CLinkA a Coreferential Links Annotator

Constantin Orăsan

School of Humanities, Languages and Social Sciences Stafford Street University of Wolverhampton Wolverhampton, WV1 1SB United Kingdom in6093@wlv.ac.uk

Abstract

The annotation of coreferential chains in a text is a difficult task, which requires a lot of concentration. Given its complexity, without an appropriate tool it is very difficult to produce high quality coreferentially annotated corpora. In this paper we discus the requirements for developing a tool for helping the human annotator in this task. The annotation scheme used by our program is derived from the one proposed by MUC-7 Coreference Task Annotation, but is not restricted only to that one. Using a very simple language the user is able to define his/her own annotation scheme. The tool has a user-friendly interface and is language and platform independent.

1. Introduction

The term corpus has been used to designate a body of naturally-occurring (authentic) language data which can be used as a basis for linguistic research (Leech, 1997). There are many ways to design a corpus. It can consist of written texts and/or spoken texts from general type of texts, or it can represent only a particular language or language variety. More recently the term corpus was applied more and more for body of language that exists in electronic format. The main reason for this is the explosion of information available online, making it easier to build a corpus for a particular purpose by downloading relevant texts from the web.

A corpus can be used as it is for linguistic research, but it becomes more valuable by annotating it. Depending on the purpose of the corpus, different aspects of the text can be marked. We can talk about a very simple annotation in which only the paragraph and sentence boundaries are marked. In the present day this could be considered an obligatory marking for every corpus. The next level of annotation can be considered the grammatical tagging, where to each word a tag, indicating its grammatical category, is associated. These kinds of corpora are widespread and available in many languages. A more complicated level of annotation is the syntactic mark-up where full or partial parsing trees are marked for each proposition. This is the second most widespread type of the corpus annotation and is in a rapid development. The meaning of the words in a text can be marked using semantic tagging. Given the difficulty of the task only few corpora that mark the meaning of the words are available. The development of corpora that mark discoursal entities in the texts has just started. Other aspects of the text can be marked (e.g. prosodic, pragmatic and stylistic annotation), but for various reasons they might be directions for future development, rather than current trends in corpus annotation.

Once a corpus is annotated it can be used for different purposes. One can extract different information relevant for his/her research. The advantage of this approach, emphasised by the research in corpus linguistics, consists in using real language instead of artificial constructions. Given the high costs involved in annotating corpus, it can be reused by other research for the same type of research or a different one.

The annotation of a corpus can be done by humans. After a training period, during which he/she is instructed which entities in the text have to be marked and in which way, he/she starts to annotate the text. This approach is very expensive and in some degree interpretative given that it is the product of the human mind. Another method is to use a program for marking the entities in the text, and then to correct the annotation. This approach works in the cases for part-of-speech tagging and in some degree for parsing, but is not appropriate for more advanced levels of annotation (e.g. semantic and discoursal annotation) because either there are no programs for doing this, or their results are very poor. Therefore, the only solution that remains is to find a way to help the human annotator in his task. In this paper we present a tool designed for annotating a text at discoursal level, more precisely, for marking the coreferential chains within the text. The remaining part of the paper is structured as follows. In the next section we discuss about the purpose of coreferential annotation and different annotation schemes proposed. In section three the requirements for an annotation tool are presented. Section four presents the annotation tool that we developed. In the final section we present conclusions and further ways of developing the tool.

2. Coreferential annotation

Two entities are coreferential if they refer to the same entity in the discourse (Haliday, 1976). Their purpose is to give cohesion and coherence to a text. A coreferential chain links together multiple expressions designating a given entity. Usually these expressions are nouns, noun phrases or pronouns, but they can also be verbs, clauses or parts of the discourse. In the MUC-7 Coreference Task Definition (Hirchman, 1997), a text annotated with coreferential links is seen as "a kind of hyperlinked version of a text, where the links connect the mentions of a given entity".

The existence of coreferentially-annotated corpora is very important for NLP tasks. They were used for training machine learning algorithms (Aone, 1995), (Connolly, 1997) and statistical algorithms (Ge, 1998) for coreference resolution. The availability of a coreferentially annotated corpus proved useful for automatic evaluation of an anaphora resolution system (Mitkov, 1999) and for optimising its results (Orăsan, 2000).

Not only the NLP community can take advantage of the coreferentially annotated corpora. Other possible users are the linguists that can use them for investigating the coreference phenomenon from different view points on real texts. The learners of a foreign language can benefit from these corpora for understanding the way in which the coreference is used in that language.

Some researchers suggested that discourse annotation is a particularly ill-defined concept because different researchers suggested different aspects to be marked in a text. Even when the same aspects of the discourse were marked, different annotation schemes and granularity were used. This is true also for annotating coreferential relations. In some cases it was chosen to annotate only anaphoric relations instead of full coreference chains. The elements that are marked can be different. One can decide to mark not only the anaphoric relations between the pronouns and NPs, but also between pronouns and clauses and discourse segments. However, there are only few texts in which the coreferential links are annotated, so it is not possible to say which is the most appropriate annotation scheme. Each annotation scheme has been developed with a special purpose in mind.

One of the first projects for building an anaphoric treebank was the one in Lancaster between 1989-91. The idea was to add to a treebank, markers for showing the anaphoric relations (Garside, 1997a). The scheme does not attempt to mark all the possible theoretical justifiable distinctions and tends to be theoretically fairly neutral. It was designed for marking in particular the relations between a proform and an antecedent and marks both the anaphoric relations and cataphoric ones. Other relations that can be marked using this scheme are: NP coreference, indirect definite anaphora, textually recoverable ellipsis, etc. An antecedent is marked in round brackets and given a unique number. Two items, which are numbered and bracketed in the same way, are considered coreferential. The anaphoric relation between a proform and its antecedent is marked using *<REF=id_referent*. In the case of cataphoric relations > symbol is used. The schema can be used even in the cases when more than an antecedent is referred like in the following example:

(1 A man 1) and (2 a woman 2) appeared on the stage. <REF=1 He was dressed with ... <REF=1,2 They were

In this case *they* refers to both *a man* and *a woman*. In the cases when the human annotator cannot decide if an entity is the correct antecedent or not, he can use a question mark before the ID of that entity to signal the ambiguity.

Another annotation scheme is the one developed by Marco de Rocha (de Rocha, 1996) for spoken texts in English and Portuguese. It reflects the widely accepted view in discourse analysis and text linguistics that the topic of discourse tends to be the preferred antecedent for a given anaphoric expression. The text is divided in segments according to the topic continuity. Whenever the topic is changed, a new segment starts. The next step is the annotation of the anaphoric expressions, for each specifying its type, the type of its antecedent, the position of its antecedent and the type of knowledge used for processing it. This annotation scheme has the advantages that attempts to encode information about the relations between anaphora and the topic of discourse and is used with spoken dialogue for more that one language. The big disadvantage of this annotation scheme is that it does not use a widely-accepted text encoding format and the annotated text is quite difficult to be understood.

Most of the researchers involved in building anaphoric annotated corpora use the scheme proposed in MUC-7 Coreference Task Definition (Hirchman, 1997) or a scheme derived from it. In this annotation scheme, it is considered that it is more important to obtain a high degree of inter-annotator agreement than to cover every possible type of coreferential relation. It marks only the "IDENTITY" relations between two noun phrases. It does not include coreference among clauses, discourse segments, or other kinds of relations like set/subset, part/whole, etc. The main reason for this is the fact that the scheme was designed for evaluating the output of automatic systems for coreference resolution and other types of entities than noun phrases, or relations than identity, are too difficult to be tackled by computers.

The scheme uses SGML (Standard Generalised Markup Language) tags to annotate anaphoric expressions in the texts. Given the great extent of SGML usage, the corpora annotated in this way can be easily used for other research. The annotation scheme makes a distinction between the initial mention of an entity in the discourse and the rest of the elements from the coreferential chain. The initial mention is marked with <COREF ID="###">, where ID is a unique number which identifies the entity. It has to be emphasised that it is not like in the scheme proposed by Garside where all the noun phrases from a chain were using the same ID. In this case each entity has its own ID. For some schemes, derived from this one, use for identifying the entities a combination of letters (usually two) followed by a number.

An element from the chain is marked using <COREF ID="#" TYPE="IDENT" REF="#"> tag. The type always has the *IDENT* value, for indicating that the relation between the entities is of identity. This attribute was included for future development of the scheme when other coreferential relations might be included. The REF attribute indicates which entity is coreferential with the current one. For our system we decided that REF value is the ID of the first element from the chain.

A COREF tag also has other attributes introduced for automatic evaluation of the automatic systems and which are optional. The "MIN" attribute is used in the answer key to indicate the minimum string that the system under evaluation must include in the COREF tag in order to be considered correct. The "STATUS" attribute is used for marking optional answers. The only value of this attribute is "OPT" (optional). An evaluation system will not score a string which is marked unless it was correctly resolved.

Given that the main reason of our tool is developing corpora in which the anaphoric relations are marked for training and evaluating results of the anaphora resolution systems, we decided to use the annotation scheme proposed by (Hirchman, 1997).

3. Requirements of a tool for coreference annotation

An annotation software is a software tool which intends to aid or undertake the task of corpus annotation. Two general methods were proposed for text annotation. The first one uses a more or less specialised editor for marking. For simple annotation tasks one can use a general-purpose editor like Notepad or Word for Windows, or Emacs for Unix, but when annotating big corpora or using complicated annotation schemes, like the coreference annotation scheme, a specialised editor has to be used. The reason for this is because a complicated annotation scheme requires a lot of concentration from the human analyst and it is very easy to introduce errors.

The second one is an automatic annotation system that offers the possibility of hand correcting its output. Given the difficulty of coreference annotation it is not possible to have a good system for this task. It is possible instead to have some heuristics which help the human annotator proposing the entities to be annotated or propose elements to be added to a coreferential chain.

The process of coreferential annotation of a text is a labour-intensive task, requiring a lot of human work. In order to help the human annotator it is necessary to provide him with a tool which makes possible to quickly identify the entities in the discourse and the relations between them. A good graphical interface offers the human annotator trouble-free and efficient interaction with the annotated text.

The tool has to be easy to be used; with a minimum time to learn how it works. It also has to hide the unnecessary details. For example in some cases the coreferential annotation is made on a text which already has some markup. If this markup does not help the coreferential annotation it has to be hidden from the human annotator.

In most of the cases the human annotators are linguists with little knowledge about annotation schemes and computers, or none. Therefore, the editor has to be designed in such a way that the humans provide the necessary information in a very simple way, using the mouse and the keyboard and the task of saving the file in the appropriate format remains to be done by the tool. In some cases the annotation schemes have to follow different recommendations (e.g. TEI). The tool can be designed in such a way that it follows the recommendations and does not allow illegal constructions.

More and more projects compare different languages from different point of view using parallel corpora. When two or more teams work for building the corpus, they have to make sure that the same annotation scheme is used. It is known that for different implementation reasons each tool imposes some constraints. Using the same tool for all languages results in fewer constraints on the resulting corpora. Therefore one of the desiderata of the annotation programs is to be language independent. In some cases the same persons annotate for more than one language, thus it is better if they can use the same tool for all the languages without being needing to learn a different tool for each language.

4. The Coreferential Links Annotator

In this section we present the tool developed by us for annotating the coreferential links in a parallel corpus for English and Spanish. At the time we started to develop this tool, in June 1999, we were aware of only two tools which could have been used. One is XANADU (Garside, 1997b), the tool developed in Lancaster for annotating discourse relations. We found that the annotation scheme used by this program is not appropriate for our purposes and the program runs only under Unix. Given that in our laboratory we have both machines with Windows and Unix we needed a system that can be run on both.

The other alternative was to use Alembic Workbench (Day, 1998) a tool developed by MITRE Corporation for annotating texts with a large variety of SGML tag types. The tool is developed in Tcl/Tk, and therefore, it is possible to run it both on Windows and Unix. Given that the tool is under development, the version which we tried to use, was not able to save the annotated files in the Windows version and it had some problems with embedded tags. Also, because the tool is designed for general annotation, rather than specifically for coreferential annotation we found that the coreferential annotation process was quite slow.

Meanwhile, another tool for annotating the coreference relations was developed by DeCristofaro (1999), but we have not evaluated it.

Having in mind the general requirements for an annotation scheme and the particular aspects of our project we started to develop our own tool. For ensuring language independence for it, we decided to implement it in Java aware of its support of Unicode. Moreover, the tool being developed in Java, it is platform independent, being possible to run it on any system which has a Java virtual machine. The tool was tested on three languages which use Latin characters: English, Spanish and Romanian and one which uses Cyrillic characters: Bulgarian. At present the tool is used for developing corpora for English, Spanish and Bulgarian.

The annotation scheme used by CLinkA is similar to the one proposed by MUC-7 Coreference Annotation Task and is based on SGML. However, the scheme is not restricted to that one. The user can customise the annotation scheme according to his/her need.

The program used two types of tags: for marking the initial mention of a chain and for marking the remaining elements from the chain. To each tag the following types of attributes can be added:

- counters, which identify uniquely each tag and are generated automatically by the program. (e.g.<COREF ID="10">). At this point in the process of developing the program, we consider introducing an additional feature for the program for having a prefix formed by letters
- references for identifying the entity referred by that item. The value of this attribute is the ID of the first element of the chain. (e.g. <COREF ID="11" REF="10">)
- attributes which have the value specified by the human annotator like attribute "MIN". When one marks an entity from the text with a tag containing this attribute, the program stops and asks for the value of this attribute

<COREF ID="0"><COREF ID="1">Macintosh Performa</COREF> User's Guide</COREF>

1 Getting Started

<COREF ID="2">The illustration on <COREF ID="3">the facing page</COREF></COREF> shows <COREF ID="4">all the equipment you will need to set up <COREF ID="5" TYPE="IDENT" REF="1">your computer</COREF> and begin using <COREF ID="6" TYPE="IDENT" REF="1">it</COREF></COREF> . Place <COREF ID="7" TYPE="IDENT" REF="1">your equipment</COREF> on <COREF ID="8">a sturdy , flat surface</COREF> near <COREF ID="9">a grounded wall outlet</COREF> . Before following <COREF ID="10">the setup instructions in <COREF ID="11">this chapter</COREF> . Before following <COREF ID="10">the setup instructions in <COREF ID="11">this chapter</COREF> , you may want to read <COREF ID="12">"Arranging your Office " in <COREF ID="13">Appendix A</COREF> , in <COREF ID="14">the section on <COREF ID="15">health-related information</COREF></COREF> , for <COREF ID="16">tips on adjusting <COREF ID="17">your work furniture</coreF></coreF> , for <COREF ID="16">the setup using <COREF ID="17">TYPE="IDENT" REF="1">the computer</coreF></coreF> , for</coreF> , for

Figure 1. An example of text annotated with CLinkA

 strings which always have the same value like TYPE. This type of attribute is very similar with the previous one and was introduced for speeding up the process of annotation. In this way the human does not need to introduce the value of the attribute when it is known from the beginning.

An additional feature was included in the annotation scheme. In many languages the phenomena of zero anaphora are quite common and for some research area it is important to be able to mark the position of them without adding artificial words to the text. The solution was to add a special tag which marks the position of this zero pronoun.

Figure 1 presents an example of a text annotated using our program. In this case the MUC-7 annotation scheme was used for tagging the corpus.

A user-friendly interface was designed for helping the annotation process (Figure 2). For speeding up the process the screen was split in two parts. In the left side the text which is going to be annotated is displayed and in the right side the list of initial elements of the chains. For marking an element, the annotator has to select first the type of tag he wants to assign to that element. The next step is to mark with the mouse the boundaries of the element and press ENTER. If the element is marked as being part of a coreferential chain, the human annotator has to indicate to which chain it belongs. This can be done in two very simple ways: by selecting an entity from the chain (either the initial mention or any element for the chain) or by selecting an element from the list in the right side of the screen. Different colours are used for different types of marked element. In this way it is very easy to identify them visually without being needing to display the SGML tags.

Each time an element is marked as being the initial mention of a coreferential chain it is added to this list. In this way the annotator does not need to scroll through the text to find the antecedent of an element. Reports from the annotators proved that the existence of this list is very useful. Initially, we planed to implement a "smart" list in which the elements which are used very frequently are always visible on the list, but by asking the opinion of people which were asked to make the annotation we discovered that changing the position of the elements in the list would be confusing. Instead, we introduced a special list in which the annotator can insert the most frequent elements.

In order to help the human annotator some semiautomatic marking can be introduced. The first, and the simplest, way is to include identical strings in the same chain. This is not true in all the cases, but especially when is about string formed from more than a word could speed up the annotation. Every time when an identical string is found the human annotator is asked to decide if the string is to be included or not in the chain. Another way would be to automatically propose which entities are going to be marked. If the text that is going to be tagged contains tags that mark the NPs, they can be used for proposing the entities. More complex methods could be introduced, but they require more linguistic information and major changes into the program. It should be emphasised that at this point the program does not include any module which uses any knowledge about the language of the text. In this way the program remains language independent. One way would be to allow the user to link the program to different NLP tools specific for each language through an API (Application Programming Interface). Given that not all the users have programming knowledge we decided that the language specific information to be specified in the original text using SGML markup. Therefore, if the input text has the noun phrases and pronoun marked the program can propose automatically the entities to be marked. The same with the tags which have the attribute MIN; if the head of the noun phrases is indicated in the input text, then this information can be proposed to the user as the value for the attribute.

Given the inevitable errors introduced by the humans during the tagging procedure and the subjectivity of this task, it is recommended that each text to be annotated by two or more persons. The program that we developed offers the possibility to compare, using the graphical interface, the two annotations in order to compare them. In this graphical environment the marking of each annotator is displayed in another colour in this way being possible to have a quick visual comparison of the two annotations. At this point we consider introducing two measures for computing the similarity between the two annotation schemes. The first one computes the agreement between the two sets which are marked using the following formula:

$$\mu = \frac{2 * C}{A + B}$$

where A is the number of items marked by the first annotator, B is the number of items marked the second on and C is the number of items which were marked by both annotators. If they marked the same items the measure is one, otherwise is less then one.

The second measure is for computing the feature-value agreement between annotators. It considers only the items which were marked by both annotators and indicates how many times the annotators used the same tags and the same value for the attribute for them. The measure uses the kappa statistic (Carletta, 1996):

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

where P(A) is the proportion of times the annotators agree and P(E) is the proportion of times that we would expect the annotators to agree by chance.

5. Conclusions

In this paper we presented annotation software for marking coreferential chains in the texts which, according to the annotators reports, is easier to be used than existing systems. The main strengths of the system are the user friendliness of the interface which helps the human to concentrate to the annotation process without being distracted by details related to the annotation scheme, the language independence which makes it possible to build parallel corpora and platform independence which allows the program to be run on any system which has a Java virtual machine.

6. Acknowledgements

I would like to thank to Violeta Sotirova for testing the program thoroughly.

This program was developed as part of the "Development of an integrated corpus-based anaphora resolution system for Spanish and English" project between University of Wolverhampton, UK and University of Alicante, Spain.

7. References

- Aone, C. and Bennett, S. W. (1995): Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies in *Proceedings of the 33rd Annual Meeting on the ACL*, pp. 122 – 129
- Carletta, J. (1996): Assessing agreement on classification tasks: The kappa statistic, *Computational Linguistics*, 22(2): 249–254
- Connolly D, Burger J. D. and Day D. S., (1997), "A machine learning approach to anaphoric reference", in Daniel Jones and Harold Somers (eds) *New Methods in Language Processing*, UCL Press Limited
- Day. D, Aberdeen J., Caskey S, Hirschman L, Robinson P, Vilain M. (1998): Alembic Workbench Corpus Development Tool, in *Proceedings of the First International Conference on Language Resource & Evaluation*, Granada, Spain, pp. 1021 – 1028

- DeCristofaro J., Strube M. and McCoy K.F. (1999): Building a Tool for Annotating Reference in discourse, in *Proceedings of the Workshop on The Relation of Discourse/Dialogue Structure and Reference*, University of Maryland, College Park, USA, 21 June 1999.
- Garside R., Fligelstone S and Botley S. (1997a): Discourse Annotation: Anaphoric Relations in Corpora in Garside R., Leech G. and McEnery A. (eds) Corpus Annotation: Linguistic Information from Computer Text Corpora, Addison Wesley Longman, pp. 66 – 84
- Garside R, Rayson P. (1997b): Higher-Level Annotation Tools, in in Garside R., Leech G. and McEnery A. (eds) Corpus Annotation: Linguistic Information from Computer Text Corpora, Addison Wesley Longman, pp. 179 – 193
- Ge, N, Hale, J and Charniak E. (1998): A Statistical Approach to Anaphora Resolution, in *Proceedings of* the Sixth Workshop on Very Large Corpora, COLING-ACL '98, pp. 161 – 170
- Halliday, M.A.K and Hasan R. (1976): *Cohesion in English*, Longman Group Ltd.
- Hirchman L. (1997): MUC-7 Coreference Task Definition at http://www.muc.saic.com/proceedings/ co_task.pdf
- Leech, G (1997): Introducing corpus annotation, in Garside R., Leech G. and McEnery A. (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Addison Wesley Longman, pp. 1–19
- Mitkov R., Orasan C. and Evans R., (1999): The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting, in *Proceedings* of Corpora and NLP: Reflecting on Methodology Workshop, TALN'99
- Orasan C., Evans R. and Mitkov R. (2000): Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms in *Proceedings of NLP 2000 Conference*, Patras, Greece.
- De Rocha, M (1996): A corpus-based study of anaphora in English and Protuguese, in Bortley, S. P. and McEnery A. M. (eds.) Corpus-based and computational approaches to discourse anaphora, London: UCL Press

_	oreferentia						
File	Options	Tags	Help				
3							
			a User's Guid	le			ID=1 Macintosh Performa ID=0 Macintosh Performa User's Gu
1 G	etting St	arted					ID=3 the facing page ID=2 The illustration on the facing pa
					pment you will need to		ID=4 all the equipment you will need
				I outlet. Before follow			ID=8 a sturdy, flat surface
					ranging your Office" in		ID=9 a grounded wall outlet ID=11 this chapter
				alth-related informati			ID=10 the setup instructions in this d
	isting your work furniture so that you're comfortable when using th puter.						ID=13 Appendix A
						ID=15 health-related information	
Plug	gging in the computer						ID=14 the section on health-related i
				nto a wall socket and before yo			ID=12 "Arranging your Office " in App ID=17 your work furniture
	thing else to your Macintosh, carefully read all the setup instructions in chapter. Then, begin by plugging in the computer according to the tructions that follow. The plug grounds the computer and protects it from ctrical damage while you are setting up.						ID=16 tips on adjusting your work fu
						ID=21 a wall socket	
					,		ID=22 anything else ID=26 the instructions that follow
v	r Macinh	ash co	mes with here	identical power cords	You can use either		ID=20 the instructions that follow ID=27 The plug
				u may need the other			ID=30 electrical damage
mon	itor.			-			ID=32 two identical power cords
1 M	ake sure your Macintosh is turned off						ID=36 your monitor ID=39 yourself
		,					ID=38 both yourself and the compute
	o protect both yourself and the computer from electrical hazards, the						ID=41 electrical hazards
	cintosh should remain turned off until you finish connecting its parts. as the bottom of the on/off switch, marked with the O symbol, to turn the				ų.	ID=43 its parts	
			FID-12>section				1.1
	G:\projects/uses0.3\data\a4 <c00ehf ehf="EHF ID" id="8" type="IDHF"></c00ehf>						

Figure 2. A screenshot of CLinkA