SemEval-2010 Task 1: Coreference Resolution in Multiple Languages

Marta Recasens, Toni Martí, Mariona Taulé
Centre de Llenguatge i Computació (CLiC)
University of Barcelona
Gran Via de les Corts Catalanes 585
08007 Barcelona
{mrecasens, amarti, mtaule}
@ub.edu

Lluís Màrquez, Emili Sapena
TALP Research Center,
Technical University of Catalonia
Jordi Girona Salgado 1-3
08034 Barcelona
{lluism, esapena}
@lsi.upc.edu

Abstract

This paper presents the task 'Coreference Resolution in Multiple Languages' to be run in SemEval-2010 (5th International Workshop on Semantic Evaluations). This task aims to evaluate and compare automatic coreference resolution systems for three different languages (Catalan, English, and Spanish) by means of two alternative evaluation metrics, thus providing an insight into (i) the portability of coreference resolution systems across languages, and (ii) the effect of different scoring metrics on ranking the output of the participant systems.

1 Introduction

Coreference information has been shown to be beneficial in many NLP applications such as Information Extraction (McCarthy and Lehnert, 1995), Text Summarization (Steinberger et al., 2007), Question Answering (Morton, 2000), and Machine Translation. In these systems, there is a need to identify the different pieces of information that refer to the same discourse entity in order to produce coherent and fluent summaries, disambiguate the references to an entity, and solve anaphoric pronouns.

Coreference is an inherently complex phenomenon. Some of the limitations of the traditional rule-based approaches (Mitkov, 1998) could be overcome by machine learning techniques, which allow automating the acquisition of knowledge from annotated corpora.

This task will promote the development of linguistic resources –annotated corpora[1]– and machine-learning techniques oriented to coreference resolution. In particular, we aim to evaluate and compare coreference resolution systems in a multilingual context, including Catalan, English, and Spanish languages, and by means of two different evaluation metrics.

By setting up a multilingual scenario, we can explore to what extent it is possible to implement a general system that is portable to the three languages, how much language-specific tuning is necessary, and the significant differences between Romance languages and English, as well as those between two closely related languages such as Spanish and Catalan. Besides, we expect to gain some useful insight into the development of multilingual NLP applications.

As far as the evaluation is concerned, by employing B-cubed (Bagga and Baldwin, 1998) and CEAF (Luo, 2005) algorithms we can consider both the advantages and drawbacks of using one or the other scoring metric. For comparison purposes, the MUC score will also be reported. Among others, we are interested in the following questions: Which evaluation metric provides a more accurate picture of the accuracy of the system performance? Is there a strong correlation between them? Can

[1] Corpora annotated with coreference are scarce, especially for languages other than English.

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 70–75,
Boulder, Colorado, June 2009.

statistical systems be optimized under both metrics at the same time?

The rest of the paper is organized as follows. Section 2 describes the overall task. The corpora and the annotation scheme are presented in Section 3. Conclusions and final remarks are given in Section 4.

2 Task description

The SemEval-2010 task 'Coreference Resolution in Multiple Languages' is concerned with automatic coreference resolution for three different languages: Catalan, English, and Spanish.

2.1 Specific tasks

Given the complexity of the coreference phenomena, we will concentrate only in two tractable aspects, which lead to the two following subtasks for each of the languages:

i) Detection of full coreference chains, composed by named entities, pronouns, and full noun phrases (NPs).
ii) Pronominal resolution, i.e. finding the antecedents of the pronouns in the text.

[Els beneficiaris de [pensions de [viudetat]$_3$]$_2$]$_1$ podran conservar [la paga]$_4$ encara_que [Ø]$_5$ es tornin a casar si [Ø]$_6$ compleixen [una sèrie de [condicions]$_8$]$_7$, segons [el reial decret aprovat ahir pel [Consell_de_Ministres]$_{10}$]$_9$.

[La nova norma]$_{11}$ afecta [els perceptors d' [una pensió de [viudetat]$_{13}$]$_{12}$ [que]$_{14}$ contreguin [matrimoni]$_{15}$ a_partir_de [l' 1_de_gener_del_2002]$_{16}$]$_{17}$.

[La primera de [les condicions]$_{18}$]$_{19}$ és tenir [més de 61 anys]$_{20}$ o tenir reconeguda [una incapacitat permanent [que]$_{22}$ inhabiliti per a [tota [professió]$_{24}$ o [ofici]$_{25}$]$_{23}$]$_{21}$.

[La segona]$_{26}$ és que [la pensió]$_{27}$ sigui [la principal o única font d' [ingressos del [pensionista]$_{30}$]$_{29}$]$_{28}$, i sempre_que [l' import anual de [la mateixa pensió]$_{32}$]$_{31}$ representi , com_a_mínim , [el 75% del [total dels [ingressos anuals del [pensionista]$_{36}$]$_{35}$]$_{34}$]$_{33}$.

The example in Figure 1 illustrates the two subtasks.[2] Given a text in which NPs are identified and indexed (including elliptical subjects, represented as Ø), the goal of (i) is to extract all coreference chains: 1–5–6–30–36, 9–11, and 7–18; while the goal of (ii) is to identify the antecedents of pronouns 5 and 6, which are 1 and 5 (or 1), respectively. Note that (b) is a simpler subtask of (a) and that for a given pronoun there can be multiple antecedents (e.g. both 1 and 5 are correct antecedents for 6).

We restrict the task to solving 'identity' relations between NPs (coreference chains), and between pronouns and antecedents. Nominal predicates and appositions as well as NPs with a non-nominal antecedent (discourse deixis) will not been taken into consideration in the recognition of coreference chains (see Section 3.1 for more information about decisions concerning the annotation scheme).

Although we target at general systems addressing the full multilingual task, we will allow taking part on any subtask of any language in order to promote participation.

[The beneficiaries of [[spouse's]$_3$ pensions]$_2$]$_1$ will be able to keep [the payment]$_4$ even if [they]$_5$ remarry provided that [they]$_6$ fulfill [a series of [conditions]$_8$]$_7$, according to [the royal decree approved yesterday by [the Council of Ministers]$_{10}$]$_9$.

[The new rule]$_{11}$ affects [the recipients of [a [spouse's]$_{13}$ pension]$_{12}$ [that]$_{14}$ get married after [January_1_,_2002]$_{16}$]$_{17}$.

[The first of [the conditions]$_{18}$]$_{19}$ is being older [than 61 years old]$_{20}$ or having [an officially recognized permanent disability [that]$_{22}$ makes one disabled for [any [profession]$_{24}$ or [job]$_{25}$]$_{23}$]$_{21}$.

[The second one]$_{26}$ requires that [the pension]$_{27}$ be [the main or only source of [the [pensioner's]$_{30}$ income]$_{29}$]$_{28}$, and provided that [the annual amount of [the pension]$_{32}$]$_{31}$ represents, at least, [75% of [the total [yearly income of [the pensioner]$_{36}$]$_{35}$]$_{34}$]$_{33}$.

Figure 1. NPs in a sample from the Catalan training data (left) and the English translation (right).

[2] The example in Figure 1 is a simplified version of the annotated format. See Section 2.2 for more details.

2.2 Evaluation

2.1.1 Input information

The input information for the task will consist of: word forms, lemmas, POS, full syntax, and semantic role labeling. Two different scenarios will be considered regarding the source of the input information:

i) In the first one, *gold standard* annotation will be provided to participants. This input annotation will correctly identify all NPs that are part of coreference chains. This scenario will be only available for Catalan and Spanish.
ii) In the second, state-of-the-art automatic linguistic analyzers for the three languages will be used to generate the input annotation of the data. The matching between the automatically generated structure and the real NPs intervening in the chains does not need to be perfect in this setting.

By defining these two experimental settings, we will be able to check the performance of coreference systems when working with perfect linguistic (syntactic/semantic) information, and the degradation in performance when moving to a more realistic scenario with noisy input annotation.

2.1.2 Closed/open challenges

In parallel, we will also consider the possibility of differentiating between closed and open challenges, that is, when participants are allowed to use strictly the information contained in the training data (closed) and when they make use of some external resources/tools (open).

2.1.3 Scoring measures

Regarding evaluation measures, we will have specific metrics for each of the subtasks, which will be computed by language and overall.

Several metrics have been proposed for the task of coreference resolution, and each of them presents advantages and drawbacks. For the purpose of the current task, we have selected two of them – B-cubed and CEAF – as the most appropriate ones. In what follows we justify our choice.

The MUC scoring algorithm (Vilain et al., 1995) has been the most widely used for at least two reasons. Firstly, the MUC corpora and the MUC scorer were the first available systems. Secondly, the MUC scorer is easy to understand and implement. However, this metric has two major weaknesses: (i) it does not give any credit to the correct identification of singleton entities (chains consisting of one single mention), and (ii) it intrinsically favors systems that produce fewer coreference chains, which may result in higher F-measures for worse systems.

A second well-known scoring algorithm, the ACE value (NIST, 2003), owes its popularity to the ACE evaluation campaign. Each error (a missing element, a misclassification of a coreference chain, a mention in the response not included in the key) made by the response has an associated cost, which depends on the type of entity (e.g. person, location, organization) and on the kind of mention (e.g. name, nominal, pronoun). The fact that this metric is entity-type and mention-type dependent, and that it relies on ACE-type entities makes this measure inappropriate for the current task.

The two measures that we are interested in comparing are B-cubed (Bagga and Baldwin, 1998) and CEAF (Luo, 2005). The former does not look at the links produced by a system as the MUC algorithm does, but looks at the presence/absence of mentions for each entity in the system output. Precision and recall numbers are computed for each mention, and the average gives the final precision and recall numbers.

CEAF (Luo, 2005) is a novel metric for evaluating coreference resolution that has already been used in some published papers (Ng, 2008; Denis and Baldridge, 2008). It mainly differs from B-cubed in that it finds the best one-to-one entity alignment between the gold and system responses before computing precision and recall. The best mapping is that which maximizes the similarity over pairs of chains. The CEAF measure has two variants: a mention-based, and an entity-based one. While the former scores the similarity of two chains as the absolute number of common mentions between them, the latter scores the relative number of common mentions.

Luo (2005) criticizes the fact that a response with all mentions in the same chain obtains 100% B-cubed recall, whereas a response with each mention in a different chain obtains 100% B-cubed

precision. However, precision will be penalized in the first case, and recall in the second case, each captured by the corresponding F-measure. Luo's entity alignment might cause that a correctly identified link between two mentions is ignored by the scoring metric if that entity is not aligned. Finally, as far as the two CEAF metrics are concerned, the entity-based measure rewards alike a correctly identified one-mention entity and a correctly identified five-mention entity, while the mention-based measure takes into account the size of the entity.

Given this series of advantages and drawbacks, we opted for including both B-cubed and CEAF measures in the final evaluation of the systems. In this way we will be able to perform a meta-evaluation study, i.e. to evaluate and compare the performance of metrics with respect to the task objectives and system rankings. It might be interesting to break B-cubed and CEAF into partial results across different kinds of mentions in order to get a better understanding of the sources of errors made by each system. Additionally, the MUC metric will also be included for comparison purposes with previous results.

Finally, for the setting with automatically generated input information (second scenario in Section 2.1.1), it might be desirable to devise metric variants accounting for partial matches of NPs. In this case, capturing the correct NP head would give most of the credit. We plan to work in this research line in the near future.

Official scorers will be developed in advance and made available to participants when posting the trial datasets. The period in between the release of trial datasets and the start of the full evaluation will serve as a test for the evaluation metrics. Depending on the feedback obtained from the participants we might consider introducing some improvements in the evaluation setting.

3 AnCora-CO corpora

The corpora used in the task are AnCora-CO, which are the result of enriching the AnCora corpora (Taulé et al., 2008) with coreference information. AnCora-CO is a multilingual corpus annotated at different linguistic levels consisting of 400K words in Catalan[3], 400K words in Spanish[2], and 120K words in English. For the purpose of the task, the corpora are split into a training (85%) and test (15%) set. Each file corresponds to one newspaper text.

AnCora-CO consists mainly of newspaper and newswire articles: 200K words from the Spanish and Catalan versions of *El Periódico* newspaper, and 200K words from the EFE newswire agency in the Spanish corpus, and from the ACN newswire agency in the Catalan corpus. The source corpora for Spanish and Catalan are the AnCora corpora, which were annotated by hand with full syntax (constituents and functions) as well as with semantic information (argument structure with thematic roles, semantic verb classes, named entities, and WordNet nominal senses). The annotation of coreference constitutes an additional layer on top of the previous syntactic-semantic information.

The English part of AnCora-CO consists of a series of documents of the Reuters newswire corpus (RCV1 version).[4] The RCV1 corpus does not come with any syntactic nor semantic annotation. This is why we only count with automatic linguistic annotation produced by statistical taggers and parsers on this corpus.

Although the Catalan, English, and Spanish corpora used in the task all belong to the domain of newspaper texts, they do not form a three-way parallel corpus.

3.1 Coreference annotation

The annotation of a corpus with coreference information is highly complex due to (i) the lack of information in descriptive grammars about this topic, and (ii) the difficulty in generalizing the insights from one language to another. Regarding (i), a wide range of units and relations occur for which it is not straightforward to determine whether they are or not coreferent. Although there are theoretical studies for English, they cannot always be extended to Spanish or Catalan since coreference is a very language-specific phenomenon, which accounts for (ii).

In the following we present some of the linguistic issues more problematic in relation to coreference annotation, and how we decided to deal with them in AnCora-CO (Recasens, 2008). Some of them are language dependent (1); others concern

[3] Freely available for research purposes from the following URL: http://clic.ub.edu/ancora

[4] Reuters Corpus RCV1 is distributed by NIST at the following URL: http://trec.nist.gov/data/reuters/reuters.html

the internal structure of the mentions (2), or the type of coreference link (3). Finally, we present those NPs that were left out from the annotation for not being referential (4).

1. Language-specific issues
   - Since Spanish and Catalan are pro-drop languages, elliptical subjects were introduced in the syntactic annotation, and they are also annotated with coreference.
   - Expletive *it* pronouns, which are frequent in English and to a lesser extent in Spanish and Catalan are not referential, and so they do not participate in coreference links.
   - In Spanish, clitic forms for pronouns can merge into a single word with the verb; in these cases the whole verbal node is annotated for coreference.
2. Issues concerning the mention structure
   - In possessive NPs, only the reference of the thing possessed (not the possessor) is taken into account. For instance, *su libro* 'his book' is linked with a previous reference of the same book; the possessive determiner *su* 'his' does not constitute an NP on its own.
   - In the case of conjoined NPs, three (or more) links can be encoded: one between the entire NPs, and additional ones for each of the constituent NPs. AnCora-CO captures links at these different levels.
3. Issues concerning types of coreference links
   - Plural NPs can refer to two or more antecedents that appear separately in the text. In these cases an entity resulting from the addition of two or more entities is created.
   - Discourse deixis is kept under a specific link tag because not all coreference resolution systems can handle such relations.
   - Metonymy is annotated as a case of identity because both mentions pragmatically corefer.
4. Non-referential NPs
   - In order to be linguistically accurate (van Deemter and Kibble, 2000), we distinguish between referring and attributive NPs: while the first point to an entity, the latter express some of its properties. Thus, attributive NPs like apposition and predicative phrases are not treated as identity coreference in AnCora-CO (they are kept distinct under the 'predicative link' tag).
   - Bound anaphora and bridging reference go beyond coreference and so are left out from consideration.

The annotation process of the corpora is outlined in the next section.

3.2 Annotation process

The Ancora coreference annotation process involves: (a) marking of mentions, and (b) marking of coreference chains (entities).

(a) Referential full NPs (including proper nouns) and pronouns (including elliptical and clitic pronouns) are the potential mentions of a coreference chain.

(b) In the current task only identity relations (coreftype="ident") will be considered, which link referential NPs that point to the same discourse entity. Coreferent mentions are annotated with the attribute *entity*. Mentions that point to the same entity share the same entity number. In Figure 1, for instance, *el reial decret aprovat ahir pel Consell_de_Ministres* 'the royal decree approved yesterday by the Council of Ministers' is entity="entity9" and *la nova norma* 'the new rule' is also entity="entity9" because they corefer. Hence, mentions referring to the same discourse entity all share the same entity number.

The corpora were annotated by a total of seven annotators (qualified linguists) using the AnCoraPipe annotation tool (Bertran et al., 2008), which allows different linguistic levels to be annotated simultaneously and efficiently. AnCoraPipe supports XML in-line annotations.

An initial reliability study was performed on a small portion of the Spanish AnCora-CO corpus. In that study, eight linguists annotated the corpus material in parallel. Inter-annotator agreement was computed with Krippendorff's alpha, achieving a result above 0.8. Most of the problems detected were attributed either to a lack of training of the coders or to ambiguities that are left unresolved in the discourse itself. After carrying out this reliability study, we opted for annotating the corpora in a two-stage process: a first pass in which all mention attributes and coreference links were coded, and a second pass in which the already annotated files were revised.

4 Conclusions

The SemEval-2010 multilingual coreference resolution task has been presented for discussion. Firstly, we aim to promote research on coreference resolution from a learning-based perspective in a multilingual scenario in order to: (a) explore portability issues; (b) analyze language-specific tuning requirements; (c) facilitate cross-linguistic comparisons between two Romance languages and between Romance languages and English; and (d) encourage researchers to develop linguistic resources – annotated corpora – oriented to coreference resolution for other languages.

Secondly, given the complexity of the coreference phenomena we split the coreference resolution task into two (full coreference chains and pronominal resolution), and we propose two different scenarios (gold standard vs. automatically generated input information) in order to evaluate to what extent the performance of a coreference resolution system varies depending on the quality of the other levels of information.

Finally, given that the evaluation of coreference resolution systems is still an open issue, we are interested in comparing different coreference resolution metrics: B-cubed and CEAF measures. In this way we will be able to evaluate and compare the performance of these metrics with respect to the task objectives and system rankings.

Acknowledgments

This research has been supported by the projects Lang2World (TIN2006-15265-C06), TEXT-MESS (TIN2006-15265-C04), OpenMT (TIN2006-15307-C03-02), AnCora-Nom (FFI2008-02691-E), and the FPU grant (AP2006-00994) from the Spanish Ministry of Education and Science, and the funding given by the government of the Generalitat de Catalunya.

References

Bagga, Amit and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of Language Resources and Evaluation Conference*.

Bertran, Manuel, Oriol Borrega, Marta Recasens, and Bàrbara Soriano. 2008. AnCoraPipe: A tool for multilevel annotation, *Procesamiento del Lenguaje Natural*, n. 41: 291-292, SEPLN.

Denis, Pascal and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2008)*.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. *Proceedings of HLT/NAACL 2005*.

McCarthy Joseph and Wendy Lehnert. 1995. Using decision trees for coreference resolution. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*.

Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, and 17th International Conference on Computational Linguistics (COLING-ACL98)*.

Morton, Thomas. 2000. Using coreference for question answering. *Proceedings of the 8th Text REtrieval Conference (TREC-8)*.

Ng, Vincent. 2008. Unsupervised models for coreference resolution. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2008)*.

NIST. 2003. In *Proceedings of ACE 2003 workshop*. Booklet, Alexandria, VA.

Recasens, Marta. 2008. *Towards Coreference Resolution for Catalan and Spanish*. Master Thesis. University of Barcelona.

Steinberger, Josef, Massimo Poesio, Mijail Kabadjov, and Karel Jezek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43:1663–1680.

Taulé, Mariona, Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel corpora with coreference information for Spanish and Catalan. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2008)*.

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

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