Coreference resolution in dialogues in English and Portuguese

Marco Rocha
Universidade Federal de Santa Catarina
Departamento de Língua e Literatura Vernáculas
Centro de Comunicação e Expressão
88040-900 Florianópolis - Brazil
marcor@cce.ufsc.br

Abstract

This paper introduces a methodology to analyse and resolve cases of coreference in dialogues in English and Portuguese. A four-attribute annotation to analyse cases of anaphora was used to analyse a sample of around three thousand cases in each language collected in dialogue corpora. The information thus gathered was analysed by means of exploratory and model-building statistical procedures. A probabilistic model was then built on the basis of aggregate combinations of categories across the four attributes. This model, in combination with direct observation of cases, was used to build an antecedent-likelihood theory, which is at present being organised as a decision tree for the purpose of testing with a view for automatic annotation and subsequent resolution of coreference cases in dialogues in both languages. It is thought that the findings could be extended to Spanish, Italian and possibly French.

Introduction

The problem of anaphora resolution has received a great deal of attention in theoretical linguistics, psycholinguistics and also in natural language processing. Perhaps as an inevitable consequence of such a large body of work related to the subject, the term **anaphora** has been used to mean a varying range of phenomena.

Approaches that build on the concept of cohesion ties (Halliday and Hasan 1976) analyse anaphoric relations within a broad framework of discourse or textual cohesion. As a result, the notion of anaphora, which had been initially linked quite closely to the older concept of pronominalisation, has been expanded to include all referring expressions with some form of antecedent either explicitly introduced in the text or inferable from it.

In an earlier study, Webber (1979) had already widened the scope of anaphoric relations, by including nonpronominal noun phrases which refer back to antecedents in the discourse; the so-called one-anaphora; and verb-phrase deletions. Gradually, the distinction between anaphoric and coreference relations became less and less relevant in approaches concerned with robust implementation of systems with a capacity for anaphora resolution. The present study follows the same sort of approach.

Therefore, the term **coreference** in the present study is used to refer to all pronominal forms, anaphoric nonpronominal noun phrases, one anaphora, numerals when used as heads of noun phrases, prepositional phrases used as responses to questions or statements, responses to questions in general (including **yes**, **no** and short answers using auxiliaries), **so** anaphora, **do**-phrase anaphora and whatever other elements in dialogues were thought to be referring expressions with an identifiable antecedent.

The next section describes the annotation scheme used to analyse the coreference cases. The third section presents the antecedent-likelihood (henceforth, AL) theory, which is the way information collected by means of the annotation was organised so as to be used to resolve new cases of coreference in other dialogues. The decision trees which are to be built on the basis of the AL theory are explained in the subsequent section, whereas the final

section concludes with a discussion of results obtained so far and an analysis of future developments.

2 The annotation of coreference cases

The operational routine of data collection was simply to search manually for tokens which coreferred in dialogue samples. Samples consisted of full dialogues for Portuguese but not for English, as a result of the sampling technique used in the English corpus (the London-Lund). Whenever a case of coreference was found, it was classified. according to four attributes, namely: type of anaphor; type of antecedent; topical role of the antecedent; and processing strategy.

The first attribute refers to the word or phrase which triggers the anaphoric link, that is, the visible item that requires the retrieval of another element in the text for semantic interpretation. Concepts such as zero pronouns or empty categories are not used in the classification. Thus, the anaphor is invariably a phonetically realised item, and a verb without a phonetically realised subject is classified as an **anaphoric verb**. Although such verbal forms are rare in English, they are fairly common in spoken Portuguese. The same approach is used for transitive verbs without a phonetically realised object, which are also frequent.

The type of antecedent concerns primarily the implicit/explicit dichotomy. Typically anaphoric words, such as *it* and *that*, may occur in nonreferential uses – for instance, the 'prop' it (Quirk et al. 1985). Thus, a third category, **nonreferential**, was used to classify these cases. Although these are not cases of coreference *strictu sensu*, it was thought important to include them, so that they could be identified when it came to implementation. Some tokens of pronouns with a vague antecedent identifiable by means of inference based on discourse information were classified as **discourse implicit** antecedents.

The attribute named as the **topical role of the antecedent** classifies the antecedent of a given coreference case according to categories which assign a saliency status to discourse entities (typically noun phrases) in a dialogue. These categories include a **discourse topic** for the dialogue; a **segment topic** for every stretch of dialogue in which the topic is considered to be the same, according to specific procedures; a **subsegment topic**, if further division within a segment is needed for the appropriate modelling of topicality; and both global and local **thematic elements**, which are salient discourse entities related to the topics above mentioned. As antecedents may also be discourse chunks of varying length, these same categories were used to classify such antecedents as **predicates of** a given topical role thought to be the dominant entity within the discourse chunk.

The aim of this attribute is to use the often mentioned relationship between topicality and coreference (see Grosz and Sidner 1986) for operational purposes. This classification does not claim to be the actual key for the modelling of topicality in dialogues from a psycholinguistic point of view. It does claim, however, to be a useful tool for the resolution of particularly hard cases of coreference, in which the antecedent is not the nearest syntactically appropriate candidate, as will be shown in section 3. The topical roles are assigned on the basis of frequency, distribution and order of appearance. This information is used in conjunction with an adaptation for dialogues of Hoey's method (Hoey 1991) to establish patterns of lexis. Procedures were thus defined for the assignment of the topical roles above mentioned to the various discourse entities in a dialogue.

The fourth attribute is the **processing strategy**, which is an attempt to classify the resolution path according to informational demands seen as the most essential for the processing at hand. The processing strategy was included in the annotation scheme as a way of enriching the classification model, uncovering distinctions which might remain unnoticed if only the type of anaphor were to be specified. The plain assignment of a type of anaphor based on word classes would ignore distinctions in the processing required for the resolution of anaphors of the same type. On the other hand,

subsuming processing information in the classification used for the type of anaphor would disrupt the intended link of the latter to phonetically realised forms in a strict way.

The annotation is entered between brackets in the order previously presented, beginning with the type of anaphor and ending with the processing strategy. The code for each one of the properties is delimited by semicolons. An example of annotated text is shown below.

(1)
B: well I think probably what Captain Kay *(FNP; ex_222; dthel; LR;)* must have said was a will is legal if it's *(SP; ex_224; dthel; FtC;)* witnessed on the back of an envelope

The first token of coreference is the anaphoric nonpronominal noun phrase *Captain Kay*, which has been previously introduced in the dialogue. The type of anaphor is classified as **FNP**, for **full noun phrase**; the next slot defines the type of antecedent as **explicit (ex_)** and assigns a number for the referent according to order of appearance in the dialogue **(222)**. The topical role of the antecedent is considered to be of a **discourse thematic element**. This means, thus, that *Captain Kay* is a fairly frequent discourse entity not only in a specific stretch of discourse, but throughout the dialogue, being, therefore, closely associated to the discourse topic. As the reference to *Captain Kay* is identified by means of verbatim repetition of the noun form under which it appeared for the first time in the dialogue, the processing strategy is defined as **lexical repetition (LR)**.

The subsequent anaphoric *it* refers to the first syntactically appropriate candidate looking backwards. Having Hobbs' (1986) naive algorithm as a reference, a primary **first-candidate** processing strategy was established under the code **FtC**. An extension of this primary strategy is the **first-candidate chain (FtCCh)**, for cases in which Hobbs' naive algorithm finds another anaphor for antecedent. This sort of chain is crucially important in dialogues, as demonstrated by Biber (1992). An example is given below[1].

(2)
B: and I went down this morning to talk to the American Embassy on the off chance that the State Department might be you know able to finance a bit of travelling in the States and they can't they've *(SP; ex_13; st; FtCCh;)* got priority on vice-chancellors and uh English schoolteachers

The second token of *they* refers to the first one, which, eventually, links both anaphors to the referent *State Department*. The two first-candidate processing strategies, together with resolutions relying on syntactic parallelism, were grouped under the umbrella category named **syntactic processes**.

As the analysis of anaphora cases found in the corpus proceeded, a number of other categories for the classification of the processing strategy came up. These included, for instance, **collocational knowledge (CK)**, for cases in which the basic information required for processing was thought to derive from the use of anaphors within crystallised phrases, such as *that is to say*. Example (3) is one of those cases.

(3)
B: the bibliography has gone about as far as I can take it on my own that *(De; ex_10; p_st; CK;)* is to say er in order to complete it I will have to visit the major resources in the United States and uh several in Europe

[1] Annotation for other cases of coreference is omitted.

By collecting these phrases in association with each type of anaphor, a collocation list of anaphoric terms was built for each one of the types, with a resolution procedure attached, which was designed on the basis of corpus data observation. This list was subsequently used as an ancillary routine in the AL theory, as will be shown later.

Several forms of **lexical knowledge**, assigned to cases in which the antecedents were identified chiefly by means of semantic information contained in the anaphor, were also identified, such as part-whole relationshps. In example (4), *monies* refers to *finances* by means of information conveyed by the lexical semantics contained in the lexical item itself, but not by means of plain repetition. Thus, the classification used is **lexical signalling (LS)**, one of the categories within the umbrella category **lexical knowledge**, along with **lexical repetition**.

(4)
B: and uh - you know my own personal finances are
A: well sure
B: it's just out
A: but you have applied er for *monies (FNP; im_12; st; LS;)* I keep hearing wherever I go

Finally, a category named as **discourse knowledge** was used to classify cases in which the resolution required full processing of combined bits of discourse information. These four broader categories, including the essentially syntactic information required for the first-candidate strategies, grouped more fine-grained subclassifications in all cases, except for collocational knowledge. Thus, the umbrella categories were used to perform a statistical analysis using the data collected by means of manual annotation. However, the more detailed classification was retained in the actual annotation of the sample. The same approach was used in the other attributes.

Frequencies for each category were then used in three different statistical procedures: a chi-square test; a measure of association; and the model-building variety of loglinear analysis. Chi-square tests with the attributes considered two by two showed statistical significance in all measurements ($p < 0.00005$) in both languages. The Goodman and Kruskal tau was used to measure association between attributes two by two. Association was shown to be high (over 0.30) between the processing strategy and the other three attributes, but relatively low (under 0.30) between these three attributes measured two by two. The loglinear analysis revealed that interactions considering three of the attributes were significant whenever the processing strategy was one of the three. The opposite was true when it was not. These results were true for both languages with minor variations.

The statistical analysis showed thus that the classification model was adequate to represent the anaphora world. Moreover, it became clear that the attribute named as **processing strategy** yielded the highest information gain, acting as a link between the type of anaphor and the other two attributes which classify the antecedent. Therefore, the type of anaphor in itself, which could be mapped from POS tags or, in some cases, skeleton parsing (see Mitkov 1997), only became truly useful information for the resolution of the anaphoric reference when associated to the definition of a processing strategy. This made of course psycholinguistic sense, as it is not difficult to infer from corpus data that the same anaphor (such as *it* or *that*) may appear in contexts that lead to distinct processing demands for their resolution.

3 The antecedent-likelihood theory

The AL theory is made up of a series of entries for each type of anaphor. Entries contain

instructions organised in an algorithm-like form to check the applicability of all possible processing strategies, relying on information taken from the training set. The initial information considered is the probability of occurrence for each processing strategy and the two other attributes. As a result, some categories included in the general classification model are never checked because there are no tokens in the training set associating them to the type of anaphor in question. The subject pronoun entry is shown below.

Subject pronoun
global probability = 0.247
Category probabilities

process. strat.	type antec.	topical role
FtCCh = 0.458	explicit=0.886	dthel= 0.341
FtC = 0.232	implicit=0.058	st= 0.188
CK = 0.155	NR= 0.049	sst = 0.156
DK = 0.090	dim=0.008	dt = 0.055
ScRf = 0.047		thel = 0.110
Pl = 0.012		fdv = 0.056
Dx = 0.003		p_st = 0.038
SK = 0.001		p_sst= 0.024
SetMb = 0.001		p_dthel=0.014
SetCr = 0.001		p_dt = 0.005
		p_thel= 0.005
		sithel= 0.004

The table with the category probabilities[2] defines the likelihood of categories in the three other variables being assigned to tokens of the anaphor type described in the entry, having the total number of tokens for the type of anaphor – not the full sample – as a reference. The first column specifies the probabilities for the categories which define the processing strategy, while the second column shows the figures for the type of antecedent, and the third column lists the topical roles of antecedents with the respective numbers. In order to make the table visually compact, most of the categories are listed using the code specified for the annotation of the sample.

[2] Categories cannot be fully described in this paper for reasons of space. The essential features have been presented though.

Some processing information can be directly derived from the table of category probabilities. Categories which are not listed in the columns of the variables they belong to were not used to classify any tokens of the anaphor type, and thus can be left out of the processing. This may mean, for instance, that the processing need not be concerned with implicit antecedents for a given type of anaphor, because there are no tokens classified as such. Another possibility is that no tokens have been classified as being processed on the basis of collocational knowledge, and thus there is no point in checking the collocation list in search of matches.

The header in AL theory entries is followed by a set of instructions organised in algorithm-like form. These instructions rely on the taxonomy employed to analyse processing strategies. The choice is based both on the results of the loglinear analysis and on the nature of the variable, which is in fact a description of the way a given anaphor token is resolved. The typical instruction appears as **check ps**, **ps** being any category included in the list of possible classifications of processing strategy for the type of anaphor. This means that the processing towards resolution of an anaphor of the type described in the entry should check, at this point, whether the processing strategy specified is a possible way to identify the correct antecedent.

The typical **check ps** instruction is usually followed by a set of attached probabilities specific to the processing strategy being checked. These probabilities concern categories in the remaining two variables. Other information, such as the probability of predicate topical roles, may be added whenever this is felt to be useful. The subsequent items in a typical **check ps** instruction are **recognition** and **resolution path**. The first item contains information about features of the token itself and the immediate context in which it occurs, based on the observation of corpus data. The purpose is to guide the processing in the attempt to recognise the need for a certain type of processing strategy in order to resolve the anaphoric reference. The second item contains

information related to the actual identification of the correct antecedent.

The amount and complexity of information included in each one of the items varies with the type of anaphor and the processing strategy. In some cases, the recognition requires careful analysis, involving a number of details and check-ups. In other cases, recognising that a certain processing strategy is the adequate one is not as difficult as identifying the antecedent, as in some cases of discourse-implicit antecedents. The AL theory is built so as to permit the expansion or reduction of guidelines included as instructions or items within instructions.

In case a given processing strategy presents sufficient diversity of recognition and/or resolution patterns, the instructions may be divided into subtypes of recognition and resolution. This approach to the form of entries applies generally but not always, that is, there may be **check ps** instructions which do not include one or more of the items described above. There may also be instructions which specify actions of an unique nature for the type of anaphor or processing strategy under scrutiny.

The extract of the subject pronoun entry shown below illustrates this flexibility. The header shown above is followed by two instructions which break with the general **check ps** norm, only to return to it in the third instruction, as shown below.

- check if POS tag is Q-tag item
  - ➢ if not, go to instruction 2; if yes
  - ➢ go to tag-question entry in collocation list
  - ➢ follow resolution path in entry
- identify pronoun
  - ➢ pronoun is *he*, *she* or *they*
    - ➢ go to instruction 5
  - ➢ pronoun is *it*
    - ➢ go to instruction 4
  - ➢ pronoun is first or second person
    - ➢ go to instruction 3
- check secondary reference
  - attached probabilities
  - type of antecedent
    - explicit = 0.889
    - implicit = 0.111
  - implicit antecedents are in a chain
  - ultimate resolution by shared knowledge
  - topical roles
    - dthel = 0.750
    - st = 0.250
- recognition
  - separate from endophoric usage
  - previous move
    - verbs *say*; *ask*; *answer*; *explain*
    - subject a third person pronoun or personal name
    - simultaneous tense and person shift between utterances
  - if it is a second person pronoun
    - check identifying vocative in the utterance
- resolution path
  - select first human candidate searching backwards
    - check lexical clues
  - if there is an identifying vocative
    - select it as the antecedent
  - verbs say; ask; answer; explain
    - subject a third person pronoun or personal name
    - simultaneous tense and person shift between utterances
- check collocation list

The AL theory was manually tested on a previously analysed dialogue used as a test bed. There were 804 cases of anaphora in the testing set for English. The AL theory predicted the correct antecedent in 98.4% of the cases, which is evidently a satisfactory result. Results were also satisfactory, although not quite as good (93.5%), for Portuguese. However, the score was only obtained on the assumption that the dialogue had been POS-tagged, parsed and segmented according to topicality, using the procedures defined for each category in the attribute named as **topical role of the antecedent**. These are not minor assumptions, particularly if it is taken into account that, in real-life processing situations, these tasks would

have to be carried out during an ongoing conversation.

Nevertheless, the approach seems worth pursuing as a promising way to solve a difficult problem in the actual implementation of dialogue interfaces and in NLP in general. Thus, the attempt to transform the AL theory into an automatic procedure may be a useful way forward.

4 The decision trees for coreference resolution

The general procedure for the resolution of any anaphora case is then to check the processing strategy with the highest probability first. If anaphors classified as **determinative possessives** in the English sample are taken as an example, this strategy would be the one named as **first-candidate chain**, in which the first appropriate candidate – in syntactic terms – searching backwards is selected, although it is also an anaphor. It may be safely assumed that this anaphor has already been dealt with, as it precedes the one being resolved.

Checking a processing strategy for adequacy involves a recognition procedure specified in the entry, which, in the example considered above, would be to check the appropriateness of the first candidate. However, the probabilities indicate that there were cases in the training set in which this type of anaphor was resolved by means of **discourse knowledge**. This means that there were tokens in which the use of syntactic information only – as in Hobbs' "naive" algorithm – would lead to the identification of an incorrect antecedent.

Therefore, ways of checking whether the first appropriate candidate is actually the correct antecedent had to be devised. Two basic routines were used: selectional restrictions and association history. As formalised in Katz and Fodor (1963), selectional restrictions are semantic constraints which the sense of a given word imposes on those syntactically related to it. Thus, whenever an anaphor is linked to a verb as a complement, it is useful to check if a candidate antecedent is a good fit by using selectional restrictions.

There were cases in the training set, however, in which selectional restrictions would not detect the incorrectness of a syntactically appropriate candidate. A second kind of lexical clue was then included as a checking routine: the association history. It is unusual – although not impossible of course – that pronoun reference is used in the first instance of an association between a verb and a referent. This is even less likely in situations in which there is an established competitor with a record of tokens repeatedly associated to the verb in question.

These checking routines may signal that it is advisable to consider bypassing the first candidate on the basis of discourse information. Checking the possibility of a resolution by means of discourse knowledge usually involves a recognition procedure, which relies on topicality information. If the alternative candidate selected is one of the highly salient discourse entities, the chances that the speaker felt the listener would successfully process the reference are much higher, making the bypass of the first candidate far more likely.

The entry for **determinative possessives** is a relatively simple one, however, if compared to those for **subject pronouns** or **anaphoric demonstratives** in English or **anaphoric verbs** in Portuguese. Moreover, entries for other types of anaphor may require various forms of checking routines, which are specific to the type of anaphor in question. In spite of this highly complex and broad set of required information, it seems possible to organise it into decision trees for operational use. The notion of decision tree (as in Quinlan 1993) may have to be somewhat expanded in order to accommodate the various bits of specific information related to each type of anaphor.

At present, several different algorithms and adaptations of these algorithms are being tested in order to establish their adequacy to the task, including the well-known C4.5. A hybrid approach, in which an example-based alternative process would choose the most closely related

case in the training set and use it to resolve a new case of anaphora, is also being considered, having the TiMBL package (Tilbury 1999) as a primary reference. It is expected that initial tests will be run soon, yielding results which will be then used to gradually improve the approach and its implementation. The GATE structure (Cunnigham et al. 1995) is likely to be used as a way to organise the various required elements of linguistic information as an integrated system. At the present stage, however, the software mentioned are quoted as reference rather than firm choice.

5. Conclusion and future developments

The process of building solutions for natural language processing on the basis of corpus information may rely simply on a classification model of any kind that would enable decision trees to be created inductively. However, the direct observation of corpus tokens allows the sort of refinement that may prove crucial for the actual operational success of the model in real-life processing situations. The approach described in this paper is an attempt to find an appropriate balance between the practicality of automatically inducing decision trees out of a training set and the thoroughness that the contrastive analysis of the various cases in the corpus is likely to accomplish.

The systematisation of observed regularities in combination with statistical evidence proved very successful in dealing with the testing set of cases previously analysed for the purpose. It is also true, nevertheless, that the complexity introduced by the inclusion of a large amount of information to be taken into account during the processing makes actual implementation extremely hard. Therefore, the high score of the manual test must be seen cautiously. Future developments of the approach described in the present paper aim at testing the actual gain of dealing with a thorough account of anaphoric relations in dialogues as compared to the increased difficulty of implementation, of which the inclusion of topicality and segmentation in the model are obvious examples. It is expected that the above mentioned balance will be eventually reached, preserving the satisfactory results to an extent that offsets the undesirable processing complexity.

References

Biber, D. (1992)Using computer-based text to analyse the referential strategies of spoken and written texts. In: Jan Svartvik, editor, *Directions in corpus linguistics*, pages 215-252, Mouton de Gruyter, Berlin.

Cunningham, H., Gaizauskas, R. and Wilks, Y. (1995). *A General Architecture for Text Engineering (GATE) -- a new approach to Language Engineering R & D.* Department of Computer Science, University of Sheffield, CS-95-21.

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, K. (1999). *TiMBL: Tilburg Memory Based Learner, version 2.0, Reference Guide.* ILK Technical Report 99-01. Tilburg University.

Grosz, B. and Sidner, C. (1986). Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175-204.

Halliday, M.K. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Hobbs, J. (1986). Resolving pronoun references. In B.L. Webber, Barbara Grosz, and K. Jones, editors, *Readings in natural language processing*. Morgan Kaufmann, Palo Alto, CA.

Hoey, M. (1991). Patterns of lexis in text. Oxford University Press, Oxford.
Katz, J.J. and Fodor, J.A. (1963). The structure of a semantic theory. *Language*, 39:170-210.

Mitkov, R. (1997) How far are we from (semi-) automatic annotation of anaphoric links in corpora ? In: Ruslan Mitkov and Branimir Boguraev, editors, *Workshop on operational factors in practical, robust anaphora resolution for unrestricted texts*, pages 110-115, Madrid, July 1997, ACL/EACL.

Quinlan, JR (1987). *C4.5 : Programs for Machine Learning.* Morgan Kaufmann, Palo Alto, CA.

Quirk, R., Greenbaum, S., Svarvik, J. and Leech, G. (1985) *A comprehensive grammar of the English language*. Longman, London.

Webber, B.L. (1979). *A formal approach to discourse anaphora.* Academic Press, New York.