Anaphora Resolution Task at Evalita 2011

Massimo Poesio and Olga Uryupina

August 18, 2011

1 Task description

In the Anaphora Resolution task, systems are required to recognized *mentions* (names, pronouns, including zeroes, and nominals) in a document and cluster them into coreference chains. All the mentions in a given chain should refer to the same *entity* and vice versa.

We follow a modified version of the guidelines for the SemEval-2010 Task 1 on Multilingual Coreference Resolution [2]. Please see Section 4 for the differences in the mention alignment algorithm.

2 The LiveMemories Corpus

The LiveMemories corpus of Italian [3] 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 [1]).

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, it will be made public after the competition.

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. Possessive pronouns are treated as NPs and therefore as markables as well.

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 as follows:

- 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]

  [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.

Language-specific issues:

- 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".

Quite a bit of information is annotated about markables that was not included in the CONLL format. This includes:

- agreement features (gender, number)

- grammatical function

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

3 Data Format

The dataset follows the format of the SemEval-2010 Task 1 on the Multilingual Coreference Resolution, with one token per line and an empty line after each sentence. Additional information, extracted with the TextPro toolkit and the MALT parser, is provided in tab-separated columns.

The columns follow the following specification:

1. word identifiers in the sentence (sentence boundaries are detecting automatically, using the TextPro toolkit),
2. word forms (tokens are extracted automatically with TextPro),
3. word lemmas (gold standard manual annotation) (no gold lemmata provided),
4. word lemmas predicted by an automatic analyzer (lemmata are extracted automatically with TextPro),
5. coarse part of speech (no gold POS provided),
6. same as 5 but predicted by an automatic analyzer (POS assigned by TextPro),
7. gold morphological features (not provided)
8. automatic morphological features (extracted with TextPro; NB: this column may contain space characters)
9. for each word, the ID of the syntactic head; '0' if the word is the root of the tree (not provided),
10. same as above, but predicted by an automatic analyzer (extracted from the output of the MALT parser),
11. dependency relation labels corresponding to the dependencies described in 9 (not provided),
12. same as 11 but predicted by an automatic analyzer (extracted from the output of the MALT parser),
13. mentions – semantic types (annotated manually)
14. same as 13 but predicted by a named entity recognizer (not provided),
15. (not provided),

16. (not provided),
17. entity annotation in BIO format, no discontinuous mentions, cf. below (annotated manually)
18. entity annotation in BIOM format, includes discontinuous mentions, cf. below (annotated manually)
19. entity annotation in the SemEval format, no discontinuous mentions, cf. below (annotated manually)

Coreference is encoded in the three last columns. In the column 17, we use a variant of the BIO format to provide complex labels. We separate multiple annotations with "@". Each annotation contains the following attributes, separated by "=":

- mention id (unique within a document)
- entity id (mentions from the same coreference chain share the same id)
- mention type
- semantic type

In the column 18, we provide the same information, but we allow for discontinuous mentions (the beginning of the second and any further parts of a mention is marked with "M-"). The systems are not expected to provide discontinuous mentions in their response.

Finally, in the column 19, we provide the information on coreference chain in the SemEval format. Multiple annotations are separated with "|". Each mention is shown at its first and last token with an entity id and round brackets. Note that entity ids in the BIO and SemEval columns do not correspond.

An example annotation is shown in Table 1 (we have omitted all the non-anaphora columns and replaced "nominal" with "nom" for simplicity). This snippet contains 3 mentions: "regione di cammino di ronda", "cammino di ronda" and "il rarissimo esempio in regione di cammino di ronda". Note that some of them overlap (at tokens 8,10,11 and 12).

Ritenuto	O	O	_
famoso	O	O	_
per	O	O	_
il	B-M_19=set_25=nom=null	B-M_19=set_25=nom=null	(11
rarissimo	I-M_19=set_25=nom=null	I-M_19=set_25=nom=null	_
esempio	I-M_19=set_25=nom=null	I-M_19=set_25=nom=null	_
in	I-M_19=set_25=nom=null	I-M_19=set_25=nom=null	_
regione	I-M_19=set_25=nom=null@B-M_20=set_26=nom=gsp	I-M_19=set_25=nom=null@B-M_20=set_26=nom=gsp	(12)
di	I-M_19=set_25=nom=null	I-M_19=set_25=nom=null	_
cammino	I-M_19=set_25=nom=null@B-M_21=set_27=nom=facility	I-M_19=set_25=nom=null@B-M_21=set_27=nom=facility	(13
di	I-M_19=set_25=nom=null@I-M_21=set_27=nom=facility	I-M_19=set_25=nom=null@I-M_21=set_27=nom=facility	_
ronda	I-M_19=set_25=nom=null@I-M_21=set_27=nom=facility	I-M_19=set_25=nom=null@I-M_21=set_27=nom=facility	13)|11)
,	O	O	_

Table 1: Annotation example

The participants are expected to provide their output in the same format as the column 17. The remaining two columns (18 and 19) have only been provided for convenience (for example, to allow participants to re-use their software developed for the SemEval task).

Please note, that at the test data will not contain columns 13, 17, 18 and 19.

4 Evaluation

We will use a variant of the scorer used for the CoNLL-2011 shared task on Coreference Resolution [1]. We will provide 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 will 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").

Thus, in the snippet from Table 1, the gold mention M_{19} ("il rarissimo esempio in regione di cammino di ronda") would be aligned with system mentions "il rarissimo esempio", "esempio" 'and so on.

Minimal spans for gold mentions have been computed through a semi-automatic procedure. The current version of the training data does not provide any information on the minimal spans. We will provide it in the updated version of the training set.

5 Submission of system results

Please send your system response to Olga Uryupina (uryupina@gmail.com). The deadline is 14th October 2011, midnight (GMT+1).

Each participant has a possibility to submit a maximum of two runs. Each run should consist of a single data file and should contain at least 3 columns: word id (column 1 in the training set), token (column 2) and anaphora (column 17, should be the last one in the system run file). The systems are not expected to recognize mention types and semantic types and can use arbitrary values for the output.

The test set will be distributed on the 4th of October 2011. An improved version of the training set will be made available on September 12th. It will contain the same documents, but the annotation will be augmented with minimal

spans and some errors and inconsistencies will be corrected.

References

[1] Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland, Oregon, June 2011.

[2] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proc. SEMEVAL 2010*, Uppsala, 2010.

[3] Kepa Joseba Rodrguez, Francesca Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of the 7th International Conference on Language Ressources and Evaluation*, 2010.