A Deeper Look into Features for Coreference Resolution

Marta Recasens[1] and Eduard Hovy[2]

[1] CLiC - University of Barcelona, Gran Via 585, Barcelona, Spain
[2] Information Sciences Institute, 4676 Admiralty Way, Marina del Rey CA, USA
mrecasens@ub.edu

Abstract. All automated coreference resolution systems consider a number of features, such as head noun, NP type, gender, or number. Athough the particular features used is one of the key factors for determining performance, they have not received much attention, especially for languages other than English. This paper delves into a considerable number of pairwise comparison features for coreference, including old and novel features, with a special focus on the Spanish language. We consider the contribution of each of the features as well as the interaction between them. In addition, given the problem of class imbalance in coreference resolution, we analyze the effect of sample selection. From the experiments with TiMBL (Tilburg Memory-Based Learner) on the AnCora corpus, interesting conclusions are drawn from both a linguistic and a computational perspective.

1 Introduction

Coreference resolution, the task of identifying which mentions in a text point to the same discourse entity, has been shown to be beneficial in many NLP applications such as Information Extraction [1], Text Summarization [2], Question Answering [3], and Machine Translation. These systems need to identify the different pieces of information concerning the same referent, produce coherent and fluent summaries, disambiguate the references to an entity, and solve anaphoric pronouns.

Given that many different types of information – ranging from morphology to pragmatics – play a role in coreference resolution, machine learning approaches [4, 5] seem to be a promising way to combine and weigh the relevant factors, overcoming the limitations of constraint-based approaches [6, 7], which might fail to capture global patterns of coreference relations as they occur in real data. Learning-based approaches decompose the task of coreference resolution into two steps: (i) classification, in which a classifier is trained on a corpus to learn the probability that a pair of NPs are or are not coreferent; and (ii) clustering, in which the pairwise links identified at the first stage are merged to form distinct coreference chains.

This paper focuses on the classification stage and, in particular, on (i) the features that are used to build the feature vector that represents a pair of mentions,[3] and (ii) the selection of positive and negative training instances. The choice of the information encoded in the feature vectors is of utmost importance as they are the basis on which the machine learning algorithm learns the pairwise coreference model. Likewise, given the highly skewed distribution of coreferent vs. non-coreferent classes, we will consider whether sample selection is helpful. The more accurate the classification is, the more accurate the clustering will be.

The goal of this paper is to provide an in-depth study of the pairwise comparison stage in order to decrease as much as possible the number of errors that are passed on to the second stage of coreference resolution. Although there have been some studies in this respect [8–10], they are few, oriented to the English or Dutch language, and dependent on poorly annotated corpora. To our knowledge, no previous studies compared systematically a large number of features relying on gold standard corpora, and experiments with sample selection have been only based on small corpora. For the first time, we consider the degree of variance of the learnt model on new data sets by reporting confidence intervals for precision, recall, and F-score measures.

The paper is organized as follows. In the next section, we review previous work. In Section 3, we list our set of 47 features and argue the linguistic motivations behind them. These features are tested by carrying out different machine learning experiments with TiMBL in Section 4, where the effect of sample selection is also assessed. Finally, main conclusions are drawn in Section 5.

2 Previous Work

Be it in the form of hand-crafted heuristics or feature vectors, what kind of knowledge is represented is a key factor for the success of coreference resolution. Although theoretical studies point out numerous linguistic factors relevant for the task, computational systems usually rely on a small number of shallow features, especially after the burst of statistical approaches. In learning-based approaches, the relative importance of the factors is not manually coded but inferred automatically from an annotated corpus. Training instances for machine learning systems are feature vectors representing two mentions (m_1 and m_2) and a label ('coreferent' or 'non-coreferent') allowing the classifier to learn to predict, given a new pair of NPs, whether they do or do not corefer.

The feature set representing m_1 and m_2 that was employed in the decision tree learning algorithm of [4] has been taken as a starting point by most subsequent systems. It consists of only 12 surface-level features (all boolean except for the first): (i) sentence distance, (ii) m_1 is a pronoun, (iii) m_2 is a pronoun, (iv) string match (after discarding determiners), (v) m_2 is a definite NP, (vi) m_2 is a demonstrative NP, (vii) number agreement, (viii) WordNet semantic class agreement,[4] (ix) gender agreement, (x) both m_1 and m_2 are proper nouns (cap-

[3] This paper restricts to computing features over a pair of mentions – without considering a more global approach – hence *pairwise comparison features.*

[4] Possible semantic classes for an NP are *female, male, person, organization, location, date, time, money, percent*, and *object.*

italized), (xi) m_1 is an alias of m_2 or vice versa, and (xii) m_1 is an apposition to m_2. The strongest indicators of coreference turned out to be string match, alias, and appositive.

Ng and Cardie [5] expanded the feature set of [4] from 12 to a deeper set of 53, including a broader range of lexical, grammatical, and semantic features such as substring match, comparison of the prenominal modifiers of both mentions, animacy match, WordNet distance, whether one or both mentions are pronouns, definite, embedded, part of a quoted string, subject function, and so on. The incorporation of additional knowledge succeeds at improving performance but only after manual feature selection, which points out the importance of removing irrelevant features that might be misleading. Surprisingly, however, some of the features in the hand-selected feature set do not seem very relevant from a linguistic point of view, like string match for pronominal mentions.

More recent attempts have explored some additional features to further enrich the set of [5]: backward features describing the antecedent of the candidate antecedent [11], semantic information from Wikipedia, WordNet and semantic roles [12], and most notably, Uryupina's [8] thesis, which investigates the possibility of incorporating sophisticated linguistic knowledge into a data-driven coreference resolution system trained on the MUC-7 corpus. Her extension of the feature set up to a total of 351 nominal features (1096 boolean/continuous) leads to a consistent improvement in the system's performance, thus supporting the hypothesis that complex linguistic factors of NPs are a valuable source of information. At the same time, however, [8] recognizes that by focusing on the addition of sophisticated features she overlooked the resolution strategy and some phenomena might be over-represented in her feature set.

Bengtson and Roth [9] show that with a high-quality set of features, a simple pairwise model can outperform systems built with complex models on the ACE dataset. This clearly supports our stress on paying close attention to designing a strong, linguistically motivated set of features, which requires a detailed analysis of each feature individually as well as of the interaction between them. Some of the features we include, like modifiers match, are also tested by [9] and, interestingly, our ablation study comes to the same conclusion: almost all the features help, although some more than others.

Hoste's [10] work is concerned with optimization issues such as feature and sample selection, and she stresses their effect on classifier performance. The study we present is in line with [8–10] but introduces a number of novelties. First, the object language is Spanish, which presents some differences as far as coreference is concerned. Second, we use a different corpus, AnCora, which is twenty times as large as MUC and, unlike ACE, it includes a non-restricted set of entity types. Third, the coreference annotation of the AnCora corpus sticks to a linguistic definition of the identity relationship more accurate than that behind the MUC or ACE guidelines. Fourth, we do not rely on the (far from perfect) output of preprocessing modules but take advantage of the gold standard annotations in the AnCora corpus in order to focus on their real effect on coreference resolution.

3 Pairwise Comparison Features

The success of machine learning systems depends largely on the feature set employed. Learning algorithms need to be provided with an adequate representation of the data, that is to say, a representation that includes the "relevant" information, to infer the best model from an annotated corpus. Identifying the constraints on when two NPs can corefer is a complex linguistic problem that remains still open. Hence, there is a necessity for an in-depth study of features for coreference resolution from both a computational and a linguistic perspective. This section makes a contribution in this respect by considering a total of 47 features, making explicit the rationale behind them.

- **Classical features** (Table 1). The features that have been shown to obtain better results in previous works [4, 5, 13] capture the most basic information on which coreference depends, but form a reduced feature set that does not account for all kinds of coreference relations.
  - PRON_m$_1$ and PRON_m$_2$ specify whether the mentions are pronouns since these show different patterns of coreference, e.g., gender agreement is of utmost importance for pronouns but might be violated by non-pronouns [10].
  - HEAD_MATCH is the top classical feature for coreference, since lexical repetition is a common coreference device.
  - WORDNET_MATCH uses the Spanish EuroWordNet[5] and is true if any of the synset's synonyms of one mention matches any of the synset's synonyms of the other mention.
  - NP type plays an important role because not all NP types have the same capability to introduce an entity into the text for the first time, and not all NP types have the same capability to refer to a previous mention in the text.
  - The fact that in newspaper texts there is usually at least one person and a location about which something is said accounts for the relevance of the NE type feature, since NE types like *person* and *organization* are more likely to corefer and be coreferred than others.
  - SUPERTYPE_MATCH compares the first hypernym of each mention found in EuroWordNet.
  - As a consequence of the key role played by gender and number in anaphora resolution, GENDER_AGR and NUMBER_AGR have been inherited by coreference systems. See below, however, for finer distinctions.
  - The rationale behind QUOTES is that a mention in quotes identifies a mention that is part of direct speech, e.g., if it is a first- or second-person pronoun, its antecedent will be found in the immediate discourse.

[5] Nominal synsets are part of the semantic annotation of AnCora. EuroWordNet covers 55% of the nouns in the corpus.

Table 1. Classical features

Feature	Definition	Value
PRON_m$_1$	m$_1$ is a pronoun	true, false
PRON_m$_2$	m$_2$ is a pronoun	true, false
HEAD_MATCH	Head match	true, false, ?[a]
WORDNET_MATCH	EuroWordNet match	true, false, ?[a]
NP_m$_1$	m$_1$ NP type	common, proper, article, indefinite, possessive, relative, demonstrative, numeral, interrogative, personal, exclamative
NP_m$_2$	m$_2$ NP type	common, proper, article, indefinite, possessive, relative, demonstrative, numeral, interrogative, personal, exclamative
NE_m$_1$	m$_1$ NE type	person, organization, location, date, number, other, null
NE_m$_2$	m$_2$ NE type	person, organization, location, date, number, other, null
NE_MATCH	NE match	true, false, ?[b]
SUPERTYPE_MATCH	Supertype match	true, false, ?[a]
GENDER_AGR	Gender agreement	true, false
NUMBER_AGR	Number agreement	true, false
ACRONYM	m$_2$ is an acronym of m$_1$	true, false, ?[c]
QUOTES	m$_2$ is in quotes	true, false
FUNCTION_m$_1$	m$_1$ function	subject, d-obj, i-obj, adjunct, prep-obj, attribute, pred-comp, agent, sent-adjunct, no function
FUNCTION_m$_2$	m$_2$ function	subject, d-obj, i-obj, adjunct, prep-obj, attribute, pred-comp, agent, sent-adjunct, no function
COUNT_m$_1$	m$_1$ count	#times m_1 appears in the text
COUNT_m$_2$	m$_2$ count	#times m_2 appears in the text
SENT_DIST	Sentence distance	#sentences between m_1 and m_2
MENTION_DIST	Mention distance	#NPs between m_1 and m_2
WORD_DIST	Word distance	#words between m_1 and m_2

[a] Not applicable. This feature is only applicable if neither m$_1$ nor m$_2$ are pronominal or conjoined.

[b] Not applicable. This feature is only applicable if both mentions are NEs.

[c] Not applicable. This feature is only applicable if m$_2$ is an acronym.

Table 2. Language-specific features

Feature	Definition	Value
ELLIP_m_1	m_1 is an elliptical pronoun	true, false
ELLIP_m_2	m_2 is an elliptical pronoun	true, false
GENDER_PRON	Gender agreement restricted to pronouns	true, false, ?
GENDER_MASCFEM	Gender agreement restricted to masc./fem.	true, false, ?
GENDER_PERSON	Gender agreement restricted to persons	true, false, ?
ATTRIBa_m_1	m_1 is attributive type A	true, false
ATTRIBa_m_2	m_2 is attributive type A	true, false
ATTRIBb_m_1	m_1 is attributive type B	true, false
ATTRIBb_m_2	m_2 is attributive type B	true, false

Table 3. Corpus-specific features

Feature	Definition	Value
NOMPRED_m_1	m_1 is a nominal predicate	true, false
NOMPRED_m_2	m_2 is a nominal predicate	true, false
APPOS_m_1	m_1 is an apposition	true, false
APPOS_m_2	m_2 is an apposition	true, false
PRONTYPE_m_1	m_1 pronoun type	elliptical, 3-person, non-3-person, demonstrative, possessive, indefinite, numeric, other, ?
PRONTYPE_m_2	m_2 pronoun type	elliptical, 3-person, non-3-person, demonstrative, possessive, indefinite, numeric, other, ?
EMBEDDED	m_2 is embedded in m_1	true, false
MODIF_m_1	m_1 has modifiers	true, false
MODIF_m_2	m_2 has modifiers	true, false

Table 4. Novel features

Feature	Definition	Value
FUNCTION_TRANS	Function transition	100 different values (e.g., subject_subject, subject_d-obj)
COUNTER_MATCH	Counter match	true, false, ?
MODIF_MATCH	Modifiers match	true, false, ?
VERB_MATCH	Verb match	true, false, ?
NUMBER_PRON	Number agreement restricted to pronouns	true, false, ?
TREE-DEPTH_m_1	m_1 parse tree depth	#nodes in the parse tree from m_1 up to the top
TREE-DEPTH_m_2	m_2 parse tree depth	#nodes in the parse tree from m_2 up to the top
DOC_LENGTH	Document length	#tokens in the document

- **Language-specific features** (Table 2). There are some language-specific issues that have a direct effect on the way coreference relations occur in a language. In the case of Spanish, we need to take into account elliptical subjects, grammatical gender, and nouns used attributively.
  - There is a need to identify elliptical pronouns in Spanish because, unlike overt pronouns, they get their number from the verb, have no gender, and always appear in subject position, as shown in (1), where the elliptical subject pronoun is marked with ⊘ and with the corresponding pronoun in brackets in the English translation.

    (1) Klebánov manifestó que ⊘ no puede garantizar el éxito al cien por cien.
    'Klebánov stated that *(he)* cannot guarantee 100% success.'

  - Since Spanish has grammatical gender, two non-pronominal nouns with different gender might still corefer, e.g., *el incremento* 'the increase' (masc.) and *la subida* 'the rise' (fem.). Gender agreement is an appropriate constraint only for pronouns.
  - GENDER_MASCFEM does not consider those NPs that are not marked for gender (e.g. elliptical pronouns, companies).
  - GENDER_PERSON separates natural from grammatical gender by only comparing the gender if one of the mentions is an NE-person.[6]
  - Attributive NPs[7] are non-referential, hence non-markables. ATTRIBa and ATTRIBb identify two Spanish constructions where these NPs usually occur:

    Type A. Common, singular NPs following the preposition *de* 'of', e.g., *educación* 'education' in *sistema de educación* 'education system.'

    Type B. Proper nouns immediately following a generic name, e.g., *Mayor* 'Main' in *calle Mayor* 'Main Street'.
- **Corpus-specific features** (Table 3). The definition of coreference in the AnCora corpus differs from that of the MUC and ACE corpora in that it separates identity from other kinds of relation such as apposition, predication, or bound anaphora. This is in line with van Deemter and Kibble's [14] criticism of MUC. Predicative and attributive NPs do not have a referential function but an attributive one, qualifying an already introduced entity. They should not be allowed to corefer with other NPs. Consequently, the use we make of nominal-predicate and appositive features is the opposite to that made by systems trained on the MUC or ACE corpora [4, 13]. Besides, the fact that AnCora contains gold standard annotation from the morphological to the semantic levels makes it possible to include additional features that rely on such rich information.
  - We employ NOMPRED to filter out predicative mentions.
  - We employ APPOS to filter out attributively used mentions.

[6] Animals are not included since they are not explicitly identitifed as NEs.

[7] *Attributively* used NPs qualify another noun.

  - Gold standard syntactic annotation makes it possible to assess the efficacy of the EMBEDDED and MODIF features in isolation from any other source of error. First, a nested NP cannot corefer with the embedding one. Second, depending on the position a mention occupies in the coreference chain, it is more or less likely that it is modified.
- **Novel features** (Table 4). We suggest some novel features that we believe relevant and that the rich annotation of AnCora enables.
  - FUNCTION_TRANS is included because although FUNCTION_m_1 and FUNCTION_m_2 already encode the function of each mention separately, there may be information in their joint behaviour.[8] E.g., *subject_subject* can be relevant since two consecutive subjects are likely to corefer:

    (2) [...] explicó *Alonso, quien anunció la voluntad de Telefónica Media de unirse a grandes productoras iberoamericanas.* Por otra parte, *Alonso* justificó el aplazamiento.
    '[...] explained *Alonso, who announced the will of Telefónica Media to join large Latin American production companies.* On the other hand, *Alonso* justified the postponement.'

  - COUNTER_MATCH prevents two mentions that contain a different numeral to corefer (e.g., *134 millones de euros* '134 million euros' and *194 millones de euros* '194 million euros'), as they point to a different number of referents.
  - Modifiers introduce extra information that might imply a change in the referential scope of a mention (e.g., *las elecciones generales* 'the general elections' and *las elecciones autonómicas* 'the regional elections'). Thus, when both mentions are modified, the synonyms and immediate hypernym of the head of each modifying phrase are extracted from EuroWordNet for each mention. MODIF_MATCH is true if one of them matches between the two mentions.
  - The verb, as the head of the sentence, imposes restrictions on its arguments. In (3), the verb *participate* selects for a volitional agent, and the fact that the two subjects complement the same verb hints at their coreference link. VERB_MATCH is true if either the two verbal lemmas or any synonym or immediate hypernym from EuroWordNet match.

    (3) *Un centenar de artistas* participará en el acto [...] el acto se abrirá con un brindis en el que participarán *todos los protagonistas de la velada.*
    '*One hundred artists* will participate in the ceremony [...] the ceremony will open with a toast in which *all the protagonists of the evening gathering* will participate.'

  - NUMBER_PRON is included since non-pronominal mentions that disagree in number might still corefer.
  - DOC_LENGTH can be helpful since the longer the document, the more coreferent mentions, and a wider range of patterns might be allowed.

[8] The idea of including conjoined features is also exploited by [9, 13].

Table 5. Characteristics of the AnCora-Es datasets

	Training set	Test set
# Words	298 974	23 022
# Entities	64 421	4 893
# Mentions	88 875	6 759
# NEs	25 758	2 023
# Nominals	53 158	4 006
# Pronominals	9 959	730

4 Experimental Evaluation

This section describes our experiments with the features presented in Section 3 as well as with different compositions of the training and test data sets. We finally assess the reliability of the most appropriate pairwise comparison model.

Data. The experiments are based on the AnCora-Es corpus [15], a corpus of newspaper and newswire articles. It is the largest Spanish corpus annotated, among other levels of linguistic information, with PoS tags, syntactic constituents and functions, named entities, nominal WordNet synsets, and coreference links.[9] We split randomly the freely available labelled data into a training set of 300k words and a test set of 23k words. See Table 5 for a description.

Learning algorithm. We use TiMBL, the Tilburg memory-based learning classifier [16], which is a descendant of the k-nearest neighbor approach. It is based on analogical reasoning: the behavior of new instances is predicted by extrapolating from the similarity between (old) stored representations and the new instances. This makes TiMBL particularly appropriate for training a coreference resolution model, as the feature space tends to be very sparse and it is very hard to find universal rules that work all the time. In addition, TiMBL outputs the information gain of each feature – very useful for studies on feature selection – and allows the user easily to experiment with different feature sets by obscuring specified features. Given that the training stage is done without abstraction but by simply storing training instances in memory, it is considerably faster than other machine learning algorithms.

We select parameters to optimize TiMBL on a held-out development set. The distance metric parameter is set to overlap, and the number of nearest neighbors (k parameter) is set to 5 in Section 4.1, and to 1 in Section 4.2.[10]

[9] AnCora is freely available from `http://clic.ub.edu/ancora`.

[10] When training the model on the full feature vectors, the best results are obtained when TiMBL uses 5 nearest neighbors for extrapolation. However, because of the strong skew in the class space, in some of the hill-climbing experiments we can only use 1 nearest neighbor. Otherwise, with 5 neighbors the majority of neighbors are of the negative class for all the test cases, and the positive class is never predicted (recall=0).

Table 6. Distribution of representative and balanced data sets

	Training set		Test set	
	Representative	Balanced	Representative	Balanced
Positive instances	105 920		8 234	
Negative instances	425 942	123 335	32 369	9 399

Table 7. Effect of sample selection on performance

	Training set	Test set	P	R	F
Model A	Representative	Representative	84.73	73.44	78.68
Model B	Representative	Balanced	88.43	73.44	80.24
Model C	Balanced	Representative	66.28	80.24	72.60
Model D	Balanced	Balanced	83.46	87.32	85.34

4.1 Sample Selection

When creating the training instances, we run into the problem of class imbalance: there are many more negative examples than positive ones. Positive training instances are created by pairing each coreferent NP with all preceding mentions in the same coreference chain. If we generate negative examples for all the preceding non-coreferent mentions, which would conform to the real distribution, then the number of positive instances is only about 7% [10]. In order to reduce the vast number of negative instances, previous approaches usually take only those mentions between two coreferent mentions, or they limit the number of previous sentences from which negative mentions are taken. Negative instances have so far been created only for those mentions that are coreferent. In a real task, however, the system must decide on the coreferentiality of all mentions.

In order to investigate the impact of keeping the highly skewed class distribution in the training set, we create two versions for each data set: a representative one, which approximates the natural class distribution, and a balanced one, which results from down-sampling negative examples. The total number of negatives is limited by taking only 5 non-coreferent mentions randomly selected among the previous mentions (back to the beginning of the document). The difference is that in the balanced sample, non-coreferent mentions are selected for each coreferent mention, whereas in the representative sample they are selected for all mentions in the document. See Table 6 for statistics of the training and test sets.

Combining each training data set with each test set gives four possible combinations (Table 7) and we compute the performance of each of the models. The output of the experiments is evaluated in terms of precision (P), recall (R) and F-score (F). Although the best performance is obtained when testing the model on the balanced sample (models B and D), making a balanced test set involves knowledge about the different classes in the test set, which is not available in non-experimental situations. Therefore, being realistic, we must carry out the evaluation on a data set that follows the natural class distribution. We focus our attention on models A and C.

Down-sampling on the training set increases R but at the cost of a too dramatic decrease in P. Because of the smaller number of negative instances in the training, it is more likely for an instance to be classified as positive, which harms P and F. As observed by [10], we can conclude that down-sampling does not lead to an increase in TiMBL, and so we opt for using model A.

4.2 Feature Selection

This section considers the informativeness of the features presented in Section 3. We carry out two different feature selection experiments: (i) an ablation study, and (ii) a hill-climbing forward selection.

In the first experiment, we test each feature by running TiMBL on different subsets of the 47 features, each time removing a different one. The majority of features have low informativeness, as no single feature brings about a statistically significant loss in performance when omitted.[11] Even the removal of HEAD_MATCH, which is reported in the literature as one of the key features in coreference resolution, causes a statistically non-significant decrease of .15 in F. We conclude that some other features together learn what HEAD_MATCH learns on its own. Features that individually make no contribution are ones that filter referentiality, of the kind *ATTRIBb*_m$_2$, and ones characterising m$_1$, such as PRON_m$_1$. Finally, some features, in particular the distance and numeric measures, seem even to harm performance. However, there is a complex interaction between the different features. If we train a model that omits all features that seem irrelevant and harmful at the individual level, then performance on the test set decreases. This is in line with the ablation study performed by [9], who concludes that all features help, although some more than others.

Forward selection is a greedy approach that consists of incrementally adding new features – one at a time – and eliminating a feature whenever it causes a drop in performance. Features are chosen for inclusion according to their information gain values, as produced by TiMBL, most informative earliest. Table 8 shows the results of the selection process. In the first row, the model is trained on a single (the most informative) feature. From there on, one additional feature is added in each row; initial "-" marks the harmful features that are discarded (provide a statistically significant decrease in either P or R, and F). P and R scores that represent statistically significant gains and drops with respect to the previous feature vector are marked with an asterisk (*) and a dagger (†), respectively. Although F-score keeps rising steadily in general terms, informative features with a statistically significant improvement in P are usually accompanied by a significant decrease in R, and vice versa.

[11] Statistical significance is tested with a one-way ANOVA followed by a Tukey's post-hoc test.

Table 8. Results of the forward selection procedure

Feature vector	P	R	F	Feature vector	P	R	F
HEAD_MATCH	92.94	17.43	29.35	COUNTER_ MATCH	81.76	63.64	71.57
PRON_ m_2	57.58†	61.14*	59.30	MODIF_ m_1	81.08	64.67	71.95
ELLIP_ m_2	65.22*	53.04†	58.50	PRONTYPE_ m_1	81.70	64.84	72.30
-ELLIP_m_1	89.74*	34.09†	49.41	GENDER_ AGR	81.60	65.12	72.44
WORDNET_ MATCH	65.22	53.04	58.50	NOMPRED_ m_1	81.89	65.04	72.50
NE_ MATCH	65.22	53.04	58.50	GENDER_ PERSON	87.95*	64.78	74.61
-PRON_m_1	86.73*	38.74†	53.56	FUNCTION_ m_2	87.06	65.96	75.06
NUMBER_ PRON	69.04*	58.20*	63.16	FUNCTION_ m_1	85.88†	69.82*	77.02
-GENDER_PRON	86.64*	37.39†	52.24	QUOTES	85.83	70.11	77.18
VERB_ MATCH	80.31*	55.53†	65.66	COUNT_ m_2	85.62	70.73	77.47
SUPERTYPE_ MATCH	80.22	55.56	65.65	COUNT_ m_1	84.57	71.35	77.40
MODIF_ m_2	78.18	61.68*	68.96	NE_ m_1	83.82	72.48	77.74
NUMBER_ AGR	79.94	61.81	69.71	ACRONYM	83.99	72.46	77.80
ATTRIBb_ m_2	80.08	61.85	69.80	NE_ m_2	83.48	73.14	77.97
ATTRIBa_ m_2	80.14	61.84	69.81	NP_ m_2	82.81	73.55	77.91
ATTRIBa_ m_1	80.22	61.83	69.84	NP_ m_1	82.27	74.05	77.94
ATTRIBb_ m_1	80.23	61.82	69.83	FUNCTION_ TRANS	82.29	73.94	77.89
EMBEDDED	80.33	61.78	69.84	TREE-DEPTH_ m_2	80.54	72.98	76.57
GENDER_ MASCFEM	81.33	62.96	70.98	-TREE-DEPTH_ m_1	78.25†	72.52	75.27
APPOS_ m_1	81.46	62.96	71.02	-SENT_ DIST	78.17†	72.16	75.05
APPOS_ m_2	81.44	62.95	71.01	-DOC_ LENGTH	79.36*	70.36†	74.79
MODIF_ MATCH	81.35	63.10	71.08	MENTION_ DIST	79.52	72.10	75.63
NOMPRED_ m_2	81.38	63.37	71.26	WORD_ DIST	79.14	71.73	75.25
PRONTYPE_ m_2	81.70	63.59	71.52				

The results show several interesting tendencies. Although HEAD_MATCH is the most relevant feature, it obtains a very low R, as it cannot handle coreference relationships involving pronouns or relations between full NPs that do not share the same head. Therefore, when PRON_ m_2 is added, R is highly boosted. With only these two features, P, R and F reach scores near the 60s. The rest of the features make a small – yet important in sum – contribution. Most of the features have a beneficial effect on performance, which provides evidence for the value of building a feature vector that includes linguistically motivated features. This includes some of the novel features we argue for, such as NUMBER_PRON and VERB_MATCH. Surprisingly, distance features seem to be harmful. However, if we train again the full model with the k parameter set to 5 and we leave out the numeric features, F does not increase but goes down. Again, the complex interaction between the features is manifested.

4.3 Model Reliability

In closing this section, we would like to stress an issue to which attention is hardly ever paid: the need for computing the reliability of a model's performance. Because of the intrinsic variability in any data set, the performance of a model trained on one training set and tested on another will never be maximal. In addition to the two experiments varying feature and sample selection reported above, we actually carried out numerous other analyses of different combinations. Every change in the sample selection resulted in a change of the feature ranking produced by TiMBL. For example, starting the hill-climbing experiment with a different feature would also lead to a different result, with a different set of features deemed harmful. Similarly, changing the test set will result in different performance of even the same model. For this reason, we believe that merely reporting system performances is not enough. It should become common practice to inspect evaluations taken over different test sets and to report the model's *averaged* performance, i.e., its F, R, and P scores, each bounded by confidence intervals.

To this end, we split randomly the test set into six subsets and evaluated each output. Then we computed the mean, variance, standard deviation, and confidence intervals of the six results of each P, R, and F-score. The exact performance of our pairwise comparison model for coreference (model A in Table 7) is 81.91±4.25 P, 69.57±8.13 R, and 75.12±6.47 F.

5 Conclusion

This paper focused on the classification stage of an automated coreference resolution system for Spanish. In the pairwise classification stage, the probability that a pair of NPs are or are not coreferent was learnt from a corpus. The more accurate this stage is, the more accurate the subsequent clustering stage will be. Our detailed study of the informativeness of a considerable number of pairwise comparison features and the effect of sample selection added to the few literature [8–10] on these two issues.

We provided a list of 47 features for coreference pairwise comparison and discussed the linguistic motivations behind each one: well-studied features included in most coreference resolution systems, language-specific ones, corpus-specific ones, as well as extra features that we considered interesting to test. Different machine learning experiments were carried out using the TiMBL memory-based learner. The features were shown to be weakly informative on their own, but to support complex and unpredictable interactions. In contrast with previous work, many of the features relied on gold standard annotations, pointing out the need for automatic tools for ellipticals detection and deep parsing.

Concerning the selection of the training instances, down-sampling was discarded as it did not improve performance in TiMBL. Instead, better results were obtained when the training data followed the same distribution as the real-world data, achieving 81.91±4.25 P, 69.57±8.13 R, and 75.12±6.47 F-score. Finally, we pointed out the importance of reporting confidence intervals in order to show the degree of variance that the learnt model carries.

Acknowledgments

We are indebted to M. Antònia Martí for her helpful comments.

This research was supported by the FPU Grant (AP2006-00994) from the Spanish Ministry of Education and Science, and the Lang2World (TIN2006-15265-C06-06) and Ancora-Nom (FFI2008-02691-E/FILO) projects.

References

1. McCarthy, J.F., Lehnert, W.G.: Using decision trees for coreference resolution. In: Proceedings of IJCAI. (1995) 1050–1055
2. Steinberger, J., Poesio, M., Kabadjov, M.A., Jeek, K.: Two uses of anaphora resolution in summarization. Information Processing and Management: an International Journal **43**(6) (2007) 1663–1680
3. Morton, T.S.: Using coreference in question answering. In: Proceedings of the Text REtrieval Conference 8. (1999) 85–89
4. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics **27**(4) (2001) 521–544
5. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of ACL. (2002) 104–111
6. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20**(4) (1994) 535–561
7. Mitkov, R.: Robust pronoun resolution with limited knowledge. In: Proceedings of ACL-COLING. (1998) 869–875
8. Uryupina, O.: Knowledge Acquisition for Coreference Resolution. PhD thesis, Saarland University (2007)
9. Bengtson, E., Roth, D.: Understanding the value of features for coreference resolution. In: Proceedings of EMNLP. (2008) 294–303
10. Hoste, V.: Optimization Issues in Machine Learning of Coreference Resolution. PhD thesis, University of Antwerp (2005)
11. Yang, X., Su, J., Zhou, G., Tan, C.L.: Improving pronoun resolution by incorporating coreferential information of candidates. In: Proceedings of ACL. (2004) 127–134
12. Ponzetto, S.P., Strube, M.: Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In: Proceedings of HLT-NAACL. (2006) 192–199
13. Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S.: A mention-synchronous coreference resolution algorithm based on the Bell tree. In: Proceedings of ACL. (2004) 21–26
14. Van Deemter, K., Kibble, R.: On coreferring: Coreference in MUC and related annotation schemes. Computational Linguistics **26**(4) (2000) 629–637
15. Recasens, M., Martí, M.A.: AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation (to appear)
16. Daelemans, W., Bosch, A.V.: Memory-Based Language Processing. Cambridge University Press (2005)