An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility

Klenner, M ; Tuggener, D

An Incremental Entity-Mention Model for Coreference Resolution with Restrictive Antecedent Accessibility

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

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

Abstract

We introduce an incremental entity-mention model for coreference resolution. Our experiments show that it is superior to a non-incremental version in the same environment. The benefits of an incremental architecture are: a 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. Additionally, we have defined a simple salience measure that - coupled with the incremental model - proved to establish a challenging baseline which seems to be on par with machine learning based systems of the 2010's SemEval shared task.

1 Introduction

With notable exceptions (Luo et al., 2004; Yang et al., 2004; Daume III and Marcu, 2005; Culotta et al., 2007; Rahman and Ng, 2009; Cai and Strube, 2010; Raghunathan et al., 2010) supervised approaches to coreference resolution are often realised 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
- Skewed training sets based on pair generation mechanics which lead to classifiers biased towards negative classification
- No means to enforce global constraints such as transitivity
- Underspecification of antecedent candidates

Mention-pair systems operate in a non-incremental mode, i.e. all pairs are classified prior to the construction of the coreference sets. A clustering step is needed where, additionally, inconsistencies (e.g. transitively incompatible pairs) can be removed. This often is realised as an optimisation step, where scores derived from pairwise classification are used as weights in a decision taking process that incorporates linguistic constraints, e.g. (Finkel and Manning, 2008). Although this overcomes the limitations of the strictly local perspective of pairwise classifiers, it still suffers from the problem of unbalanced data (much more negative than positive examples are generated). The large number of candidate pairs, in general, is a problem, e.g. (Wunsch et al., 2009).

These problems can be remedied by an incremental entity-mention model, where candidate pairs are evaluated on the basis of emerging coreference sets. The amount of candidate pairs is reduced, since only one (virtual prototype) example of each coreference set needs to be compared to a new anaphor candidate[1]. Moreover, the problem of inconsistent decisions vanishes, since the virtual prototype of a coreference set bears all the known morphological and semantic information of the elements of the set. If an anaphor candidate is compatible with the prototype then it is compatible with each member of the coreference set. A clustering phase on top of the pairwise classifier no longer is needed.

[1] We are aware of the fact that, linguistically speaking, anaphoric expressions depend on previously mentioned entities (e.g. 'she' → 'Clinton'), whereas coreferent expressions do not always (e.g. 'Hillary Clinton' ... 'United States Secretary of State'). We use the terms 'anaphoric' and 'anaphora' to subsume both relations.

We have compared our incremental entity-mention model to a non-incremental mention-pair version. The memory-based learner TiMBL (Daelemans et al., 2007) was used for pairwise classification. To define a simple baseline, we adopted previous work on salience-based models for coreference resolution. It turns out that our salience measure coupled with the incremental model performs quite well, e.g. it outperformes the systems from the 2010's SemEval shared task on 'coreference resolution in multiple languages' in our own post-task evaluation.

Our system uses real preprocessing (i.e. the use of a parser (Schneider, 2008; Sennrich et al., 2009)) and extracts markables (nouns, named entities and pronouns) from the chunks based on POS tags delivered by the preprocessing pipeline.

We first introduce the incremental model, present constraints on buffer list access, discuss our filtering system and our approximation of the binding theory. We then turn to our simple salience measure initially used as a baseline. In the empirical section, the impact of the incremental entity-mention model on the number of candidate pairs is quantified and a comparison of the variants (incremental, non-incremental etc.) of our German system on the TüBa-D/Z (Naumann, 2006) is given. We also describe our post-task evaluation with the 2010's SemEval data, the results from the BioNLP shared task on coreference resolution in the biomedical domain and our results on the CoNLL 2011 shared task development set.

2 Our Incremental Entity-mention Model

Fig. 1 shows the base algorithm. Let I be the chronologically ordered list of markables, C the set of coreference sets (i.e. the coreference partition) and B a buffer where markables are stored, if they are not anaphoric (but might be valid antecedents). 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 (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 (in order to be coreferent, e.g. two pronouns must agree in person, number and gender, while two nouns, at least in German, need not necessarily agree in gender).

If an antecedent candidate is already in a coreference set (r_j), m_i is compared to the virtual prototype of the set in order to reduce underspecification. The virtual prototype bears information accumulated from all elements of the coreference set. For instance, assume a candidate pair 'Clinton ... she'. Since the gender of 'Clinton' is unspecified, the pair might or might not be a good candidate. But if 'Clinton' is part of a coreference set, let's say: {'Hillary Clinton', 'she', 'her', 'Clinton'} then we can derive the gender from the other members and are more safe in our decision. The virtual prototype here would be: singular, feminine, human.

In languages such as German, where morphological information is much more discriminatory than in English and where at the same time underspecification appears quite often (e.g. the reflexive pronoun 'sich' might refer to any third person noun phrase, be it singular or plural, masculine, feminine or neutral), this is particularly helpful.

If no compatible antecedent candidates are found, m_i is added to the buffer (Step 8). If there are compatible candidates in the candidate list Cand, the most salient $ante_i \in$ Cand (or, in the machine learning setting, the most probable) 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 restrict 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.

Although this rule is heuristic and no evaluation of the impact of different versions of such a 'discourse model' have been carried out yet, we believe that 'accessibility' of antecedent candidates along these lines is a fruitful notion. It might

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

lead to cognitively adequate models for coreference resolution, where cognitive burden determines which antecedent candidates are valid at all. Clearly, future work must start with an evaluation of our current setting.

2.2 Filtering and Training Based on Anaphora Type

There is a number of conditions not shown in the basic algorithm in Fig. 1 that define compatibility of antecedent and anaphor candidates based on POS tags: Reflexive pronouns must be bound to the subject governed by the same verb. Relative pronouns are bound to the next NP in the left context. 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. 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 (e.g. 'Hillary Clinton' ... 'Clinton'). Demonstrative NPs are mapped to nominal NPs by matching their heads (e.g. 'The recent findings' ... 'these findings'). 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'). To licence non-matching (bridging) nominal anaphora, we apply hyponymy and synonymy searches in WordNet (Fellbaum, 1998) and GermaNet (Hamp and Feldweg, 1997) respectively.

For the machine learning approaches we used the standard features of mention-pair models (e.g. (Soon et al., 2001)). We trained individual classifiers per anaphora type, i.e. for nominal anaphora, reflexive, possessive, relative and personal pronouns. We manually tuned the feature selection of each classifier. Both the mention-pair and the entity-mention model share these features and filters.

2.3 Binding Theory as a Filter

There is another principle that nicely combines with our incremental model and helps reducing the number of candidates even further: binding theory (e.g. (Büring, 2005)). We know that 'Clinton' and 'her' cannot be coreferent in the sentence 'Clinton met her'. Thus, the pair 'Clinton'-'her' need not be considered at all. Furthermore, all mentions of the 'Clinton' coreference set, say {'Hillary Clinton', she, her, 'Clinton'}, are transitively exclusive and can be discarded as antecedent candidates.

Actually, there are subtle restrictions to be captured here. We have not implemented a full-blown binding theory on top of our dependency parsers. Instead, we approximated binding restrictions by subclause detection. 'Clinton' and 'her' are in the same subclause (the main clause) and are, thus, exclusive. This is true for nouns and personal pronouns, only. Possessive and reflexive pronouns are allowed to be bound in the same subclause.

2.4 An Empirically-based Salience Measure

In the pioneer work of (Lappin and Leass, 1994), salience calculation included manually specified weights for grammatical functions (e.g. *subject* got the highest score). The distance between the candidates and other properties are

[2]To identify animacy and gender of NEs, we use a list of known first names annotated with gender information and look up Wikipedia categories to map NEs to WordNet/GermaNet synsets. To obtain animacy information for common nouns, we conduct a WordNet search.

[3]If we do not apply this restriction, too many false positives are produced - simple head matching appears to be very noisy.

also taken into account in order to determine salience. Such approaches suffered from a proper empirical justification[4]. Consequently, machine-learning approaches have replaced manually designed salience measures. Now it is the classifier that determines 'salience'.

Our salience measure is a variant of the one in (Lappin and Leass, 1994). Instead of manually specifying the weights, we derived them empirically on the basis of the coreference gold standard (for German, this is the coreference annotated treebank TüBa-D/Z ; for English, OntoNotes[5] was used). 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 a true mention. We get a hierarchical ordering of the dependency labels (*subject* > *object* > *pobject* ...) according to which antecedent candidates are ranked.

Clearly, future work will have to establish a more elaborate calculation of salience to be used for classification without machine learning. To our surprise, however, this salience measure performed quite well together with our incremental architecture.

3 Evaluation

We evaluate our system in two languages (German and English) and in two domains (newswire text and abstracts from the biomedical domain). We directly compare our incremental entity mention model to the generative mention-pair model on the basis of the German TüBa-D/Z corpus in a 5-fold cross-validation. We also investigate the competitiveness of the incremental model compared to other systems in two tasks and languages: SemEval[6] (English and German) and BioNLP[7] (English). Results of the CoNLL 2011[8] shared task development data (English) are also provided.

[4]There are notable exceptions, e.g. (Ge et al., 1998), where salience calculation is combined with statistics.

[5]http://www.bbn.com/ontonotes/

[6]http://stel.ub.edu/semeval2010-coref/

[7]https://sites.google.com/site/bionlpst/home/protein-gene-coreference-task/

[8]http://conll.bbn.com/

3.1 Reducing the Number of Candidate Pairs

Anaphora Type	Pos	Neg
Mention-pair model (**171526 instances**)		
Nouns	5626	5144
Relative pronouns	1428	2459
Reflexive pronouns	1372	728
Possessive pronouns	**5346**	**21571**
Personal pronouns	**23025**	**104827**
Total	36797	134729
Entity-mention model (**40229 instances**)		
Nouns	1776	3787
Relative pronouns	1382	2330
Reflexive pronouns	462	530
Possessive pronouns	**1416**	**8156**
Personal pronouns	**4023**	**16367**
Total	9059	31170

Figure 2: Number of training instances per anaphora type of Fold 1 of the TüBa-D/Z

Fig. 2 shows the number of training instances of the first fold (about 5'000 sentences) from the TüBa-D/Z both for the incremental and the non-incremental algorithm. Overall a huge reduction by a factor of 4 (-131297 instances, -76.55 %) can be observed when moving from the non-incremental mention-pair to the incremental entity-mention model. As we use the same filter set in all runs, no true mentions are deleted in the incremental approach. The reduction in positives results from pairing an anaphor candidate with only one virtual prototype of the coreference set it belongs to as opposed to redundantly pairing it with all members of its set. As during testing only pairs consisting of the set's virtual prototype and the anaphor candidate are considered, this is sufficient and the additional pairs are not needed. The reduction in negatives results from the same mechanism. Instead of pairing the anaphor with all mentions of a set it does not belong to, only one negative pair with the prototype is generated. Additionally, some pairs are created with compatible members from the buffer list.

The reason for the relatively minor reduction in reflexive and relative pronouns is that the search for antecedents is limited to the same sentence or even a specific (sub-) clause. On the other hand, we allow for possessive and personal pronouns a window of three sentences wherein antecedent candidates may be found. In the latter two cases, the incremental approach to pair generation has a more drastic impact on the number of training instances (-64.44%, -84.05% resp.).

3.2 TüBa-D/Z Model Comparison

We can see from the results (Fig. 3) that the incremental entity-mention model outperforms the mention-pair model. The entity-mention model with the TiMBL classifier performed best by improving recall (+ 7.01%) and losing some precision (- 0.79%) compared to the mention-pair model. To our surprise, the simple salience approach performed quite well, losing only 0.85% precision and 1.88% recall compared to its machine learning variant. Given that bridging anaphora is not resolved in the salience mode, a reduction in recall was to be expected. It still outperforms the mention-pair model that implements machine learning.

Model	F1	P	R
Mention-pair (TiMBL + ILP)	49.35	53.67	45.69
Entity-mention (TiMBL)	52.79	52.88	52.70
Entity-mention (salience)	51.41	52.03	50.82

Figure 3: CEAF scores of the 5-fold TüBa-D/Z cross-validation

Overall the results of the TüBa-D/Z evaluation are low, indicating that end-to-end coreference resolution with real preprocessing is still a difficult problem. It is important to note that we implemented a version of the CEAF metric which does not account for singletons (i.e. coreference sets with only one mention) because we believe that finding singletons is not a crucial part of the coreference resolution task and that it improves results artificially. We can see the difference of evaluating with or without singletons if we compare these results with the ones from SemEval (Fig. 5), where singletons are considered in the evaluation process. The SemEval German task also uses data from the TüBa-D/Z , allowing an approximate comparison of the results to illustrate the effects of considering singletons in evaluation. The CEAF F1-measure of our incremental model reaches 76.8% on the SemEval data (Fig. 5), while without singletons, we reach 52.79% in the TüBa-D/Z evaluation (Fig. 3).

3.3 Error Analysis

We simulated perfect resolution of the individual classifiers of the best performing system (Entity-mention(TiMBL)) from the model comparison (Fig. 4). We ran the system on the first fold (ca. 5000 sentences) of the TüBa-D/Z , resolving one type of anaphora (e.g. nominal anaphora) using gold standard information per run, while the other anaphora types were resolved by the system.This gives us an indication of the upper bounds of the system: How good would our system be, if it resolved e.g. nominal anaphora perfectly?

with filters means that only pairs that pass the filters are resolved. In the *without filters* mode, all pairs of the corresponding anaphora type are resolved correctly, disregarding filter decisions. The other anaphora types are resolved by the system in both modes. The difference in performance between the *with* and *without filtering* mode indicates how good our filters are: the smaller the difference, the better the filters (compare values horizontally). The performance difference of the individual classifiers with perfect resolution compared to the overall system performance (right column, compare vertically) indicates the difficulty of resolving that anaphora type.

For example, in the first row that indicates resolution performances of nominal anaphora we can see that we roughly lose 10% in F1 measure due to our nominal filters (72.70% - 62.61%). Compared to the actual system performance in the last row in the right column (53.86%) we see that we lose an additional 9% in F1 measure because of imperfect resolution of nominal anaphora (62.61% - 53.86%). This sums up to a total loss of 19% in F1 measure compared to system performance with perfect resolution of nominal anaphora. Compared to the minor difference of 1.8% F1 measure between perfect and imperfect resolution of reflexive pronouns (-1.5% through filtering and -0.3% through imperfect classification) the difficulty of resolving nominal anaphora becomes obvious.

3.4 SemEval 2010, BioNLP 2011 and CoNLL 2011

To get an indication of the competitiveness of our incremental approach we carried out evaluations over recent shared task data sets. The SemEval coreference task (Recasens et al., 2010) focused on coreference resolution in multiple languages and comparing different evaluation metrics. The test data for German was composed of the TüBa-D/Z whereas the English data was gathered from the OntoNotes corpus.

The main goal of the BioNLP protein/gene coreference task was to resolve non-name-containing mentions in protein/gene-interactions to their appropriate name-containing antecedents

	Without filtering			With filtering		
	F1	Precision	Recall	F1	Precision	Recall
Nouns	**72.70**	69.53	76.17	**62.61**	63.70	61.55
Personal pronouns	60.42	62.05	58.88	58.86	60.64	57.19
Relative pronouns	56.25	57.91	54.68	55.97	57.65	54.39
Possessive pronouns	56.06	57.35	54.82	55.81	57.18	54.51
Reflexive pronouns	55.68	57.11	54.32	54.16	55.64	52.77
System	-	-	-	**53.86**	54.64	53.09

Figure 4: CEAF scores for the simulation of perfect classification (upper bounds) of the individual classifiers for the first 5000 sentences of the TüBa-D/Z .

and thereby improving overall recall of interaction extraction (i.e. the main task). The test data consists of abstracts gathered from PubMed.

As the SemEval training data for English and German were not available at the time of our post-task experiments, we were only able to evaluate the salience based classification.

The SemEval coreference task offers many different settings. Since we are interested in real end-to-end coreference resolution we evaluated the *open/regular* setting, meaning that real preprocessing components are used as opposed to perfect gold standard preprocessing data. Results of the SemEval task are given in Figure 5.

Except for the (recently questioned, e.g. (Luo, 2005; Cai and Strube, 2010)) MUC metric in the English evaluation, the incremental model (incr) achieved best results throughout the SemEval experiments in both languages. All other systems that competed in the task implemented a mention-pair model. Overall, an improvement can be observed compared to the other systems, mainly in precision.

The simple salience based measure is not suited for resolving bridging anaphora. Therefore, bridging anaphora was not resolved by the system in these experiments (but still included in the evaluation) which might be a reason for the relatively low recall.

More recently, we have adapted our salience-based incremental architecture to the biomedical domain. Our results in the recent BioNLP 2011 shared task are competitive as well (see Fig. 6).

The results of our evaluation over the CoNLL 2011 shared task development set are given in Fig. 7. CEAF and BCUB scores are considerably lower compared to the SemEval results. We believe these differences originate from the updated scoring algorithms for CEAF and BCUB. They were modified for the CoNLL scorer according to suggestions by (Cai and Strube, 2010). The CoNLL scorer has stricter mention boundary handling than the SemEval scorer. Moreover, singletons were not marked in the CoNLL data.

Team	R	P	F1
A	22.18	73.26	34.05
incr	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 6: BioNLP 2011 Protein/Gene Coreference Task Results

Metric	R	P	F1
CEAFM	51.08	51.08	51.08
CEAFE	44.35	39.93	42.03
BCUB	60.91	70.69	65.44
BLANC	63.63	72.58	66.81
MUC	45.18	49.83	47.39

Figure 7: CoNLL 2011 Development Set Results

4 Related Work

The work of (Soon et al., 2001) is a prototypical and often re-implemented (baseline) model that is based on pairwise classification and machine learning. Our non-incremental mention-pair model can be seen as an adaption of this system and its features. Coreference clustering is discussed e.g. in (Denis and Baldridge, 2009; Finkel and Manning, 2008). Our mention-pair model uses the Balas algorithm for clustering as discussed in (Klenner, 2007).

Direct empirical comparison of supervised mention-pair and entity-mention models can be found in e.g. (Luo et al., 2004; Yang et al., 2004; Rahman and Ng, 2009). Only in (Rahman and Ng, 2009) a clear improvement by the entity-mention model is observed. Other supervised entity-mention models such as (Daume III and Marcu, 2005; Culotta et al., 2007; Raghunathan et al., 2010) are not directly compared to

	CEAF			MUC			BCUB			BLANC		
System	R	P	F1	R	P	F1	R	P	F1	R	P	F1
German, open regular												
bart	61.4	61.2	61.3	**61.4**	36.1	45.5	75.3	58.3	65.7	**55.9**	60.3	57.3
incr	**76.8**	**70.4**	**73.4**	50.4	**47.1**	**48.7**	**81.7**	**75.6**	**78.5**	55	**72.6**	**57.8**
English, open regular												
bart	70.1	64.3	67.1	**62.8**	52.4	57.1	74.9	67.7	71.1	55.3	73.2	57.7
corry-b	70.4	67.4	68.9	55.0	54.2	54.6	73.7	74.1	73.9	57.1	75.7	60.6
corry-c	**70.9**	67.9	69.4	54.7	55.5	55.1	73.8	73.1	73.5	57.4	63.8	59.4
corry-m	66.3	63.5	64.8	61.5	53.4	**57.2**	**76.8**	66.5	71.3	**58.5**	56.2	57.1
incr	67.6	**73**	**70.2**	34	**62.5**	44.1	66.7	**86**	**75.1**	57.1	**78.4**	**61.1**

Figure 5: Our SemEval 2010 post-task evaluation results

mention-pair models. Also, in the recent SemEval 2010 and BioNLP 2011 shared tasks no entity-mention models participated.

Our work differs from the research mentioned above as it focuses on using an incremental entity-mention architecture to impose constraints on candidate pair generation as opposed to generating cluster-level features for (machine learning-based) classification. Our hypothesis, also for future work, is that progress is possible by not only improving classifier performance but by improving other steps of the coreference resolution pipeline that lead up to the classifier, namely pair generation and antecedent candidate accessibility.

5 Conclusions

We have introduced an incremental entity-mention algorithm for coreference resolution and evaluated its impact on pair generation and the performance of architectural variants. A performance comparison of our model to systems from different shared tasks produced good results. We also discussed a simple and very fast salience-based approach that performed quite well, i.e. it outperformed all systems of the 2010's SemEval shared task.

The benefits of an incremental model are:

- due to the restricted access to potential antecedent candidates, the number of generated candidate pairs can be reduced drastically
- no additional coreference clustering is necessary
- global constraints (e.g. transitivity) are easily integrated
- underspecification of antecedent candidates can often be compensated by other members of the emerging coreference sets

Our theory on how to restrict the accessibility of antecedent candidates has proven to be (empirically) successful, as it outperformed other systems. However, we are aware of the fact that we need to explore in a more principled and empirically grounded way, what the parameters of such an evolving discourse model are. We strive for a theory whose decisions, in the best case, relate to the restrictions of human cognitive capacity.

Finally, our implementation of a binding theory is incomplete. Since binding theory provides hard restrictions, it is a crucial component of any theory on antecedent accessibility.

Web demos of the salience based system for English and German are available[9].

Acknowledgements. Our project is funded by the Swiss National Science Foundation (grant 105211-118108). We are grateful to OntoGene[10] for their help and advice regarding the BioNLP shared task.

References

Daniel Büring. 2005. *Binding Theory*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *HLT '07: The Conference of the North American Chapter of ACL; Proceedings of the Main Conference*, pages 81–88, Rochester, New York, April. Association for Computational Linguistics.

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

[10] http://www.ontogene.org/

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2007. Timbl: Tilburg memory-based learner. Technical report, Induction of Linguistic Knowledge, Tilburg University and CNTS Research Group, University of Antwerp.

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.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. In *Procesamiento del Lenguaje Natural 42*, pages 87–96, Barcelona: SEPLN.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Jenny Rose Finkel and Christopher D. Manning. 2008. Enforcing transitivity in coreference resolution. In *HLT '08: Proceedings of the 46th Annual Meeting of the ACL on Human Language Technologies*, pages 45–48, Morristown, NJ, USA. Association for Computational Linguistics.

Niye Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–171, Montreal, Canada.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Somerset, NJ, USA. Association for Computational Linguistics.

Manfred Klenner. 2007. Enforcing consistency on coreference sets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 323–328, Borovets, Bulgaria.

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 of ACL*, ACL '04, Stroudsburg, PA, USA. 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.

Karin Naumann, 2006. *Manual for the Annotation of Indocument Referential Relations*. SFS (Seminar für Sprachwissenschaft), http://www.sfs.uni-tuebingen.de/tuebadz.shtml.

Vincent Ng. 2010. Supervised noun phrase coreference research: the first fifteen years. In *Proceedings of the 48th Annual Meeting of ACL*, ACL '10, pages 1396–1411, Stroudsburg, PA, USA. 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.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, 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.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In *Proc. of the German Society for Computational Linguistics and Language Technology 2009 (GSCL 2009)*, pages 115–124, Potsdam, Germany.

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.

Holger Wunsch, Sandra Kübler, and Rachael Cantrell. 2009. Instance sampling methods for pronoun resolution. In *Proceedings of Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

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*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.