An Incremental Model for the Coreference Resolution Task of BioNLP 2011

Don Tuggener, Manfred Klenner, Gerold Schneider, Simon Clematide, Fabio Rinaldi
Institute of Computational Linguistics, University of Zurich, Switzerland
{tuggener,klenner,gschneid,siclemat,rinaldi}@cl.uzh.ch

Abstract

We introduce our incremental coreference resolution system for the BioNLP 2011 Shared Task on Protein/Gene interaction. The benefits of an incremental architecture over a mention-pair model are: a reduction of the number of candidate pairs, a means to overcome the problem of underspecified items in pair-wise classification and the natural integration of global constraints such as transitivity. A filtering system takes into account specific features of different anaphora types. We do not apply Machine Learning, instead the system classifies with an empirically derived salience measure based on the dependency labels of the true mentions. The OntoGene pipeline is used for preprocessing.

1 Introduction

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 (Klenner et al., 2010) naturally enforces transitivity constraints and thereby reduces the number of potential antecedent candidates. The system achieved good results in the BioNLP 2011 shared task (Fig. 1)

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

Figure 1: Protein/Gene Coreference Task

2 Preprocessing: The OntoGene Pipeline

OntoGene's text mining system is based on an internally-developed fast, broad-coverage, deep-syntactic parsing system (Schneider, 2008). The parser is wrapped into a pipeline which uses a number of other NLP tools. The parser is a key component in a pipeline of NLP tools (Rinaldi et al., 2010), used to process input documents. First, in a preprocessing stage, the input text is transformed into a custom XML format, and sentences and tokens boundaries are identified. The OntoGene pipeline also includes a step of term annotation and disambiguation, which are not used for the BioNLP shared task, since relevant terms are already provided in both the training and test corpora. The pipeline also includes part-of-speech taggers, a lemmatizer and a syntactic chunker.

When the pipeline finishes, each input sentence has been annotated with additional information, which can be briefly summarized as follows: sentences are tokenized and their borders are detected; each sentence and each token has been assigned an ID; each token is lemmatized; tokens which belong to terms are grouped; each term is assigned a normal-form and a semantic type; tokens and terms are then grouped into chunks; each chunk has a type (NP or VP) and a head token; each sentence is described as a syntactic dependency structure. All this information is represented as a set of predicates and stored into the Knowledge Base of the system, which can then be used by different applications, such as the OntoGene Relation Miner (Rinaldi et al., 2006) and the OntoGene Protein-Protein Interaction discovery tool (Rinaldi et al., 2008).

3 Our Incremental Model for Coreference Resolution

```
for i=1 to length(I)
    for j=1 to length(C)
        $r_j$ := virtual prototype of coreference set $C_j$
        Cand := Cand $\oplus$ $r_j$ if compatible($r_j, m_i$)
    for k= length(B) to 1
        $b_k$:= the k-th licensed buffer element
        Cand := Cand $\oplus$ $b_k$ if compatible($b_k, m_i$)
if Cand = {} then B := B $\oplus$ $m_i$
if Cand $\neq$ {} then
    $ante_i$ := most salient element of Cand
    C := augment(C,$ante_i$,$m_i$)
```

Figure 2: Incremental model: base algorithm

Proceedings of BioNLP Shared Task 2011 Workshop, pages 151–152,
Portland, Oregon, USA, 24 June, 2011. ©2011 Association for Computational Linguistics

Fig. 2 shows the base algorithm. Let I be the chronologically ordered list of NPs, C be the set of coreference sets and B a buffer, where NPs are stored, if they are not anaphoric (but might be valid antecedents). Furthermore m_i is the current NP and $\oplus$ means concatenation of a list and a single item. The algorithm proceeds as follows: a set of antecedent candidates is determined for each NP m_i (steps 1 to 7) from the coreference sets (r_j) and the buffer (b_k). A valid candidate r_j or b_k must be compatible with m_i. The definition of compatibility depends on the POS tags of the anaphor-antecedent pair. The most salient available candidate is selected as antecedent for m_i.

3.1 Restricted Accessibility of Antecedent Candidates

In order to reduce underspecification, m_i is compared to a virtual prototype of each coreference set (similar to e.g. (Luo et al., 2004; Yang et al., 2004; Rahman and Ng, 2009)). The virtual prototype bears morphologic and semantic information accumulated from all elements of the coreference set. Access to coreference sets is restricted to the virtual prototype. This reduces the number of considered pairs (from the cardinality of a set to 1).

3.2 Filtering based on Anaphora Type

Potentially co-refering NPs are extracted from the OntoGene pipeline based on POS tags. We then apply filtering based on anaphora type: Reflexive pronouns must be bound to a NP that is governed by the same verb. Relative pronouns are bound to the closest NP in the left context. Personal and possessive pronouns are licensed to bind to morphologically compatible antecedent candidates within a window of two sentences. 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[1].

3.3 Binding Theory as a Filter

We know through binding theory that 'modulator' and 'it' cannot be coreferent in the sentence *"Overexpression of protein inhibited stimulus-mediated transcription, whereas modulator enhanced it"*. Thus, the pair 'modulator'-'it' need not be considered at all. We have not yet implemented a full-blown binding theory. Instead, we check if the antecedent and the anaphor are governed by the same verb.

4 An Empirically-based Salience Measure

Our salience measure is a partial adaption of the measure from (Lappin and Leass, 1994). The salience of a NP is solely defined by the salience of the dependency label it bears. The salience of a dependency label, D, is estimated by the number of true mentions (i.e. co-refering NPs) that bear D (i.e. are connected to their heads with D), divided by the total number of true mentions (bearing any D). 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 and selected.

References

Manfred Klenner, Don Tuggener, and Angela Fahrni. 2010. Inkrementelle koreferenzanalyse für das deutsche. In *Proceedings der 10. Konferenz zur Verarbeitung Natürlicher Sprache*.

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:P. 535–561.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 968–977, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. 2006. An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3.

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. 2008. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, and Martin Romacker. 2010. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):472–480.

Gerold Schneider. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. An np-cluster based approach to coreference resolution. In *Proceedings of the 20th international conference on Computational Linguistics*.

[1] As we do not perform anaphoricity determination of nominal NPs, we do not consider bridging anaphora (anaphoric nouns that are connected to their antecedents through semantic relations and cannot be identified by string matching).