

Resolving Anaphors in Embedded Sentences

Saliha Azzam
University of Sheffield
Department of Computer Science
Regent Court
211 Portobello Street
Sheffield S1 4DP U.K.
S.Azzam@dcs.shef.ac.uk

Abstract

We propose an algorithm to resolve anaphors, tackling mainly the problem of intrasentential antecedents. We base our methodology on the fact that such antecedents are likely to occur in embedded sentences. Sidner's focusing mechanism is used as the basic algorithm in a more complete approach. The proposed algorithm has been tested and implemented as a part of a conceptual analyser, mainly to process pronouns. Details of an evaluation are given.

1 Introduction

Intrasentential antecedents, i.e., antecedents occurring in the same sentence as the anaphor, are a crucial issue for the anaphora resolution method. The main problem is to determine the constraints that intrasentential phrases must respect in anaphoric relations. These constraints are used to determine relations between a given anaphor and its antecedents. Until now, this kind of constraint has been tackled mainly in terms of syntactic aspects, see (Lappin and Leass, 1994) (Merlo, 1993) and (Hobbs, 1985). We propose to consider new kinds of criteria that combine semantic restrictions with sentence structure.

One of these criteria is, for example, the way in which the verb meaning influences the sentence structure, then the way in which the sentence structure influences the anaphoric relations between intrasentential phrases. The structure we studied is the embedded sentence structure. Indeed, an important assumption we have made is that embedded sentences favour the occurrence of intrasentential antecedents. We exploit the focusing mechanism proposed by Sidner (Sidner, 1979) (Sidner, 1981) (Sidner, 1983) extending and refining her algorithms. The algorithm is designed for anaphors generally, even if we focus mainly on pronouns in this paper. Indeed, the distinction between different kinds of anaphors is made at the level of anaphor interpretation rules. These resolution rule aspects will not be

developed here; though they have been developed in the literature, e.g., see (Carter, 1987), and (Sidner, 1981) (Sidner, 1983). We focus more on the mechanisms that handle these different kinds of rules.

We first present how intrasentential antecedents occur in embedded sentences. We recall the main ideas of the focusing approach, then expand on the main hypotheses which led the design of the anaphora resolution algorithm.

2 Intrasentential Antecedents

2.1 Embedded sentences and elementary events

An embedded sentence contains either more than one verb or a verb and derivations of other verbs (see sentence 1 with verbs *said* and *forming*).

- 1) Three of the world's leading advertising groups, Agence Havas S.A. of France, Young & Rubicam of the U.S. and Dentsu Inc. of Japan, said they are forming a global advertising joint venture.

Broadly speaking embedded sentences concern more than one fact. In sentence 1 there is the fact of saying something and that of forming a joint venture. We call such a fact an elementary event (EE hereafter). Thus an embedded sentence will contain several EEs.

Factors that influence embedded sentences are mainly semantic features of verbs. For example the verb *to say*, that takes a sentence complement favours an introduction of a new fact, i.e., "to say something", and the related fact. There are other classes of verbs such as *want to*, *hope that*, and so on. In the following, subordinate phrases, like relative or causal sentences, will be also considered as embedded ones.

2.2 Embedded sentences with intrasentential antecedents

First of all, we will distinguish the Possessive, Reciprocal and Reflexive pronouns (PRR hereafter) from the other pronouns (non-PRR hereafter).

On the basis of 120 articles, of 4 sentences on average, containing 332 pronouns altogether, we made the following assumption (1):

Assumption : non-PRR pronouns can have intrasentential antecedents, only if these pronouns occur in an embedded sentence.

The statistics below show that of 262 non-PRR pronouns, there are 244 having intrasentential antecedents, all of which occur in embedded sentences and none in a "simple" sentence. The remaining 18 non-PRR pronouns have intersentential antecedents.

Pronouns	332
non-PRR	262
With intrasentential antecedents in an embedded sentence	244
With intrasentential in a simple sentence	0
With intersentential antecedents	18

Our assumption means that, while the PRR pronouns may find their antecedents in an non embedded sentence (e.g., sentences 2 and 3) the non-PRR pronouns can not.

- 2) Vulcan made *its* initial Investment in Telescan in May, 1992.
- 3) The agencies HCM and DYR are *themselves* joint ventures.

Without jumping to conclusions, we cannot avoid making a parallel with the topological relations defined in the binding theory (Chomsky, 1980), between two coreferring phrases in the syntactic tree level. Assumption 1 redefines these relations in an informal and less rigorous way, at the semantic level, i.e., considering semantic parameters such as the type of verbs that introduce embedded sentences.

2.3 Using Sidner's Focusing Approach

To resolve anaphors one of the most suitable existing approaches when dealing with anaphor issues in a conceptual analysis process is the focusing approach proposed by Sidner. However, this mechanism is not suitable for intrasentential cases. We propose to exploit its main advantages in order to build our anaphora resolution mechanism extending it to deal also with intrasentential antecedents.

We describe the main elements of the focusing approach that are necessary to understand our method, without going into great detail, see (Sidner, 1979) (Sidner, 1981) (Sidner, 1983). Sidner proposed a methodology, modelling the way "focus" of attention and anaphor resolution influence one another. Using pronouns reflects what the speaker has focused on in the previous sentence, so that the focus is that phrase which the pronouns refer to. The resolution is organised through the following processes:

- The expected focus algorithm that selects an initial focus called the "expected focus". This selection may be "confirmed" or "rejected" in subsequent sentences. The expected focus is generally chosen on the basis of the verb semantic categories. There is a preference in terms of thematic position: the "theme" (as used by Gruber and Anderson, 1976 for the notion of the object case of a verb) is the first, followed by the goal, the instrument and the location ordered according to their occurrence in the sentence; the final item is the agent that is selected when no other role suits.
- The anaphora interpreter uses the state of the focus and a set of algorithms associated with each anaphor type to determine which element of the data structures is the antecedent. Each algorithm is a filter containing several interpretation rules (IR). Each IR in the algorithm appropriate to an anaphor suggests one or several antecedents depending on the focus and on the anaphor type.
- An evaluation of the proposed antecedents is performed using different kinds of criteria (syntax, semantics, inferential, etc.)
- The focusing algorithm makes use of data structures, i.e., the focus registers that represent the state of the focus: the current focus (CF) representation, alternate focus list (AFL) that contains the other phrases of the sentence and the focus stack (FS). A parallel structure to the CF is also set to deal with the agentive pronouns. The focusing algorithm updates the state of the focus after each sentence anaphor (except the first sentence). After the first sentence, it confirms or rejects the predicted focus taking into account the results of anaphor interpretation. In the case of rejection, it determines which phrase is to move into focus.

This is a brief example (Sidner 1983) :

- a Alfred and Zohar liked to play baseball.
 - b *They* played it every day after school before dinner.
 - c After *their* game, Alfred and Zohar had ice cream cones.
 - d *They* tasted really good.
- In a) the expected focus is "baseball" (the theme)
 - In b) "it" refers to "baseball" (CF). "they" refers to Alfred and Zohar (AF)
 - The focusing algorithm confirms the CF.
 - In d) "they" refers to "ice cream cones" in AFL.
 - The focusing algorithm decides that since no anaphor refers to the CF, the CF is stacked and

“ice cream cones” is the new CF (focus movement).

We call a basic focusing cycle the cycle that includes :

- the focusing algorithm
- followed by the interpretation of anaphors,
- then by the evaluation of the proposed antecedents.

2.4 What needs to be improved in the focusing approach?

2.4.1 Intrasentential antecedents

The focusing approach always prefers the previous sentences’ entities as antecedents to the current sentences. In fact only previous sentence entities are present in the focus registers. Thus phrases of the current sentence can not be proposed as antecedents. This problem has already been underlined, see (Carter, 1987) in particular who proposed augmenting the focus registers with the entities of the current sentence. For example in sentence 4 while the focus algorithm would propose only “John” as an antecedent for “him”, in Carter’s method “Bill” will also be proposed.

4) John walked into the room. He told Bill someone wanted to see *him*.

2.4.2 Initial Anaphors

The focusing mechanism fails in the expected focus algorithm when encountering anaphors occurring in the first sentence of a text, which we call initial anaphors, such as *They* in sentence (1). The problem with initial anaphors is that the focus registers cannot be initialised or may be wrongly filled if there are anaphors inside the first sentence of the text. It is clear that taking the sentence in its classical meaning as the unit of processing in the focusing approach, is not suitable when sentences are embedded.

We will focus on the mechanisms and algorithmic aspects of the resolution (how to fill the registers, how to structure algorithms, etc.) and not on the rule aspects, like how IRs decide to choose *Bill* and not *John* (sentence 4).

3 Our Solution

As stated above, embedded sentences include several elementary events (EEs). EEs are represented as conceptual entities in our work. We consider that such successive EEs involve the same context that is introduced by several successive short sentences. Moreover, our assumption states that when non-PRR anaphors have intrasentential antecedents, they occur in embedded sentences. Starting with these considerations, the algorithm is governed by the hypotheses expanded below.

3.1 Main hypotheses

First hypothesis : EE is the unit of processing in the basic focusing cycle.

An EE is the unit of processing in our resolution algorithm instead of the sentence. The basic focusing cycle is applied on each EE in turn and not sentence by sentence. Notice that a simple sentence coincides with its EE.

Second hypothesis : The “initial” EE of a well formed first sentence does not contain non-PRR pronouns just as an initial simple sentence cannot.

For example, in the splitting of sentence 1 into two EEs (see below), EE1 does not contain non-PRR pronouns because it is the initial EE of the whole discourse.

EE1) “Three of the world’s leading advertising groups, Agence Havas S.A. of France, Young & Rubicam of the U.S. and Dentsu Inc. of Japan, said”
EE2) “they are forming a global advertising joint venture.”

Third hypothesis : PRR pronouns require special treatment.

PRR could refer to intrasentential antecedents in simple sentences (such as in those of sentences 3 and 4). An initial EE could then contain an anaphor of the PRR type. Our approach is to add a special phase that resolves first the PRRs occurring in the initial EE before applying the expected focusing algorithm on the same initial EE. In all other cases, PRRs are treated equally to other pronouns.

This early resolution relies on the fact that the PRR pronouns may refer to the agent, as in sentence 3, as well as to the complement phrases. However the ambiguity will not be huge at this first level of the treatment. Syntactic and semantic features can easily be used to resolve these anaphors. This relies also on the fact that the subject of the initial EE cannot be a pronoun (second hypothesis).

Having mentioned this particular case of PRR in initial EE, we now expand on the whole algorithm of resolution.

3.2 The Algorithm

In the following, remember that what we called the basic focusing cycle is the following successive steps :

- applying the resolution rules,
- applying the focusing algorithm, i.e., updating the focus registers
- the evaluation of the proposed antecedents for each anaphor.

The algorithm is based on the decomposition of the sentence into EEs and the application of the basic focusing cycle on each EE in turn and not sentence by sentence.

The complete steps are given below (see also figure 1):

Step 1 Split the sentence, i.e., its semantic representation, into EEs.

Step 2 Apply the expected focus algorithm to the first EE.

Step 3 Perform the basic focusing cycle for every anaphor of all the EEs of the current sentence.

Step 4 Perform a collective evaluation (i.e., evaluation that involves all the anaphors of the sentence), when all the anaphors of the current sentence are processed.

Step 5 Process the next sentence until all the sentences are processed:

- split the sentence into EEs
- apply Step 3 then Step 4.

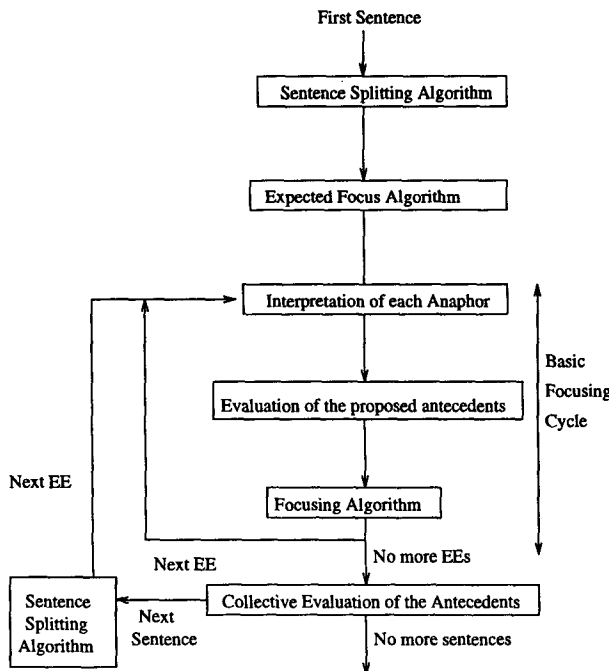


Figure 1: The Algorithm

Main Results :

1. Intrasentential antecedents are taken into account when applying the focusing algorithm. For example, in sentence 1, the intrasentential

antecedent *Bill* will be taken into account, because EE1 would be processed beforehand by the expected focusing algorithm.

2. The problem of initial anaphors is then resolved. The expected focusing algorithm is applied only on the initial EE which must not contain anaphors.

3.3 Examples and results

To illustrate the algorithm, let's consider the following sentence :

Lafarge Coppee said it would buy 10 percent in National Gypsum, the number two plasterboard company in the US, a purchase which allows it to be present on the world's biggest plasterboard market.

At the conceptual level, there are 3 EEs. They are involved respectively by the *said*, *buy*, and *allows* verbs. They correspond respectively to the following surface sentences:

EE1 "Lafarge Coppee said"

EE2 "it would buy 10 percent in National Gypsum, the number two plasterboard company in the US"

EE3 "a purchase which allows it to be present on the world's biggest plasterboard market."

Consider the algorithm :

- the expected focusing algorithm is applied to the first EE, EE1, which contains non-PRR anaphors.
- the other phases of the algorithm, i.e., the basic focusing cycle, are applied to the subsequent EEs :
 - EE2 contains only one pronoun *it*, which is resolved by the basic focusing cycle
 - *it* in EE3 will be resolved in the same way.

The anaphora resolution has been implemented as a part of a conceptual analyser (Azzam, 1995a). It dealt particularly with pronouns. It has been tested on a set of 120 news reports. We made two kinds of algorithm evaluations: the evaluation of the implemented procedure and an evaluation by hand. For the implementation the success rate of resolution was 70%. The main cases of failure are related to the non implemented aspects like the treatment of coordination ambiguities and the appositions, or other anaphoric phenomena, like ellipsis.

For the second evaluation which concerns the real evaluation of the approach, i.e., without going into the practical issues concerning implementation, the success rate was 95%. The main cases of failure were due to the cases that were not considered by the algorithm, like for example the pronouns occurring before their antecedents, i.e., cataphors. Such cases occur for example in sentences 5 and 6 pointed out

by Hobbs (Hobbs, 1985) to discuss the cases that are not handled easily in the literature.

- 5) Mary sacked out in *his* apartment before *Sam* could kick her out.
- 6) Girls who *he* has dated say that *Sam* is charming.

Our algorithm fails in resolving *his* in 5, because the algorithm searches only for the entities that precede the anaphor in the text. The same applies for *he* in 6. However improving our algorithm to process classical cases of cataphors, such as that in sentence 6, should not require major modifications, only a change in the order in which the EEs are searched.

For example, to process pronouns of the sentence 6 split into two EES (see below), the algorithm must consider EE2 before EE1. This means applying the step 2 of the algorithm to EE2, then step 3 to EE1. The sentence 5 should require specific treatment, though.

- EE1) “that *Sam* is charming”
- EE2) “Girls who *he* has dated say”

Hobbs also pointed out the cases of “picture noun” examples, as in sentences 7 and 8:

- 7) John saw a picture of him.
- 8) John’s father’s portrait of him.

In 7 our algorithm is successful, i.e., it will not identify *him* with *John* because of our previous assumption (section 2.2). However our algorithm would fail in 8 because the non-PRR pronoun *him* could refer to *John* which occurs in the same EE.

Notice that Hobbs’ (Hobbs, 1985) remark that “the more deeply the pronoun is embedded and the more elaborate the construction it occurs in, the more acceptable the non reflexive” is consistent with our assumption.

For example in the embedded sentence 9 where either the reflexive (himself) or non reflexive pronouns (him) may be used, it is more natural to make use of *him*.

- 9) John claimed that the picture of him hanging in the post office was a fraud.

4 The Conceptual Level

We comment here on the main aspects of the conceptual analysis that are related to the anaphora resolution process. They concern mainly the way of splitting embedded sentences and the problems of determining the theme and of managing the other ambiguities and the several readings.

The conceptual analyser’s strategy consists of a continuous step-by-step translation of the original natural language sentences into conceptual structures (CS hereafter). This translation uses the results of the syntactic analysis (syntactic tree). It is a progressive substitution of the NL terms located in the syntactic tree with concepts and templates of

the conceptual representation language. Triggering rules are evoked by words of the sentence and allow the activation of well-formed CS templates when the syntactico-semantic filter is unified with the syntactic tree. The values caught by the filter variables are the arguments of the CS roles, i.e., they fill the CS roles. If they are anaphors, they are considered to be unbound variables and result in unfilled roles in the CS. The anaphora resolution aims therefore at filling the unfilled roles with the corresponding antecedents.

4.1 Splitting into EEs

The splitting of a sentence in EE is done on the corresponding CS. A minimal CS is a template comprising a predicate that identifies the basic type of the represented event and a set of roles or predicate cases.

For example, the sentence “to say that they agree to form a joint venture” is represented, in a simplified way, with three templates, corresponding to the predicates:

- *move* information (from “to say”),
- *produce* an agreement (from “to agree”),
- *produce* a joint venture (from “to form”).

Given that one template at the semantic level represents an elementary event, the splitting is implicitly already done when these templates are created in the triggering phase. Indeed, the syntactico-semantic filter of the triggering rules takes into account the semantic features of words (mainly verbs) for recognising in the surface sentence those that are able to trigger an elementary event.

4.2 Determining the theme

Gruber and Anderson characterise the theme as follows: if a verb describes a change to some entity, whether of position, activity, class or possession, then the theme is the changed entity, (Gruber, 1976) and (Anderson, 1977). As Carter (Carter, 1987) demonstrated, this definition of Gruber and Anderson is sufficient to apply the focusing mechanism. This assumption is particularly apt when we dispose of a conceptual representation. Indeed, to determine the thematic roles, we established a set of thematic rules that affect for a given predicative occurrence, its thematic functions according to the predicate type, the role type and the argument’s semantic class.

4.3 Managing other ambiguities

An important aspect appears when one designs a concrete system, namely how to make other disambiguation processes cohabit. In the conceptual analyser, the general disambiguation module (GDM) deals with other ambiguities, like prepositional phrase attachment. It coordinates the treat-

ment of different kinds of ambiguities. This is necessary because the conceptual structures (CS) on which the rules are performed could be incomplete because of other types of ambiguities not being resolved. For example, if the CF of the sentence is a PP object that is not attached yet in the CS the thematic rules fail to fill the CF. The GDM ensures that every disambiguation module intervenes only if previous ambiguities have already been resolved. The process of co-ordinating ambiguity processing is fully expanded in (Azzam, 1995b).

4.4 Multiple readings

When dealing with ambiguities, another important aspect is managing multiple readings. At a certain point when the GDM calls the anaphora module to deal with a given anaphor, the status of the conceptual analysis could be characterised by the following parameters :

- The set of conceptual structures for the current reading R_i on which the resolution is performed, given that several readings could arise from previous ambiguity processing.
- The set of conceptual structures of the current sentence S_i where the anaphor occurs;
- The set of conceptual structures of the current elementary event EE_i where the anaphor occurs after the S_i splitting.
- The state of the focus (content of the registers), SF_i

The main assumption is that the anaphora resolution algorithm always applies to a single state, (R_i , S_i , EE_i , SF_i) when resolving a given anaphor (Step 3) :

- a If several antecedents are still possible after the individual evaluation of the anaphor, R_i is then duplicated, in R_{ij} , as many times as there are possibilities.
- b When performing the collective evaluation of all S_i anaphors, every inconsistent R_{ij} is suppressed.
- c The result is a set of readings (R_{ij} , S_j , EE_j , SF_j).

5 Conclusion

We have proposed a methodology to resolve anaphors occurring in embedded sentences. The main idea of the methodology is the use of other kinds of restrictions between the anaphor and its antecedents than the syntactic ones. We demonstrated that anaphors with intrasentential antecedents are closely related to embedded sentences and we showed how to exploit this data to design the anaphora resolution methodology. Mainly, we exploited Sidner's focusing mechanism, refining the classical unit of processing, that is the sentence, to

that of the elementary event. The algorithm has been implemented (in Common Lisp, Sun Sparc) to deal with pronouns as a part of a deep analyser. The main advantages of the proposed algorithm is that it is independent from the knowledge representation language used and the deep understanding approach in which it is integrated. Thus, it could be set up in any conceptual analyser, as long as a semantic representation of the text is available. Moreover Sidner's approach does not impose its own formalisms (syntactic or semantic) for its application. The improvement of the proposed algorithm requires dealing with special cases of anaphors such as cataphors and also with specific cases which are not easily handled in the literature. For example, we saw that a solution to processing cataphors could be to reconsider the order in which the conceptual structures (elementary events beforehand) are searched.

6 Acknowledgements

This work has been supported by the European Community Grant LE1-2238 (AVENTINUS project).

References

- Anderson, S.R. 1977. Formal syntax. In Wasow and Akmajian, editors, *Comment on the paper by Wasow in Culicover*. Academic Press, pages 361-376.
- Azzam, Saliha. 1995a. *Computation of Ambiguities (Anaphors and PPs) in NL texts. CLAM : The prototype*. Ph.D. thesis, Paris Sorbonne University.
- Azzam, Saliha. 1995b. Anaphors, pps and disambiguation process for conceptual analysis. In *14th International Joint Conference on Artificial Intelligence (IJCAI'95)*. San Mateo (CA): Morgan Kaufmann.
- Carter, David. 1987. *Interpreting Anaphors in natural language texts*. Chichester : Ellis Horwood.
- Gruber, J.S. 1976. *Lexical structures in syntax and semantics*. North-Holland.
- Hobbs, Jerry. 1985. Resolving pronoun references. In B. Grosz K. Sparck-Jones B. Webber, editor, *Readings in Natural Language*, volume 44. Morgan Kaufmann Publishers Los Altos California, pages 311-338.
- Lappin, S. and H.J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:535-561.
- Merlo, P. 1993. For an incremental computation of intrasentential coreference. In *the 13th International Joint Conference on Artificial Intelligence (IJCAI'93)*, pages 1216-1221. San Mateo (CA): Morgan Kaufmann.

- Sidner, C. 1979. Toward a computation of intrasentential coreference. Technical Report TR-537, MIT. Artificial Intelligence Laboratory.
- Sidner, C. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7:217-231.
- Sidner, C. 1983. Focusing in the comprehension of definite anaphora. In Brady, M and Berwick R.C, editors, *Computational Models of Discourse*. Cambridge (MA) : The MIT Press.