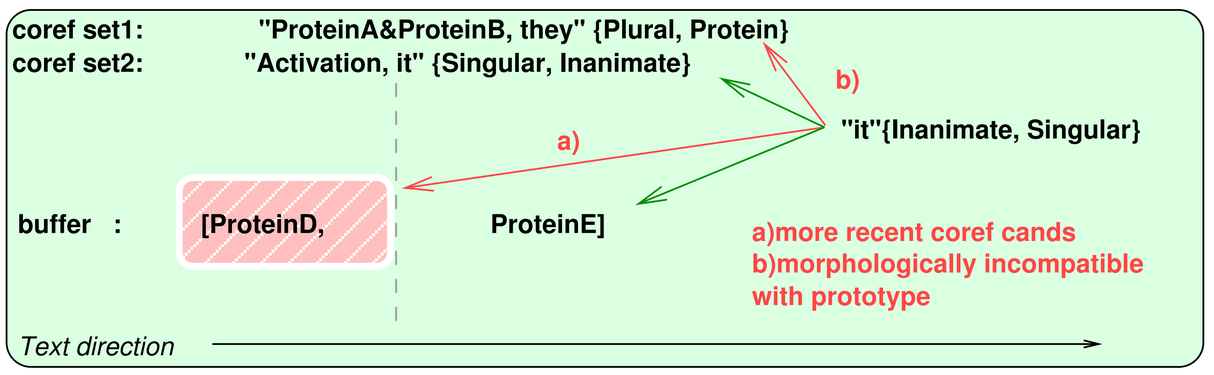
Incremental Approach

After markable extraction based on POS tags and the output of the preprocessing pipeline, articles are processed in a left-to-right manner, similar to f.e. [2]. Coreference sets are established on-the-fly. Candidate anaphors are compared to members of the waiting list and the coreference sets.

- Non-anaphoric markables are stored in the waiting list/buffer, as they might be valid antecedents for subsequent anaphors.
- Anaphoric markables are added to already established coreference sets if so determined by filters and salience.

Restrictive Antecedent Selection

We create a virtual prototype per evolving coreference set, containing morphological and semantic information accumulated from all the members of the set (f.e. [Singular, Animate etc.]). This reduces the problem of underspecified items in the mention-pair model. Furthermore, since only the virtual prototype is accessible and the actual members of a set are hidden, we do not create transitively redundant pairs.
Waiting list access is restricted if compatible and more recent coreference set candidates are available.
Antecedent candidates that are exclusive through binding theory constraints are omitted. We know f.e. that that 'modulator' and 'it' cannot be coreferent in the sentence "Overexpression of protein inhibited stimulus-mediated transcription, whereas modulator enhanced it".



Filtering based on anaphora type

- Reflexive pronouns → subject in the same (sub-)clause.
- Relative pronouns → closest markable to the left
- Possessive/personal pronouns → window of 3 sentences, antecedent with highest salience is selected
- Nouns / Named entities → different string matching filters (no non-matching candidates). Demonstrative NPs containing the lemmata 'protein' or 'gene' are licensed to bind to name containing mentions. Demonstrative NPs not containing the trigger lemmata can be resolved to string matching NPs preceding them.

Antecedent selection with a simple salience measure

Salience is calculated solely on the basis of the dependency labels of the true mentions. The salience of a dependency label, D, is estimated by the number of true mentions in the gold standard that bear D, divided by the total number of true mentions. The salience of the label 'subject' is thus calculated by:

$$\frac{Number\ of\ true\ mentions\ bearing\ subject}{Total\ number\ of\ true\ mentions}$$

We get a hierarchical ordering of the dependency labels (subject > object > pobject > ...) according to which antecedent candidates are ranked.

References

[1] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. 27, 4 (December 2001), 521-544.

[2] Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. Comput. Linguist. 20, 4 (December 1994), 535-561.

[3] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon, Ontogene in Biocreative II, Genome Biology, 9:S13, 2008.

Institute of Computational Linguistics
www.cl.uzh.ch

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

NOVARTIS

Contact: Don Tuggener
tuggener@cl.uzh.ch