Instance Sampling for Multilingual Coreference Resolution

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

Abstract

In this paper we investigate the effect of downsampling negative training instances on a multilingual memory-based coreference resolution approach. We report results on the SemEval-2010 task 1 data sets for six different languages (Catalan, Dutch, English, German, Italian and Spanish) and for four evaluation metrics (MUC, B^3, CEAF, BLANC). Our experiments show that downsampling negative training examples does not improve the overall system performance for most targeted languages and that the various evaluation metrics do not show a significantly distinct behavior across the different samples.

1 Introduction

In the last decade the research in the area of Computational Linguistics (CL) has been directed to new, flexible, efficient and most importantly automated methods for Natural Language Processing. The latter has motivated a shift from rule-based to machine-learning (ML) methods in the hope that those will lead to more robust and efficient solutions. Thus, the previously used rule-based approaches (cf. e.g. (Mitkov, 1998; Poesio et al., 2002)) to anaphora and coreference resolution (CR) have been followed by machine-learning techniques (cf. e.g. (Soon et al., 2001; Ng and Cardie, 2002b)). In general, one of the biggest disadvantages of the rule-based approaches is the fact that the created coreference resolution systems must be constantly extended in order to provide rules for yet unseen cases. Thus, whenever a new language is considered, a distinct set of rules needs to be assembled, which can hardly be completed in a reasonable time frame. Yet, approaching the CR task on a multilingual level means that the resulting coreference procedure needs to be robust and general enough to lead to good results in an unseen environment. This provides a reasonable motivation for the use of ML methods, since only those can be designed with the required flexibility by keeping efficiency in mind.

Previous work in the area (Zhekova and Kübler, 2010) developed such a robust multilingual machine-learning based CR system, UBIU (see section 3.1), which we use in our work and which is not specifically fine tuned to any of the languages it is applied to. However, achieving good and linguistically motivated results in a multilingual environment is not an easy task. For this reason, the general performance of the system must be maximally optimized so that it is able to efficiently use the little but relevant information that it is provided with.

Based on their complexity and flexibility, ML methods, as the ones used in UBIU, offer various possibilities to optimize the system performance to the given task. Such an optimization is, for example, instance sampling. Since there are contradictory opinions on whether the latter has a positive or rather negative effect on the overall coreference system performance (see section 2) and since by now there is no work on its application to a multilingual CR approach, we apply instance sampling on UBIU in this paper. We first present various approaches related to our work (section 2), further in section 3, we describe the experimental setup by introducing the CR system that we used for our experiments (section 3.1) as well as the approached investigation (section 3.2). In section 4, we present our results and, in section 5, we draw some conclusive remarks and outline a reasonable continuation and investigation of the multilingual coreference resolution approach.

2 Previous Work

In her work, Uryupina (2004) reports that in the MUC-7 (Hirschman, 1997) corpus only about 1-2% (approximate ratio of 1:48) of the instances are positive (coreferent). The same was also reported for the MUC-6 data by Ng and Cardie (2002a).

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 150–155,
Hissar, Bulgaria, 13 September 2011.

Such extremely skewed distribution of positive vs. negative examples in the training data is believed to cause difficulties for the classification process. This happens since ML approaches are influenced by the unbalanced assembly of training instances and approach a classification system that intends to partially keep the ratio that is already distorted. Hoste (2005) also comments that standard classification algorithms may show poor performance when applied to an unbalanced data set since minority classes are completely ignored by some algorithms. The latter are then not applicable on data such as the one assembled in a state-of-the-art CR tasks. However, other algorithms are able to find a reasonable trade-off between the correctly and wrongly identified minority class labels.

In order to account for the disproportionate data, multiple approaches to coreference resolution have employed instance sampling techniques (Ng and Cardie, 2002a; Uryupina, 2004; Zhao and Ng, 2007; Wunsch et al., 2009; Recasens and Hovy, 2009). One possibility for this is instead of keeping all possible instances in the training data, to randomly remove negative vectors. The latter can be also excluded via a statistically or linguistically motivated algorithm that is applied until an optimal ratio for the task is reached. Once this is done, the data can be used by the classifier. Another possibility to reach a normalized ratio is by mining more positive instances in the data such as the approach presented by Ng and Cardie (2002a).

In their work, Wunsch et al. (2009) compare different instance sampling techniques with different classifiers on the task of anaphora resolution on a single language – German. They report that all applied methods lead to an improvement of the overall system performance independently of the type of the classifier (memory-based learner, decision trees, maximum entropy learner). Better system performance from the use of instance sampling is also reported by Uryupina (2004). However, both improvements, as the authors discuss, are a result of increased recall and drastically decreased precision. In her PhD thesis, Hoste (2005) shows that downsampling negative examples leads to an unacceptable trade-off between recall and precision. The latter was recently confirmed in (Recasens and Hovy, 2009) where the authors conclude that while using a memory-based classifier, downsampling negative instances for training does not lead to an improvement of the overall performance.

All distinct methods for instance sampling were employed in different CR systems. Some of them were completely ML based, others used a hybrid approach to the task. Moreover, none of the systems was able to test the exact same sampling technique on more than one language and on more than one evaluation metric. This makes it hard to gain an objective overview of when and how instance sampling, and specifically downsampling of negative examples in the training data, influences the overall performance of a CR system. If we consider the findings as in (Wunsch et al., 2009; Ng and Cardie, 2002a; Uryupina, 2004) we can expect that using downsampling will significantly increase the performance of a multilingual memory-based coreference resolution system. However, if we favor the theories in (Hoste, 2005; Recasens and Hovy, 2009) we can only expect a change in the overall system performance gained by an unacceptable trade-off between system precision and recall.

Our assumption is that instance sampling can lead to a significant and well balanced improvement in the overall performance for systems that use hybrid approaches and are thus highly tuned for specific languages. Such systems make use of explicit rules that are language specific and often hand-crafted (in various stages of the CR process, e.g. preprocessing, postprocessing, etc.). Those rules are generally accurate on their own and lead to good performance overall. Thus, systems that make use of such rules can only benefit if the ML component favors a classification system with a higher rate for positive answers. The system that we use for our experiments is exclusively ML based and constructed in an exceptionally general way such that it can be easily applied to diverse new languages without much additional effort.

3 Experimental Setup

In order to evaluate the influence of instance sampling on a multilingual CR approach, which to our knowledge has not yet been attempted, we investigated its effect in the setting defined by the SemEval-2010 task 1 (Recasens et al., 2010). In the following section, we will first shortly introduce the employed coreference resolution system (see section 3.1) and then present the design of the experiments that we conducted (see section 3.2).

3.1 UBIU

The coreference resolution system, UBIU (Zhekova and Kübler, 2010), that we used in our work was initially designed for the multilingual CR task (Recasens et al., 2010). The prevailing purpose for the use and further development of UBIU is to gain more insight into the problems that occur when the CR task is extended from the use of only one language to multiple ones. For this reason, UBIU is structured in a way that allows for a quick and easy integration of a new language, given that the provided data is formatted in the style used by SemEval-2010 (Recasens et al., 2010).

The coreference resolution pipeline in UBIU starts with a basic preprocessing step of the data in which only insignificant formatting and restructuring of the data is conducted. Further, an important step is approached – mention identification. During this step, the relevant UBIU module extracts the nominal/pronominal phrases that are further considered in the coreference process. The system stores the mention boundaries and extracts the syntactic heads of the phrases, which are further passed to the next system module responsible for the feature extraction. The latter follows the mention-pair model that uses a subset of the features presented by Rahman and Ng (2009) (as listed in (Zhekova and Kübler, 2010)) to create feature vectors that are passed to the next module in the system. The same process is executed for both the training and the test set, which leads to their transformation from the original data format to a format represented by feature vectors. Both training and tests sets are then further used by the next module in the UBIU pipeline.

For the actual coreference classification, UBIU implements a ML approach and is thus structured around the idea of memory-based learning (MBL) (Daelemans and van den Bosch, 2005). The MBL learner that is used for classification is TiMBL (Daelemans et al., 2007). In general, a MBL classifier makes use of a similarity metric in order to identify the most similar examples (the k nearest neighbors (k-nn)) in the training data to the example that has been currently classified in the test data. Based on the classes that those k-nn instances have, a decision for the yet unlabeled vector can be made. Once labeled, the references between the syntactic heads of the phrases and the actual boundaries of the phrases is restored in a postprocessing step and the final coreference chains of clustered coreferent phrases are created.

3.2 Experiments

We conduct six different experiments on all six languages (Catalan, Dutch, English, German, Italian and Spanish) and show the results for all four evaluation metrics (MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005), BLANC (Recasens and Hovy, 2011)). For each language, we used as training data the development set provided by the SemEval-2011 task 1 corpora. As test data we employed the official test set from the task. The system performance that we report is different from the one that was reported during UBIU's participation in the task (Recasens et al., 2010) as a result of various improvements on the system and the use of a subset of the actual training data. For scoring, we employed the software provided by task 1. Each separate run of the system used different ratio between the positive and negative examples in the training process. The base ratio for all languages that was observed in the development set when derived in a context window of three sentences is as follows: Catalan – 1:25; Dutch – 1:14; English – 1:26; German – 1:31; Italian – 1:45; Spanish – 1:24. We further explored the following five ratios: 1:10, 1:7, 1:5, 1:4, 1:2. In order to achieve the downsampled sets we use an approach based on random removal of negative instances.

4 Results

In the current section, we discuss the final results of the system (listed in table 1) that the multilingual coreference resolution system UBIU achieved for all six experimental runs. In order to gain more insight into the actual effect of the sampling approach on the classification system, in section 4.1, we also report the distribution of positive vs. negative examples in the test sets that have already been classified. We then divide and report our observations in three different classes: differences in system performance across the various evaluation metrics (presented in section 4.2), differences in system performance across the various languages (introduced in section 4.3) and differences in system performance across both language families (accounted for in section 4.4).

		MUC			B^3			CEAF-M			CEAF-E			BLANC		
	train ratio	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
C	1:25	14.14	**30.78**	**19.38**	53.31	**69.12**	**60.20**	**54.44**	**49.20**	**51.69**	**70.23**	46.31	**55.81**	50.65	**62.19**	49.15
D	1:14	02.65	04.48	03.33	23.71	**20.58**	**22.04**	28.75	09.35	14.11	**49.71**	06.39	11.33	**50.00**	50.21	27.71
E	1:26	11.76	**32.06**	17.21	62.79	**75.32**	**68.49**	**62.70**	**57.98**	**60.25**	**76.50**	**55.81**	**64.54**	50.41	**61.87**	49.30
G	1:31	14.04	**26.65**	**18.39**	50.67	**51.31**	**50.99**	**52.80**	**44.24**	**48.14**	**59.88**	42.47	**49.70**	**50.06**	**56.02**	44.19
I	1:45	04.31	**24.06**	**07.31**	35.70	**56.86**	**43.86**	**37.89**	**41.16**	**39.46**	**46.80**	**38.29**	**42.12**	**50.02**	**59.00**	42.98
S	1:24	15.00	**30.49**	**20.11**	54.82	**70.32**	**61.61**	**55.72**	**52.71**	**54.17**	**70.93**	50.65	**59.10**	50.71	**60.71**	49.74
C	1:10	17.25	16.73	16.99	53.87	53.20	53.53	48.16	43.52	45.73	55.93	48.44	51.92	50.59	52.89	49.64
D		04.35	04.32	04.33	24.13	18.30	20.81	28.58	09.30	14.03	45.96	06.57	11.50	49.99	49.72	27.98
E		19.54	15.84	17.50	64.01	56.86	60.22	53.27	49.26	51.19	58.23	57.09	57.66	50.68	53.09	50.25
G		**17.29**	14.16	15.57	**51.16**	42.60	46.49	48.77	40.86	44.47	51.05	**43.10**	46.74	50.01	50.52	44.36
I		05.39	10.18	07.05	35.80	50.21	41.80	34.98	37.99	36.42	41.17	38.12	39.58	49.98	49.38	43.13
S		18.18	20.36	19.21	55.51	59.09	57.24	52.22	49.40	50.77	61.55	53.21	57.07	50.95	56.17	50.42
C	1:7	17.77	15.71	16.68	53.99	50.73	52.31	47.02	42.49	44.64	53.29	48.84	50.97	50.64	52.84	49.78
D		05.87	04.89	05.34	24.45	17.33	20.28	28.82	09.38	14.15	44.43	06.70	11.65	49.99	49.84	28.13
E		20.78	15.34	**17.65**	64.13	54.37	58.85	52.05	48.13	50.01	55.44	57.12	56.27	50.82	53.66	50.48
G		16.17	11.95	13.74	51.03	40.70	45.28	47.15	39.51	43.00	48.78	42.73	45.55	49.99	49.76	44.41
I		**05.42**	08.58	06.64	35.78	48.48	41.17	34.03	36.97	35.44	39.49	38.03	38.75	50.00	50.24	43.27
S		20.26	18.54	19.36	56.01	53.28	54.61	50.32	47.61	48.93	56.47	54.54	55.49	**51.12**	55.28	50.82
C	1:5	18.91	15.45	17.00	54.23	47.94	50.89	46.14	41.69	43.80	50.84	**49.26**	50.04	**50.66**	52.78	49.86
D		09.09	**05.57**	06.91	25.40	15.24	19.05	**30.52**	**09.93**	**14.99**	43.08	07.42	12.66	**50.00**	**50.32**	28.07
E		19.90	15.57	17.47	63.95	55.68	59.53	53.42	49.40	51.33	57.34	**57.24**	57.29	50.76	54.03	50.32
G		16.29	10.93	13.08	51.05	39.06	44.25	46.05	38.59	41.99	46.82	42.66	44.64	49.99	49.80	44.49
I		05.21	07.21	06.05	35.79	46.66	40.51	33.22	36.08	34.59	37.81	37.77	37.79	49.99	49.67	43.28
S		18.80	14.62	16.45	55.68	47.30	51.15	46.21	43.71	44.92	49.19	53.60	51.30	51.04	53.83	50.79
C	1:4	18.70	13.67	15.79	54.14	43.93	48.50	43.64	39.43	41.43	46.21	49.12	47.62	**50.66**	52.33	50.00
D		09.71	05.40	**06.94**	25.45	14.36	18.36	30.43	09.90	14.94	41.33	**07.53**	**12.73**	**50.00**	50.28	28.11
E		22.08	14.56	17.55	64.35	50.60	56.65	49.64	45.91	47.70	51.44	56.82	54.00	50.78	52.98	50.48
G		16.44	10.09	12.50	51.10	37.12	43.00	44.48	37.27	40.55	44.14	41.96	43.02	49.96	49.32	44.53
I		05.36	07.26	06.17	**35.82**	46.32	40.40	32.98	35.82	34.34	37.43	37.63	37.53	49.98	49.58	43.29
S		20.86	15.67	17.90	56.08	45.83	50.44	45.53	43.07	44.26	47.87	53.76	50.64	51.08	53.64	50.88
C	1:2	**20.71**	12.23	15.38	**54.65**	34.36	42.19	37.54	33.93	35.64	34.84	47.42	40.16	50.60	51.55	**50.13**
D		**11.72**	04.52	06.52	**25.96**	10.18	14.63	28.84	09.38	14.16	29.86	07.49	11.98	49.98	49.69	**28.69**
E		**23.89**	12.65	16.54	**64.74**	42.09	51.01	43.55	40.28	41.85	41.57	55.10	47.39	**50.83**	52.18	**50.71**
G		16.80	08.26	11.08	51.06	31.84	39.22	39.87	33.41	36.36	37.06	40.25	38.59	49.93	49.17	**44.79**
I		04.02	03.54	03.76	35.64	38.98	37.24	28.88	31.37	30.07	29.95	35.97	32.69	50.00	50.00	**43.51**
S		**21.44**	13.09	16.26	**56.26**	36.29	44.12	39.11	37.00	38.02	36.91	52.41	43.31	51.07	52.45	**51.02**

Table 1: System performance over all languages (C(atalan), D(utch), E(nglish), G(erman), I(talian) and S(panish)) and sampling variations.

4.1 Test Set Distribution

In table 2, we list the various distributions of the positive vs. negative examples in both training and test sets of each sample. The base distribution of examples in the train data for all languages is as presented in section 3.2. The figures show that memory-based learning is highly sensitive to the distribution of positive vs. negative examples in the data. It approaches a classification system that ensures a distribution of the instances in the final outcome that is to some extent proportionate to the training ratio of both classes. Yet, this does not ensure that a positively classified instance is correctly labeled, which motivates our investigation of the system performance in the various samples.

train	test					
	Catalan	Dutch	English	German	Italian	Spanish
base	1:66.15	1:55.71	1:66.93	1:63.26	1:126.14	1:67.66
1:10	1:18.85	1:36.48	1:13.06	1:16.77	1:36.12	1:23.78
1:7	1:15.04	1:27.58	1:11.28	1:13.97	1:28.31	1:15.92
1:5	1:12.30	1:17.51	1:12.42	1:11.26	1:23.06	1:12.22
1:4	1:9.49	1:13.52	1:8.88	1:9.65	1:21.95	1:10.50
1:2	1:4.35	1:4.71	1:5.11	1:5.82	1:9.09	1:5.43

Table 2: Distribution of positive vs. negative examples in the train and already classified test set.

4.2 Differences Across Metrics

Considering the results displayed in table 1 there are several significant differences in system performance across the samples in respect to the evaluation metrics that were used to evaluate it.

From all four metrics only MUC and B^3 show a distinctive change in recall when the sample of negative examples in the training set reduces and in particular when it reaches a ratio of 1:2. The differences for B^3 are not surprisingly high, but the MUC metric shows an exceedingly boosted performance. The latter, we assume, is due to one of MUC's most important shortcomings, namely the fact that overmerged entities are not punished but rather rewarded by the metric. In a training setting, in which only 2 negative examples are used for each positive one, the classifier is bound to return a high number of positive instances, thus leading to highly overmerged coreference chains. Both variants of the CEAF metric do not show an improvement in recall for all different samples apart from the CEAF-M variant with respect to Dutch, which has best recall in a sample 1:5. Similar to CEAF, the BLANC metric also reaches best recall

train	Catalan	Dutch	English	German	Italian	Spanish
base	47.25	15.70	51.96	42.28	35.15	48.95
1:10	43.57	15.73	47.36	39.47	33.60	46.94
1:7	42.88	15.91	46.65	38.40	33.05	45.84
1:5	42.32	16.34	47.19	37.69	32.44	42.92
1:4	40.67	16.22	47.28	36.72	32.35	42.82
1:2	36.70	15.20	41.50	34.01	29.45	38.55

Table 3: Average system performance over all languages and sampling variations.

train	Romance	Germanic
base	43.78	36.65
1:10	41.37	34.19
1:7	40.59	33.65
1:5	39.22	33.74
1:4	38.61	33.41
1:2	34.90	30.24

Table 4: Average system performance over both language families and sampling variations.

values for most of the languages in the original examples ratio. Moreover, the differences in scores for which different ratios performed better are relatively small.

With respect to precision, the behavior of most metrics is quite similar. Apart from CEAF-E, for which precision does not show a clear pattern, all metrics reach the highest precision scores for all languages in the base example distribution.

From the given precision and recall figures, it is not surprising that the final F-scores of most metrics are also highest for the original distribution of positive vs. negative training examples. What is surprising here is that the BLANC metric reaches highest scores in the 1:2 train ratio for which neither the precision nor the recall perform best. This, we assume, is due to the more complex way of calculating BLANC's final score, which as Recasens and Hovy (2011) discuss puts equal emphasis on coreference and non-coreference links. Yet, the improvement in scores is, as an average over all languages, less than 1%, which we do not consider noteworthy.

On the basis of those observations, we can conclude that instance sampling does not lead to a considerable improvement of the CR system performance for most of the four evaluation metrics. The only relatively higher figures were reached by MUC's and B^3's recall as well as for BLANC's final scores. Our assumption is that the high concentration of positively labeled examples lead to overmerged entities for which the evaluation metrics reach better recall, but this does not necessarily lead to an overall better performance.

4.3 Differences Across Languages

Since in this evaluation approach we are more interested into how the given change in the training ratio influences the overall performance of the system per language and not each separate metric, we use the scores (listed in table 3) that are achieved by the average calculation of the F-score for each separate language. It is surprising to see that for all languages, apart from Dutch, there is no improvement on the overall performance of the system for any of the artificially created samples. For Dutch, the averaged F-score rises slightly but gradually for the samples 1:10, 1:7 and 1:5, where for the latter sample the classifier reaches an averaged performance of 16.34% as compared to its performance in the base distribution – 15.70%. Again, this is not an exceedingly high improvement of system performance. However, a possible explanation for the fact that instance sampling reaches better results only for Dutch might be triggered by its outlier nature and considerably low overall performance. On the basis of that, we can assume that instance sampling can be more advantageous for less efficient memory-based classifiers than for the high performance ones. Yet, the change in scores might also be based on the variations across the annotation schemes of the different languages. In order to determine the exact reason, further investigation on the topic is needed.

4.4 Differences Across Language Families

A multilingual coreference resolution system as UBIU is hard to design in a way in which it will be able to perform optimally for each newly introduced language. Thus, it is reasonable to assume that system generalizations and respectively optimizations will be more sensible if based around the concept of the language family and not the separate language. Accordingly, we attempt a further generalization of the system performance that allows us to note the differences in the classification output for the Romance and Germanic language families. In table 4, we report the averaged results. Yet, the classifier performance curves across the samples formed on the basis of the two language families and not on the separate languages again do not show a significant variation from one another. Both performance types gradually decrease for each sample, which shows that there are no specific differences among language families that can be captured by an instance sampling approach.

5 Conclusion and Future Work

In the current paper, we presented our results from an instance sampling approach applied on a memory-based coreference resolution system. The novelty of our work lies in the investigation and employment of the sampling procedure in a multilingual environment that, to our knowledge, has not yet been explored. We show that despite the intermediate differences in precision and recall over the four evaluation metrics their overall F-scores are highest for the base sample distribution. Our hypothesis is that when trained on a sample with high concentration of positive examples, classifiers attempt the classification process in a way that keeps the ratio of positive vs. negative examples proportionate in their output. This leads to overmerged entities for which some metrics reach better recall, yet this does not necessarily lead to a boosted overall performance because of the generally lower precision. However, the increase of performance for one of the languages, Dutch, shows that instance sampling can be advantageous to some languages. Based on the language family we did not observe a considerable variation in the system performance. On account of our results, we believe that coreference resolution approaches should further concentrate more on the integration of new and novel linguistic information as well as world knowledge rather than on technical and statistical system optimization.

Acknowledgment

We 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 *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.

Walter Daelemans and Antal van den Bosch. 2005. *Memory Based Language Processing*. Cambridge University Press.

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, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Lynette Hirschman. 1997. MUC-7 Coreference Task Definition.

Véronique Hoste. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, University of Antwerp.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT '05*, Morristown, USA.

Ruslan Mitkov. 1998. Robust Pronoun Resolution with Limited Knowledge. In *Proceedings of ACL/COLING 1998*, Montreal, Canada.

Vincent Ng and Claire Cardie. 2002a. Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules. In *Proceedings of EMNLP 2002*.

Vincent Ng and Claire Cardie. 2002b. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of ACL 2002*, Philadelphia, PA.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. 2002. Acquiring Lexical Knowledge For Anaphora Resolution. In *Proceedings of LREC 2002*, Las Palmas, Gran Canaria.

Altaf Rahman and Vincent Ng. 2009. Supervised Models for Coreference Resolution. In *Proceedings of EMNLP 2009*, Singapore.

Marta Recasens and Eduard Hovy. 2009. A Deeper Look into Features for Coreference Resolution. In *Proceedings of DAARC 2009*.

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

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 SemEval 2010*, Uppsala, Sweden.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4).

Olga Uryupina. 2004. Linguistically Motivated Sample Selection for Coreference Resolution. In *Proceedings of DAARC 2004*, Sao Miguel, Portugal.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings MUC 1995*, Columbia, MD.

Holger Wunsch, Sandra Kübler, and Rachael Cantrell. 2009. Instance Sampling Methods for Pronoun Resolution. In *Proceedings of RANLP 2009*, Borovets, Bulgaria.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. In *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic.

Desislava Zhekova and Sandra Kübler. 2010. UBIU: A Language-Independent System for Coreference Resolution. In *Proceedings of SemEval 2010*, Uppsala, Sweden.