Anaphora Resolution Exercise: An overview

Constantin Orăsan, Dan Cristea, Ruslan Mitkov, António Branco

University of Wolverhampton, "Alexandru-Ioan Cuza" University, University of Wolverhampton, University of Lisbon
Wolverhampton, Iasi, Wolverhampton, Lisbon
United Kingdom, Romania, United Kingdom, Portugal
C.Orasan@wlv.ac.uk, dcristea@info.uaic.ro, R.Mitkov@wlv.ac.uk. antonio.branco@di.fc.ul.pt

Abstract

Evaluation campaigns have become an established way to evaluate automatic systems which tackle the same task. This paper presents the first edition of the Anaphora Resolution Exercise (ARE) and the lessons learnt from it. This first edition focused only on English pronominal anaphora and NP coreference, and was organised as an exploratory exercise where various issues were investigated. ARE proposed four different tasks: pronominal anaphora resolution and NP coreference resolution on a predefined set of entities, pronominal anaphora resolution and NP coreference resolution on raw texts. For each of these tasks different inputs and evaluation metrics were prepared. This paper presents the four tasks, their input data and evaluation metrics used. Even though a large number of researchers in the field expressed their interest to participate, only three institutions took part in the formal evaluation. The paper briefly presents their results, but does not try to interpret them because in this edition of ARE our aim was not about finding why certain methods are better, but to prepare the ground for a fully-fledged edition.

1. Introduction

Anaphora is the linguistic phenomenon of pointing back to a previously mentioned item in the text. *Anaphora resolution* is the process of resolving an anaphoric expression to the expression it refers to. If the antecedent and the anaphor have the same referent in the real world they are *coreferential* (Mitkov, 2002). The process of building chains of coreferential entities is called *coreference resolution*. Both anaphora resolution and coreference resolution are vital to a number of applications such as machine translation, automatic summarisation, question answering and information extraction, but despite the large number of automatic methods developed during the last few years it is quite difficult to compare their results due to the fact that they usually use different evaluation methods on different corpora.

Evaluation campaigns have become an established way to evaluate the results of systems which tackle the same task. In line with this, the Anaphora Resolution Exercise (ARE)[1] was organised with the general objective of encouraging researchers to develop discourse anaphora resolution systems and evaluate them in a common and consistent environment. The first edition of the Anaphora Resolution Exercise was held in conjunction with the 6^{th} Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2007)[2] and focused only on English pronominal anaphora and NP coreference. This initial edition was used mainly as an exploratory exercise where various issues were investigated and to gain experience on how to set up the first fully-fledged edition. This paper presents an overview of the first Anaphora Resolution Exercise, and it is structured as follows: Section 2. briefly presents similar evaluation exercises in the fields of anaphora and coreference as well as evaluations in other fields. The settings of the Anaphora Resolution Exercise (ARE) are presented in detail in Section 3, followed by the evaluation results in Section 4. The paper finishes with conclusions and highlights future directions for the evaluation exercise.

2. Related work

The need to have a consistent evaluation for anaphora and coreference has been repeatedly highlighted by researchers in the field (Barbu and Mitkov, 2001; Mitkov, 2000) but the only evaluation conferences which focused explicitly on coreference and anaphora were the Message Understanding Conferences (MUC)[3] which included a coreference track. The Automatic Content Extraction (ACE)[4] also focuses to a certain extent on anaphora and coreference resolution, but the participants are required to identify only certain types of relations present between a predefined set of entities. In contrast to MUC and ACE, the long term goals of ARE are to identify a large set of linguistically motivated relations between a variety of entities, in different languages.

Evaluation conferences are common in other domains. SUMMAC (Mani et al., 1998) and DUC (TIDES, 2000) were organised in the field of automatic summarisation, TREC conferences[5] evaluate different aspects of text retrieval, whilst CLEF campaigns[6] attempt the similar tasks in a multilingual environment.

3. The Anaphora Resolution Exercise

As mentioned above, the first edition of the exercise gave us the opportunity to explore various issues that need to be tackled during the organisation of such an event. The tasks proposed to participants in this edition focused only on English anaphora and coreference. This section presents the four tasks organised in the first edition, the data used for evaluation and the evaluation metrics employed in the exercise.

[1] http://clg.wlv.ac.uk/evants/ARE/
[2] http://daarc2007.di.fc.ul.pt/
[3] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
[4] http://www.nist.gov/speech/tests/ace/index.htm
[5] http://trec.nist.gov/
[6] http://www.clef-campaign.org/

3.1. Tasks at ARE

The tasks proposed in this exercise addressed different problems related to anaphora and coreference resolution. The purpose of the exercise was to give participants the chance to test their systems in different settings. On the one hand, we wanted to be able to compare fully automatic systems that need to implement all the steps involved in the resolution processes, including the identification of referential expression and their candidates. On the other hand, we wanted to be able to assess only the resolution process when the entities that need to be resolved and their candidates are known by the systems. As can be seen these goals largely address the difference made between the evaluation of algorithms and evaluation of systems (Mitkov, 2001).

3.1.1. Task 1: Pronominal anaphora resolution on pre-annotated texts

The main purpose of the first task was to evaluate to what extent pronoun resolution algorithms work. In order to achieve this, participants were provided with documents in which the NPs were annotated. Among these NPs, some of them had an attribute which indicated that they are referential and have to be resolved. In this task only referential pronouns had this attribute (i.e. pleonastic pronouns were not annotated). The participants had to identify for each pronominal referential expression any antecedent from the list of annotated NPs. By proposing this task, we wanted to focus on how correctly pronoun resolution algorithms can select the antecedent for a pronoun from a list of known candidates and are not influenced by wrongly identified candidates.

3.1.2. Task 2: Coreferential chains resolution on pre-annotated texts

The second task of the exercise was similar to the first one in that it did not require participants to determine which expressions are referential. In contrast to Task 1, in this task participants had to process all the referential expressions, not only the pronouns, and cluster them together in coreferential chains. The input documents were annotated with information about which NPs are part of a coreferential chain (i.e. definite descriptions which were not part of a coreferential chain or non-referential pronouns were not annotated). This task was proposed in order to learn how correctly participants can cluster known entities in coreferential chains without being influenced by wrongly identified candidates.

3.1.3. Task 3: Pronominal anaphora resolution on raw texts

In most application oriented situations, anaphora and coreference resolution systems need to do more than just find the antecedent of a pronoun or cluster coreferential noun phrases together. They also need to identify which entities have to be resolved and which NPs should be considered in the process. The last two tasks address exactly this problem. In the third task of ARE, participants were given unannotated texts and they had to identify which pronouns need to be resolve and find any of their antecedents. The purpose of this task was to assess the performance of fully automatic pronoun resolution systems.

3.1.4. Task 4: Coreferential chains resolution on raw texts

In the fourth task of ARE, participants had to identify full coreferential chains in unannotated texts. This task is similar to the third task in that it does require participants to address all the steps involved in the resolution process, including the identification of referential NPs. The purpose of this task was to assess the performance of fully automatic coreference resolvers.

As can be seen, the main difference between the four tasks is that Tasks 1 and 2 evaluate the resolution algorithms on an almost perfect input (i.e. input in which the entities to be resolved are known)[7], whilst Tasks 3 and 4 simulate application oriented situations where there is no guarantee that the entities to be resolved can be correctly identified. Tasks 1 and 3 focus on pronominal anaphora resolution and require that for each referential pronoun an antecedent is determined, whilst Tasks 2 and 4 address the problem of coreference resolution where entities that refer to the same thing in real world need to be clustered together.

3.2. The data

The data used in ARE was derived from the coreferentially annotated corpus presented in (Hasler et al., 2006). This corpus contains newspaper articles extracted from the Reuters corpus (Rose et al., 2002) and totalises over 55,000 words. Due to the fact that the project which built this corpus also investigated the phenomenon of cross-document coreference, the corpus contains five clusters of related documents: Bukavu bombing, Peru hostages, Tajikistan hostages, Israel suicide bomb and China-Taiwan hijack. Four of these clusters were released as training data and the fifth one was used in the testing stage.

The corpus annotation process, described in (Hasler et al., 2006), involved first identification of all the markables (NPs) in a text regardless of whether they were coreferential or not. For the coreferential relations two types of tags were annotated: **coref** and **ucoref**. The **coref** tag was used where there was no doubt that one entity corefers with another, whilst **ucoref** marked that an annotator was relatively sure of coreference but there is an element of uncertainty. For ARE, all the coreferential links marked as **ucoref** were ignored as it was considered that if a human cannot decide with high certainty about a coreferential link, a computer should not have to determine it.

The original annotation made a distinction between different relations and types of relations between entities. The relations marked IDENTITY, SYNONYMY, GENERALISATION and SPECIALISATION between entities. Indirect anaphora, such as the *the house ... the door* relation, was not annotated. The list of possible types of relations included NP, COPULAR, APPOSITION,

[7]The input is not really perfect because the preprocessing tools employed by participants can still introduce some errors. However, there are no errors at the stage where the entities which have to be resolved are identified.

BRACKETED TEXT, SPEECH PRONOUN and OTHER. For ARE the only relations kept were IDENTITY, SYNONYMY, GENERALISATION and SPECIALISATION, all between NPs. Any other relation was removed from the corpus before it was released to the participants.

For each task a different type of input was prepared in order to facilitate evaluation.[8]

3.2.1. Input data for Task 1

The input for the first task marked all the entities which are candidates for pronouns. The input was encoded using XML and looked like in the following example:

```
<p>
  <entity id="2">Israeli-PLO relations</entity>
  have hit
  <entity id="3">a new low</entity>
  with
  <entity id="4">the Palestinian Authority</entity>
  saying
  <entity id="5">Israel</entity>
  is wrong to think
  <entity id="6">it</entity>
  can treat
  <entity id="7">the Authority</entity>
  like
 <entity id="8">a client militia</entity>
 .
</p>
```

All the entities which are candidates for pronouns were marked using the **entity** tag and were assigned an unique ID. In addition to the annotated text, the participants were given the pronouns they needed to resolve using XML format as well:

```
<anaphoric_pronouns>
  <pronoun id="6" value="it"/>
  <pronoun id="91" value="they"/>
  <pronoun id="83" value="He"/>
  <pronoun id="159" value="he"/>
  <pronoun id="167" value="his"/>
</anaphoric_pronouns>
```

3.2.2. Input data for Task 2

The input of Task 2 marked all the entities which can be involved in coreferential chains using the **entity** tag. Unique IDs identified each of these entities.

```
<p>
  At
  <entity id="18">
    an emergency summit in
    <entity id="19">Toronto</entity>
  </entity>
  ,
  <entity id="20">
    the leaders of
    <entity id="21">both nations</entity>
  </entity>
  agreed to push for
  <entity id="22">
    direct talks with
    <entity id="23">the rebels</entity>
  </entity>
  , even_though
  <entity id="24">they</entity>
  ruled out
  <entity id="229">
    <entity id="26">the guerrillas'</entity>
    non-negotiable demand --
    <entity id="27">
      freedom for
      <entity id="28">
        <entity id="29">their</entity>
      jailed comrades
      </entity>
    </entity>
  </entity>
  .
</p>
```

3.2.3. Input data for Tasks 3 and 4

For these two tasks the entities which need to be resolved were not indicated and therefore no annotation was necessary for them. However, in order to facilitate the evaluation process, the input of these two tasks is not plain text. Spaces and punctuation were preceded in the texts by the **node** tag so that snippets of texts can be easily identified. An example of a text is:

```
<p>
  <node id="26"/>
  Japan
  <node id="27"/>
  and
  <node id="28"/>
  Peru
  <node id="29"/>
  on
  <node id="30"/>
  Saturday
  <node id="31"/>
  took
  <node id="32"/>
  a
  <node id="33"/>
  tough
  <node id="34"/>
  stand
  <node id="35"/>
  on
  <node id="36"/>
  rebel
  <node id="37"/>
  demands
  <node id="38"/>
  in
```

Using this format, it is possible to indicate that the entity *Japan and Peru* starts at node 26 and finishes at node 28.[9]

3.3. Evaluation methods

The measures used to evaluate the results of the participants were success rate, precision, recall and f-measure. Given that in the tasks 3 and 4 the participants did not receive the list of entities to be resolved, the evaluation had to consider both the correctness of the resolution and how accurately the entities were identified. In light of this, an overlap measure was calculated in order to find out to what extent the entities identified by the system matched those in the gold standard. The rest of this section presents the evaluation measures for each task:

3.3.1. Evaluation method for Task 1

Given that the purpose of Task 1 was to pair referential pronouns with any of their antecedents, the evaluation metric used for this task was *success rate* defined as the number of correctly resolved anaphoric pronouns divided by the total number of anaphoric pronouns to be resolved. In computing the number of correctly resolved

[8] More detailed information about the input formats can be found at http://clg.wlv.ac.uk/events/ARE/

[9] Each word is precedes by a **node** tag which assigns the word an ID. For this reason for the entity *Japan and Peru* the end position is 28 and not 29.

anaphoric pronouns, we counted for each pronoun one of the following scores:

- 0 when the pronoun was not correctly resolved
- 0.5 when the program resolves correctly the pronoun to another pronoun, but the pronoun selected as the antecedent was not correctly resolved. This case was introduce to acknowledge the fact that the program could select a correct antecedent, although there is no correct information about this antecedent (i.e. the meaning of the referred pronoun was not correctly determined)
- 1 when the pronoun is correctly resolved to a non-pronominal entity from the chain. If a pronoun is selected as the antecedent, then there is at least one antecedent in the co-reference chain which is non-pronominal.

3.3.2. Evaluation method for Task 2

The evaluation metrics used in Task 2 were precision, recall and f-measure, as defined in the MUC-6 coreference scoring scheme (Vilain et al., 1995). In this scheme, coreferential chains returned by a system and from the gold standard are transformed sets of equivalence classes which are then compared using standard precision and recall measures. Missing links influence the recall score, whilst superfluous links indicate precision errors.

3.3.3. Evaluation methods for Tasks 3 and 4

Evaluation of Tasks 3 and 4 is more difficult because it requires measuring both how successfully the tasks were completed and how correctly the entities involved in the tasks were identified. Due to the fact that in Tasks 3 and 4 the participants were not given the entities to resolve in many cases the entities identified by the automatic systems did not match those in the gold standard. As a result, an *overlap measure* was introduced to indicate how successfully the automatic system can identify the entities (Cristea and Postolache, 2006). The overlap between two entities E_1 and E_2 is defined as:

$$overlap(E_1, E_2) = \frac{length(overlap\,string)}{max(length(E_1), length(E_2)}$$

where:

- $length(overlapstring)$ represents the length in words of the string resulting from the overlap of E_1 and E_2, if this overlap is possible, otherwise 0
- $max(length(E_1), length(E_2))$ represents the longest of the two entities

The value of this overlap metric is always between 0 an 1. For example the overlap between *the government of Zair* and *Zair's government* is 0 whereas the overlap between *the government of Zair* and *the government* is 0.5.

The evaluation metrics for Task 3 were modified versions of precision, recall and f-measure which take into account the overlap between the resolved entities and those in the gold standard. Given that in this task the pronouns to be resolved are not indicated in the input file, non-referential pronouns need to be filtered out. This made necessary the use precision and recall instead of simple success rate as in Task 1.
To calculate precision and recall the following formulae were used:

$$Precision = \frac{Score\,of\,correctly\,resolved\,pronouns}{Number\,of\,pronouns\,attempted\,to\,resolve}$$

$$Recall = \frac{Score\,of\,correctly\,resolved\,pronouns}{Number\,of\,pronouns\,in\,the\,gold\,standard}$$

where the score of correctly resolved pronouns is calculated as:

$$Score = sum(overlap(E_{automatic}, E_{gold}))$$

where:

- $E_{automatic}$ is the entity determined by the resolver
- E_{gold} is an entity from the gold standard that maximises the overlap score

As in the Task 1, if a pronoun is resolved to another pronoun the score is 1 if there is there is at least one antecedent in the co-reference chain which is non-pronominal, 0.5 if there is no non-pronominal element in the chain or one of the pronouns in the chain is not correctly resolved, and 0 if it is not correctly resolved.

Task 4 uses a modified version the MUC-6 scores where the overlap between entities identified automatically and those in the gold standard is taken into consideration when the score is calculated.

4. Results

Even though the evaluation exercise generated quite a bit of interest in the research community, only three institutions participated in the formal evaluation: University of Karlsruhe, Germany; Alexandru Ioan Cuza University, Iasi, Romania, and University of Wolverhampton, UK. In order to minimise the amount of tuning which can be done on the testing data, the participants had only 48 hours to submit their results from the moment they have downloaded the data. All the communication was done via a specially designed web page. Each participant received detailed evaluation results, but the figures reported in this paper have been anonymised as promised to the participants. The main reason for this was that in this first edition we were less interested to find out why a method performed better. Instead we wanted to offer the participants an environment where they can test their systems, and gain experience to set up the first fully-fledged edition.
Tables 1 and 2 present the evaluation results. As can be seen in the tables, some of the tasks received more interest than others. For example there was only one submission for Task 2, but three submissions for Task 4. As a result of this we can conclude that the participants found Task 4 more important and realistic.

Method	Success rate
Baseline	33.16%
S1	43.07%
S2	71.55%

Table 1: The results of task 1

Method	Precision	Recall	F-measure
Task 2			
S1	53.01%	45.72%	48.32%
Task 3			
Baseline	23.44%	19.09%	19.52%
S1	25.41%	8.23%	11.65%
S2	65.45%	65.45 %	65.45%
Task 4			
Baseline	45.89%	60.18%	50.28%
S1	32.87%	13.09%	17.68%
S2	49.08%	27.68%	33.93%
S3	62.34%	56.07%	58.02%

Table 2: The results of the tasks 2, 3 and 4

5. Conclusions and future plans

This paper has presented the first edition of the Anaphora Resolution Exercise. Four tasks were proposed to participants: pronominal anaphora resolution on a predefined set of entities, NP coreference on a predefined set of entities, pronominal anaphora resolution on raw texts and NP coreference on raw texts. The task which received most of the attention is the NP coreference on raw text where three different runs were submitted.

As already mentioned this first edition was exploratory in order to identify issues raised by such an exercise. For the future, two clear directions have emerged: tackle other types of anaphoric expressions such as indirect anaphora and evaluate systems which process language other than English. However, the feasibility of both of these tasks depends very much on the availability of annotated data. The evaluation methods need also more attention. Many researchers consider the MUC evaluation scheme too generous. In the future we plan to employ more evaluation measures. The overlap measure seems to lead to some unexpected results and therefore will need to be improved.

6. Acknowledgements

We would like to thank Laura Hasler for the help with the data preparation and Iustina Ilisei for implementing the evaluation measures. We would also like to thank all the participants and the DAARC2007 organisers without whom ARE would not have been possible.

7. References

Cătălina Barbu and Ruslan Mitkov. 2001. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 34 – 41, Toulouse, France, July 9 – 11.

Dan Cristea and Oana Postolache. 2006. Anaphora resolution: framework, creation of resources, and evaluation. In Svetla Koeva and Mila Dimitrova-Vulchanova, editors, *Proceedings of the Fifth International Conference on Formal Approaches to South Slavic and Balkan Languages*, pages 1 – 5, Sofia, Bulgaria, October 18 – 20.

Laura Hasler, Constantin Orăsan, and Karin Naumann. 2006. NPs for Events: Experiments in Coreference Annotation. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006)*, pages 1167 – 1172, Genoa, Italy, 24 – 26 May.

Inderjeet Mani, Therese Firmin, David House, Michael Chrzanowski, Gary Klein, Lynette Hirshman, Beth Sundheim, and Leo Obrst. 1998. The TIPSTER SUMMAC text summarisation evaluation: Final report. Technical Report MTR 98W0000138, The MITRE Corporation.

Ruslan Mitkov. 2000. Towards more comprehensive evaluation in anaphora resolution. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume III, pages 1309 – 1314, Athens, Greece.

Ruslan Mitkov. 2001. Outstanding issues in anaphora resolution. In Al. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 110–125. Springer.

Ruslan Mitkov. 2002. *Anaphora resolution*. Longman.

Tony G. Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 827 – 833, Las Palmas de Gran Canaria, May.

TIDES. 2000. Translingual Information Detection, Extraction and Summarization (TIDES). http://www.darpa.mil/iao/TIDES.htm.

M. Vilain, J. Burger, J. Aberdeen, D. Connoly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45 – 52, San Francisco, California, USA.