SemEval-2010 Task 1: Coreference Resolution in Multiple Languages

Marta Recasens[⋆] **Lluís Màrquez**[†] **Emili Sapena**[†] **M. Antònia Martí**[⋆]
Mariona Taulé[⋆] **Véronique Hoste**[‡] **Massimo Poesio**[◇] **Yannick Versley**[⋆⋆]

⋆: CLiC, University of Barcelona, {mrecasens,amarti,mtaule}@ub.edu
†: TALP, Technical University of Catalonia, {lluism,esapena}@lsi.upc.edu
‡: University College Ghent, veronique.hoste@hogent.be
◇: University of Essex/University of Trento, poesio@essex.ac.uk
⋆⋆: University of Tübingen, versley@sfs.uni-tuebingen.de

Abstract

This paper presents the SemEval-2010 task on *Coreference Resolution in Multiple Languages*. The goal was to evaluate and compare automatic coreference resolution systems for six different languages (Catalan, Dutch, English, German, Italian, and Spanish) in four evaluation settings and using four different metrics. Such a rich scenario had the potential to provide insight into key issues concerning coreference resolution: (i) the portability of systems across languages, (ii) the relevance of different levels of linguistic information, and (iii) the behavior of scoring metrics.

1 Introduction

The task of coreference resolution, defined as the identification of the expressions in a text that refer to the same discourse entity (1), has attracted considerable attention within the NLP community.

(1) *Major League Baseball* sent *its* head of security to Chicago to review the second incident of an on-field fan attack in the last seven months. *The league* is reviewing security at all ballparks to crack down on spectator violence.

Using coreference information has been shown to be beneficial in a number of NLP applications including Information Extraction (McCarthy and Lehnert, 1995), Text Summarization (Steinberger et al., 2007), Question Answering (Morton, 1999), and Machine Translation. There have been a few evaluation campaigns on coreference resolution in the past, namely MUC (Hirschman and Chinchor, 1997), ACE (Doddington et al., 2004), and ARE (Orasan et al., 2008), yet many questions remain open:

- To what extent is it possible to implement a general coreference resolution system portable to different languages? How much language-specific tuning is necessary?
- How helpful are morphology, syntax and semantics for solving coreference relations? How much preprocessing is needed? Does its quality (perfect linguistic input versus noisy automatic input) really matter?
- How (dis)similar are different coreference evaluation metrics—MUC, B-CUBED, CEAF and BLANC? Do they all provide the same ranking? Are they correlated?

Our goal was to address these questions in a shared task. Given six datasets in Catalan, Dutch, English, German, Italian, and Spanish, the task we present involved automatically detecting full coreference chains—composed of named entities (NEs), pronouns, and full noun phrases—in four different scenarios. For more information, the reader is referred to the task website.[1]

The rest of the paper is organized as follows. Section 2 presents the corpora from which the task datasets were extracted, and the automatic tools used to preprocess them. In Section 3, we describe the task by providing information about the data format, evaluation settings, and evaluation metrics. Participating systems are described in Section 4, and their results are analyzed and compared in Section 5. Finally, Section 6 concludes.

2 Linguistic Resources

In this section, we first present the sources of the data used in the task. We then describe the automatic tools that predicted input annotations for the coreference resolution systems.

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

		Training			Development			Test	
	#docs	#sents	#tokens	#docs	#sents	#tokens	#docs	#sents	#tokens
Catalan	829	8,709	253,513	142	1,445	42,072	167	1,698	49,260
Dutch	145	2,544	46,894	23	496	9,165	72	2,410	48,007
English	229	3,648	79,060	39	741	17,044	85	1,141	24,206
German	900	19,233	331,614	199	4,129	73,145	136	2,736	50,287
Italian	80	2,951	81,400	17	551	16,904	46	1,494	41,586
Spanish	875	9,022	284,179	140	1,419	44,460	168	1,705	51,040

Table 1: Size of the task datasets.

2.1 Source Corpora

Catalan and Spanish The AnCora corpora (Recasens and Martí, 2009) consist of a Catalan and a Spanish treebank of 500k words each, mainly from newspapers and news agencies (El Periódico, EFE, ACN). Manual annotation exists for arguments and thematic roles, predicate semantic classes, NEs, WordNet nominal senses, and coreference relations. AnCora are freely available for research purposes.

Dutch The KNACK-2002 corpus (Hoste and De Pauw, 2006) contains 267 documents from the Flemish weekly magazine Knack. They were manually annotated with coreference information on top of semi-automatically annotated PoS tags, phrase chunks, and NEs.

English The OntoNotes Release 2.0 corpus (Pradhan et al., 2007) covers newswire and broadcast news data: 300k words from The Wall Street Journal, and 200k words from the TDT-4 collection, respectively. OntoNotes builds on the Penn Treebank for syntactic annotation and on the Penn PropBank for predicate argument structures. Semantic annotations include NEs, words senses (linked to an ontology), and coreference information. The OntoNotes corpus is distributed by the Linguistic Data Consortium.[2]

German The TüBa-D/Z corpus (Hinrichs et al., 2005) is a newspaper treebank based on data taken from the daily issues of “die tageszeitung” (taz). It currently comprises 794k words manually annotated with semantic and coreference information. Due to licensing restrictions of the original texts, a taz-DVD must be purchased to obtain a license.[2]

Italian The LiveMemories corpus (Rodríguez et al., 2010) will include texts from the Italian Wikipedia, blogs, news articles, and dialogues (MapTask). They are being annotated according to the ARRAU annotation scheme with coreference, agreement, and NE information on top of automatically parsed data. The task dataset included Wikipedia texts already annotated.

The datasets that were used in the task were extracted from the above-mentioned corpora. Table 1 summarizes the number of documents (docs), sentences (sents), and tokens in the training, development and test sets.[3]

2.2 Preprocessing Systems

Catalan, Spanish, English Predicted lemmas and PoS were generated using FreeLing[4] for Catalan/Spanish and SVMTagger[5] for English. Dependency information and predicate semantic roles were generated with JointParser, a syntactic-semantic parser.[6]

Dutch Lemmas, PoS and NEs were automatically provided by the memory-based shallow parser for Dutch (Daelemans et al., 1999), and dependency information by the Alpino parser (van Noord et al., 2006).

German Lemmas were predicted by TreeTagger (Schmid, 1995), PoS and morphology by RFTagger (Schmid and Laws, 2008), and dependency information by MaltParser (Hall and Nivre, 2008).

Italian Lemmas and PoS were provided by TextPro,[7] and dependency information by MaltParser.[8]

[2] Free user license agreements for the English and German task datasets were issued to the task participants.

[3] The German and Dutch training datasets were not completely stable during the competition period due to a few errors. Revised versions were released on March 2 and 20, respectively. As to the test datasets, the Dutch and Italian documents with formatting errors were corrected after the evaluation period, with no variations in the ranking order of systems.

[4] http://www.lsi.upc.es/ nlp/freeling

[5] http://www.lsi.upc.edu/ nlp/SVMTool

[6] http://www.lsi.upc.edu// xlluis/?x=cat:5

[7] http://textpro.fbk.eu

[8] http://maltparser.org

3 Task Description

Participants were asked to develop an automatic system capable of assigning a discourse entity to every mention,[9] thus identifying all the NP mentions of every discourse entity. As there is no standard annotation scheme for coreference and the source corpora differed in certain aspects, the coreference information of the task datasets was produced according to three criteria:

- Only NP constituents and possessive determiners can be mentions.
- Mentions must be referential expressions, thus ruling out nominal predicates, appositives, expletive NPs, attributive NPs, NPs within idioms, etc.
- Singletons are also considered as entities (i.e., entities with a single mention).

To help participants build their systems, the task datasets also contained both gold-standard and automatically predicted linguistic annotations at the morphological, syntactic and semantic levels. Considerable effort was devoted to provide participants with a common and relatively simple data representation for the six languages.

3.1 Data Format

The task datasets as well as the participants' answers were displayed in a uniform column-based format, similar to the style used in previous CoNLL shared tasks on syntactic and semantic dependencies (2008/2009).[10] Each dataset was provided as a single file per language. Since coreference is a linguistic relation at the discourse level, documents constitute the basic unit, and are delimited by “#begin document ID” and “#end document ID” comment lines. Within a document, the information of each sentence is organized vertically with one token per line, and a blank line after the last token of each sentence. The information associated with each token is described in several columns (separated by “\t” characters) representing the following layers of linguistic annotation.

ID (column 1). Token identifiers in the sentence.

Token (column 2). Word forms.

ID	Token	Intermediate columns	Coref
1	Major	...	(1
2	League	...	_
3	Baseball	...	1)
4	sent	...	_
5	its	...	(1)|(2
6	head	...	_
7	of	...	_
8	security	...	(3)|2)
9	to	...	_
...	...	...	...
27	The	...	(1
28	league	...	1)
29	is	...	_

Table 2: Format of the coreference annotations (corresponding to example (1) in Section 1).

Lemma (column 3). Token lemmas.

PoS (column 5). Coarse PoS.

Feat (column 7). Morphological features (PoS type, number, gender, case, tense, aspect, etc.) separated by a pipe character.

Head (column 9). ID of the syntactic head (“0” if the token is the tree root).

DepRel (column 11). Dependency relations corresponding to the dependencies described in the Head column (“sentence” if the token is the tree root).

NE (column 13). NE types in open-close notation.

Pred (column 15). Predicate semantic class.

APreds (column 17 and subsequent ones). For each predicate in the Pred column, its semantic roles/dependencies.

Coref (last column). Coreference relations in open-close notation.

The above-mentioned columns are “gold-standard columns,” whereas columns 4, 6, 8, 10, 12, 14, 16 and the penultimate contain the same information as the respective previous column but automatically predicted—using the preprocessing systems listed in Section 2.2. Neither all layers of linguistic annotation nor all gold-standard and predicted columns were available for all six languages (underscore characters indicate missing information).

The coreference column follows an open-close notation with an entity number in parentheses (see Table 2). Every entity has an ID number, and every mention is marked with the ID of the entity it refers to: an opening parenthesis shows the beginning of the mention (first token), while a closing parenthesis shows the end of the mention (last

[9] Following the terminology of the ACE program, a *mention* is defined as an instance of reference to an object, and an *entity* is the collection of mentions referring to the same object in a document.

[10] http://www.cnts.ua.ac.be/conll2008

token). For tokens belonging to more than one mention, a pipe character is used to separate multiple entity IDs. The resulting annotation is a well-formed nested structure (CF language).

3.2 Evaluation Settings

In order to address our goal of studying the effect of different levels of linguistic information (preprocessing) on solving coreference relations, the test was divided into four evaluation settings that differed along two dimensions.

Gold-standard versus **Regular setting.** Only in the gold-standard setting were participants allowed to use the gold-standard columns, including the last one (of the test dataset) with true mention boundaries. In the regular setting, they were allowed to use only the automatically predicted columns. Obtaining better results in the gold setting would provide evidence for the relevance of using high-quality preprocessing information. Since not all columns were available for all six languages, the gold setting was only possible for Catalan, English, German, and Spanish.

Closed versus **Open setting.** In the closed setting, systems had to be built strictly with the information provided in the task datasets. In contrast, there was no restriction on the resources that participants could utilize in the open setting: systems could be developed using any external tools and resources to predict the preprocessing information, e.g., WordNet, Wikipedia, etc. The only requirement was to use tools that had not been developed with the annotations of the test set. This setting provided an open door into tools or resources that improve performance.

3.3 Evaluation Metrics

Since there is no agreement at present on a standard measure for coreference resolution evaluation, one of our goals was to compare the rankings produced by four different measures. The task scorer provides results in the two mention-based metrics B^3 (Bagga and Baldwin, 1998) and CEAF-ϕ_3 (Luo, 2005), and the two link-based metrics MUC (Vilain et al., 1995) and BLANC (Recasens and Hovy, in prep). The first three measures have been widely used, while BLANC is a proposal of a new measure interesting to test.

The mention detection subtask is measured with recall, precision, and F_1. Mentions are rewarded with 1 point if their boundaries coincide with those of the gold NP, with 0.5 points if their boundaries are within the gold NP including its head, and with 0 otherwise.

4 Participating Systems

A total of twenty-two participants registered for the task and downloaded the training materials. From these, sixteen downloaded the test set but only six (out of which two task organizers) submitted valid results (corresponding to nine system runs or variants). These numbers show that the task raised considerable interest but that the final participation rate was comparatively low (slightly below 30%).

The participating systems differed in terms of architecture, machine learning method, etc. Table 3 summarizes their main properties. Systems like BART and Corry support several machine learners, but Table 3 indicates the one used for the SemEval run. The last column indicates the external resources that were employed in the open setting, thus it is empty for systems that participated only in the closed setting. For more specific details we address the reader to the system description papers in Erk and Strapparava (2010).

5 Results and Evaluation

Table 4 shows the results obtained by two naive baseline systems: (i) SINGLETONS considers each mention as a separate entity, and (ii) ALL-IN-ONE groups all the mentions in a document into a single entity. These simple baselines reveal limitations of the evaluation metrics, like the high scores of CEAF and B^3 for SINGLETONS. Interestingly enough, the naive baseline scores turn out to be hard to beat by the participating systems, as Table 5 shows. Similarly, ALL-IN-ONE obtains high scores in terms of MUC. Table 4 also reveals differences between the distribution of entities in the datasets. Dutch is clearly the most divergent corpus mainly due to the fact that it only contains singletons for NEs.

Table 5 displays the results of all systems for all languages and settings in the four evaluation metrics (the best scores in each setting are highlighted in bold). Results are presented sequentially by language and setting, and participating systems are ordered alphabetically. The participation of systems across languages and settings is rather irregular,[11] thus making it difficult to draw firm conclu-

[11] Only 45 entries in Table 5 from 192 potential cases.

	System Architecture	ML Methods	External Resources
BART (Broscheit et al., 2010)	Closest-first with entity-mention model (English), Closest-first model (German, Italian)	MaxEnt (English, German), Decision trees (Italian)	GermaNet & gazetteers (German), I-Cab gazetteers (Italian), Berkeley parser, Stanford NER, WordNet, Wikipedia name list, U.S. census data (English)
Corry (Uryupina, 2010)	ILP, Pairwise model	SVM	Stanford parser & NER, WordNet, U.S. census data
RelaxCor (Sapena et al., 2010)	Graph partitioning (solved by relaxation labeling)	Decision trees, Rules	WordNet
SUCRE (Kobdani and Schütze, 2010)	Best-first clustering, Relational database model, Regular feature definition language	Decision trees, Naive Bayes, SVM, MaxEnt	—
TANL-1 (Attardi et al., 2010)	Highest entity-mention similarity	MaxEnt	PoS tagger (Italian)
UBIU (Zhekova and Kübler, 2010)	Pairwise model	MBL	—

Table 3: Main characteristics of the participating systems.

sions about the aims initially pursued by the task. In the following, we summarize the most relevant outcomes of the evaluation.

Regarding languages, English concentrates the most participants (fifteen entries), followed by German (eight), Catalan and Spanish (seven each), Italian (five), and Dutch (three). The number of languages addressed by each system ranges from one (Corry) to six (UBIU and SUCRE); BART and RelaxCor addressed three languages, and TANL-1 five. The best overall results are obtained for English followed by German, then Catalan, Spanish and Italian, and finally Dutch. Apart from differences between corpora, there are other factors that might explain this ranking: (i) the fact that most of the systems were originally developed for English, and (ii) differences in corpus size (German having the largest corpus, and Dutch the smallest).

Regarding systems, there are no clear "winners." Note that no language-setting was addressed by all six systems. The BART system, for instance, is either on its own or competing against a single system. It emerges from partial comparisons that SUCRE performs the best in *closed*×*regular* for English, German, and Italian, although it never outperforms the CEAF or B^3 singleton baseline. While SUCRE always obtains the best scores according to MUC and BLANC, RelaxCor and TANL-1 usually win based on CEAF and B^3. The Corry system presents three variants optimized for CEAF (Corry-C), MUC (Corry-M), and BLANC (Corry-B). Their results are consistent with the bias introduced in the optimization (see English:*open*×*gold*).

Depending on the evaluation metric then, the rankings of systems vary with considerable score differences. There is a significant positive correlation between CEAF and B^3 (Pearson's $r = 0.91$, $p < 0.01$), and a significant lack of correlation between CEAF and MUC in terms of recall (Pearson's $r = 0.44$, $p < 0.01$). This fact stresses the importance of defining appropriate metrics (or a combination of them) for coreference evaluation.

Finally, regarding evaluation settings, the results in the *gold* setting are significantly better than those in the *regular*. However, this might be a direct effect of the mention recognition task. Mention recognition in the regular setting falls more than 20 F_1 points with respect to the gold setting (where correct mention boundaries were given). As for the *open* versus *closed* setting, there is only one system, RelaxCor for English, that addressed the two. As expected, results show a slight improvement from *closed*×*gold* to *open*×*gold*.

6 Conclusions

This paper has introduced the main features of the SemEval-2010 task on coreference resolution.

	CEAF			MUC			B^3			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
SINGLETONS: Each mention forms a separate entity.												
Catalan	61.2	61.2	61.2	0.0	0.0	0.0	61.2	100	75.9	50.0	48.7	49.3
Dutch	34.5	34.5	34.5	0.0	0.0	0.0	34.5	100	51.3	50.0	46.7	48.3
English	71.2	71.2	71.2	0.0	0.0	0.0	71.2	100	83.2	50.0	49.2	49.6
German	75.5	75.5	75.5	0.0	0.0	0.0	75.5	100	86.0	50.0	49.4	49.7
Italian	71.1	71.1	71.1	0.0	0.0	0.0	71.1	100	83.1	50.0	49.2	49.6
Spanish	62.2	62.2	62.2	0.0	0.0	0.0	62.2	100	76.7	50.0	48.8	49.4
ALL-IN-ONE: All mentions are grouped into a single entity.												
Catalan	11.8	11.8	11.8	100	39.3	56.4	100	4.0	7.7	50.0	1.3	2.6
Dutch	19.7	19.7	19.7	100	66.3	79.8	100	8.0	14.9	50.0	3.2	6.2
English	10.5	10.5	10.5	100	29.2	45.2	100	3.5	6.7	50.0	0.8	1.6
German	8.2	8.2	8.2	100	24.8	39.7	100	2.4	4.7	50.0	0.6	1.1
Italian	11.4	11.4	11.4	100	29.0	45.0	100	2.1	4.1	50.0	0.8	1.5
Spanish	11.9	11.9	11.9	100	38.3	55.4	100	3.9	7.6	50.0	1.2	2.4

Table 4: Baseline scores.

The goal of the task was to evaluate and compare automatic coreference resolution systems for six different languages in four evaluation settings and using four different metrics. This complex scenario aimed at providing insight into several aspects of coreference resolution, including portability across languages, relevance of linguistic information at different levels, and behavior of alternative scoring metrics.

The task attracted considerable attention from a number of researchers, but only six teams submitted their final results. Participating systems did not run their systems for all the languages and evaluation settings, thus making direct comparisons between them very difficult. Nonetheless, we were able to observe some interesting aspects from the empirical evaluation.

An important conclusion was the confirmation that different evaluation metrics provide different system rankings and the scores are not commensurate. Attention thus needs to be paid to coreference evaluation. The behavior and applicability of the scoring metrics requires further investigation in order to guarantee a fair evaluation when comparing systems in the future. We hope to have the opportunity to thoroughly discuss this and the rest of interesting questions raised by the task during the SemEval workshop at ACL 2010.

An additional valuable benefit is the set of resources developed throughout the task. As task organizers, we intend to facilitate the sharing of datasets, scorers, and documentation by keeping them available for future research use. We believe that these resources will help to set future benchmarks for the research community and will contribute positively to the progress of the state of the art in coreference resolution. We will maintain and update the task website with post-SemEval contributions.

Acknowledgments

We would like to thank the following people who contributed to the preparation of the task datasets: Manuel Bertran (UB), Oriol Borrega (UB), Orphée De Clercq (U. Ghent), Francesca Delogu (U. Trento), Jesús Giménez (UPC), Eduard Hovy (ISI-USC), Richard Johansson (U. Trento), Xavier Lluís (UPC), Montse Nofre (UB), Lluís Padró (UPC), Kepa Joseba Rodríguez (U. Trento), Mihai Surdeanu (Stanford), Olga Uryupina (U. Trento), Lente Van Leuven (UB), and Rita Zaragoza (UB). We would also like to thank LDC and die tageszeitung for distributing freely the English and German datasets.

This work was funded in part by the Spanish Ministry of Science and Innovation through the projects TEXT-MESS 2.0 (TIN2009-13391-C04-04), OpenMT-2 (TIN2009-14675-C03), and KNOW2 (TIN2009-14715-C04-04), and an FPU doctoral scholarship (AP2006-00994) held by M. Recasens. It also received financial support from the Seventh Framework Programme of the EU (FP7/2007-2013) under GA 247762 (FAUST), from the STEVIN program of the Nederlandse Taalunie through the COREA and SoNaR projects, and from the Provincia Autonoma di Trento through the LiveMemories project.

	Mention detection			CEAF			MUC			B^3			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
Catalan															
closed×gold															
RelaxCor	100	100	100	70.5	70.5	**70.5**	29.3	77.3	42.5	68.6	95.8	**79.9**	56.0	81.8	59.7
SUCRE	100	100	100	68.7	68.7	68.7	54.1	58.4	**56.2**	76.6	77.4	77.0	72.4	60.2	**63.6**
TANL-1	100	96.8	98.4	66.0	63.9	64.9	17.2	57.7	26.5	64.4	93.3	76.2	52.8	79.8	54.4
UBIU	75.1	96.3	84.4	46.6	59.6	52.3	8.8	17.1	11.7	47.8	76.3	58.8	51.6	57.9	52.2
closed×regular															
SUCRE	75.9	64.5	69.7	51.3	43.6	47.2	44.1	32.3	**37.3**	59.6	44.7	51.1	53.9	55.2	**54.2**
TANL-1	83.3	82.0	82.7	57.5	56.6	**57.1**	15.2	46.9	22.9	55.8	76.6	**64.6**	51.3	76.2	51.0
UBIU	51.4	70.9	59.6	33.2	45.7	38.4	6.5	12.6	8.6	32.4	55.7	40.9	50.2	53.7	47.8
open×gold															
open×regular															
Dutch															
closed×gold															
SUCRE	100	100	100	58.8	58.8	**58.8**	65.7	74.4	**69.8**	65.0	69.2	**67.0**	69.5	62.9	**65.3**
closed×regular															
SUCRE	78.0	29.0	42.3	29.4	10.9	15.9	62.0	19.5	**29.7**	59.1	6.5	11.7	46.9	46.9	**46.9**
UBIU	41.5	29.9	34.7	20.5	14.6	**17.0**	6.7	11.0	8.3	13.3	23.4	**17.0**	50.0	52.4	32.3
open×gold															
open×regular															
English															
closed×gold															
RelaxCor	100	100	100	75.6	75.6	**75.6**	21.9	72.4	33.7	74.8	97.0	**84.5**	57.0	83.4	61.3
SUCRE	100	100	100	74.3	74.3	74.3	68.1	54.9	**60.8**	86.7	78.5	82.4	77.3	67.0	**70.8**
TANL-1	99.8	81.7	89.8	75.0	61.4	67.6	23.7	24.4	24.0	74.6	72.1	73.4	51.8	68.8	52.1
UBIU	92.5	99.5	95.9	63.4	68.2	65.7	17.2	25.5	20.5	67.8	83.5	74.8	52.6	60.8	54.0
closed×regular															
SUCRE	78.4	83.0	80.7	61.0	64.5	**62.7**	57.7	48.1	**52.5**	68.3	65.9	**67.1**	58.9	65.7	**61.2**
TANL-1	79.6	68.9	73.9	61.7	53.4	57.3	23.8	25.5	24.6	62.1	60.5	61.3	50.9	68.0	49.3
UBIU	66.7	83.6	74.2	48.2	60.4	53.6	11.6	18.4	14.2	50.9	69.2	58.7	50.9	56.3	51.0
open×gold															
Corry-B	100	100	100	77.5	77.5	77.5	56.1	57.5	56.8	82.6	85.7	84.1	69.3	75.3	**71.8**
Corry-C	100	100	100	77.7	77.7	**77.7**	57.4	58.3	57.9	83.1	84.7	83.9	71.3	71.6	71.5
Corry-M	100	100	100	73.8	73.8	73.8	62.5	56.2	**59.2**	85.5	78.6	81.9	76.2	58.8	62.7
RelaxCor	100	100	100	75.8	75.8	75.8	22.6	70.5	34.2	75.2	96.7	**84.6**	58.0	83.8	62.7
open×regular															
BART	76.1	69.8	72.8	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	79.8	76.4	78.1	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	79.8	76.4	78.1	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	79.8	76.4	78.1	66.3	63.5	64.8	61.5	53.4	**57.2**	76.8	66.5	71.3	58.5	56.2	57.1
German															
closed×gold															
SUCRE	100	100	100	72.9	72.9	72.9	74.4	48.1	**58.4**	90.4	73.6	81.1	78.2	61.8	**66.4**
TANL-1	100	100	100	77.7	77.7	**77.7**	16.4	60.6	25.9	77.2	96.7	**85.9**	54.4	75.1	57.4
UBIU	92.6	95.5	94.0	67.4	68.9	68.2	22.1	21.7	21.9	73.7	77.9	75.7	60.0	77.2	64.5
closed×regular															
SUCRE	79.3	77.5	78.4	60.6	59.2	**59.9**	49.3	35.0	**40.9**	69.1	60.1	**64.3**	52.7	59.3	**53.6**
TANL-1	60.9	57.7	59.2	50.9	48.2	49.5	10.2	31.5	15.4	47.2	54.9	50.7	50.2	63.0	44.7
UBIU	50.6	66.8	57.6	39.4	51.9	44.8	9.5	11.4	10.4	41.2	53.7	46.6	50.2	54.4	48.0
open×gold															
BART	94.3	93.7	94.0	67.1	66.7	**66.9**	70.5	40.1	**51.1**	85.3	64.4	**73.4**	65.5	61.0	**62.8**
open×regular															
BART	82.5	82.3	82.4	61.4	61.2	**61.3**	61.4	36.1	**45.5**	75.3	58.3	**65.7**	55.9	60.3	**57.3**
Italian															
closed×gold															
SUCRE	98.4	98.4	98.4	66.0	66.0	**66.0**	48.1	42.3	**45.0**	76.7	76.9	**76.8**	54.8	63.5	**56.9**
closed×regular															
SUCRE	84.6	98.1	90.8	57.1	66.2	**61.3**	50.1	50.7	**50.4**	63.6	79.2	**70.6**	55.2	68.3	**57.7**
UBIU	46.8	35.9	40.6	37.9	29.0	32.9	2.9	4.6	3.6	38.4	31.9	34.8	50.0	46.6	37.2
open×gold															
open×regular															
BART	42.8	80.7	55.9	35.0	66.1	45.8	35.3	54.0	**42.7**	34.6	70.6	46.4	57.1	68.1	**59.6**
TANL-1	90.5	73.8	81.3	62.2	50.7	**55.9**	37.2	28.3	32.1	66.8	56.5	**61.2**	50.7	69.3	48.5
Spanish															
closed×gold															
RelaxCor	100	100	100	66.6	66.6	66.6	14.8	73.8	24.7	65.3	97.5	**78.2**	53.4	81.8	55.6
SUCRE	100	100	100	69.8	69.8	**69.8**	52.7	58.3	**55.3**	75.8	79.0	77.4	67.3	62.5	**64.5**
TANL-1	100	96.8	98.4	66.9	64.7	65.8	16.6	56.5	25.7	65.2	93.4	76.8	52.5	79.0	54.1
UBIU	73.8	96.4	83.6	45.7	59.6	51.7	9.6	18.8	12.7	46.8	77.1	58.3	52.9	63.9	54.3
closed×regular															
SUCRE	74.9	66.3	70.3	56.3	49.9	52.9	35.8	36.8	**36.3**	56.6	54.6	55.6	52.1	61.2	**51.4**
TANL-1	82.2	84.1	83.1	58.6	60.0	**59.3**	14.0	48.4	21.7	56.6	79.0	**66.0**	51.4	74.7	**51.4**
UBIU	51.1	72.7	60.0	33.6	47.6	39.4	7.6	14.4	10.0	32.8	57.1	41.6	50.4	54.6	48.4
open×gold															
open×regular															

Table 5: Official results of the participating systems for all languages, settings, and metrics.

References

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi. 2010. TANL-1: coreference resolution by parse analysis and similarity clustering. In *Proceedings of SemEval-2*.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC Workshop on Linguistic Coreference*, pages 563–566.

Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodríguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. 2010. BART: A multilingual anaphora resolution system. In *Proceedings of SemEval-2*.

Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. 1999. Memory-based shallow parsing. In *Proceedings of CoNLL 1999*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program – Tasks, data, and evaluation. In *Proceedings of LREC 2004*, pages 837–840.

Katrin Erk and Carlo Strapparava, editors. 2010. *Proceedings of SemEval-2*.

Johan Hall and Joakim Nivre. 2008. A dependency-driven parser for German dependency and constituency representations. In *Proceedings of the ACL Workshop on Parsing German (PaGe 2008)*, pages 47–54.

Erhard W. Hinrichs, Sandra Kübler, and Karin Naumann. 2005. A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 13–20.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 Coreference Task Definition – Version 3.0. In *Proceedings of MUC-7*.

Véronique Hoste and Guy De Pauw. 2006. KNACK-2002: A richly annotated corpus of Dutch written text. In *Proceedings of LREC 2006*, pages 1432–1437.

Hamidreza Kobdani and Hinrich Schütze. 2010. SUCRE: A modular system for coreference resolution. In *Proceedings of SemEval-2*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP 2005*, pages 25–32.

Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of IJCAI 1995*, pages 1050–1055.

Thomas S. Morton. 1999. Using coreference in question answering. In *Proceedings of TREC-8*, pages 85–89.

Constantin Orasan, Dan Cristea, Ruslan Mitkov, and António Branco. 2008. Anaphora Resolution Exercise: An overview. In *Proceedings of LREC 2008*.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *Proceedings of the International Conference on Semantic Computing (ICSC 2007)*, pages 517–526.

Marta Recasens and Eduard Hovy. in prep. BLANC: Implementing the Rand Index for Coreference Evaluation.

Marta Recasens and M. Antònia Martí. 2009. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, DOI:10.1007/s10579-009-9108-x.

Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. 2010. Anaphoric annotation of Wikipedia and blogs in the Live Memories Corpus. In *Proceedings of LREC 2010*, pages 157–163.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2010. RelaxCor: A global relaxation labeling approach to coreference resolution for the SemEval-2 Coreference Task. In *Proceedings of SemEval-2*.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*, pages 777–784.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management: an International Journal*, 43(6):1663–1680.

Olga Uryupina. 2010. Corry: A system for coreference resolution. In *Proceedings of SemEval-2*.

Gertjan van Noord, Ineke Schuurman, and Vincent Vandeghinste. 2006. Syntactic annotation of large corpora in STEVIN. In *Proceedings of LREC 2006*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of MUC-6*, pages 45–52.

Desislava Zhekova and Sandra Kübler. 2010. UBIU: A language-independent system for coreference resolution. In *Proceedings of SemEval-2*.