

Evalita 2011: Anaphora Resolution Task

Olga Uryupina¹ and Massimo Poesio^{1,2}

¹ University of Trento

² University of Essex

uryupina@gmail.com,

poesio@essex.ac.uk

Abstract. This paper discusses the Anaphora Resolution task proposed as a track at the Evalita-2011 evaluation campaign for the Italian language. The annotation guidelines have been designed to cover a large variety of linguistic phenomena related to anaphora/coreference. We describe the annotation scheme, the evaluation methodology and the participants' results.

Keywords: Anaphora Resolution, Coreference Resolution, EvalIta 2011, Computational Linguistics, evaluation, Italian

1 Introduction

Anaphora Resolution is a vital prerequisite for a variety of high-level Natural Language Processing tasks, such as, for example, Document Summarization or Question Answering. For the past two decades, it has received a lot of attention from the computational linguistics community, leading to the development of robust and complex approaches to the task. Most of this algorithms, however, have been created for and applied to the English data.

In the present report, we discuss the Anaphora Resolution task proposed as a track at the Evalita-2011 evaluation campaign for the Italian language. This task provides several additional challenges compared to the evaluation on the English data: first, several phenomena, for example, zero pronouns, require language-specific treatment and second, preprocessing information for Italian might be less accurate and therefore, a coreference resolution system should be able to cope with a noisy input.

The annotation guidelines used for the task have been designed to cover a large variety of linguistic phenomena related to coreference. Thus, unlike in the ACE-style annotation schemes (for example, the one used for the Evalita-2009 LEDR track [1]), our guidelines do not restrict the scope of annotation to specific semantic types of entities. This makes the current dataset more challenging for automatic processing.

In the next section, we provide a definition of the Anaphora Resolution task. In Section 3 we describe the annotation guidelines and provide some corpus statistics. In Section 4 we discuss the evaluation methodology and provide information on the results obtained by the participants.

2 Task description

In the Anaphora Resolution task, we measure a system’s ability to correctly recognize *mentions* of the same real-world *entity* within a given document.

In the provided data, we distinguish between several types of mentions: names, nominals (NPs) and pronouns, including zeroes. We also encode the information on the semantic class of a mention (cf. Table 1 for details). The participants, however, are not expected to submit any classification of mentions into their types and semantic classes.

To annotate chains within a document, we follow a modified version of the guidelines for the SemEval-2010 Task 1 on Multilingual Coreference Resolution [2]. The following section gives a brief overview of the scheme. We have extended the SemEval-2010 Task 1 guidelines to provide *minimal spans* for each mention, computed semi-automatically. Similar to the ACE-style evaluation, we use minimal spans to provide better alignment of system and gold mentions.

3 The LiveMemories Corpus

The LiveMemories corpus of Italian [3], used for the current task, is being annotated according to guidelines derived from the guidelines for the ARRAU corpus (in English) and the VENEX corpus (in Italian). The guidelines differ considerably from those used for the Evalita-2009 LEDR task and are more similar to those of OntoNotes (CoNLL-2010 shared task [4]).

The corpus has been originally annotated in the MMAX format and then converted to the CoNLL format to be used for Evalita 2011. Some information is only available in the original MMAX dataset.

There is one markable for every NP (not only referring NPs) and, unlike in Evalita-2009, there are no restrictions to anaphoric links between mentions of entities of a certain type. The full list of used semantic types, along with the correspondent corpus statics, is provided in Table 1.

An attribute of the markable specifies the logical form value of the NP:

- non-referring, i.e., not introducing a discourse entity - a subordinate attribute specifies whether expletive, idiom, predicative, quantifier, or coordination (see below)
- referring, which can be in turn discourse new or discourse old

This information is not encoded explicitly in the CoNLL format, but can be inferred from the lack of semantic type for a given mention.

The position taken on the most commonly controversial issues is described below.

Predicative NPs are not anaphorically linked with a mention of the entity of which the predication is made: e.g., in (the Italian version of):

- (1) [the broadest measure of trade], known as [the current account]

Table 1. Corpus statistics for the training set: mentions of different semantic types

abstract	7211	(24.0%)
person	5072	(16.9%)
- (non-referring markable)	4263	(14.2%)
concrete	2798	(9.33%)
gsp	2710	(9.03%)
time	2485	(8.28%)
location	2283	(7.61%)
organization	1469	(4.89%)
facility	1351	(4.50%)
numeral	244	(0.81%)
animate	71	(0.23%)
unmarked	19	(0.06%)
unknown	6	(0.02%)
total	29982	

[the current account] is not anaphorically linked with [the broadest measure of trade]

In case of plural reference to multiple antecedents introduced by singular NPs, split antecedents are marked both when the two NPs are not coordinated, as in:

- (2) [Giovanni]_i incontro' [Giuseppe]_j. [I due ragazzi]_{i,j} andarono al cinema.

and - more controversially - when they are coordinated:

- (3) [Giovanni]_i e [Giuseppe]_j si incontrarono. [I due ragazzi]_{i,j} andarono al cinema.

(i.e., the coordinated NP is not marked as antecedent).

In the MMAX format, this information is encoded as split antecedents. In the CoNLL format, it has been omitted.

In the MMAX version of the corpus discontinuous markables are used for cases of coordination in which a single modifier modifies two heads with disjoint reference, as in:

- (4) studenti e docenti dell'Universita' di Trento

In this example, a discontinuous markable is created for “[studenti .. dell'Universita' di Trento]”.

In the CoNLL format, we provide both discontinuous markables (column 18) and their simplified versions (columns 17,19). The simplified version for “[studenti .. dell'Universita' di Trento]” in our example is “[studenti]”. The systems are not expected to produce discontinuous markables.

Incorporated clitics are marked on the verb, and a special tag is used to indicate the type of clitic. E.g., in

(5) [Giovanni]_i e' un seccatore. Non [dargli]_i retta.

the verb “dargli” is treated as a markable and linked to “Giovanni”.

Such markables receive mention type “verbale”.

Several markable attributes markables have been eliminated from the dataset in the CONLL format. They include:

- agreement features (gender, number)
- grammatical function

The full MMAX annotation will be made available to groups who are interested after the competition.

4 Evaluation

4.1 Scoring metrics

We use a variant of the scorer developed for the CoNLL-2011 shared task on Coreference Resolution [4]. It provides all the 5 metrics commonly used in the coreference community: MUC, B3, CEAF- ϕ_3 , CEAF- ϕ_4 and Blanc. Following the practice established at CoNLL-2011, we rely on the average of MUC, B3 and CEAF- ϕ_4 to rank the systems.

We have modified the CoNLL scorer to allow for partial alignment between system and gold mentions according to the MUC/ACE guidelines. If a system mention includes a minimal span of a gold mention and is included in its maximal span, the two get aligned and the system receives no penalty. The maximal span corresponds to the annotated mention boundaries, and the minimal span – to the semantic head for nominals and to the NE part for proper names. For example, “sul lago” has a minimal span “lago”. This is a notable difference from the SemEval alignment algorithm, where the syntactic head was considered to be a minimal span (“sul lago” would have a head “sul”).

4.2 System results

Three participants have initially registered for the task. Unfortunately, only one group, the University of Pisa, has submitted their runs. Table 2 shows the scores obtained by the group’s submissions.

Table 2. Official results: F-scores for 5 different metrics.

	MUC	B3	CEAF _m	CEAF _e	BLANC
UniPisa, run 1	26.36	83.79	72.99	78.89	55.94
UniPisa, run 2	25.07	83.64	72.53	78.38	55.8

5 Conclusion

In this report, we have presented an overview of the Anaphora Resolution track at the Evalita-2011 evaluation campaign.

We have expected this task to be challenging: it involves complex language-specific modeling and also requires a robust approach that can deal with the noise from various preprocessing modules. This complexity can explain the fact that the task has been too difficult for the participants: only one of the three registered teams has submitted their runs.

References

1. Bartalesi Lenzi, V., Sprugnoli, R.: EVALITA 2009: Description and results of the local entity detection and recognition (LEDR) task. In: Proceedings of Evalita-2009. (2009)
2. Recasens, M., Màrquez, L., Sapena, E., Martí, M.A., Taulé, M., Hoste, V., Poesio, M., Versley, Y.: Semeval-2010 task 1: Coreference resolution in multiple languages. In: Proceedings of SEMEVAL 2010, Uppsala (2010)
3. Rodriguez, K.J., Delogu, F., Versley, Y., Stemle, E., Poesio, M.: Anaphoric annotation of wikipedia and blogs in the live memories corpus. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. (2010)
4. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011), Portland, Oregon (2011)