Procesamiento del Lenguaje Natural, Revista nº 42, marzo de 2009, pp. 97-104 recibido 15-01-09, aceptado 02-03-09

AQA: a multilingual Anaphora annotation scheme for Question Answering*

AQA: Un modelo de anotación anafórico multilingüe para Búsqueda de Respuestas

E. Boldrini[1], M. Puchol-Blasco[1], B. Navarro[1], P. Martínez-Barco[1] and C. Vargas-Sierra[2]
[1] Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
[2] Departamento de Filología Inglesa
Universidad de Alicante
Alicante, Spain
{eboldrini, marcel, borja, patricio}@dlsi.ua.es and chelo.vargas@ua.es

Resumen: En este trabajo presentamos AQA, un modelo multilingüe de anotación de expresiones anafóricas, ideado para ser utilizado en Aprendizaje Automático para mejorar los sistemas de Búsqueda de Repuestas. Con este modelo se ha anotado la colección de preguntas-respuestas del CLEF 2008, concretamente en los idiomas español, italiano e inglés. AQA está inspirado en el meta-modelo MATE, ajustado a nuestras necesidades. Con AQA se especifica la relación entre la anáfora y su antecedente (que puede ser directa o indirecta), las agrupaciones por tópico y cambios de subtópico, así como diferentes tipos de anáforas (pronominal, adverbial, superficial, descripciones definidas y elipsis). Se ha realizado una anotación ciega entre dos anotadores más un árbitro que decide en caso de desacuerdo. Los resultados de la evaluación muestran un 87% de acuerdo entre los anotadores. Algunos problemas de anotación serán expuestos en el trabajo. Nuestra finalidad es ampliar este modelo a otras lenguas y otros corpus, y aplicarlo finalmente en el desarrollo de un sistema de resolución de la anáfora en preguntas-respuestas multilingüe basado en técnicas de aprendizaje automático para mejorar la interacción hombre-máquina.
Palabras clave: resolución anáfora, corpus multilingüe, Aprendizaje Automático, acuerdo de anotación, interacción, sistemas de Búsqueda de Respuestas.

Abstract: This paper presents AQA, a multilingual anaphora annotation scheme that can be applied in Machine Learning for the improvement of Question Answering systems. It has been used to annotate the collection of CLEF 2008 in Spanish, Italian and English. AQA is inspired by the MATE meta-model, which has been adjusted to our needs. By using AQA we specify the relationshiop between the anaphora and its antecedent, cases of topic and subtopic, and we label different types of anaphoric expressions. A blind annotation was carried out by two annotators, and a referee for solving cases of disagreement. The results of the evaluation show an 87% level of inter-annotator agreement. Some annotation problems will be reported in this paper. Our aim is to extend this model to other languages, and to apply it to the development of an Anaphora Resolution system based on Machine Learning techinques in order to improve a real human machine-interaction.
Keywords: anaphora resolution, multilingual corpora, Machine Learning, inter-annotator agreement, interaction, Question Answering systems.

ISSN 1135-5948

1 Introduction

This paper presents AQA, a multilingual anaphora annotation scheme that can be applied to every question-answer corpus (QA) with cases of anaphora.

In recent years there has been a growing interest in the creation of anaphora annotation schemes, especially for English. In this context, it is worth mentioning the UCREL anaphora annotation scheme (Fligelstone, 1992), developed at Lancaster University. The SGML-based MUC annotation scheme (Hirschman y Chincho, 1998), created for MUC-7, focused on anaphora for Information Extraction task, and other annotation schemes based on MUC are by Mitkov et al. (2000) or by Navarro (2007), among others. Proposals for other languages could also be found. To mention but a few examples, we find proposals for French (Popescu-Belis and Robba (1997); Tutin et al. (2000)); for Spanish and Catalan (Recasens et al. (2007)); or for Basque (Aduriz et al. (2007)).

As it is well-known, the MATE/GNOME meta-scheme by Massimo Poesio (2004) can be adjusted to meet different needs and goals. AQA annotation scheme is inspired by this meta-model.

The problem of anaphora resolution in dialogues and/or in QA series has been explored in several works (Martínez-Barco y Palomar, 2001; Jain et al., 2004; Negri y Kouylekov, 2007). However, as far as we know, little work has been reported on anaphora resolution in QA series in a multilingual framework[1].

In this paper, we focus on this subject. We have developed a multilingual anaphora annotation scheme in order to label the QA corpus of CLEF 2008 in Spanish, Italian, and English, aiming at using this annotated corpus for the application of Machine Learning (ML) techniques in the development of anaphora resolution systems. Our final goal is to achieve an anaphora resolution system for collection of multilingual questions and answers capable of providing a more realistic interaction between the user and the system.

The remainder of this paper is organized as follows: Section 2 describes the principles we adopted for the annotation. Sections 3 and 4 present the main aspects of the annotation scheme, the tag set developed and an analysis of problematic cases. Sections 5 and 6 illustrate the evaluation and the results, and finally conclusions from the study are discussed in section 6.

2 Principles

The design of an annotation scheme involves a number of decisions that are crucial for the final result of its performance. The approach pursued with AQA is based on the next general principles:

1. AQA scheme is specific for QA texts. The behaviour of anaphoric and coreferencial expressions in question-answering and, in general, in dialogues, is different from narrative texts. In fact, the dialogue structure (QA structure) has significant influence on anaphoric relations, and, especially, where the antecedent is located. In this sense, the antecedent of a specific anaphoric expression in a question could be located at the same question, at previous questions or at previous anwers (Negri y Kouylekov, 2007).

2. AQA scheme has been created *ad hoc* for multilingual applications. Indeed, our objective is to develop the same annotation scheme for different languages to have the possibility to employ it in multilingual QA systems. At present, the working languages in the project are English, Spanish and Italian.

3. With AQA annotation scheme we focus on the highest computational efficiency. Our final aim is to develop an anaphora resolution system for multilingual QA based on ML techniques. Consequently, the design of the specific scheme for ML has always been taken into account.

4. With AQA annotation scheme we are looking for a broad applicability. In this sense, we do not follow any specific linguistic theory about anaphoric relations. Instead, we assume a standard point of

* This paper has been supported by the following projects: "Question Answering Learning technologies in a multiLingual and Multimodal Environment QALL-ME" (FP6 IST-033860), "Intelligent, Interactive and Multilingual Text Mining based on Human Language Technologies, TEXT-MESS" (TIN2006-15265-C06-01), by the Generalitat Valenciana throught the research grant BFPI06/182, and by the grant BII2008-7898717 of the University of Alicante.

[1] About multilingual question-answering, see CLEF campaign at http://clef-campaign.org/

view about the anaphoric phenomenon (Mitkov, 2002).

The first step of our work consisted in deciding what had to be annotated, and in creating the resulting markup scheme. In the next section the main aspects of the markup scheme are presented.

3 Markup scheme and tags

The anaphoric elements that are manually specified are the following:

- the anaphora type: we label pronominal, superficial, and adverbial anaphora, as well as some cases of ellipsis (elliptical subject, elliptical object, and nominal phrases with nominal complement but with elliptical head) and definite descriptions.
- the relation type between anaphoric expression and its direct or bridging antecedent. Thanks to the link between the anaphora and its antecedent we are able to detect all the coreference chains throughout the corpus.
- the topic change in a set of questions. We decided to detect the beginning and the end of each topic and subtopic. Questions grouped together share the same topic. However, we also observed some cases of subtopic in the same group.

The tags created to build up our model are the following:

- `<t></t>` *(topic)*: the function of this tag is to group questions about the same topic.
- `<subt></subt>` *(subtopic)*: this tag is used to mark the cases of topic change in the same group of questions.
- `<q></q>` *(question)*: this tag indicates the question/answer pair. It has the ID attribute, which identifies the pair.
- `<de></de>` *(discourse entity)*: discourse entities (antecedents) are detected by assigning to the `ant="ref"` attribute of each anaphora the same ID attribute of its antecedent.
- `<link></link>` *(anaphora)*: the anaphora element includes all the information about the anaphora. The available attributes for this tag are the following:
  - `rel="dir|indir"` *(direct or bridging)*: this element indicates the relationship between the anaphora and its antecedent: direct *(dir)* or bridging *(indir)*.
  - `status="ok|no"` *(sure or uncertain)*: by inserting this attribute the annotator marks his/her (un)certainty with respect to a given annotation.
  - `type="pron|sup|adv|elips|dd"`: this attribute specifies the type of anaphora, i.e., pronominal, adverbial, superficial. It is also used for ellipsis or definite description.
  - `ref="n1"`: for indicating the number of the discourse entity *(de)* the anaphora is referring to.
  - `ant="q|a"` *question or answer*: this tag specifies if the antecedent is in the question or in the answer. If the answer does not appear in the corpus, but the antecedent is within the answer, the `ant="ref"` tag will not appear. The antecedent is marked only with the tag `ant="a"`.
  - `refq="q1"`: the question-answer pair in which the anaphora antecedent is situated. It will correspond to a specific `q id` labelled in the corpus.

Figure 1 shows a group of questions annotated using AQA. Some of these tags and a case of subtopic change can be observed.

4 Some problematic cases

4.1 Antecedent detection

Anaphora annotation is a difficult task with a poor level of inter-annotator agreement (Mitkov, 2002). One of the main complex aspects is the ambiguity for the antecedent detection. In fact, there are cases in which more than one discourse entity could be the antecedent of an anaphoric expression.

In the CLEF 2008 QA corpus there are many cases in which the antecedent can be labelled in the question, but also in the answer. In these cases, the annotators always mark the antecedent closest to the anaphoric

```
<t>
  <q id="q538">
    What was the name of the plane used by
    <de id="n52">John Paul II</de> in
    <link rel="indir" status="ok" ant="q"
    refq="q538" type="dd" ref="n52"> his
    travel</link> to the USA in 1995?
  </q>
  <subt>
    <q id="q539">
      What instrument did Niccol Paganini
      play?
    </q>
  </subt>
</t>
```

Figure 1: Sample of the QA corpus CLEF 2008 annotated with AQA scheme.

expression. However, if the corpus does not contain the answer (as in CLEF 2008 QA corpus), questions are given priority, as we work only with a collection of queries. When the annotators cannot find the antecedent of the anaphora under analysis in one of the questions of the collection, they will be forced to label the antecedent in the answer, although it does not appear explicitly in the corpus.

4.2 World knowledge

In order to label the anaphora and its antecedent properly, the annotators must activate sometimes their world knowledge. The problem may arise when it is not possible to know if annotators have the necessary world and cultural knowledge to detect the correct antecedent.

For example, in this case,

```
<t>
  <q id="q404">
    Which was <de id="n2">the "gordo" in the
    1995 Christmas</de>?
  </q>
  <q id="q405">
    Which was <link rel="indir" status="no"
    type="dd" ref="n2" ant="q" refq="q404">
    the prize</link>?
  </q>
</t>
```

"the prize" is the definite description of "gordo", but if the annotators do not know that in Spain the "gordo" is a typical Christmas lottery prize (and not Santa Claus or a "fat" men[2]), they will not be able to detect the correct antecedent for this anaphora.

[2]The literal translation of "gordo" in English is "fat".

It is not an easy task to deal with these cases of ambiguity arising from a lack of pragmatic or cultural knowledge. As a consequence, they are the main cause of mistakes during the annotation.

4.3 Collective nouns

We also detect some cases of collective nouns, which are singular nouns referring to a plural concept. The problem here is that the anaphora does not always match up in number with its antecedent, and this situation could produce cases of ambiguity. Annotators must apply semantic criteria and common sense in order to detect the correct antecedent.

In this example:

```
<t>
  <q id="q432">
    What is <de id="n18">the starring cast
    </de> of the film Beetlejuice?
  </q>
  <q id="q433">
    Who of <link rel="dir" status="ok"
    type="pron" ref="n18" ant="q" refq="q432">
    them</link> is the main character?
  </q>
</t>
```

As the previous example shows, the pronominal anaphora "them" is referring to the "starring cast": "them" is plural and "the starring cast" is singular. The relation between them is correct, since the starring cast is a collective noun that refers to the group of actors who are performing in a movie.

4.4 Doubtful position of the antecedent

We also detected cases in which the antecedent recognition could be ambiguous, because the annotator has to choose between multiple options.

Let us see an example:

```
<t>
  <q id="q465">
    What transport was used in <de id="n36">the Kon-Tiki
    Expedition</de>?
  </q>
  <q id="q466">
    How many people crewed <link rel="dir"
    status="ok" type="pron" ref="n36" ant="q"
    refq="q465">it</link>?
  </q>
</t>
```

The annotator does not know whether the antecedent of "it" is the "transport" or "the

Kon-Tiki Expedition". In fact this pronoun does not provide any information regarding its genre.

As we have just mentioned, the general rule is to select the closest antecedent to the anaphora, which in this case is "the Kon-Tiki Expedition".

4.5 Nested antecedent

The problems mentioned in this subsection and in the next one do not represent special cases of difficulty, but they could produce ambiguity when specifying the correct size of the antecedent.

There are cases in which we have an antecedent inside another one, and they are referring to two different anaphors. The next example shows this specific case:

```
<t>
  <q id="q427">
    Who were <de id="n14">the founders of <de
    id="n15">Magnum Photos</de> </de>?
  </q>
  <q id="q428">
    In what year did <link rel="dir"
    status="ok" ant="q" refq="q427"
    type="pron" ref="n14">they</link> found
    <link rel="dir" status="ok" type="pron"
    ref="n15" ant="q" refq="q427">it</link>?
  </q>
</t>
```

The antecedent of "them" is "the founders of Magnum Photos", while the antecedent for "it" is only "Magnum Photos".

4.6 An anaphora inside an antecedent of another one

There are cases in which the anaphoric element has to be annotated inside the antecedent of an anaphora that has another antecedent. For example:

```
<t>
  <q id="q434">
    What is <de id="n19">a censer</de>?
  </q>
  <q id="q435">
    What name is given to <de id="n20"> <link
    rel="dir" status="no" type="pron"
    ref="n19" ant="q" refq="q434">the one
    </link> of the Cathedral of Santiago de
    Compostela </de>?
  </q>
  <q id="q436">
    How much does <link rel="dir" status="ok"
    type="pron" ref="n20" ant="q" refq="q435">
    it</link> weight?
  </q>
</t>
```

Finally, we would like to mention a specific problem in the Italian and Spanish corpus: the clitic pronouns. They appear attached to the verb. When clitic pronouns are detected, we do not separate the verb from the pronoun.

5 Evaluation

In order to know the quality of this annotation scheme, we have developed a pilot evaluation, manually annotating the CLEF multilingual QA corpus. There are 600 questions in the corpus, each one translated into English (200), Italian (200) and Spanish (200). At the current state of the project, these results are preliminary. In the near future, our aim is to annotate a larger corpus.

A blind annotation was carried out by two annotators. After this process, we evaluated the inter-annotator agreement independently for each aspect of anaphoric annotation and language. Finally we calculated the general agreement. The evaluation aspects we took into consideration are the following:

1. topic boundary;
2. anaphora detection;
3. anaphora attributes; and
4. antecedent recognition.

5.1 Measures used

The measures used to calculate the inter-annotator agreement are the kappa value (when static classes are present), and the observed agreement (when non static classes are present). Kappa is computed according to Cohen method (Cohen, 1960; Carletta, 1996; Artstein y Poesio, 2008):

$$k = \frac{P(A)-P(E)}{1-P(E)}$$

where $P(A)$ is the observed agreement among annotators, and $P(E)$ the probability that annotators agree by chance.

5.2 Topic boundary evaluation

Topic boundary can be seen as a binary classification. For each question the class "n" is assigned to mark a new topic, while the class "s" is employed when the question is about the same topic as the previous query. Taking into account these premises, Table 1 shows the contingency table and the kappa measure.

	Spanish		Italian		English	
A1/A2	**S**	**N**	**S**	**N**	**S**	**N**
S	62	0	62	0	61	0
N	0	138	0	138	1	138
Kappa	1		1		0.988	

Table 1: Contingency table for topic boundary evaluation.

5.3 Anaphora detection

Anaphora detection has not specific classes for using kappa measure. As a consequence, only the observed agreement among the annotators can be extracted. The anaphora detection agreement is presented in Table 2. The acronyms used in this table mean: A1: anaphors detected by annotator 1; A2: anaphors detected by annotator 2; AA: anaphors detection agreement; DAB: different anaphora boundary, that is, anaphors that coincide in the two corpora, but having different content.

	Spanish	Italian	English
A1	70	69	67
A2	70	69	68
AA	70	69	67
DAB	1	1	0

Table 2: Anaphora detection agreement.

5.4 Anaphora attributes

Once the anaphora has been detected, the method used for anaphora attribute evaluation is the kappa statistic. The results of the anaphora detection agreement are: 70 anaphors in Spanish, 69 in Italian, and 67 in English.

Regarding the antecedent attribute, Q is used when the antecedent is detected in the question, while A is used when the antecedent is in the answer. Table 3 presents the contingency table for this attribute.

	Spanish		Italian		English	
A1/A2	**Q**	**A**	**Q**	**A**	**Q**	**A**
Q	64	0	62	0	61	0
A	0	6	0	7	0	6
Kappa	1		1		1	

Table 3: Contingency table for antecedent attribute evaluation.

The anaphora type was lebelled taking into consideration 5 attributes: Elipsis (Elips), Pronominal (Pron), Adverbial (Adv), Superficial (Sup) and Definite Description (DD). The results for the type attribute are shown at Table 4.

	Spanish		Italian		English	
	A1	**A2**	**A1**	**A2**	**A1**	**A2**
Elips	33	33	32	32	3	3
Pron	13	15	13	13	42	42
Adv	1	1	2	2	1	1
Sup	1	0	0	0	0	0
DD	22	21	22	22	21	21
P(A)	0.97		1		1	
Kappa	0.955		1		1	

Table 4: Anaphora type agreement.

We also evaluated the agreement obtained regarding the relation attribute. In this case, it is possible to choose between two options; the first one is D (direct relation), while the second is I (indirect relation). Table 5 illustrates the results.

	Spanish		Italian		English	
A1/A2	**D**	**I**	**D**	**I**	**D**	**I**
D	52	0	51	0	52	0
I	4	14	1	17	2	13
Kappa	0.838		0.961		0.909	

Table 5: Contingency table for relation attribute evaluation.

5.5 Antecedent recognition

Antecedent recognition has no fixed classes for using kappa measure, and as a consequence, the observed agreement among the annotators should be extracted. The antecedent recognition agreement is presented in Table 6. The acronyms used in this table mean: TAA: total antecedents into the answer; TAQ: total antecedents into the question; ASQ: anaphors pointing the same questions, it means, refq agreement; and ADB: antecedents with different boundary.

	Spanish	Italian	English
TAA	6	7	6
TAQ	64	62	61
ASQ	64	62	61
ADB	2	3	1

Table 6: Antecedent recognition agreement.

5.6 General agreement

The general agreement is showed in Table 7. In this evaluation, all the aforementioned at-

tributes have been considered: first column shows the amount of anaphors detected, and second column the amount of anaphors with exact agreement. Finally, the average for all languages is calculated as general agreement.

	Total	**Agreement**	**%**
Spanish	70	60	0.857
Italian	69	60	0.869
English	67	59	0.880
Average			**0.868**

Table 7: General agreement.

Surprisingly, all these results show a high level of agreement between two annotators in all aspects evaluated.

With these results we can conclude that the annotation scheme has been well designed, and its application to this multilingual QA corpus has been carried out correctly. However, as we said before, these results are only preliminary. Probably, the ambiguity level of this corpus is not too high, thus we will apply the same annotation scheme to a larger corpus, with more languages, more anaphoric expressions, and more cases of ambiguity.

In this case, the results are promising, and they indicate that the project is progressing successfully.

6 Conclusion and Future Work

In this paper we have presented AQA, an anaphora annotation scheme for the manual annotation of multilingual QA corpora. With this scheme we mark different types of anaphors, the relationship between anaphora and its antecedent, and the groups of questions with the same topic.

The main purpose of this scheme is to develop an anaphora resolution system based in ML techniques in order to improve the interaction between the user and the QA system and, in this way, establishing a dialogue between them. In fact, by using AQA, a ML system will be able to extract many features capable of detecting the correct antecedent for each anaphora.

As we can conclude from the evaluation results, we reached a considerable inter-annotator agreement rate. However, our intention is to apply the scheme to other collections of questions and other languages to check AQA reliability.

As we mentioned in the previous section, we carried out the research with three languages involved. This multilingualism offers some advantages, but it is also a source of complexity. The main advantage is that the corpus shows cases in which the anaphoric relation is the same in different languages, so we can extract cross-linguistic features for anaphora resolution. However, using different languages may cause problems. In fact, languages are very complex and different from each other. Working with a parallel corpus does not provide any guarantee of similarity between them: there are cases in which the same query is different in the three languages, and the annotator should take into account these differences in order to annotate the corpus properly.

In any case, as Future Work, we will apply the AQA annotation scheme to a larger corpus with more texts written in more languages in order to check its reliability, and, finally, to improve a multilingual anaphora resolution system for QA.

Bibliografía

Aduriz, I., K. Ceberio, y A. Díaz de Ilarraza. 2007. Pronominal Anaphora in Basque: Annotation issues for later computational treatment. En A. Branco, editor, *Anaphora: Analysis, Algorithms and Applications. 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007*, volumen 4410 de *Selected Papers. Lecture Notes in Computer Science*, Lagos Portugal.

Artstein, R. y M. Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computationl Linguistics*, 34(4):555–596.

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

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Fligelstone, S. 1992. Developing a Scheme for Annotating Text to Show Anaphoric relations. En G. Leitner, editor, *New Direction in English Language Corpora: Methodology, Results, Software Developments*. Mouton de Gruyter, Berlin, páginas 153–170.

Hirschman, L. y N. Chincho. 1998. Muc-7 coreference task definition (version 3.0). En *Proceedings of Message Understanding Conference (MUC-7)*.

Jain, P., M. Mital, S. Kumar, A. Mukerjee, y A. Raina. 2004. Anaphora resolution in multi-person dialogues. En Michael Strube y Candy Sidner, editores, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, páginas 47–50, Cambridge, Massachusetts, USA.

Martínez-Barco, P. y M. Palomar. 2001. Computational Approach to Anaphora Resolution in Spanish Dialogues. *Journal of Artificial Intelligence Research*, 15:263–287.

Mitkov, R. 2002. *Anaphora Resolution*. Longman.

Mitkov, R., R. Evans, C. Orasan, C. Barbu, L. Jones, y V. Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. En *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, Lancaster.

Navarro, B. 2007. *Metodología, construcción y explotación de corpus anotados semántica y anafóricamente*. Ph.D. tesis, University of Alicante, Alicante.

Negri, M. y M. Kouylekov. 2007. 'Who Are We Talking About?' Tracking the Referent in a Question Answering Series. En A. Branco, editor, *Anaphora: Analysis, Algorithms and Applications. 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007*, volumen 4410 de *Selected Papers. Lecture Notes in Computer Science*, Lagos Portugal.

Poesio, M. 2004. Discourse annotation and semantic annotation in the gnome corpus. En *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, páginas 72–79, Barcelona.

Popescu-Belis, A. y I. Robba. 1997. Cooperation between pronoun and reference resolution for unrestricted texts. En *Proceedings of the ACL'97/EACL'97 workshop on Operational Factor in Practical, Robust Anaphora Resolution*, Madrid.

Recasens, M., M.A. Martí, y M. Taulé. 2007. Text as a Scene: Discourse deixis and Bridging relations. *Procesamiento del Lenguaje Natural*, 39:205–212.

Tutin, A., F. Trouilleux, C. Clouzot, E. Gaussier, A. Zaenen, S. Rayot, y G. Antoniadis. 2000. Anotating a large corpus with anaphoric links. En *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, Lancaster.