

UBIU for Multilingual Coreference Resolution in OntoNotes

Desislava Zhekova Sandra Kübler Joshua Bonner Marwa Ragheb Yu-Yin Hsu

Indiana University
Bloomington, IN, USA

{dzhekova, skuebler, jebonner, mragheb, hsuy}@indiana.edu

Abstract

The current work presents the participation of UBIU (Zhekova and Kübler, 2010) in the CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes (Pradhan et al., 2012). Our system deals with all three languages: Arabic, Chinese and English. The system results show that UBIU works reliably across all three languages, reaching an average score of 40.57 for Arabic, 46.12 for Chinese, and 48.70 for English. For Arabic and Chinese, the system produces high precision, while for English, precision and recall are balanced, which leads to the highest results across languages.

1 Introduction

Multilingual coreference resolution has been gaining considerable interest among researchers in recent years. Yet, only a very small number of systems target coreference resolution (CR) for more than one language (Mitkov, 1999; Harabagiu and Maiorano, 2000; Luo and Zitouni, 2005). A first attempt at gaining insight into the comparability of systems on different languages was accomplished in the SemEval-2010 Task 1: Coreference Resolution in Multiple Languages (Recasens et al., 2010). Six systems participated in that task, UBIU (Zhekova and Kübler, 2010) among them. However, since systems participated across the various languages rather irregularly, Recasens et al. (2010) reported that the data points were too few to allow for a proper comparison between different approaches. Further significant issues concerned system portability across

the various languages and the respective language tuning, the influence of the quantity and quality of diverse linguistic annotations as well as the performance and behavior of various evaluation metrics.

The CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes (Pradhan et al., 2011) targeted unrestricted CR, which aims at identifying nominal coreference but also event coreference, within an English data set from the OntoNotes corpus. Not surprisingly, attempting to include such event mentions had a detrimental effect on overall accuracy, and the best performing systems (e.g., (Lee et al., 2011)) did not attempt event anaphora. The current shared task extends the task definition to three different languages (Arabic, Chinese and English), which can prove challenging for rule-based approaches such as the best performing system from 2011 (Lee et al., 2011).

In the current paper, we present UBIU, a memory-based coreference resolution system, and its results in the CoNLL-2012 Shared Task. We give an overview of UBIU in Section 2. In Section 3, we present the system results, after which Section 4 lays out some conclusive remarks.

2 UBIU

UBIU (Zhekova and Kübler, 2010) is a coreference resolution system designed specifically for a multilingual setting. As shown by Recasens et al. (2010), multilingual coreference resolution can be approached by various machine learning methods since machine learning provides a possibility for robust abstraction over the variation of language phenomena and specificity. Therefore, UBIU employs

a machine learning approach, memory-based learning (MBL) since it has proven to be a good solution to various natural language processing tasks (Daelemans and van den Bosch, 2005). We employ TiMBL (Daelemans et al., 2010), which uses k nearest neighbour classification to assign class labels to the targeted instances. The classifier settings we used were determined by a non-exhaustive search over the development data and are as follows: the *IB1* algorithm, similarity is computed based on weighted overlap, gain ratio is used for the relevance weights and the number of nearest neighbors is set to $k=3$ (cf. (Daelemans et al., 2010) for an explanation of the system parameters).

In UBIU, we use a pairwise mention model (Soon et al., 2001; Broscheit et al., 2010) since this model has proven more robust towards multiple languages (Wunsch, 2009) than more elaborate ones. We concentrate on nominal coreference resolution, i.e. we ignore the more unrestricted cases of event coreference. Below, we describe the modules used in UBIU in more detail.

2.1 Preprocessing

The preprocessing module oversees the proper formatting of the data for all modules applied in later stages during coreference resolution. During preprocessing, we use the speaker information, if provided, and replace all 1st person singular pronouns from the token position with the information provided in the speaker column and adjust the POS tag correspondingly.

2.2 Mention Detection

Mention detection is the process of detecting the phrases that are potentially coreferent and are thus considered candidates for the coreference process. Mention detection in UBIU is based on the parse and named entity information provided by the shared task. This step is crucial for the overall system performance, and we aim for high recall at this stage. Singleton mentions that are added in this step can be filtered out in later stages. However, if we fail to detect a mention in this stage, it cannot be added later. We predict a mention for each noun phrase and named entity provided in the data. Additionally, we extract mentions for possessive pronouns in English as only those did not correspond to a noun phrase in

	MD		
	R	P	F ₁
Arabic	97.13	19.06	31.87
Chinese	98.33	31.64	47.88
English	96.73	30.75	46.67

Table 1: Mention detection (development set).

the syntactic structure provided by the task. In Arabic and Chinese, possessives are already marked as noun phrases.

The system results on mention detection on the development set are listed in Table 1. The results show that we reach very high recall but low precision, as intended. The majority of the errors are due to discrepancies between noun phrases and named entities on the one hand and mentions on the other. Furthermore, since we do not target event coreference, we do not add mentions for the verbs in the data, which leads to a reduction of recall.

In all further system modules, we represent a mention by its head, which is extracted via heuristic methods. For Arabic, we select the first noun or pronoun while for Chinese and English, we extract the the pronoun or the last noun of a mention unless it is a common title. Additionally, we filter out mentions that correspond to types of named entities that in a majority of the cases in the training data are not coreferent (i.e. cardinals, ordinals, etc.).

One problem with representing mentions mostly by their head is that it is difficult to decide between the different mention spans of a head. Since automatic mentions are considered correct only if they match the exact span of a gold mention, we include all identified mention spans for every extracted head for classification, which can lead to losses in evaluation. For example, consider the instance from the development set in (1): the noun phrase *the Avenue of Stars* is coreferent and thus marked as a gold mention (key 7). UBIU extracts two different spans for the same head *Avenue*: *the Avenue* (MD 3) and *the Avenue of Stars* (MD 5).

	token	POS	parse	key	MD	output
(1)	the	DT	(NP(NP*	(7	(3 5	(9
	Avenue	NNP	*)	-	3)	9)
	of	IN	(PP*	-	-	-
	Stars	NNPS	(NP*))	7)	(4 5)	-

Both mention spans are passed to the coreference resolver, together with additional features (i.e. men-

	MD	MUC	B ³	CEAF _E	Average
	F ₁	F ₁	F ₁	F ₁	F ₁
long	100.0	100.0	100.0	100.0	100.0
short	50.00	0	66.66	66.66	44.44

Table 2: The scores for the short example in (1).

tion length, head modification, etc.) that will allow the resolver to distinguish between the spans. The classifier decides that the shorter mention is coreferent and that the longer mention is a singleton. In order to show the effect of this decision, we assume that there is one coreferent mention to *key* 7. We consider the two possible spans and show the respective scores in Table 2. The evaluation in Table 2 shows that providing the correct coreference link but the wrong, short mention span, *the Avenue*, has considerable effects to the overall performance. First, as defined by the task, the mention is ignored by all evaluation metrics leading to a decrease in mention detection and coreference performance. Moreover, the fact that this mention is ignored means that the second mention becomes a singleton and is not considered by MUC either, leading to an F₁ score of 0. This example shows the importance of selecting the correct mention span.

2.3 Singleton Classification

A singleton is a mention which corefers with no other mention, either because it does not refer to any entity or because it refers to an entity with no other mentions in the discourse. Because singletons comprise the majority of mentions in a discourse, their presence can have a substantial effect on the performance of machine learning approaches to CR, both because they complicate the learning task and because they heavily skew the proportion in the training data towards negative instances, which can bias the learner towards assuming no coreference relation between pairs of mentions. For this reason, information concerning singletons needs to be incorporated into the CR process so that such mentions can be eliminated from consideration.

Boyd et al. (2005), Ng and Cardie (2002), and Evans (2001) experimented with machine learning approaches to detect and/or eliminate singletons, finding that such a module provides an improvement in CR performance provided that the classifier

#	Feature Description
1	the depth of the mention in the syntax tree
2	the length of the mention
3	the head token of the mention
4	the POS tag of the head
5	the NE of the head
6	the NE of the mention
7	PR if the head is premodified, PO if it is not; UN otherwise
8	D if the head is in a definite mention; I otherwise
9	the predicate argument corresponding to the mention
10	left context token on position token -3
11	left context token on position token -2
12	left context token on position token -1
13	left context POS tag of token on position token -3
14	left context POS tag of token on position token -2
15	left context POS tag of token on position token -1
10	right context token on position token +1
11	right context token on position token +2
12	right context token on position token +3
13	right context POS tag of token on position token +1
14	right context POS tag of token on position token +2
15	right context POS tag of token on position token +3
16	the syntactic label of the mother node
17	the syntactic label of the grandmother node
18	a concatenation of the labels of the preceding nodes
19	C if the mention is in a PP; else I

Table 3: The features used by the singleton classifier.

does not eliminate non-singletons too frequently. Ng (2004) additionally compared various feature- and constraint-based approaches to incorporating singleton information into the CR pipeline. Feature-based approaches integrate information from the singleton classifier as features while constraint-based approaches filter singletons from the mention set. Following these works, we include a k nearest neighbor classifier for singleton mentions in UBIU with 19 commonly-used features described below. However, unlike Ng (2004), we use a combination of the feature- and constraint-based approaches to incorporate the classifier’s results.

Each training/testing instance represents a noun phrase or a named entity from the data together with features describing this phrase in its discourse. The list of features is shown in Table 3. The instances that are classified by the learner as singletons with a distance to their nearest neighbor below a threshold (i.e., half the average distance observed in the training data) are filtered from the mention set, and are thus not considered in the pairwise coreference classification. For the remainder of the mentions, the class that the singletons classifier has assigned to the instance is used as a feature in the coreference classifier. Experiments on the development set showed

		MD	MUC	B ³	CEAF _E	Average
		F ₁	F ₁	F ₁	F ₁	F ₁
Arabic	+SC	58.36	34.75	58.26	37.39	43.47
	-SC	56.12	34.96	58.52	36.05	43.18
Chinese	+SC	52.30	42.70	61.11	32.86	45.56
	-SC	50.40	41.19	60.96	32.47	44.87
English	+SC	67.38	53.20	59.23	34.90	49.11
	-SC	65.55	51.57	59.18	34.38	48.38

Table 4: Evaluation of using (+SC) or not (-SC) the singleton classifier in UBIU on the development set.

that the most important features across all languages are the POS tag of the head word, definiteness, and the mother node in the syntactic representation. Information about head modification is helpful for English and Arabic, but not for Chinese.

The results of using the singleton classifier in UBIU on the development set are shown in Table 4. They show a moderate improvement for all evaluation metrics and all languages, with the exception of MUC and B³ for Arabic. The most noticeable improvement can be observed in mention detection, which gains approx. 2% in all languages. A manual inspection of the development data shows that the version using the singleton classifier extracts a slightly higher number of coreferent mentions than the version without. However, the reduction of mentions that are never coreferent, which was the main goal of the singleton classifier, is also present in the version without the classifier, so that the results of the classifier only have a minimal influence on the final results.

2.4 Coreference Classification

Coreference classification is the process in which all identified mentions are paired up and features are extracted to build feature vectors that represent the mention pairs in their context. Each mention is represented in the feature vector by its syntactic head. The vectors for the pairs are then used by the memory-based learner TiMBL.

As anaphoric mentions, we consider all definite phrases; we then create a pair for each anaphor with each mention preceding it within a window of 10 (English, Chinese) or 7 (Arabic) sentences. We consider a shorter window of sentences for Arabic because of its NP-rich syntactic structure and its longer sentences, which leads to an increased number of possible mention pairs. The set of features that we

use, listed in Table 5, is an extension of the set by Rahman and Ng (2009). Before classification, we apply a morphological filter, which excludes vectors that disagree in number or gender (applied only if the respective information is provided or can be deduced from the data).

Both the anaphor and the antecedent carry a label assigned to them by the singletons classifier. Yet, we consider as anaphoric only the heads of definite mentions. Including a feature representing the class assigned by the singletons classifier for each anaphor triggers a conservative learner behavior, i.e., fewer positive classes are assigned. Thus, to account for this behavior, we ignore those labels for the anaphor and include only one feature (no. 25 in Table 5) in the vector for the antecedent.

2.5 Postprocessing

In postprocessing, we create the equivalence classes of mentions that were classified as coreferent and

#	Feature Description
1	m_j - the antecedent
2	m_k - the mention (further m.) to be resolved
3	C if m_j is a pronoun; else I
4	C if m_k is a pronoun; else I
5	the concatenated values of feature 3 and feature 4
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 are the same string; else I
9	C if both are non-pronominal and are the same string; else I
10	C if both are pronouns; I if neither is a pronoun; else U
11	C if both are proper nouns; I if neither is; else U
12	C if both m. have the same speaker; I if they do not
13	C if both m. are the same named entity; I if they are not and U if they are not assigned a NE
14	token distance between m_j and m_k
15	sentence distance between m_j and m_k
16	normalised levenstein distance for both m.
17	PR if m_j is premodified, PO if it is not; UN otherwise
18	PR if m_k is premodified, PO if it is not; UN otherwise
19	the concatenated values for feature 17 and 18
20	D if m_j is in a definite m.; I otherwise
21	C if m_j is within the subject; I-within an object; U otherwise
22	C if m_k is within the subject; I-within an object; U otherwise
23	C if neither is embedded in a PP; I otherwise
24	C if neither is embedded in a NP; I otherwise
25	C if m_j has been classified as singleton; I otherwise
26	C if both are within ARG0-ARG4; I-within ARGM; else U
27	C if m_j is within ARG0-ARG4; I-within ARGM; else U
28	C if m_k is within ARG0-ARG4; I-within ARGM; else U
29	concatenated values for features 27 and 28
30	the predicate argument label for m_j
31	the predicate argument label for m_k
32	C if both m. agree in number; else I
33	C if both m. agree in gender; else I

Table 5: The features used by the coreference classifier.

		MD			MUC			B ³			CEAF _E			Average
		R	P	F ₁	R	P	F ₁	R	P	F ₁	R	P	F ₁	F ₁
Automatic Mention Detection														
auto	Arabic	27.54	80.34	41.02	19.64	62.13	29.85	41.91	90.72	57.33	56.79	24.81	34.53	40.57
	Chinese	35.12	72.52	47.32	31.19	57.97	40.56	49.49	77.65	60.45	45.92	25.24	32.58	44.53
	English	65.78	68.49	67.11	54.28	52.79	53.52	62.26	54.90	58.35	33.52	34.96	34.22	48.70
gold	Arabic	28.00	82.21	41.78	15.47	45.92	23.15	39.22	84.86	53.65	55.10	24.22	33.65	36.82
	Chinese	37.84	74.84	50.27	33.95	60.29	43.44	50.95	77.28	61.41	46.68	26.13	33.50	46.12
	English	66.05	69.62	67.79	54.45	53.59	54.02	61.66	55.62	58.48	33.82	34.65	34.23	48.91
Gold Mention Boundaries														
auto	Arabic	27.48	75.53	40.29	18.75	56.47	28.16	42.67	89.25	57.74	55.53	25.36	34.82	40.24
	Chinese	36.97	73.98	49.30	32.09	58.30	41.39	49.43	77.38	60.32	46.35	25.71	33.07	44.93
	English	66.45	70.91	68.61	54.96	54.67	54.82	61.85	55.60	58.56	34.38	34.67	34.53	49.30
gold	Arabic	28.06	82.39	41.87	15.56	46.18	23.28	39.23	84.95	53.67	55.10	24.20	33.63	36.86
	Chinese	37.89	74.79	50.30	33.93	60.19	43.39	50.87	77.27	61.35	46.62	26.13	33.49	46.08
	English	65.82	71.72	68.65	54.68	55.51	55.09	61.22	56.59	58.82	34.85	34.04	34.44	49.45
Gold Mentions														
auto	Arabic	100	100	100	42.48	80.36	55.58	50.87	89.69	64.92	71.96	34.52	46.66	55.72
	Chinese	100	100	100	42.02	79.57	55.00	50.22	80.81	61.94	60.27	27.08	37.37	51.44
	English	100	100	100	68.38	78.11	72.92	63.04	58.60	60.74	52.64	37.10	43.53	59.06
gold	Arabic	100	100	100	45.58	73.27	56.20	52.27	82.35	63.95	70.17	37.54	48.91	56.35
	Chinese	100	100	100	44.12	80.89	57.10	51.79	80.53	63.04	60.37	27.69	37.96	52.70
	English	100	100	100	68.54	78.10	73.01	63.14	58.63	60.80	52.84	37.44	43.83	59.21

Table 6: UBIU system performance in the shared task.

insert the appropriate class/entity IDs in the data, removing mentions that constitute a class on their own – singletons. We bind all pronouns (except the ones that were labeled as singletons by the singleton classifier) that were not assigned an antecedent to the last seen subject and if such is not present to the last seen mention. We consider all positively classified instances in the clustering process.

3 Evaluation

The results of the final system evaluation are presented in Table 6. Comparing the results for mention detection (MD) on the development set (see Table 1, which shows MD before the resolution step) and the final test set (Table 6, showing MD after resolution and the deletion of singletons), we encounter a reversal of precision and recall tendencies (even though the results are not fully comparable since they are based on different data sets). This is due to the fact that during mention detection, we aim for high recall, and after coreference resolution, all mentions identified as singletons by the system are excluded from the answer set. Thus mentions that are coreferent in the key set but wrongly classified in the answer set are removed, leading to a decrease in recall. With regard to MD precision, a considerable increase is recorded, showing that the majority of the mentions that the system indicates as coreferent

have the correct mention spans. Additionally, the problem of selecting the correct span (as described in Section 2) is another factor that has a considerable effect on precision at that stage – mentions that were accurately attached to the correct coreference chain are not considered if their span is not identical to the span of their counterparts in the key set.

Automatic Mention Detection In the first part in Table 6, we show the system scores for UBIU’s performance when no mention information is provided in the data. We report both gold (using gold linguistic annotations) and auto (using automatically annotated data) settings. A comparison of the results shows that there are only minor differences between them with gold outperforming auto apart from Arabic for which there is a drop of 3.75 points in the gold setting. However, the small difference between all results shows that the quality of the automatic annotation is good enough for a CR system and that further improvements in the quality of the linguistic information will not necessarily improve CR.

If we compare results across languages, we see that Arabic has the lowest results. One of the reasons for this decreased performance can be found in the NP-rich syntactic structure of Arabic. This leads to a high number of identified mentions and in combination with the longer sentence length to a higher

number of training/test instances. Another reason for the drop in performance for Arabic can be found in the lack of annotations expected by our system (named entities and predicted arguments) that were not provided by the task due to time constraints and the accuracy of the annotations. Further, Arabic is a morphologically rich language for which only the simplified standard POS tags were provided and not the gold standard ones that contain much richer and thus more helpful morphology information.

The results for Chinese and English are relatively close. We can also see that the $CEAF_E$ results are extremely close, with a difference of less than 1%. MUC, in contrast, shows the largest differences with more than 30% between Arabic and English in the gold setting. It is also noteworthy that the results for English show a balance between precision and recall while both Arabic and Chinese favor precision over recall in terms of mention detection, MUC, and B^3 . The reasons for this difference between languages need to be investigated further.

Gold Mention Boundaries The results for this set of experiments is based on a version of the test set that contains the gold boundaries of all mentions, including singletons. Thus, we use these gold mention boundaries instead of the ones generated by our system. These experiments give us an insight on how well UBIU performs on selecting the correct boundaries. Since we do not expect the system’s selection to be perfect, we would expect to see improved system performance given the correct boundaries. The results are shown in the second part of Table 6. As for using automatically generated mentions the tendencies in scores between gold and auto linguistic annotations are kept. A further comparison of the overall results between the two settings also shows only minor changes. The only exception is the auto setting for Arabic, for which we see drop in MD precision of approximately 5%. This also results in lower MUC and B^3 precision and $CEAF_E$ recall. The reasons for this drop in performance need to be investigated further. The fact that most results for both auto and gold settings change only slightly shows that having information about the correct mention boundaries is not very helpful. Thus, the system seems to have reached its optimal performance on selecting mention boundaries given the

information that it has.

Gold Mentions The last set of experiments is based on a version of the test set that contains the gold mentions, i.e., all mentions that are coreferent, but without any information about the identity of the coreference chains. The results of this set of experiments gives us information about the quality of the coreference classifier. The results are shown in the third part of Table 6. Using gold parses leads to only minor improvement of the overall system performance, yet, in that case all languages, including Arabic, show consistent increase of results. Altogether, there is a major improvement of the scores in MD, MUC, and $CEAF_E$. The B^3 scores only show minor improvements, resulting from a slight drop in precision across languages. The results also show considerably higher precision than recall for MUC and B^3 , and higher recall for $CEAF_E$. This means that the coreference decisions that the system makes are highly reliable but that it still has a preference for treating coreferent mentions as singletons.

A comparison across languages shows that providing gold mentions has a considerable positive effect on the system performance for Arabic since for that setting Chinese leads to lower overall scores. We assume that this is again due to the NP-rich syntactic structure of Arabic and the fact that providing the mentions decreases drastically the number of mentions the system works with and has to choose from during the resolution process.

4 Conclusion and Future Work

We presented the UBIU system for coreference resolution in a multilingual setting. The system performed reliably across all three languages of the CoNLL 2012 shared task. For the future, we are planning an in-depth investigation of the performance of the mention detection module and the singleton classifier, as well as in investigation into more complex models for coreference classification than the mention pair model.

Acknowledgments

This work is based on research supported by the US Office of Naval Research (ONR) Grant #N00014-10-1-0140. We would also like to thank Kiran Kumar for his help with tuning the system.

References

- Adriane Boyd, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential it: A machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, FeatureEng '05, pages 40–47, Ann Arbor, MI.
- Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanolini. 2010. BART: A Multilingual Anaphora Resolution System. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 104–107, Uppsala, Sweden.
- Walter Daelemans and Antal van den Bosch. 2005. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.
- Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2010. TiMBL: Tilburg Memory Based Learner, version 6.3, reference guide. Technical Report ILK 10-01, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.
- Richard Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45 – 57.
- Sanda M. Harabagiu and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of ANLP 2000*, Seattle, WA.
- Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford’s multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, OR.
- Xiaoqiang Luo and Imed Zitouni. 2005. Multi-Lingual Coreference Resolution with Syntactic Features. In *Proceedings of HLT/EMNLP 2005*, Vancouver, Canada.
- Ruslan Mitkov. 1999. Multilingual anaphora resolution. *Machine Translation*, 14(3-4):281–299.
- Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings COLING '02*, pages 1–7, Taipei, Taiwan.
- Vincent Ng. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Barcelona, Spain.
- 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.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.
- Ataf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP*, pages 968–977, Singapore.
- 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*, pages 1–8, 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):521–544.
- 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 the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala, Sweden.