Online Evaluation of Coreference Resolution.

Susanne Salmon-Alt, Laurent Romary, Andrei Popescu-Belis, Loïs Rigouste

Online Evaluation of Coreference Resolution

Andrei Popescu-Belis
ISSCO/TIM/ETI, University of Geneva
40, bd du Pont d'Arve
CH-1211 Geneva 4 – Switzerland
andrei.popescu-belis@issco.unige.ch

Loïs Rigouste
ENST
F-75012 Paris – France
rigouste@enst.fr

Susanne Salmon-Alt
ATILF-CNRS
F-54XXX Nancy – France
susanne.alt@loria.fr

Laurent Romary
LORIA / Langue et Dialogue
F-54XXX Nancy – France
laurent.romary@loria.fr

Abstract

Not needed for submission.

Keywords (max. 5)

Evaluation of coreference resolution, online resources, comparison of metrics
[Until here: 67 words]

1. Modelling coreference resolution

The detection of coreference links between referring expressions (REs) is an essential step in automatic text understanding. To evaluate how well a program or a human perform on this task, we need more than comparing links one by one, since the same understanding of a text could be derived from different sets of links. Indeed, what matters is that REs are understood as referring to the correct conceptual entities in the real world.

1.1. Varieties of coreference

REs may engage in a variety of referring relations – traditionally called anaphora or coreference. Anaphora is a relation between an antecedent RE and an anaphoric RE: the referent of the latter is determined by the referent of the former. Identity-of-sense, identity-of-sense, and bridging are traditional types of anaphora. Coreference is a relation between two REs that have the same referent, and may or may not coincide with an anaphoric relation.

1.2. Theoretical approach to coreference resolution

Evaluation requires an accurate definition of the task. We view anaphora as an asymmetric link between the antecedent and the anaphor. Since there is the correct antecedent may not be unique, we believe that evaluation of anaphora resolution requires knowledge of all coreference (identity) links. Therefore, we focus here on coreference.

Coreference resolution consists in find the correct coreference links between REs (strictly speaking, identifying REs is not part of coreference resolution). Coreference links are transitive, therefore they generate equivalence classes, each class containing all REs that point to ("refer to") the same entity. Coreference resolution amounts to finding the correct equivalence classes, no matter what links are used to construct them.

Non identity coreference is qualitatively different. For instance, in bridging or associative anaphora, the bridging relation holds in fact between two entities, not between two REs (it is a conceptual relation). Therefore, the construction of such relations should be evaluated after coreference resolution, using the equivalence classes, not individual REs. The metrics below do not tackle this problem.

1.3. Targeted application domain

Evaluation metrics for coreference resolution have two applications. They can be used to compare the performance of a program on this task, given a correct answer defined by human judges. But they serve also to measure agreement between human judges (inter-annotator agreement), which is often not perfect. The value of inter-annotator agreement is an upper bound on this performance expected from systems.

2. Data representation of coreference evaluation

2.1. Annotation model

Based on the general principles currently adopted within ISO TC37/SC4, and on the investigation we made of previous work on reference coding, a meta-model for structural constraints on any reference annotation and a core set of essential data categories has been proposed. The important features of the meta-model are:

- stand-off annotation to account for annotating different linguistic levels of the same primary data and for comparing different annotations for the same linguistic level;
- autonomous 'link' level for representing unambiguously and simultaneously disjoint targets (universe entities or antecedents) as well as more than one referential link for the same referring expression;
- reified markable on the reference level to take into account any type of information as input for reference annotation (surface strings, morphological entities, syntactic chunks, gesture representations, universe entities,...), for adding any type of information on markables (often user-defined and heterogeneous) and for recursive structure on markables.

2.2. Conversion to pivot format

Conversion tools (namely, XSLT scripts) have been designed to convert various existing formats for coreference to a pivot format based on the annotation model above. An example of a coreference relation between two REs (or markables) is provided below:

```
<?xml version="1.0" encoding="UTF-8" ?>
```

```
<struct type="reference_annotation_collection">
  <struct type="markable" id="markable_1">
    <feat type="sourcetext" target="word11..word12">The man</feat>
  </struct>
  <struct type="markable" id="markable_2">
    <feat type="sourcetext" target="word18">he</feat>
  </struct>
  <struct type="reflink" id="reflink_1">
    <feat type="reflinktype">identity</feat>
    <feat type="source" target="markable_2" />
    <feat type="target" target="markable_1" />
  </struct>
</struct>
```

XSLT stylesheets are provided to convert from eight existing formats for coreference annotation to the pivot format. In addition, a "repair" script is provided to clean source files that do not fully respect the XML syntax.

2.3. Online evaluation interface

The evaluation interface, implemented in Perl (using CGI and XML) allows users to evaluate their systems over the Internet, by providing in one of the supported annotation formats a "key" (correct file) and a "response" (containing the performance to evaluate, from a system or a second human annotator), using an upload button on the interface's homepage.

Initial conversion to the pivot format (a pre-processing option) outputs messages allowing users to check that the right rules were applied. The resulting pivot files can be displayed. If the option "synchronized re-indexing of markable IDs" is selected, then the key and response pivot files are scanned for markables (REs), and these are re-indexed and sorted in ascending order of their 'target' attributes, so that in the two files the same IDs point to the same markable (complex recursive markables are also supported).

Then, starting from synchronized pivot files, the scores are computed from the markable IDs only, first building equivalence classes (partitions) from markables in the key and the response. If the markables (REs) declared in both files are the same, scores are computed by applying the comparison functions that take two partitions of the same set as input (defined below in section 3). If the key and response REs differ, several matching strategies are proposed to the user (cf. 2.4 below). Finally the scores of the five implemented metrics are displayed.

2.4. Synchronisation of markables

If the markables in the key do not match exactly those in the response, several options are in theory available. One is to argue that the markable identification task is distinct from coreference resolution, and should be evaluated separately. Another option is to synchronise the key and response markable sets so that they become identical (of course with different coreference links), and this can be done in four ways, by using, respectively, the intersection, the union, the key or the response sets. The four score sets (for coreference) are then all displayed in the evaluation interface, along with a precision/recall score on the RE identification task. As shown below, substantial differences may appear between the four score sets.

3. Overview of evaluation measures

3.1. Formalism: partitions and projections

A unified formal framework describing the various evaluation metrics has been defined. A key notion is that the REs in a text are partitioned by the various referents in classes of coreferent REs (equivalence classes). If an entity is referred to just once in the text, the corresponding RE forms a singleton class. Evaluating coreference resolution amounts to comparing two partitions of the same set of REs. Note that other referential links, such as whole/part are better formalized as links between classes rather than REs.

A useful notion is the projection of a class, for instance from the key, onto the response partition: the projection is the set of all intersections of the key class with response classes. The number of projections of a class varies between 1 and the size of the class. Conversely, response classes can also be projected. Intuitively, the closer the partitions are, the smaller the size of projections is.

3.2. Implemented metrics

Since the first attempt to define an evaluation measure for coreference at the MUC-6 conference, other proposals attempted to improve existing measures. The five implemented in our interface are:

- the MUC measure (M) computes the numbers of missing and superfluous coreference links in the response depending only on the equivalence classes (the MUC count is in fact too indulgent);
- the B^3 measure (B) also defines recall and precision, but compute it per RE, then average the values to obtain global scores (the scores are lower than MUC when many REs are unduly grouped, but still well above 0%);
- the kappa factor (K) can be also applied to coreference, especially to measure inter-annotator agreement. However, it is computed by estimating the probability of agreement by chance using a series of assumptions that are subject to discussion. Kappa is less indulgent than MUC, but bears less information (one score vs. two);
- the core-DE (discourse entity) measure (C) is based on the construction of core-DEs, that is, the program's view (response) of each correct DE, and counts missing REs as recall errors. It was proved that these scores are lower than MUC on any response;
- the mutual information measure (H) is based on the analogy with communication channels and the notion of mutual referring information. Recall and precision measure, respectively, irrelevant information gains and loss of information.

Recall and precision for the M-B-C and H measures vary from 0 to 1. The K-score varies from –1 to +1: +1 for perfect agreement, 0 for random agreement, –1 and less for negative statistical correlation.

3.3. Advantages and drawbacks

The measures have various advantages and drawbacks, i.e. they do not always reflect accurately the "quality" of a response, being often quite "lenient". Several criteria can be defined to assess the properties of a measure (meta-evaluation). Our evaluation interface displays all the five scores: their concordant variation is a good sign of reliability. Not all existing measures were implemented, e.g., "descriptive specificity", a version of (C). Also, the evaluation of anaphoric links must be dealt with separately.

4. Exploitation of the online evaluator

The evaluator was developed and tested in an ongoing project about corpora annotated with coreference links. A series of texts annotated by two evaluators were available, as well as various constructed examples on which scores were previously computed by hand. Some of the texts were annotated for specific phenomena (certain types of coreference), therefore they are not always typical of coreference. We compared in particular the four strategies proposed when the RE set differs between key and response.

For instance, on one text only coreferences induced by definite anaphora links were annotated. Here, despite different RE sets between annotators, the scores do not vary significantly with the RE combination strategy. Given the reduced number of links (150 for 350 REs) and the fact that classes have 1 or 2 REs (a very particular situation), the (M) score is low while the (H) score is high, since singletons are preserved.

On another set of texts, only coreferences induced by demonstrative anaphora links were annotated. Here, in some cases, the differences in RE sets induced significant variation among the four strategies. Also, in some cases, the (H) scores are much lower than the (M) scores, which shows that (H) and (M) are not always comparable. Examples can be found where the (K) score is lower than -1, which is against its original definition. This shows that the calculation of (K) for coreference resolution should probably be revised.

5. Conclusion

The coreference evaluator, now available online, will be an essential resource for coreference studies, especially on large corpora, where manual evaluation is impossible. Since many annotation formats and evaluation metrics are supported, the evaluator should be flexible enough for many categories of users. Further work should extend it towards non-identity coreference and anaphora resolution.