University of Zurich[UZH]

An Incremental Model for Coreference Resolution with Restrictive Antecedent Accessibility

Manfred Klenner
Don Tuggener

General Approach

Our incremental entity-mention model addresses 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
- Use of a virtual prototype per coreference set to reduce underspecification of antecedent candidates

Intended to be a baseline, the system is light-weighted:

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

Incremental Entity-mention Model

We use the Pro3Gres dependency parser to obtain morphological and syntactic information. Markables are extracted based on the POS tags. We perform Named Entity Classification with a large name list containing gender information. For Animacy Detection we use WordNet.

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, 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. (A new coreference set is opened if the antecedent is from the waiting list.)

Filtering based on anaphora type

- Reflexive pronouns → subject
- Relative pronouns → closest markable to the left
- Possessive/personal pronouns → Morphologically compatible with antecedent candidate, window of 3 sentences, antecedent with highest salience is selected
- Nouns / Named entities → different string matching filters (no non-matching candidates)

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 antecedents are ranked.

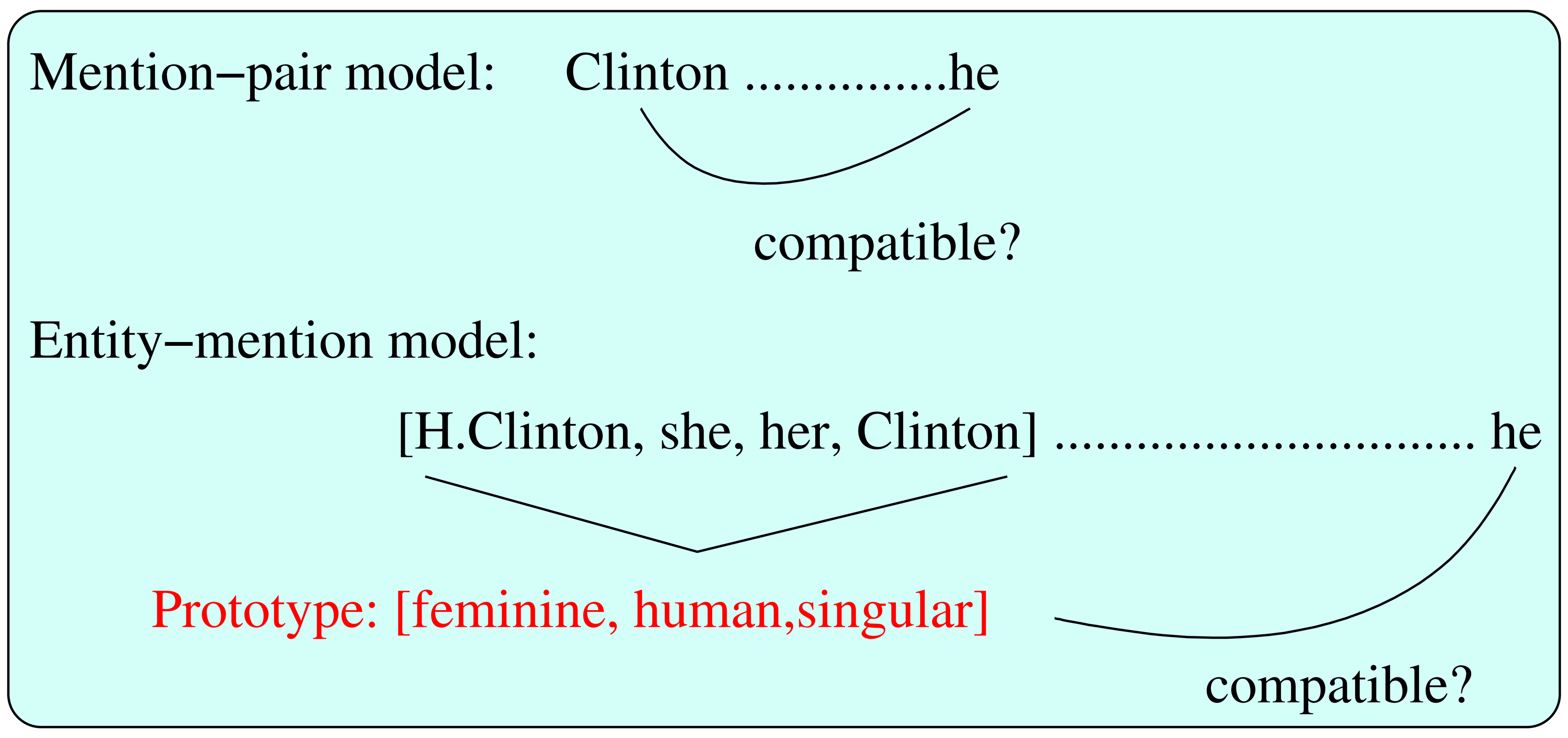
References

[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.

Restrictive Antecedent Accessibility

We create a virtual prototype per evolving coreference set, containing morphological and semantic information accumulated from all the members of the set. This reduces the problem of underspecified items in the mention-pair model. For instance, assume a candidate pair 'Clinton' ↔ 'he'.



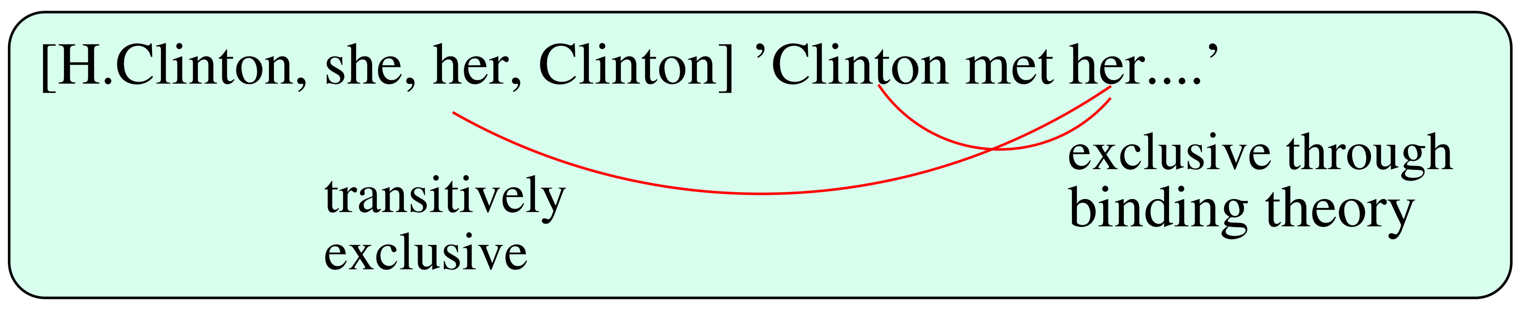
Furthermore, since only the virtual prototype is accessible and the actual members of a set are hidden, we do not create transitively redundant pairs. In the example above, assume 'Clinton' ↔ 'she':

> Mention-pair model: access all compatible antecedents, creates four pairs
> 'Clinton' ↔ 'she' | 'her' ↔ 'she' | 'she' ↔ 'she' |
> 'Hillary Clinton' ↔ 'she'
>
> Entity-mention model, coreference set established, only prototype subsuming other mentions is accessible:
> Prototype [feminine, human, singular] ↔ 'she'

The number of considered pairs is reduced from the cardinality of a set to 1. We observed a huge reduction of candidate pairs by a good $\frac{3}{4}$ when moving from the non-incremental mention-pair to the incremental entity-mention model.

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 that 'Clinton' and 'her' can't be coreferent in the sentence 'Clinton met her'. Thus, the pair 'Clinton'-'her' need not be considered at all and we know all mentions of the 'Clinton' coreference set are transitively exclusive.



CoNLL open track results

Our system was outperformed by the two other participants (58.31, 55.71 (we, 51.77)) in the open regular task.

This comes as no surprise since we do not treat non-matching non-pronominal anaphora at all. Compared to the results in the closed section we are in the middle field, despite of this shortcoming.

Metric	R	P	F1
MUC	49.04	50.71	49.86
BCUBED	61.70	68.61	64.97
CEAF (M)	50.03	50.03	50.03
CEAF (E)	41.28	39.70	40.48
BLANC	66.05	73.90	69.05

Institute of Computational Linguistics
www.cl.uzh.ch

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

Contact: Don Tuggener
tuggener@cl.uzh.ch