Coreference Resolution across Corpora: Languages, Coding Schemes, and Preprocessing Information

Marta Recasens
CLiC - University of Barcelona
Gran Via 585
Barcelona, Spain
mrecasens@ub.edu

Eduard Hovy
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey CA, USA
hovy@isi.edu

Abstract

This paper explores the effect that different corpus configurations have on the performance of a coreference resolution system, as measured by MUC, B^3, and CEAF. By varying separately three parameters (language, annotation scheme, and preprocessing information) and applying the same coreference resolution system, the strong bonds between system and corpus are demonstrated. The experiments reveal problems in coreference resolution evaluation relating to task definition, coding schemes, and features. They also expose systematic biases in the coreference evaluation metrics. We show that system comparison is only possible when corpus parameters are in exact agreement.

1 Introduction

The task of coreference resolution, which aims to automatically identify the expressions in a text that refer to the same discourse entity, has been an increasing research topic in NLP ever since MUC-6 made available the first coreferentially annotated corpus in 1995. Most research has centered around the rules by which mentions are allowed to corefer, the features characterizing mention pairs, the algorithms for building coreference chains, and coreference evaluation methods. The surprisingly important role played by different aspects of the corpus, however, is an issue to which little attention has been paid. We demonstrate the extent to which a system will be evaluated as performing differently depending on parameters such as the corpus language, the way coreference relations are defined in the corresponding coding scheme, and the nature and source of preprocessing information.

This paper unpacks these issues by running the same system—a prototype entity-based architecture called CISTELL—on different corpus configurations, varying three parameters. First, we show how much language-specific issues affect performance when trained and tested on English and Spanish. Second, we demonstrate the extent to which the specific annotation scheme (used on the same corpus) makes evaluated performance vary. Third, we compare the performance using gold-standard preprocessing information with that using automatic preprocessing tools.

Throughout, we apply the three principal coreference evaluation measures in use today: MUC, B^3, and CEAF. We highlight the systematic preferences of each measure to reward different configurations. This raises the difficult question of why one should use one or another evaluation measure, and how one should interpret their differences in reporting changes of performance score due to 'secondary' factors like preprocessing information.

To this end, we employ three corpora: ACE (Doddington et al., 2004), OntoNotes (Pradhan et al., 2007), and AnCora (Recasens and Martí, 2009). In order to isolate the three parameters as far as possible, we benefit from a 100k-word portion (from the TDT collection) that is common to both ACE and OntoNotes. We apply the same coreference resolution system in all cases. The results show that a system's score is not informative by itself, as different corpora or corpus parameters lead to different scores. Our goal is not to achieve the best performance to date, but rather to expose various issues raised by the choices of corpus preparation and evaluation measure and to shed light on the definition, methods, evaluation, and complexities of the coreference resolution task.

The paper is organized as follows. Section 2 sets our work in context and provides the motivations for undertaking this study. Section 3 presents the architecture of CISTELL, the system used in the experimental evaluation. In Sections 4, 5,

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1423–1432,
Uppsala, Sweden, 11-16 July 2010.

and 6, we describe the experiments on three different datasets and discuss the results. We conclude in Section 7.

2 Background

The bulk of research on automatic coreference resolution to date has been done for English and used two different types of corpus: MUC (Hirschman and Chinchor, 1997) and ACE (Doddington et al., 2004). A variety of learning-based systems have been trained and tested on the former (Soon et al., 2001; Uryupina, 2006), on the latter (Culotta et al., 2007; Bengtson and Roth, 2008; Denis and Baldridge, 2009), or on both (Finkel and Manning, 2008; Haghighi and Klein, 2009). Testing on both is needed given that the two annotation schemes differ in some aspects. For example, only ACE includes singletons (mentions that do not corefer) and ACE is restricted to seven semantic types.[1] Also, despite a critical discussion in the MUC task definition (van Deemter and Kibble, 2000), the ACE scheme continues to treat nominal predicates and appositive phrases as coreferential.

A third coreferentially annotated corpus—the largest for English—is OntoNotes (Pradhan et al., 2007; Hovy et al., 2006). Unlike ACE, it is not application-oriented, so coreference relations between all types of NPs are annotated. The identity relation is kept apart from the attributive relation, and it also contains gold-standard morphological, syntactic and semantic information.

Since the MUC and ACE corpora are annotated with only coreference information,[2] existing systems first preprocess the data using automatic tools (POS taggers, parsers, etc.) to obtain the information needed for coreference resolution. However, given that the output from automatic tools is far from perfect, it is hard to determine the level of performance of a coreference module acting on gold-standard preprocessing information. OntoNotes makes it possible to separate the coreference resolution problem from other tasks.

Our study adds to the previously reported evidence by Stoyanov et al. (2009) that differences in corpora and in the task definitions need to be taken into account when comparing coreference resolution systems. We provide new insights as the current analysis differs in four ways. First, Stoyanov et al. (2009) report on differences between MUC and ACE, while we contrast ACE and OntoNotes. Given that ACE and OntoNotes include some of the same texts but annotated according to their respective guidelines, we can better isolate the effect of differences as well as add the additional dimension of gold preprocessing. Second, we evaluate not only with the MUC and B^3 scoring metrics, but also with CEAF. Third, all our experiments use true mentions[3] to avoid effects due to spurious system mentions. Finally, including different baselines and variations of the resolution model allows us to reveal biases of the metrics.

Coreference resolution systems have been tested on languages other than English only within the ACE program (Luo and Zitouni, 2005), probably due to the fact that coreferentially annotated corpora for other languages are scarce. Thus there has been no discussion of the extent to which systems are portable across languages. This paper studies the case of English and Spanish.[4]

Several coreference systems have been developed in the past (Culotta et al., 2007; Finkel and Manning, 2008; Poon and Domingos, 2008; Haghighi and Klein, 2009; Ng, 2009). It is not our aim to compete with them. Rather, we conduct three experiments under a specific setup for comparison purposes. To this end, we use a different, neutral, system, and a dataset that is small and different from official ACE test sets despite the fact that it prevents our results from being compared directly with other systems.

3 Experimental Setup

3.1 System Description

The system architecture used in our experiments, CISTELL, is based on the incrementality of discourse. As a discourse evolves, it constructs a model that is updated with the new information gradually provided. A key element in this model are the entities the discourse is about, as they form the discourse backbone, especially those that are mentioned multiple times. Most entities, however, are only mentioned once. Consider the growth of the entity *Mount Popocatépetl* in (1).[5]

[1] The ACE-2004/05 semantic types are person, organization, geo-political entity, location, facility, vehicle, weapon.

[2] ACE also specifies entity types and relations.

[3] The adjective *true* contrasts with *system* and refers to the gold standard.

[4] Multilinguality is one of the focuses of SemEval-2010 Task 1 (Recasens et al., 2010).

[5] Following the ACE terminology, we use the term *mention* for an instance of reference to an object, and *entity* for a collection of mentions referring to the same object. Entities

(1) We have an update tonight on [this, the volcano in Mexico, they call El Popo]$_{m3}$ ... As the sun rises over [Mt. Popo]$_{m7}$ tonight, the only hint of the fire storm inside, whiffs of smoke, but just a few hours earlier, [the volcano]$_{m11}$ exploding spewing rock and red-hot lava. [The fourth largest mountain in North America, nearly 18,000 feet high]$_{m15}$, erupting this week with [its]$_{m20}$ most violent outburst in 1,200 years.

Mentions can be pronouns ($m20$), they can be a (shortened) string repetition using either the name ($m7$) or the type ($m11$), or they can add new information about the entity: $m15$ provides the supertype and informs the reader about the height of the volcano and its ranking position.

In CISTELL,[6] discourse entities are conceived as 'baskets': they are empty at the beginning of the discourse, but keep growing as new attributes (e.g., name, type, location) are predicated about them. Baskets are filled with this information, which can appear within a mention or elsewhere in the sentence. The ever-growing amount of information in a basket allows richer comparisons to new mentions encountered in the text.

CISTELL follows the learning-based coreference architecture in which the task is split into classification and clustering (Soon et al., 2001; Bengtson and Roth, 2008) but combines them simultaneously. Clustering is identified with basket-growing, the core process, and a pairwise classifier is called every time CISTELL considers whether a basket must be clustered into a (growing) basket, which might contain one or more mentions. We use a memory-based learning classifier trained with TiMBL (Daelemans and Bosch, 2005). Basket-growing is done in four different ways, explained next.

3.2 Baselines and Models

In each experiment, we compute three baselines (1, 2, 3), and run CISTELL under four different models (4, 5, 6, 7).

1. ALL SINGLETONS. No coreference link is ever created. We include this baseline given the high number of singletons in the datasets, since some evaluation measures are affected by large numbers of singletons.
2. HEAD MATCH. All non-pronominal NPs that have the same head are clustered into the same entity.
3. HEAD MATCH + PRON. Like HEAD MATCH, plus allowing personal and possessive pronouns to link to the closest noun with which they agree in gender and number.
4. STRONG MATCH. Each mention (e.g., m_{11}) is paired with previous mentions starting from the beginning of the document (m_1–m_{11}, m_2–m_{11}, etc.).[7] When a pair (e.g., m_3–m_{11}) is classified as coreferent, additional pairwise checks are performed with all the mentions contained in the (growing) entity basket (e.g., m_7–m_{11}). Only if *all* the pairs are classified as coreferent is the mention under consideration attached to the existing growing entity. Otherwise, the search continues.[8]
5. SUPER STRONG MATCH. Similar to STRONG MATCH but with a threshold. Coreference pairwise classifications are only accepted when TiMBL distance is smaller than 0.09.[9]
6. BEST MATCH. Similar to STRONG MATCH but following Ng and Cardie (2002)'s best link approach. Thus, the mention under analysis is linked to the *most confident* mention among the previous ones, using TiMBL's confidence score.
7. WEAK MATCH. A simplified version of STRONG MATCH: not all mentions in the growing entity need to be classified as coreferent with the mention under analysis. A single positive pairwise decision suffices for the mention to be clustered into that entity.[10]

3.3 Features

We follow Soon et al. (2001), Ng and Cardie (2002) and Luo et al. (2004) to generate most of the 29 features we use for the pairwise model. These include features that capture information from different linguistic levels: textual strings (head match, substring match, distance, frequency), morphology (mention type, coordination, possessive phrase, gender match, number match), syntax (nominal predicate, apposition, relative clause, grammatical function), and semantic match (named-entity type, is-a type, supertype).

containing one single mention are referred to as *singletons*.

[6] 'Cistell' is the Catalan word for 'basket.'

[7] The opposite search direction was also tried but gave worse results.

[8] Taking the first mention classified as coreferent follows Soon et al. (2001)'s first-link approach.

[9] In TiMBL, being a memory-based learner, the closer the distance to an instance, the more confident the decision. We chose 0.09 because it appeared to offer the best results.

[10] STRONG and WEAK MATCH are similar to Luo et al. (2004)'s entity-mention and mention-pair models.

For Spanish, we use 34 features as a few variations are needed for language-specific issues such as zero subjects (Recasens and Hovy, 2009).

3.4 Evaluation

Since they sometimes provide quite different results, we evaluate using three coreference measures, as there is no agreement on a standard.

- MUC (Vilain et al., 1995). It computes the number of links common between the true and system partitions. Recall (R) and precision (P) result from dividing it by the minimum number of links required to specify the true and the system partitions, respectively.
- B^3 (Bagga and Baldwin, 1998). R and P are computed for each mention and averaged at the end. For each mention, the number of common mentions between the true and the system entity is divided by the number of mentions in the true entity or in the system entity to obtain R and P, respectively.
- CEAF (Luo, 2005). It finds the best one-to-one alignment between true and system entities. Using true mentions and the ϕ_3 similarity function, R and P are the same and correspond to the number of common mentions between the aligned entities divided by the total number of mentions.

4 Parameter 1: Language

The first experiment compared the performance of a coreference resolution system on a Germanic and a Romance language—English and Spanish—to explore to what extent language-specific issues such as zero subjects[11] or grammatical gender might influence a system.

Although OntoNotes and AnCora are two different corpora, they are very similar in those aspects that matter most for the study's purpose: they both include a substantial amount of texts belonging to the same genre (news) and manually annotated from the morphological to the semantic levels (POS tags, syntactic constituents, NEs, WordNet synsets, and coreference relations). More importantly, very similar coreference annotation guidelines make AnCora the ideal Spanish counterpart to OntoNotes.

[11]Most Romance languages are pro-drop allowing zero subject pronouns, which can be inferred from the verb.

Datasets Two datasets of similar size were selected from AnCora and OntoNotes in order to rule out corpus size as an explanation of any difference in performance. Corpus statistics about the distribution of mentions and entities are shown in Tables 1 and 2. Given that this paper is focused on coreference between NPs, the number of mentions only includes NPs. Both AnCora and OntoNotes annotate only multi-mention entities (i.e., those containing two or more coreferent mentions), so singleton entities are assumed to correspond to NPs with no coreference annotation.

Apart from a larger number of mentions in Spanish (Table 1), the two datasets look very similar in the distribution of singletons and multi-mention entities: about 85% and 15%, respectively. Multi-mention entities have an average of 3.9 mentions per entity in AnCora and 3.5 in OntoNotes. The distribution of mention types (Table 2), however, differs in two important respects: AnCora has a smaller number of personal pronouns as Spanish typically uses zero subjects, and it has a smaller number of bare NPs as the definite article accompanies more NPs than in English.

Results and Discussion Table 3 presents CISTELL's results for each dataset. They make evident problems with the evaluation metrics, namely the fact that the generated rankings are contradictory (Denis and Baldridge, 2009). They are consistent across the two corpora though: MUC rewards WEAK MATCH the most, B^3 rewards HEAD MATCH the most, and CEAF is divided between SUPER STRONG MATCH and BEST MATCH.

These preferences seem to reveal weaknesses of the scoring methods that make them biased towards a type of output. The model preferred by MUC is one that clusters many mentions together, thus getting a large number of correct coreference links (notice the high R for WEAK MATCH), but

	AnCora	OntoNotes
Pronouns	14.09	17.62
Personal pronouns	2.00	12.10
Zero subject pronouns	6.51	–
Possessive pronouns	3.57	2.96
Demonstrative pronouns	0.39	1.83
Definite NPs	37.69	20.67
Indefinite NPs	7.17	8.44
Demonstrative NPs	1.98	3.41
Bare NPs	33.02	42.92
Misc.	6.05	6.94

Table 2: Mention types (%) in Table 1 datasets.

		#docs	#words	#mentions	#entities (e)	#singleton e	#multi-mention e
AnCora	Training	955	299,014	91,904	64,535	54,991	9,544
	Test	30	9,851	2,991	2,189	1,877	312
OntoNotes	Training	850	301,311	74,692	55,819	48,199	7,620
	Test	33	9,763	2,463	1,790	1,476	314

Table 1: Corpus statistics for the large portion of OntoNotes and AnCora.

	MUC			B^3			CEAF
	P	R	F	P	R	F	P / R / F
AnCora - Spanish							
1. ALL SINGLETONS	–	–	–	100	73.32	84.61	73.32
2. HEAD MATCH	55.03	37.72	44.76	91.12	79.88	**85.13**	75.96
3. HEAD MATCH + PRON	48.22	44.24	46.14	86.21	80.66	83.34	76.30
4. STRONG MATCH	45.64	51.88	48.56	80.13	82.28	81.19	75.79
5. SUPER STRONG MATCH	45.68	36.47	40.56	86.10	79.09	82.45	**77.20**
6. BEST MATCH	43.10	35.59	38.98	85.24	79.67	82.36	75.23
7. WEAK MATCH	45.73	65.16	**53.75**	68.50	87.71	76.93	69.21
OntoNotes - English							
1. ALL SINGLETONS	–	–	–	100	72.68	84.18	72.68
2. HEAD MATCH	55.14	39.08	45.74	90.65	80.87	**85.48**	76.05
3. HEAD MATCH + PRON	47.10	53.05	49.90	82.28	83.13	82.70	75.15
4. STRONG MATCH	47.94	55.42	51.41	81.13	84.30	82.68	78.03
5. SUPER STRONG MATCH	48.27	47.55	47.90	84.00	82.27	83.13	78.24
6. BEST MATCH	50.97	46.66	48.72	86.19	82.70	84.41	**78.44**
7. WEAK MATCH	47.46	66.72	**55.47**	70.36	88.05	78.22	71.21

Table 3: CISTELL results varying the corpus language.

also many spurious links that are not duly penalized. The resulting output is not very desirable.[12] In contrast, B^3 is more P-oriented and scores conservative outputs like HEAD MATCH and BEST MATCH first, even if R is low. CEAF achieves a better compromise between P and R, as corroborated by the quality of the output.

The baselines and the system runs perform very similarly in the two corpora, but slightly better for English. It seems that language-specific issues do not result in significant differences—at least for English and Spanish—once the feature set has been appropriately adapted, e.g., including features about zero subjects or removing those about possessive phrases. Comparing the feature ranks, we find that the features that work best for each language largely overlap and are language independent, like head match, is-a match, and whether the mentions are pronominal.

5 Parameter 2: Annotation Scheme

In the second experiment, we used the 100k-word portion (from the TDT collection) shared by the OntoNotes and ACE corpora (330 OntoNotes documents occurred as 22 ACE-2003 documents, 185 ACE-2004 documents, and 123 ACE-2005 documents). CISTELL was trained on the same texts in both corpora and applied to the remainder. The three measures were then applied to each result.

Datasets Since the two annotation schemes differ significantly, we made the results comparable by mapping the ACE entities (the simpler scheme) onto the information contained in OntoNotes.[13] The mapping allowed us to focus exclusively on the differences expressed on both corpora: the types of mentions that were annotated, the definition of identity of reference, etc.

Table 4 presents the statistics for the OntoNotes dataset merged with the ACE entities. The mapping was not straightforward due to several problems: there was no match for some mentions due to syntactic or spelling reasons (e.g., *El Popo* in OntoNotes vs. *Ell Popo* in ACE). ACE mentions for which there was no parse tree node in the OntoNotes gold-standard tree were omitted, as creating a new node could have damaged the tree.

Given that only seven entity types are annotated in ACE, the number of OntoNotes mentions is al-

[12] Due to space constraints, the actual output cannot be shown here. We are happy to send it to interested requesters.

[13] Both ACE entities and types were mapped onto the OntoNotes dataset.

		#docs	#words	#mentions	#entities (e)	#singleton e	#multi-mention e
OntoNotes	Training	297	87,068	22,127	15,983	13,587	2,396
	Test	33	9,763	2,463	1,790	1,476	314
ACE	Training	297	87,068	12,951	5,873	3,599	2,274
	Test	33	9,763	1,464	746	459	287

Table 4: Corpus statistics for the aligned portion of ACE and OntoNotes on gold-standard data.

	MUC			B^3			CEAF
	P	R	F	P	R	F	P / R / F
OntoNotes scheme							
1. ALL SINGLETONS	–	–	–	100	72.68	84.18	72.68
2. HEAD MATCH	55.14	39.08	45.74	90.65	80.87	**85.48**	76.05
3. HEAD MATCH + PRON	47.10	53.05	49.90	82.28	83.13	82.70	75.15
4. STRONG MATCH	46.81	53.34	49.86	80.47	83.54	81.97	76.78
5. SUPER STRONG MATCH	46.51	40.56	43.33	84.95	80.16	82.48	76.70
6. BEST MATCH	52.47	47.40	49.80	86.10	82.80	84.42	**77.87**
7. WEAK MATCH	47.91	64.64	**55.03**	71.73	87.46	78.82	71.74
ACE scheme							
1. ALL SINGLETONS	–	–	–	100	50.96	67.51	50.96
2. HEAD MATCH	82.35	39.00	52.93	95.27	64.05	**76.60**	66.46
3. HEAD MATCH + PRON	70.11	53.90	60.94	86.49	68.20	76.27	68.44
4. STRONG MATCH	64.21	64.21	64.21	76.92	73.54	75.19	**70.01**
5. SUPER STRONG MATCH	60.51	56.55	58.46	76.71	69.19	72.76	66.87
6. BEST MATCH	67.50	56.69	61.62	82.18	71.67	76.57	69.88
7. WEAK MATCH	63.52	80.50	**71.01**	59.76	86.36	70.64	64.21

Table 5: CISTELL results varying the annotation scheme on gold-standard data.

most twice as large as the number of ACE mentions. Unlike OntoNotes, ACE mentions include premodifiers (e.g., *state* in *state lines*), national adjectives (e.g., *Iraqi*) and relative pronouns (e.g., *who, that*). Also, given that ACE entities correspond to types that are usually coreferred (e.g., people, organizations, etc.), singletons only represent 61% of all entities, while they are 85% in OntoNotes. The average entity size is 4 in ACE and 3.5 in OntoNotes.

A second major difference is the definition of coreference relations, illustrated here:

(2) [This] was [an all-white, all-Christian community that all the sudden was taken over ... by different groups].

(3) [[Mayor] John Hyman] has a simple answer.

(4) [Postville] now has 22 different nationalities ... For those who prefer [the old Postville], Mayor John Hyman has a simple answer.

In ACE, nominal predicates corefer with their subject (2), and appositive phrases corefer with the noun they are modifying (3). In contrast, they do not fall under the identity relation in OntoNotes, which follows the linguistic understanding of coreference according to which nominal predicates and appositives express properties of an entity rather than refer to a second (coreferent) entity (van Deemter and Kibble, 2000). Finally, the two schemes frequently disagree on borderline cases in which coreference turns out to be especially complex (4). As a result, some features will behave differently, e.g., the appositive feature has the opposite effect in the two datasets.

Results and Discussion From the differences pointed out above, the results shown in Table 5 might be surprising at first. Given that OntoNotes is not restricted to any semantic type and is based on a more sophisticated definition of coreference, one would not expect a system to perform better on it than on ACE. The explanation is given by the ALL SINGLETONS baseline, which is 73–84% for OntoNotes and only 51–68% for ACE. The fact that OntoNotes contains a much larger number of singletons—as Table 4 shows—results in an initial boost of performance (except with the MUC score, which ignores singletons). In contrast, the score improvement achieved by HEAD MATCH is much more noticeable on ACE than on OntoNotes, which indicates that many of its coreferent mentions share the same head.

The systematic biases of the measures that were observed in Table 3 appear again in the case of

MUC and B^3. CEAF is divided between BEST MATCH and STRONG MATCH. The higher value of the MUC score for ACE is another indication of its tendency to reward correct links much more than to penalize spurious ones (ACE has a larger proportion of multi-mention entities).

The feature rankings obtained for each dataset generally coincide as to which features are ranked best (namely NE match, is-a match, and head match), but differ in their particular ordering.

It is also possible to compare the OntoNotes results in Tables 3 and 5, the only difference being that the first training set was three times larger. Contrary to expectation, the model trained on a larger dataset performs just slightly better. The fact that more training data does not necessarily lead to an increase in performance conforms to the observation that there appear to be few general rules (e.g., head match) that systematically govern coreference relationships; rather, coreference appeals to individual unique phenomena appearing in each context, and thus after a point adding more training data does not add much new generalizable information. Pragmatic information (discourse structure, world knowledge, etc.) is probably the key, if ever there is a way to encode it.

6 Parameter 3: Preprocessing

The goal of the third experiment was to determine how much the source and nature of preprocessing information matters. Since it is often stated that coreference resolution depends on many levels of analysis, we again compared the two corpora, which differ in the amount and correctness of such information. However, in this experiment, entity mapping was applied in the opposite direction: the OntoNotes entities were mapped onto the automatically preprocessed ACE dataset. This exposes the shortcomings of automated preprocessing in ACE for identifying all the mentions identified and linked in OntoNotes.

Datasets The ACE data was morphologically annotated with a tokenizer based on manual rules adapted from the one used in CoNLL (Tjong Kim Sang and De Meulder, 2003), with TnT 2.2, a trigram POS tagger based on Markov models (Brants, 2000), and with the built-in WordNet lemmatizer (Fellbaum, 1998). Syntactic chunks were obtained from YamCha 1.33, an SVM-based NP-chunker (Kudoh and Matsumoto, 2000), and parse trees from Malt Parser 0.4, an SVM-based parser (Hall et al., 2007).

Although the number of words in Tables 4 and 6 should in principle be the same, the latter contains fewer words as it lacks the null elements (traces, ellipsed material, etc.) manually annotated in OntoNotes. Missing parse tree nodes in the automatically parsed data account for the considerably lower number of OntoNotes mentions (approx. 5,700 fewer mentions).[14] However, the proportions of singleton:multi-mention entities as well as the average entity size do not vary.

Results and Discussion The ACE scores for the automatically preprocessed models in Table 7 are about 3% lower than those based on OntoNotes gold-standard data in Table 5, providing evidence for the advantage offered by gold-standard preprocessing information. In contrast, the similar—if not higher—scores of OntoNotes can be attributed to the use of the annotated ACE entity types. The fact that these are annotated not only for proper nouns (as predicted by an automatic NER) but also for pronouns and full NPs is a very helpful feature for a coreference resolution system.

Again, the scoring metrics exhibit similar biases, but note that CEAF prefers HEAD MATCH + PRON in the case of ACE, which is indicative of the noise brought by automatic preprocessing.

A further insight is offered from comparing the feature rankings with gold-standard syntax to that with automatic preprocessing. Since we are evaluating now on the ACE data, the NE match feature is also ranked first for OntoNotes. Head and is-a match are still ranked among the best, yet syntactic features are not. Instead, features like NP type have moved further up. This reranking probably indicates that if there is noise in the syntactic information due to automatic tools, then morphological and syntactic features switch their positions.

Given that the noise brought by automatic preprocessing can be harmful, we tried leaving out the grammatical function feature. Indeed, the results increased about 2–3%, STRONG MATCH scoring the highest. This points out that conclusions drawn from automatically preprocessed data about the kind of knowledge relevant for coreference resolution might be mistaken. Using the most successful basic features can lead to the best results when only automatic preprocessing is available.

[14] In order to make the set of mentions as similar as possible to the set in Section 5, OntoNotes singletons were mapped from the ones detected in the gold-standard treebank.

		#docs	#words	#mentions	#entities (e)	#singleton e	#multi-mention e
OntoNotes	Training	297	80,843	16,945	12,127	10,253	1,874
	Test	33	9,073	1,931	1,403	1,156	247
ACE	Training	297	80,843	13,648	6,041	3,652	2,389
	Test	33	9,073	1,537	775	475	300

Table 6: Corpus statistics for the aligned portion of ACE and OntoNotes on automatically parsed data.

	MUC			B^3			CEAF
	P	R	F	P	R	F	P / R / F
OntoNotes scheme							
1. ALL SINGLETONS	–	–	–	100	72.66	84.16	72.66
2. HEAD MATCH	56.76	35.80	43.90	92.18	80.52	**85.95**	76.33
3. HEAD MATCH + PRON	47.44	54.36	50.66	82.08	83.61	82.84	74.83
4. STRONG MATCH	52.66	58.14	55.27	83.11	85.05	84.07	**78.30**
5. SUPER STRONG MATCH	51.67	46.78	49.11	85.74	82.07	83.86	77.67
6. BEST MATCH	54.38	51.70	53.01	86.00	83.60	84.78	78.15
7. WEAK MATCH	49.78	64.58	**56.22**	75.63	87.79	81.26	74.62
ACE scheme							
1. ALL SINGLETONS	–	–	–	100	50.42	67.04	50.42
2. HEAD MATCH	81.25	39.24	52.92	94.73	63.82	**76.26**	65.97
3. HEAD MATCH + PRON	69.76	53.28	60.42	86.39	67.73	75.93	**68.05**
4. STRONG MATCH	58.85	58.92	58.89	73.36	70.35	71.82	66.30
5. SUPER STRONG MATCH	56.19	50.66	53.28	75.54	66.47	70.72	63.96
6. BEST MATCH	63.38	49.74	55.74	80.97	68.11	73.99	65.97
7. WEAK MATCH	60.22	78.48	**68.15**	55.17	84.86	66.87	59.08

Table 7: CISTELL results varying the annotation scheme on automatically preprocessed data.

7 Conclusion

Regarding evaluation, the results clearly expose the systematic tendencies of the evaluation measures. The way each measure is computed makes it biased towards a specific model: MUC is generally too lenient with spurious links, B^3 scores too high in the presence of a large number of singletons, and CEAF does not agree with either of them. It is a cause for concern that they provide contradictory indications about the core of coreference, namely the resolution models—for example, the model ranked highest by B^3 in Table 7 is ranked *lowest* by MUC. We always assume evaluation measures provide a 'true' reflection of our approximation to a gold standard in order to guide research in system development and tuning.

Further support to our claims comes from the results of SemEval-2010 Task 1 (Recasens et al., 2010). The performance of the six participating systems shows similar problems with the evaluation metrics, and the singleton baseline was hard to beat even by the highest-performing systems.

Since the measures imply different conclusions about the nature of the corpora and the preprocessing information applied, should we use them now to constrain the ways our corpora are created in the first place, and what preprocessing we include or omit? Doing so would seem like circular reasoning: it invalidates the notion of the existence of a true and independent gold standard. But if apparently incidental aspects of the corpora can have such effects—effects rated quite differently by the various measures—then we have no fixed ground to stand on.

The worrisome fact that there is currently no clearly preferred and 'correct' evaluation measure for coreference resolution means that we cannot draw definite conclusions about coreference resolution systems at this time, unless they are compared on exactly the same corpus, preprocessed under the same conditions, and all three measures agree in their rankings.

Acknowledgments

We thank Dr. M. Antònia Martí for her generosity in allowing the first author to visit ISI to work with the second. Special thanks to Edgar Gonzàlez for his kind help with conversion issues.

This work was partially supported by the Spanish Ministry of Education through an FPU scholarship (AP2006-00994) and the TEXT-MESS 2.0 Project (TIN2009-13391-C04-04).

References

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

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP 2008*, pages 294–303, Honolulu, Hawaii.

Thorsten Brants. 2000. TnT – A statistical part-of-speech tagger. In *Proceedings of ANLP 2000*, Seattle, WA.

Aron Culotta, Michael Wick, Robert Hall, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Proceedings of HLT-NAACL 2007*, pages 81–88, Rochester, New York.

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

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

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.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Jenny Rose Finkel and Christopher D. Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of ACL-HLT 2008*, pages 45–48, Columbus, Ohio.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP 2009*, pages 1152–1161, Singapore. Association for Computational Linguistics.

Johan Hall, Jens Nilsson, Joakim Nivre, Gülsen Eryigit, Beáta Megyesi, Mattias Nilsson, and Markus Saers. 2007. Single malt or blended? A study in multilingual parser optimization. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL 2007*, pages 933–939.

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

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of HLT-NAACL 2006*, pages 57–60.

Taku Kudoh and Yuji Matsumoto. 2000. Use of support vector learning for chunk identification. In *Proceedings of CoNLL 2000 and LLL 2000*, pages 142–144, Lisbon, Portugal.

Xiaoqiang Luo and Imed Zitouni. 2005. Multi-lingual coreference resolution with syntactic features. In *Proceedings of HLT-EMNLP 2005*, pages 660–667, Vancouver.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of ACL 2004*, pages 21–26, Barcelona.

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

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, pages 104–111, Philadelphia.

Vincent Ng. 2009. Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of NAACL-HLT 2009*, pages 575–583, Boulder, Colorado.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of EMNLP 2008*, pages 650–659, Honolulu, Hawaii.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *Proceedings of ICSC 2007*, pages 517–526, Washington, DC.

Marta Recasens and Eduard Hovy. 2009. A Deeper Look into Features for Coreference Resolution. In S. Lalitha Devi, A. Branco, and R. Mitkov, editors, *Anaphora Processing and Applications (DAARC 2009)*, volume 5847 of *LNAI*, pages 29–42. Springer-Verlag.

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.

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 Fifth International Workshop on Semantic Evaluations (SemEval 2010)*, Uppsala, Sweden.

Wee M. Soon, Hwee T. Ng, and Daniel C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of ACL-IJCNLP 2009*, pages 656–664, Singapore.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL 2003*, pages 142–147. Edmonton, Canada.

Olga Uryupina. 2006. Coreference resolution with and without linguistic knowledge. In *Proceedings of LREC 2006*.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

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, San Francisco.