An Incremental Model for Coreference Resolution with Restrictive Antecedent Accessibility

Manfred Klenner
Institute of Computational Linguistics
University of Zurich
klenner@cl.uzh.ch

Don Tuggener
Institute of Computational Linguistics
University of Zurich
tuggener@cl.uzh.ch

Abstract

We introduce an incremental model for coreference resolution that competed in the CoNLL 2011 shared task (open regular). We decided to participate with our *baseline* model, since it worked well with two other datasets. The benefits of an incremental over a mention-pair architecture are: a drastic reduction of the number of candidate pairs, a means to overcome the problem of underspecified items in pairwise classification and the natural integration of global constraints such as transitivity. We do not apply machine learning, instead the system uses an empirically derived salience measure based on the dependency labels of the true mentions. Our experiments seem to indicate that such a system already is on par with machine learning approaches.

1 Introduction

With notable exceptions (Luo et al., 2004; Yang et al., 2004; Daume III and Marcu, 2005; Culotta et al., 2007; Klenner, 2007; Rahman and Ng, 2009; Klenner and Ailloud, 2009; Cai and Strube, 2010; Raghunathan et al., 2010) supervised approaches to coreference resolution are often realized by pairwise classification of anaphor-antecedent candidates. A popular and often reimplemented approach is presented in (Soon et al., 2001). As recently discussed in (Ng, 2010), the so called mention-pair model suffers from several design flaws which originate from the locally confined perspective of the model:

- Generation of (transitively) redundant pairs, as the formation of coreference sets (coreference clustering) is done after pairwise classification
- Thereby generation of skewed training sets which lead to classifiers biased towards negative classification
- No means to enforce global constraints such as transitivity
- Underspecification of antecedent candidates

These problems can be remedied by an incremental entity-mention model, where candidate pairs are evaluated on the basis of the emerging coreference sets. A clustering phase on top of the pairwise classifier no longer is needed and the number of candidate pairs is reduced, since from each coreference set (be it large or small) only one mention (the most representative one) needs to be compared to a new anaphor candidate. We form a 'virtual prototype' that collects information from all the members of each coreference set in order to maximize 'representativeness'. Constraints such as transitivity and morphological agreement can be assured by just a single comparison. If an anaphor candidate is compatible with the virtual prototype, then it is by definition compatible with all members of the coreference set.

We designed our system to work purely with a simple, yet empirically derived salience measure. It turned out that it outperformed (for German and English, using CEAF, B-cubed and Blanc) the systems from the 2010's SemEval shared task[1] on 'coreference resolution in multiple languages'. Only with the more and more questioned (Luo, 2005; Cai and

[1] We have carried out a post task evaluation with the data provided on the SemEval web page.

Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 81–85,
Portland, Oregon, 23-24 June 2011.

Strube, 2010) MUC measure our system performed worse (at least for English). Our system uses real preprocessing (i.e. a dependency parser (Schneider, 2008)) and extracts markables (nouns, named entities and pronouns) from the chunks and based on POS tags delivered by the preprocessing pipeline. Since we are using a parser, we automatically take part in the *open regular session*. Please note that the dependency labels are the only additional information being used by our system.

2 Our Incremental Model

Fig. 1 shows the basic algorithm. Let I be the chronologically ordered list of markables, C be the set of coreference sets (i.e. the coreference partition) and B a buffer, where markables are stored, if they are not found to be anaphoric (but might be valid antecedents, still). Furthermore m_i is the current markable 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 markable m_i (steps 1 to 7) from the coreference sets and the buffer. 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 (in order to be coreferent, e.g. two pronouns must agree in person, number and gender etc.).

In order to reduce underspecification, m_i is compared to a virtual prototype of each coreference set. The virtual prototype bears information accumulated from all elements of the coreference set. For instance, assume a candidate pair 'she .. Clinton'. Since the gender of 'Clinton' is unspecified, the pair might or might not be a good candidate. But if there is a coreference set already including 'Clinton', let's say: {'Hilary Clinton', her, she} then we know the gender from the other members and are more save in our decision. The virtual prototype here would be something like: singular, feminine, human.

From the set of candidates, Cand, the most salient $ante_i \in$ Cand is selected (step 10) and the coreference partition is augmented (step 11). If $ante_i$ comes from a coreference set, m_i is added to that set. Otherwise ($ante_i$ is from the buffer), a new set is formed, $\{ante_i, m_i\}$, and added to the set of coreference sets.

2.1 Restricted Accessibility of Antecedent Candidates

As already discussed, access to coreference sets is restricted to the virtual prototype - the concrete members are invisible. This reduces the number of considered pairs (from the cardinality of a set to 1).

Moreover, we also restrict the access to buffer elements: if an antecedent candidate, r_j, from a coreference set exists, then elements from the buffer, b_k, are only licensed if they are more recent than r_j. If both appear in the same sentence, the buffer element must be more salient in order to get licensed.

2.2 Filtering based on Anaphora Type

There is a number of conditions not shown in the basic algorithm from Fig. 1 that define compatibility of antecedent and anaphor candidates based on POS tags. Reflexive pronouns must be bound in the subclause they occur, more specifically to the subject governed by the same verb. Personal and possessive pronouns are licensed to bind to morphologically compatible antecedent candidates (named entities, nouns[2] and pronouns) within a window of three sentences.

We use the information given by CoNLL input data to identify 'speaker' and the person adressed by 'you'. 'I' refers to one of the coreference sets whose speaker is the person who, according to the CoNLL data, is the producer of the sentence. 'You' refers to the producer of the last sentence not being produced by the current 'speaker'. If one didn't have access to these data, it would be impossible to correctly identify the reference of 'I', since turn taking is not indicated in the pure textual data.

As we do not use machine learning, we only apply string matching techniques to match nominal NPs and leave out bridging anaphora (i.e. anaphoric nouns that are connected to their antecedents through a semantic relation such as hyponymy and cannot be identified by string matching therefore). Named entities must either match completely or the antecedent must be longer than one token and all tokens of the anaphor must be contained in the antecedent (to capture relations such

[2]To identify animacy and gender of NEs we use a list of known first names annotated with gender information. To obtain animacy information for common nouns we conduct a WordNet lookup.

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

Figure 1: Incremental Model: Base Algorithm

as 'Hillary Clinton ... Clinton'). Demonstrative NPs are mapped to nominal NPs by matching their heads. Definite NPs match with noun chunks that are longer than one token[3] and must be contained completely without the determiner (e.g. 'Recent events ... the events'). From the candidates that pass these filters the most salient one is selected as antecedent. If two or more candidates with equal salience are available, the closest one is chosen.

2.3 Binding Theory as a Filter

There is another principle that help reduce the number of candidates even further: binding theory. We know that 'He' and 'him' cannot be coreferent in the sentence 'He gave him the book'. Thus, the pair 'He'-'him' need not be considered at all. Actually, there are subtle restrictions to be captured here. We have not implemented a full-blown binding theory on top of our dependency parser, yet. Instead, we approximated binding restrictions by subclause detection. 'He' and 'him' in the example above are in the same subclause (the main clause) and are, thus, exclusive. This is true for nouns and personal pronouns, only. Possesive and reflexive pronouns are allowed to be bound in the same subclause.

2.4 An Empirically-based Salience Measure

Since we look for a simple and fast salience measure and do not apply machine learning in our baseline system, our measure is solely based on the grammatical functions (given by the dependency labels) of the true mentions. Grammatical functions have played a major role in calculating salience, especially in rule based system such as (Hobbs, 1976; Lappin and Leass, 1994; Mitkov et al., 2002; Siddharthan, 2003). Instead of manually specifying the weights for the dependency labels like (Lappin and Leass, 1994), we derived them empirically from the coreference CoNLL 2011 gold standard (training data). The salience of a dependency label, D, is estimated by the number of true mentions in the gold standard that bear D (i.e. are connected to their heads with D), divided by the total number of true mentions. The salience of the label *subject* is thus calculated by:

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

For a given dependency label, this fraction indicates how strong is the label a clue for bearing an antecedent. This way, we get a hierarchical ordering of the dependency labels (subject $>$ object $>$ pobject $>$ ...) according to which antecedents are ranked. Clearly, future work will have to establish a more elaborate calculation of salience. To our surprise, however, this salience measure performed quite well, at least together with our incremental architeture.

3 Evaluation

The results of our evaluation over the CoNLL 2011 shared task development set are given in Fig. 2 (development set) and 3 (official results on the test set).

The official overall score of our system in the open regular setting is 51.77.

Our results are mediocre. There are several rea-

[3] If we do not apply this restriction too many false positives are produced.

Metric	R	P	F1
CEAFM	49.73	49.73	49.73
CEAFE	44.26	37.70	40.72
BCUB	59.17	71.66	66.06
BLANC	62.70	72.74	64.82
MUC	42.20	49.21	45.44

Figure 2: CoNLL 2011 Development Set Results

Metric	R	P	F1
CEAFM	50.03	50.03	50.03
CEAFE	41.28	39.70	40.48
BCUB	61.70	68.61	64.97
BLANC	66.05	73.90	69.05
MUC	49.04	50.71	49.86

Figure 3: CoNLL 2011 Test Set Results

sons for that. First and foremost, the scorer requires chunk extensions to match perfectly. That is, even if the head of an antecedent is found, this does not count if the chunk extension of that noun phrase was not correctly identified. Since chunks do not play a major role in depencendy parsing, our approximation might be faulty[4]. Another shortcomming are nominal anaphora that can not be identified by string matching (e.g. Obama ... The president). Our simple salience-based approach does not cope at all with this type of anaphora.

4 Related Work

(Ng, 2010) discusses the entity-mention model which operates on emerging coreference sets to create features describing the relation of an anaphor candidate and established coreference sets. (Luo et al., 2004) implemented such a model but it performed worse than the mention-pair model. (Yang et al., 2004) presented an incremental model which used some coreference set specific features, namely introducing the number of mentions in a set as a feature besides checking for morphological compatibility with all mentions in a set. They also report that the set size feature only marginally improves or in some combinations even worsens system performance. (Daume III and Marcu, 2005) introduced a wide range of set specific features, capturing set count, size and distribution amongst others, in a joint model for the ACE data.

All the above mentioned systems use an incremental model to generate features describing the emerging coreference sets and the anaphor candidate. In contrast, we use an incremental architecture to control pair generation in order to prevent generation of either redundant or irrelevant pairs.

5 Conclusions

We have introduced an incremental model for coreference resolution based on an empirically derived salience measure that is meant as a simple and very fast baseline system. We do not use machine learning, nor do we resolve more complex nominal anaphora such as 'Obama ... The president' (but we handle those that can be resolved by simple pattern matching, e.g. Hilary Clinton .. Clinton). Given these restrictions, our system performed well.

The central idea of our approach is that the evolving coreference sets should restrict the access to antecedent candidates in a twofold way: by use of virtual prototypes that accumulate the properties of all members of a coreference set (e.g. wrt. animacy), but also by restricting reachable buffer elements (i.e. yet unattached markables).

The benefits of our incremental model are:

- due to the restricted access to antecedent candidates, the number of generated candidate pairs can be reduced drastically[5]
- no coreference clustering phase is needed
- the problem of underspecification that exists for any pair-wise model can be compensated by a virtual prototype that accumulates the properties of the elements of a coreference set

These benefits are independent of the underlying classification scheme, be it a simple salience-based one or a more advanced machine learning one. The work presented here thus would like to opt for further research based on incremental architectures. Web demos for English and German are available[6].

[4]Especially Asiatic names pose problems to our parser, quite often the extensions could not get correctly fixed.

[5]We observed a reduction over 75% in some experiments when moving from a mention-pair to an incremental entity-mention model.

[6]http://kitt.cl.uzh.ch/kitt/coref/

References

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGdial 2010 Conference: The 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 81–88, Rochester, New York, April. Association for Computational Linguistics.

Hal Daume III and Daniel Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 97–104, Morristown, NJ, USA. Association for Computational Linguistics.

Jerry R. Hobbs. 1976. Pronoun resolution. Technical Report 76-1, Research Report, Department of Computer Sciences, City College, City University of New York.

Manfred Klenner and Etienne Ailloud. 2009. Optimization in Coreference Resolution Is Not Needed: A Nearly-Optimal Zero-One ILP Algorithm with Intensional Constraints. In *Proc. of the EACL*.

Manfred Klenner. 2007. Enforcing consistency on coreference sets. In *In Recent Advances in Natural Language Processing (RANLP)*, pages 323–328.

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

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*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

Ruslan Mitkov, Richard Evans, and Constantin Orasan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 168–186, London, UK. Springer-Verlag.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Stroudsburg, PA, USA. 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.

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

Advaith Siddharthan. 2003. Resolving pronouns robustly: Plumbing the depths of shallowness. In *Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*.

Wee M. Soon, Hwee T. Ng, and Daniel. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, December.

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*.