Singletons and Coreference Resolution Evaluation

Sandra Kübler
Indiana University
skuebler@indiana.edu

Desislava Zhekova
University of Bremen
zhekova@uni-bremen.de

Abstract

This paper presents an empirical study on the influence of singletons on the evaluation of coreference resolution systems. We present results on two English data sets used in the SEMEVAL 2010 shared task 1 and the CONLL 2011 shared task using the scorers of both shared tasks. We show that singletons, both in the gold standard and in the system output, have an immense impact on the overall evaluation – in an experiment where the coreference resolution results remain unchanged over the different settings.

1 Introduction

In the last decade, the task of Coreference Resolution has become an important enterprise in Natural Language Processing. At the same time, the need for proper benchmarking increased over time. In the last year, two major shared tasks were concerned with coreference resolution: the SEMEVAL 2010 task 1 "Coreference Resolution in Multiple Languages" (Recasens et al., 2010) and the CONLL shared task 2011 "Modeling Unrestricted Coreference in OntoNotes" (Pradhan et al., 2011). Both shared tasks introduced a new element into the definition of coreference resolution: The detection of mentions. Previous to these shared tasks, the availability of gold standard mentions was often assumed, and research concentrated on the resolution of coreference relationships between mentions. (e.g. (Luo et al., 2004; Denis and Baldridge, 2007)).

However, in many approaches to coreference resolution, the problem is even more restricted, and the coreference resolution component expects only such mentions that are coreferent in the present context, i.e. no singletons are present in the data. "Singleton" is a cover term for mentions that are never coreferent, such as in `in general` or `on the contrary`, and mentions that are potentially coreferent but occur only once in a document. If the extraction of mentions is part of the task definition, then filtering singletons is generally necessary since methods for mention identification often overgenerate and produce all noun phrases (NPs), including all singletons. *Twinless mentions* (Stoyanov et al., 2009) are mentions that have been identified by a coreference resolution system but are not included in the gold data, or vice versa. Twinless mentions can lead to considerable changes in overall system performance, and Stoyanov et al. (2009) report that at that time B^3 was not prepared to handle them. For the CONLL shared task, the metrics were updated to obtain "better alignment for B^3 and CEAF so that the gold standard set and the system output have the same number of mentions" (p.c. S. Pradhan). In this paper, we investigate how the presence of singletons in either gold standard or in the system output influences the results. We compare the English data sets of the SEMEVAL and the CONLL shared task and the two versions of the scorer used there.

A simple solution was chosen by Rahman and Ng (2011), who remove twinless mentions that the coreference resolution system identifies as singletons with the motivation that the system should be rewarded for identifying the mentions as a whole, and can still be punished for their incorrectly resolved coreference. Yet, this approach is only applicable when the gold standard answers are available for evaluation. It can be used to address shortcomings of the evaluation metrics and to gain a more

Proceedings of Recent Advances in Natural Language Processing, pages 261–267,
Hissar, Bulgaria, 12-14 September 2011.

	CEAF			MUC			B^3			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
SINGLETONS	71.2	71.2	71.2	0.0	0.0	0.0	71.2	100	83.2	50.0	49.2	49.6
ALL-IN-ONE	10.5	10.5	10.5	100	29.2	45.2	100	3.5	6.7	50.0	0.8	1.6

Table 1: Baseline scores for the English data set in the SEMEVAL task 1.

	IM			MUC			B^3			$CEAF_E$			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
SEMEVAL scorer															
ABC, DE	100	100	100	66.7	66.7	66.7	73.3	73.3	73.3	80.0	80.0	80.0	58.3	58.3	58.3
ABC, DE, Y	100	83.3	90.9	66.7	66.7	66.7	73.3	61.1	66.7	80.0	53.3	64.0	51.8	52.3	49.8
ABC, DE, X	83.3	100	90.9	66.7	66.7	66.7	61.1	73.3	66.6	53.3	80.0	64.0	58.3	58.3	58.3
CONLL scorer															
ABC, DE	100	100	100	66.7	66.7	66.7	73.3	73.3	73.3	80.0	80.0	80.0	58.3	58.3	58.3
ABC, DE, Y	100	83.3	90.9	66.7	66.7	66.7	73.3	73.3	73.3	80.0	80.0	80.0	58.3	58.3	58.3
ABC, DE, X	100	100	100	66.7	66.7	66.7	77.8	77.8	77.8	86.7	86.7	86.7	65.9	65.9	65.9

Table 2: Coreference scores on an artificial example.

objective overview of the system coreference performance. But it is not possible in a real world system.

The remainder of the paper is structured as follows: Section 2 discusses coreference evaluation metrics and their behavior in the presence of singletons. Section 3 describes the English data sets from the shared tasks, which we use for our investigation, and section 4 gives a short description of the coreference resolution system that we use. In section 5, we investigate the influence of singletons in the gold standard and system sets for both data sets, and in section 6, we investigate how the presence and treatment of pronoun singletons influences scoring results on the CONLL data set.

2 Coreference Evaluation

Apart from the open research question how to distinguish singletons from coreferent mentions, there is the question how the standard evaluation metrics, MUC (Vilain et al., 1995), B^3(Bagga and Baldwin, 1998), CEAF (Luo, 2005), and BLANC (Recasens and Hovy, 2011), react to the presence of singletons in the data. Recasens et al. (2010) present two baselines, one in which every mention in the data set is considered a singleton, and one in which all mentions are grouped into one chain. The singleton baseline reaches high scores for the metrics CEAF and B^3, with an overall performance of above 70% for English. The MUC metric, on the other hand, is not at all sensitive to the existence of singleton mentions. Yet, for the second baseline, in which all mentions were linked to one single entity, the MUC metric reported the highest results. Table 1 shows the results for both baselines.

Let us consider a small artificial example, in which the gold standard contains two coreference chains, `A-B-C` and `D-E` and the system erroneously attached `A` to the chain `D-E`. Then, we introduce one singleton in the gold standard, `X` and one in the system output `Y`. Since, the metrics in the CONLL shared task were modified to handle singletons (cf. section 1) we use both versions of the scorer, the SEMEVAL scorer and the CONLL scorer. The results are presented in table 2. This example shows that with the SEMEVAL scorer, all metrics but MUC, are sensitive to singletons in the system output and in the gold standard data. However, the presence of a singleton (Y) in the system output leads to a decrease in the results while an additional singleton (X) in the gold standard increases results although the system output is unchanged. With the CONLL scorer, all metrics are insensitive to singletons in the system output. An additional singleton in the gold standard still increases scores for B^3, $CEAF_E$ (mention-based CEAF), and BLANC. Overall, this scorer leads to higher system results.

However, the above example is a small, artificial example. It remains unclear how the results change in real world situations in which a large number of coreference chains provide the grounds for many types of errors. For this reason, we empirically investigate the influence of singletons on the English

```
1 0 By IN (TOP(S(PP* - - - Speaker#1 * (ARGM-TMP* (ARGM-TMP*   -     -
1 1 1940 CD (NP*)) - - - Speaker#1 (DATE) *) *)               (29)  (1)
1 2 , , * - - - Speaker#1 * * *                                -     -
1 3 China NNP (NP(NP(NP* - - - Speaker#1 (GPE) (ARG0* (ARG0*  (31   (2)|(3|(4|(5
1 4 's POS *) - - - Speaker#1 * * *                            31)  5)
1 5 War NNP *) - - - Speaker#1 (EVENT) * *                     -    (6)|4)
1 6 of IN (PP* - - - Speaker#1 * * *                           -     -
1 7 Resistance NNP (NP(NP*) - - - Speaker#1 (ORG) * *          -    (7)|(8|(9)
1 8 against IN (PP* - - - Speaker#1 * * *                      -     -
1 9 Japan NNP (NP*))))) - - - Speaker#1 (GPE) *) *)           (72)  (10)|(11)|8)|3)
1 10 had VBD (VP* have 03 - Speaker#1 * (V*) *                 -    (12)
1 11 entered VBN (VP* enter 01 1 Speaker#1 * * (V*)            -    (13)
1 12 a DT (NP* - - - Speaker#1 * * (ARG1*                      -    (14
1 13 stalemate NN *))) - - - Speaker#1 * * *)                  -    14)
1 14 . . *)) - - - Speaker#1 * * *                             -     -
```

Table 3: An example sentence from the CoNLL shared task data set.

data sets of the SemEval and the CoNLL shared task. We investigate different strategies of handling singletons, and their influence on results of a robust coreference resolution system, UBIU.

3 The Shared Task English Data Sets

Both shared tasks for coreference resolution in the last year, the SemEval 2010 task 1 (Recasens et al., 2010) and the CoNLL shared task 2011 (Pradhan et al., 2011), included an English data set, based on OntoNotes (Hovy et al., 2006). However, both data sets differ in the texts selected for their assembly as well as in the annotations on the gold standard. We discuss these differences below.

3.1 The SemEval English Data Set

The SemEval task 1 (Recasens et al., 2010) aimed at the evaluation and comparison of coreference resolution systems in a multilingual environment targeting six languages (Catalan, Dutch, English, German, Italian, Spanish). The main focus of the task was on system portability across different languages and the importance of various linguistic annotations for the system performance for all languages.

All data sets contained linguistic annotation at the morphological, syntactic, and semantic levels, including both gold standard and automatic annotations. The task description defined that only NP constituents and possessive pronouns were considered mentions; nominal predicates, appositives, expletive NPs, attributive NPs, and NPs within idioms were not considered mentions. The task description also specified that singletons were included in the data annotations since they represent coreference chains containing a single mention.

3.2 The CoNLL 2011 Shared Task Data Set

The CoNLL 2011 shared task (Pradhan et al., 2011) was defined as modeling unrestricted coreference. This shared task focused on English as its only language, and it also used the OntoNotes corpus as its basis. The task definition specifies that names, nominal mentions, and pronouns are considered mentions. Additionally, verbs that are coreferent with a noun phrase are marked as mentions. Singletons are not considered mentions. The annotation in the data set included POS tags, syntactic information, semantic role labeling, and WordNet information and corpus-based number and gender information.

Table 3 shows an example sentence from the CoNLL shared task data set with automatic annotations. Here, mention (72), `Japan` is coreferent with the mention `the enemy's` in the following sentence. Since in contrast to the SemEval data set, singletons are not annotated as mentions, noun phrases such as `China's War of Resistance` are not annotated as mentions. The last column in the example is not from the data set but is generated by UBIU (see below).

4 UBIU

UBIU (Zhekova and Kübler, 2010) was developed as a multilingual coreference resolution system. For such a task, a robust approach is necessary to make the system applicable for a variety of languages. Pronoun resolution results for German show that a mention pair model gives higher results than more complex architectures (Wunsch, 2009), thus we

use a mention-pair approach, in combination with TiMBL (Daelemans et al., 2007), a memory-based learner that labels the feature vectors from the test set based on the k nearest neighbors in the training data. Based on a non-exhaustive parameter optimization on the development set, we use the *IB1* algorithm, weighted overlap as similarity metric, and gain ratio for weighting. The number of nearest neighbors is $k = 3$. The classifier is preceded by a mention extractor, which identifies possible mentions, and a feature extractor to gather the information required for classification in the form of vector features.

The mention extractor uses POS, syntactic, and lemma information that was provided in the CoNLL data set. An example of its output for the example sentence is given in the last column of table 3. Syntactic information is used to assign a mention to each of the noun phrases existing according to that annotation. Additionally, possessive pronouns and proper nouns, which are selected based on POS information are assigned a separate mention. Since verbs can be coreferent, additional mentions are included for each verb with a predicate lemma.

The feature extractor creates a feature vector for each possible pair of a mention and all its possible antecedents in a context of 3 sentences. Since mentions are represented by their syntactic head, the module uses a heuristic to select the rightmost noun in a noun phrase. However, since postmodifying prepositional phrases may be present in the mention, the noun may not be followed by a preposition.

Initially, UBIU used a wide set of features (Zhekova and Kübler, 2010), which constitutes a subset of the features by Rahman and Ng (2009). Our experiments in the CoNLL 2011 shared task (Zhekova and Kübler, 2011) showed that adding additional information, such as WordNet or number/gender information, does not improve performance for our system when applied on the CoNLL data set. For this reason, we use the basic feature set shown in table 4.

Another important step is to separate singleton mentions from coreferent ones since only the latter are annotated in OntoNotes. Our mention extractor overgenerates in that it extracts all possible mentions, and only after classification, the system can decide which mentions are singletons.

#	Feature Description
1	m_j - the antecedent
2	m_k - the mention to be resolved
3	Y if m_j is a pronoun; else N
4	number - S(ingular) or P(lural)
5	Y if m_k is a pronoun; else N
6	C if the m. are the same string; else I
7	C if one m. is a substring of the other; else I
8	C if both m. are pronominal and the same string; else I
9	C if the m. are non-pronominal and the same string; else I
10	C if m. are pronominal and either the same pronoun or differ only w.r.t. case; NA if at least one is not pronominal; else I
11	C if the m. agree in number; I if they disagree; NA if the number for one or both mentions cannot be determined
12	C if both m. are pronouns; I if neither are pronouns; else NA
13	C if both m. are Prop. N.; I if neither are Prop. N; else NA
14	sentence distance between the mentions

Table 4: The pool of features used in the base feature set.

5 Singletons in the SemEval and CoNLL Data Sets

In this section, we investigate the influence of singletons on the evaluation of UBIU. Since the system's coreference resolution performs below the state of the art systems, we assume that a wide range of errors will be present in the system output. We compare the system performance based on the data sets from the shared tasks, and we evaluate the system output with the two versions of the scorer from the shared tasks. For both data sets, we train UBIU on the training data. For the SemEval data, we test on the test set, for the CoNLL set, we use the development set since the gold standard annotation for the test set is not available yet. Overall, we have four different settings for the experiment w.r.t. singletons:

1. G+S/S+S: Singletons are included in the gold standard (i.e. training and test data) and in the system output.
2. G+S/S-S: Singletons are included in the gold standard but are removed in the system output.
3. G-S/S+S: Singletons are removed from the gold standard but not from the system output.
4. G-S/S-S: Singletons are removed from the gold standard and from the system output.

The coreference resolution information in the system data remains the same over all settings, the only changes made to the data sets concern the singletons. Since the CoNLL data set does not include singletons, we can only evaluate the last two settings for this data set. The results of these evaluations are

	IM			MUC			B^3			$CEAF_E$			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
SEMEVAL data – SEMEVAL scorer															
G+S/S+S	88.12	81.54	84.70	19.82	62.80	30.13	64.10	79.01	70.78	79.96	57.46	66.87	50.79	78.51	50.03
G+S/S-S	14.32	92.86	24.81	19.82	63.75	30.24	7.06	74.01	12.89	3.99	44.13	7.32	60.35	74.01	62.54
G-S/S+S	71.23	10.00	17.54	24.86	6.13	9.83	42.87	11.49	18.13	53.91	2.89	5.49	50.03	51.83	17.36
G-S/S-S	56.93	12.52	20.53	24.86	6.14	9.85	28.52	8.93	13.60	39.22	6.34	10.92	50.12	52.51	20.46
SEMEVAL data – CONLL scorer															
G+S/S+S	87.72	81.18	84.32	19.77	62.64	30.05	73.92	96.24	83.62	91.02	71.51	80.10	53.48	78.78	55.90
G+S/S-S	14.04	91.10	24.34	19.77	63.59	30.16	73.91	96.41	83.68	91.09	71.45	80.09	53.48	79.50	55.91
G-S/S+S	45.38	6.37	11.17	12.61	3.11	4.99	86.92	43.79	58.24	20.53	39.42	27.00	50.36	50.19	50.22
G-S/S-S	37.90	8.33	13.66	12.61	3.11	5.00	86.25	42.22	56.69	20.55	42.20	27.64	51.11	50.57	50.72
CONLL data – SEMEVAL scorer															
G-S/S+S	96.55	18.55	31.12	31.25	25.12	27.85	38.07	17.06	23.57	61.98	3.66	6.91	50.01	51.63	22.85
G-S/S-S	65.16	40.16	49.69	33.87	27.29	30.23	26.94	31.86	29.20	46.04	17.09	24.93	50.84	65.01	38.33
CONLL data – CONLL scorer															
G-S/S+S	95.11	18.27	30.66	30.59	24.58	27.26	68.11	64.25	66.12	34.16	36.88	35.47	53.44	59.15	54.80
G-S/S-S	62.71	38.66	47.83	30.59	24.65	27.30	67.06	62.65	64.78	34.19	40.16	36.94	54.10	60.29	55.67

Table 5: System results with and without singletons on the SEMEVAL and CONLL data.

shown in table 5. Overall, there are considerable differences in the results, ranging in F-score from 4.99 in the SEMEVAL data set with the G-S/S+S setting and the MUC metric of the CONLL scorer to 83.68 in the same data set with the G+S/S-S setting and the B^3 metric of the CONLL scorer. This is disconcerting given that there is no difference in system quality, but simply in the representation of singletons. The differences between settings within a single metric are similarly extreme: B^3's F-score, for example, ranges from 70.78 to 12.89, on the same data set using the same scorer, the only difference is the presence of singletons in the system output.

A comparison of the scores for mention identification (IM) shows that the scorer version has a considerable influence on the results on the SEMEVAL data set: In the G-S/S+S setting, recall decreases from 71.23% to 45.38%. In the CONLL data set, this effect is also present, but to a lesser degree: The F-score decreases from 31.12 to 30.66 in the same setting. Any setting with a difference in the presence of singletons between gold standard and system output results in extreme differences in precision and recall. When singletons are present in the system output but not in the gold standard, recall is boosted; precision profits from the presence of singletons in the gold standard. The fact that UBIU obtains higher IM scores on the CONLL data set may be due to the strategy for mention detection, which was developed explicitly for the CONLL data set.

Contrary to our expectation that MUC will remain constant across the 4 settings, there is a significant decrease in F-score on the SEMEVAL data set between the settings in which the gold standard contains singletons and the one where it does not. The F-scores drop from approximately 30 to 9. Additionally, while there is no significant difference between the settings in which there are no singletons in the gold standard for the SEMEVAL set, the CONLL set shows a deterioration of approximately 3 percent points from G-S/S+S to G-S/S-S for the SEMEVAL scorer. The B^3 results of the SEMEVAL scorer closely model mention quality. Additionally, the results of the CONLL scorer are significantly higher than those by the SEMEVAL scorer. In the G-S/S-S setting, for example, the F-score ranges from 13.60 to 56.69 on the SEMEVAL data and from 29.20 to 64.78 on the CONLL data. $CEAF_E$ and BLANC show similar trends.

A comparison of UBIU on the two data sets shows that based on the majority of the metrics, the CONLL shared task was the easier of the two. All of the results for the CONLL set are higher than for the SEMEVAL set, with the only exception of MUC for the G-S/S-S setting. This is surprising given that the CONLL task also included verbal coreference, which should be a challenge for a system whose features were developed for nominal coreference. However, the CONLL training set was also more extensive with 2374 documents, in comparison to 322 documents in the SEMEVAL training set.

	IM			MUC			B^3			$CEAF_E$			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
SEMEVAL scorer															
AllS	96.55	18.55	31.12	31.25	25.12	27.85	38.07	17.06	23.57	61.98	3.66	6.91	50.01	51.63	22.85
NoS	58.86	38.42	46.50	33.87	27.29	30.23	25.13	29.62	27.19	40.56	17.26	24.22	50.86	63.97	37.61
PronS	65.16	40.16	49.69	33.87	27.29	30.23	26.94	31.86	29.20	46.04	17.09	24.93	50.84	65.01	38.33
AttP	70.52	28.69	40.78	28.70	12.35	17.27	26.94	16.03	20.10	40.54	14.29	21.13	50.51	57.15	32.64
CONLL scorer															
AllS	95.11	18.27	30.66	30.59	24.58	27.26	68.11	64.25	66.12	34.16	36.88	35.47	53.44	59.15	54.80
NoS	56.44	36.84	44.59	30.59	24.65	27.30	67.06	62.65	64.78	34.19	40.16	36.94	54.10	60.29	55.67
PronS	62.71	38.66	47.83	30.59	24.65	27.30	67.06	62.65	64.78	34.19	40.16	36.94	54.10	60.29	55.67
AttP	67.76	27.56	39.18	25.68	11.05	15.45	75.97	42.30	54.34	21.44	42.02	28.39	52.56	52.19	52.36

Table 6: System results with varying treatment of pronouns.

6 Pronominal Singletons in the System Output

Here, we have a closer look at pronoun singletons in the system output. We include all types of anaphoric pronouns in our investigation, i.e. personal, reflexive, demonstrative, and possessive pronouns. Relative and indefinite pronouns are not annotated as mentions in the data and thus excluded from our study. Since most of the pronouns are inherently anaphoric, we know that, apart from expletive pronouns, they must be part of a coreference chain. We examine the effect of singleton pronouns on the scorers' results.

We use the CONLL data set for this study since it does not contain singletons. This means, the expectation for the system is that it does not include singletons in the answers. On the system side, we investigate the following four settings:

1. AllS: In this setting, singletons are not filtered out, i.e. all mentions for pronouns, NPs, names, verbs, etc. remain in the final system.
2. NoS: This setting filters out all singletons, i.e. all mentions that were marked by the mention extractor but for which the coreference resolution module did not find any coreferring mentions, are deleted from the system answers.
3. PronS: This is similar to the NoS setting, but here all the pronominal singletons remain in the answers. I.e. the filter deletes all NP mentions, but does not delete any pronoun mentions.
4. AttP: In the final setting, singleton pronouns are attached to an antecedent. I.e. the system enforces coreference for all pronouns. If the coreference resolution module does not find an antecedent for the pronoun, a heuristic enforces coreference to the closest preceding mention. As in the NoPron setting, all singletons that do not consist of a pronoun are deleted.

The results of the system performance given the above settings are shown in table 6. Similar to the findings in section 5, there is a difference between the scores achieved by the SEMEVAL scorer and the CONLL scorer. The CONLL MUC scores are somewhat lower while the CONLL B^3, $CEAF_E$, and BLANC scores are higher by a wide margin to maximally 2.8 times the original F-score.

The mention quality (IM) shows the expected results: For the AllS setting, the system reaches a very high recall of 96.55/95.11%, but at the same time a very low precision, which also results in the lowest F-score. Since all the singletons are included in the system answer, a high number of mentions are found, but many of the identified mentions are twinless singletons. When we exclude all singletons in the NoS setting, recall reaches its lowest value, but precision profits so that the F-score is higher overall than the AllS score. Forcing the pronouns into a coreference relation has a positive influence on recall, which increases to 70.52/67.76%, but a negative influence on precision, which decreases to 28.78/27.56%. These results show that adding the pronouns and their coreferent mentions has a positive influence on recall but the missing separation of expletive pronouns from anaphoric ones has a detrimental effect on precision.

MUC, which should not be sensitive to singletons in the system answers, shows the same scores for the settings with no singletons (NoS) and with only

pronominal singletons. Given the CoNLL scorer, all metrics show the same scores for the NoS and PronS settings, thus they are insensitive towards the presence of non-pronominal singleton. However, for the setting with all singletons, all scores based on the SemEval scorer are considerably lower than for the settings without singletons or with only pronominal singletons. The reason for this difference is unclear at this point and needs to be investigated further.

7 Conclusion and Future Work

In this paper, we investigated the influence of singletons in the gold standard as well as in the system output on coreference resolution evaluation. We have shown that all metrics are affected by the presence of singletons in the gold standard. Especially in a setting in which both the gold standard and the system output contain singletons, the evaluation scores of both versions of the scorer are artificially boosted. However, the presence of singletons in the system output also has an effect on evaluation, but to a considerably lesser degree. This means that a system may not always be rewarded for having a reliable filter for singletons. Including singletons in the training data is a necessary step towards more realistic settings. However, including singletons in the gold standard for evaluation artificially boosts results.

Acknowledgment

This work is based on research supported by the US Office of Naval Research (ONR) Grant #N00014-10-1-0140. We also gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition (Project I5-DiaSpace).

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC Workshop on Linguistic Coreference*, Granada, Spain.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2007. TiMBL: Tilburg memory based learner – version 6.1 – reference guide. Technical Report ILK 07-07, ILK-CL, Tilburg University.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of HLT-NAACL 2007*, Rochester, NY.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT-NAACL*, New York, NY.

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 ACL*, Barcelona, Spain.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP*, Vancouver, Canada.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of CoNLL 2011*, Portland, OR.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP*, Singapore.

Altaf Rahman and Vincent Ng. 2011. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *JAIR*, 40:469–521.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand Index for coreference evaluation. *Natural Language Engineering*.

Marta Recasens, Lluìs Márquez, 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 SemEval*, Uppsala, Sweden.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of ACL-AFNLP*, Singapore.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of MUC-6*, Columbia, MD.

Holger Wunsch. 2009. *Rule-Based and Memory-Based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. Ph.D. thesis, Universität Tübingen.

Desislava Zhekova and Sandra Kübler. 2010. UBIU: A language-independent system for coreference resolution. In *Proceedings of SemEval*, Uppsala, Sweden.

Desislava Zhekova and Sandra Kübler. 2011. UBIU: A Robust System for Resolving Unrestricted Coreference. In *Proceedings of CoNLL: Shared Task*, Portland, OR.