Online Evaluation of Coreference Resolution

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

Loïs Rigouste
ENST
46, rue Barrault
F-75634 Paris Cedex 13
France
rigouste@enst.fr

Susanne Salmon-Alt
ATILF-CNRS
44, avenue de la Libération
B.P. 30687
F-54063 Nancy Cedex
France
susanne.alt@loria.fr

Laurent Romary
LORIA
Campus Scientifique
BP 239
F-54506 Vandoeuvre-lès-Nancy Cedex
France
laurent.romary@loria.fr

Abstract

This paper presents the design of an online evaluation service for coreference resolution in texts. We argue that coreference, as an equivalence relation between referring expressions (RE) in texts, should be properly distinguished from anaphora and has therefore to be evaluated separately. The annotation model for coreference is based on links between REs. The program presented in this article compares two such annotations, which may be the output of coreference resolution tools or of human judgement. In order to evaluate the agreement between the two annotations, the evaluator first converts the input annotation format into a pivot format, then abstracts equivalence classes from the links and provides five scores representing in different ways the similarity between the two partitions: MUC, B3, Kappa, Core-discourse-entity, and Mutual-information. Although we consider that the identification of REs (i.e. the elements of the partition) should not be part of coreference resolution properly speaking, we propose several solutions for the frequent case when the input files do not agree on the elements of the text to consider as REs.

Introduction

Reference resolution is an essential step in automatic text understanding. Therefore, the detection of coreference and anaphoric links between referring expressions (REs) has been a constantly active research topic in the past decades. Despite several evaluation campaigns, such as the MUC series (MUC-6, 1995; MUC-7, 1998), there seems to be no general agreement on a standard evaluation measure for this type of task. There has been however progress on the annotation model for this problem and resources annotated for coreference (Poesio, Bruneseaux and Romary, 1999; Salmon-Alt, 2001; Salmon-Alt and Romary, 2004), and on the formalization of evaluation metrics (Popescu-Belis, 2000, 2003; Mitkov, 2001).

In this paper, we build upon previous results in the specification of coreference resolution tasks, their annotation and evaluation, in order to propose an online service for the evaluation of coreference resolution. Starting with an overview of the various aspects of coreference resolution (section 1), we proceed to discuss the various issues related to annotation formats from the point of view of an evaluation service (section 2). The implemented evaluation measures are described in section 3, and an outline of the use of the evaluator is given in section 4.

1. Scope of Coreference Evaluation

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 performs on this task, we need more than comparing links one by one, since equivalent understandings of a text could be derived from different sets of links between REs. Indeed, what matters is that REs are understood as referring to the correct conceptual entities in the real world.

1.1. Coreference and Anaphora

REs may engage in a variety of referential and/or semantic relations – traditionally called anaphora or coreference, depending on the scope of the theoretical framework or the intended application. However, as clearly stated in Van Deemter and Kibble (2000), these two relations have different properties and should be carefully distinguished.

Coreference holds between two REs that have the same referent (*a dog … the animal*). In this definition, coreference is symmetrical, reflexive and transitive and therefore an equivalence relation. On the contrary, anaphora is a relation of interpretational dependency between an antecedent RE and an anaphoric RE. In general, one considers that the referent of the latter is determined by knowledge inferred from the former. In this definition, anaphora may coincide with coreference (*a dog … the animal*), but covers also many cases where the referents of the anaphor and the antecedent are not identical. Classical cases are bridging or associative anaphora (*a dog … its owner*), while the inclusion of more complex interpretational dependency such as in "identity-of-sense" relations, "function/value" relations, predicative nominals or bound anaphora are subject of discussion (Hirschman, 1997; Van Deemter and Kibble, 2000; Davies and Poesio, 2000; Salmon-Alt, 2001; Mitkov, 2002).

1.2. Coreference Resolution Task

Evaluation requires an accurate definition of the task. In our view, coreference resolution consists in finding the correct coreference links between REs, i.e. links between expressions referring to the same extra-linguistic entity. As defined in the previous section, coreference links are transitive. Therefore they generate equivalence classes, each class containing all REs that point to ("refer to") the same referent. As a result, coreference resolution amounts to finding the correct equivalence classes, no matter what links are used to construct them.

```
<?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">coreference</feat>
    <feat type="source" target="markable_2" />
    <feat type="target" target="markable_1" />
  </struct>
</struct>
```

Figure 1. Annotation of a coreference relation between two markables in pivot format.

More complex anaphoric or referential relations are qualitatively different. For instance, in certain bridging or associative anaphora, the relation holds 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 the individual REs. More generally, for anaphora in the sense of our definition (i.e. an asymmetric link to the antecedent), the correct antecedent may not be unique, therefore we believe that the correct evaluation of anaphora resolution must rely on knowledge of all coreference links. As a consequence, we focus in the following on the evaluation of coreference relations only, and do not tackle the issue of evaluating non-coreferential anaphoric relations.

1.3. Targeted Application Domain

Evaluation metrics for coreference resolution have three applications. First, they can be used to compare the performance of a program on this task, given a correct annotation (*key* or *gold-standard*) defined by human judges. Second, they serve 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 the performance expected from systems. Third, as follows from the previous section, coreference evaluation should precede any serious evaluation of more complex interpretational dependency relations, such as bridging anaphora.

2. Data Representation and Processing

2.1. Annotation Model

Based on the general principles currently adopted by ISO TC37/SC4, and on our investigation on previous work on reference coding, a meta-model for structural constraints on any reference annotation and a core set of data categories have been proposed (Salmon-Alt and Romary, 2004). By its generality, the meta-model subsumes previous proposals for coreference annotation (Poesio, Bruneseaux and Romary, 1999; Davies and Poesio, 2000). 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. As shown in Figure 1, the stand-off annotation is realized by pointers (*target* attributes of markable elements) to the primary source, i.e. a reference file containing uniquely identified primary linguistic units ("words");
- *autonomous markable elements* on the reference level to take into account any type of information as input for reference annotation (surface strings, morphological entities, syntactic chunks, etc.), to add any type of information on markables (often user-defined and heterogeneous) and to allow recursive structure on markables. Figure 1 shows two simple markable elements, one for the RE "the man" and another for "he";
- *autonomous link elements* for representing unambiguously and simultaneously disjoint targets (universe entities or antecedents) as well as more than one referential link for the same referring expression. Figure 1 gives an example of a coreference link between markable 1 and markable 2.

2.2. Data Conversion and Preprocessing

Computing the coreference evaluation scores relies on data fulfilling the following conditions:

- the input files must be well-formed XML files;
- the annotation format should fit the requirement of a pivot format, corresponding to the above introduced annotation model;
- in both files, markables pointing to the same primary data ("words") must have the same ID.

First, in case the input files do not fully respect the XML syntax (case of SGML files, for example), a "repair" script is provided, trying to clean them automatically, for instance by adding the XML header, and quotes around attributes.

Second, to make the evaluation tool compatible with current coreference encoding practice, conversion tools (XSL stylesheets) have been designed to convert various existing formats for coreference to a pivot format based on the annotation model described above. The supported schemes are MUC (Hirschman, 1997), MATE (Davies and Poesio, 2000), MMAX (Müller and Strube, 2001) and several proprietary formats used formerly in the authors' projects. The conversion to the pivot format outputs messages allowing users to check that the correct rules were applied. The resulting pivot files can be displayed.

Third, the tool proposes a synchronized re-indexing of markable IDs. This option is necessary in case the annotators (human or machine) had the possibility of creating or deleting markables during the annotation of links[1]. If this option 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. As a result, markables with the same IDs in both files point to the same primary source elements (complex recursive markables are also supported).

2.3. Synchronization of Markables

All the measures we implemented rely on the hypothesis that the set of markables in both files is the same (see note 1). In case the markables do not match this constraint, 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 synchronize the key and response markable sets so that they become identical (of course with different coreference links). However, there is no "perfect synchronisation". Therefore, we propose to the users four possibilities: the intersection of the markable sets, or their union, or simply the key set, or the response set. The four sets of scores 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.

2.4. Online Evaluation Interface

The evaluation interface, implemented in Perl (using CGI and XML) allows users to evaluate their annotations 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 (see *http://ananas.loria.fr*). Some scores are symmetric with respect to the key and the response, i.e. they do not change if the files are switched: for instance, *kappa* or the *f-measures*. Recall and precision scores are switched too if the files are switched.

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 sets of markables (REs) declared in both files are the same, the 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 applied (cf. 2.3). Finally the scores of the five implemented metrics are displayed.

[1] Theoretically, the markables in both files should be same, since the identification of markables is often supposed to be an input for the linking procedure. However, this requirement is not always applicable in practice (for a critical discussion, see Van Deemter and Kibble, 2000) and some of the current annotation tools (for example MMAX, Müller and Strube, 2001) do not require a separation of these two tasks.

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 set of REs or markables in a text is partitioned by the various referents in equivalence classes of coreferent REs. 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 interpretational links, such as *whole/part* are better formalized as links between classes rather than between 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 number of projections is; when the partitions are identical, each equivalence class projects onto exactly one class (itself).

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 metrics 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). Missing links are equated to recall errors, while superfluous links count as precision errors, two measures inspired from information retrieval (Salton and McGill, 1983);
- the B^3 measure (B) also defines recall and precision (Bagga and Baldwin, 1998), but computes it *per RE*, then averages 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) (Krippendorff, 1980; Carletta, 1996) can be also applied to coreference, especially to measure inter-annotator agreement (Passonneau, 1997). 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. These scores are lower than MUC on any response (Popescu-Belis, 2000);
- the mutual information measure (H) is based on the analogy with communication channels and the notion of mutual referring information (Popescu-Belis, 2000). 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 meta-evaluation criteria can be defined to assess the properties of a measure (Popescu-Belis, 2000). Our evaluation interface displays all the five scores: their *concordant variation* is a good sign of reliability. However, not all existing measures were implemented, e.g., "descriptive specificity", a version of (C). Also, the evaluation of bridging or non coreferential pronominal anaphora must be dealt with separately – a consequence of our theoretical choices.

4. Discussion 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. However, some of the texts were annotated for specific phenomena (only certain types of expressions), therefore they are not always typical. 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, computed according to (Passonneau, 1997) is lower than -1, which is against its original definition (Krippendorff, 1980). 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.

Acknowledgments

The work presented here was supported by the CNRS (France) through the ANANAS project "Annotation Anaphorique pour l'Analyse Sémantique de Corpus", *http://www.atilf.fr/ananas/*.

References

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. Proceedings LREC'98 Workshop on Linguistic Coreference. Granada, Spain.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22(2), p. 249–254.

Davies S., Poesio M. (2000). MATE Deliverable 1.1, Chapter 3: Coreference. Available at: *http://www.cogsci.ed.ac.uk/~poesio/MATE/coreference.html*.

Hirschman, L. (1997). MUC-7 Coreference Task Definition. The MITRE Corporation.

Krippendorff, K. (1980). Content Analysis: An Introduction to Its Methodology. Sage Publications.

Mitkov, R. (2001). Towards a more consistent evaluation and comprehensive evaluation of anaphora resolution algorithms and systems. Applied Artificial Intelligence 15(3), p. 253–276.

Mitkov, R. (2002) Anaphora Resolution. Studies in Language and Linguistics. Longman.

MUC-6 (1995) Proceedings of the Sixth Message Understanding Conference. Morgan Kaufmann.

MUC-7 (1998) Proceedings of the Seventh Message Understanding Conference. Available at *http://www.itl.nist.gov/iad/894.02/related_projects/muc/*.

Müller Ch., Strube M. (2001). Annotating Anaphoric and Bridging Relations with MMAX. Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

Passonneau, R. J. (1997). Applying reliability metrics to co-reference annotation. Technical Report CUCS-017-97, Dept. of Computer Science, Columbia University.

Poesio, M., Bruneseaux, F., and Romary, L. (1999). The MATE meta-scheme for coreference in dialogues in multiple language. Proceedings of the ACL Workshop on Standards for Discourse Tagging, Maryland.

Popescu-Belis, A. (1998). How corpora with annotated coreference links improve anaphora and reference resolution. Proceedings LREC'98, vol. 1, pp. 567–572. Granada, Spain.

Popescu-Belis, A. (2000). Évaluation numérique de la résolution de la référence : critiques et propositions. T.A.L. : Traitement automatique des langues, 40(2), p. 117–146.

Popescu-Belis, A. (2003). "Evaluation-Driven Design of a Robust Reference Resolution System". Natural Language Engineering, 9(2), p. 1–26.

Salmon-Alt, S. (2001). Entre corpus et théorie: l'annotation (co-)référentielle. T.A.L. : Traitement automatique des langues, 42(2), p. 459–486.

Salmon-Alt, S., and Romary, L. (2004). RAF: Towards a Reference Annotation Framework. Proceedings of LREC 2004, Lisbon.

Salton, G. and McGill, M. J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Van Deemter, K., and Kibble, R. (2000). On Coreferring: Coreference in MUC and Related Annotation Schemes. Computational Linguistics, 26(4), p. 629–637.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D. and Hirschman, L. (1995). A Model-Theoretic coreference scoring scheme Proceedings of MUC-6, pp. 45–52. Morgan Kaufmann.