A KNOWLEDGE-BASED AND SUBLANGUAGE-ORIENTED MODEL FOR ANAPHORA RESOLUTION

Ruslan Mitkov

Machine Translation Unit
School of Computer Science
University of Science Malaysia
11800 Penang, Malaysia
Fax (60-4) 873335
Email ruslan@cs.usm.my

ABSTRACT

The paper discusses a new knowledge-based and sublanguage-oriented model for anaphora resolution, which integrates syntactic, semantic, discourse, domain and heuristical knowledge for the sublanguage of computer science. Special attention is paid to a new approach for tracking the center of a discourse segment, which plays an important role in proposing the most likely antecedent to the anaphor in case of ambiguity.

INTRODUCTION

Anaphora resolution is a complicated problem in computational linguistics. Considerable research has been done by computational linguists ([Carbonell & Brown 88], [Dahl & Ball 90], [Frederking & Gehrke 87], [Hayes 81], [Hobbs 78], [Ingria & Stallard 89], [Rich & LuperFoy 88], [Robert 89]), but no complete theory has emerged which offers a resolution procedure with success guaranteed. All approaches developed - even if we restrict our attention to pronominal anaphora, which we will do throughout this paper - from purely syntactic ones to highly semantic and pragmatic ones, only provide a partial treatment of the problem.

Given the complexity of the problem, we think that to secure a comparatively successful handling of anaphora resolution one should adhere to the following principles: 1) restriction to a domain (sublanguage) rather than focus on a particular natural language as a whole; 2) maximal use of linguistic information integrating it into a uniform architecture by means of partial theories. Some more recent treatments of anaphora ([Carbonell & Brown 88], [Rich & LuperFoy 88]) do express the idea of "multi-level approach", or "distributed architecture", but their ideas a) do not seem to capture enough discourse and heuristical knowledge and b) do not concentrate on and investigate a concrete domain, and thus risk being too general. We have tried nevertheless to incorporate some of their ideas into our own proposals.

THE ANAPHORA RESOLUTION MODEL

Our anaphora resolution model integrates modules containing different types of knowledge - syntactic, semantic, domain, discourse, heuristical and common sense/world knowledge. All the modules share a common representation of the current discourse.

The syntactic module, for example, knows that the anaphor and antecedent must agree in number, gender and person. It checks if the c-command constraints hold and establishes disjoint reference. In cases of syntactic parallelism, it prefers the noun phrase with the same syntactic role as the anaphor, as the most probable antecedent. It knows when cataphora is possible and can indicate syntactically topicalized noun phrases, which are more likely to be antecedents than non-topicalized ones.

The semantic module checks for semantic consistency between the anaphor and the possible antecedent. It filters out semantically incompatible candidates following the current verb semantics or the animacy of the candidate. In cases of semantic parallelism, it prefers the noun phrase, having the same semantic role as the anaphor, as a most likely antecedent. Finally, it generates a set of possible antecedents whenever necessary.

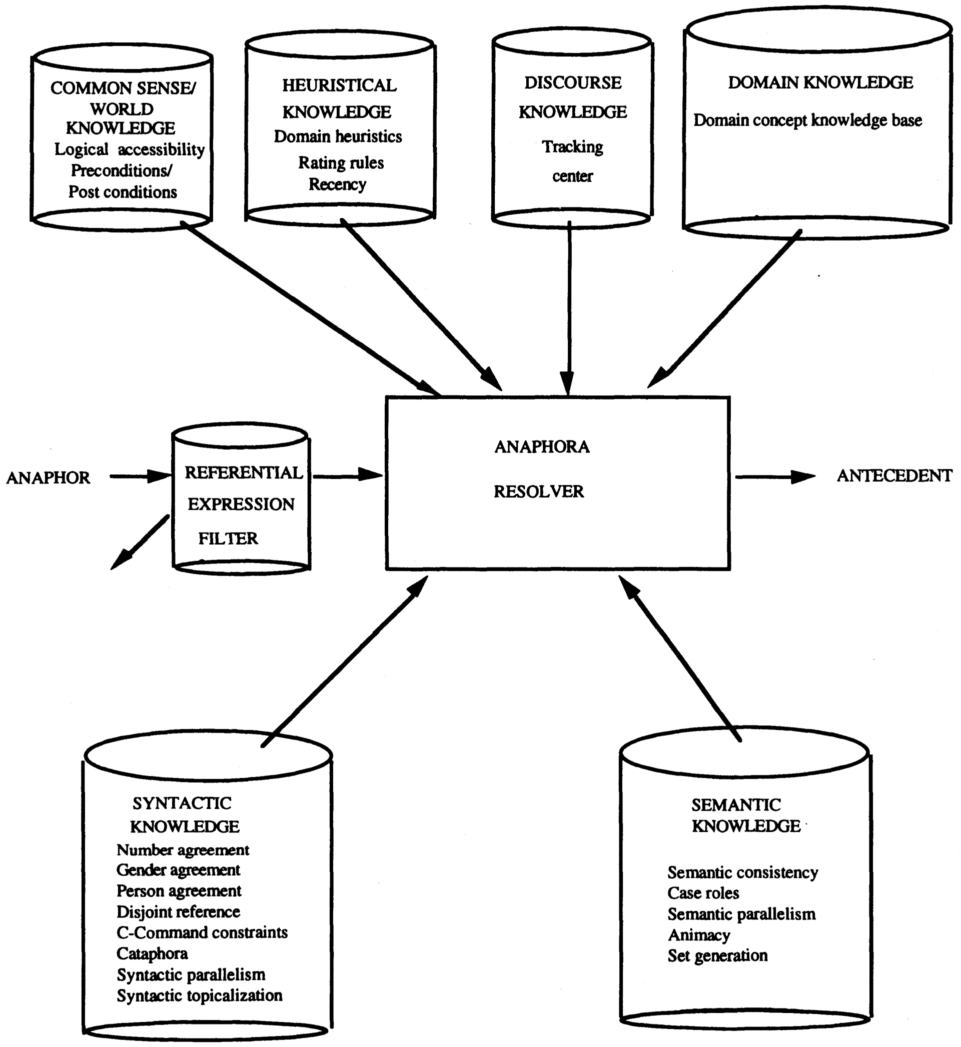
The domain knowledge module is practically a knowledge base of the concepts of the domain considered and the discourse knowledge module knows how to track the center of the current discourse segment.

The heuristical knowledge module can sometimes be helpful in assigning the antecedent. It has a set of useful rules (e.g. the antecedent is to be located preferably in the current sentence or in the previous one) and can forestall certain impractical search procedures.

The use of common sense and world knowledge is in general commendable, but it requires a huge knowledge base and set of inference rules. The present version of the model does not have this module implemented; its development, however, is envisaged for later stages of the project.

The syntactic and semantic modules usually filter the possible candidates and do not propose an antecedent (with the exception of syntactic and semantic parallelism). Usually the proposal for an antecedent comes from the domain, heuristical, and discourse modules. The latter plays an important role in tracking the center and proposes it in many cases as the most probable candidate for an antecedent.

The general structure of our anaphora resolution model is:



THE NEED FOR DISCOURSE CRITERIA

Although the syntactic and semantic criteria for the selection of an antecedent are already very strong they are not always sufficient to discriminate among a set of possible candidates. Moreover, they serve more as filters to eliminate unsuitable candidates than as proposers of the most likely candidate. Additional criteria are therefore needed.

As an illustration, consider the following text.

> Chapter 3 discusses these additional or auxiliary storage devices, which are similar to our own domestic tape cassettes and record discs. Figure 2 illustrates their connection to the main central memory.

In this discourse segment neither the syntactic, nor the semantic constraints can eliminate the ambiguity between "storage devices", "tape cassettes" or "record discs" as antecedents for "their", and thus cannot turn up a plausible antecedent from among these candidates. A human reader would be in a better position since he would be able to identify the central concept, which is a primary candidate for pronominalization. Our hypothesis is that the center of a sentence is the prime candidate for pronominal reference. Such a hypothesis has also been expressed by other researchers ([Allen87]), who use the alternative notion of 'focus'.

Following this hypothesis, and recognizing "storage devices" as the center, an anaphora resolution model would not have problems in picking up the center of the previous sentence ("storage devices") as antecedent for "their".

We see now that the main problem which arises is the tracking of the center in a discourse segment. Certain ideas and algorithms for tracking focus ([Allen87]) or center [Brennan et al.87] have been proposed, provided that one knows the focus or center of the first sentence in the segment. However, they do not try to identify this center. Our approach determines the most probable center of the first sentence, and then tracks it all the way through the segment, correcting the proposed algorithm at each step.

TRACKING THE CENTER IN THE SUBLANGUAGE OF COMPUTER SCIENCE

Our approach is sublanguage-oriented and has been developed on the basis of an examination of numerous computer science texts.

Before we present an informal discussion of the algorithm, we will present briefly some results which we obtained from our empirical observations and which helped us develop efficient sublanguage-dependent heuristics for tracking the center in the sublanguage of computer science.

1) First, we found that the subject is a primary candidate for center (in about 73% of the cases). The second most likely center would be the object (25%) and the third most likely one the verb phrase as a whole (2%).

2) Moreover, a noun phrase representing a domain concept is much more likely to be a center than a noun phrase which does not represent a domain concept.

3) Certain verbs like {discuss, illustrate, summarize, examine, describe, define, show...} select the object as a preferred center.

4) The occurrence of subjects like "chapter", "section", "table", personal pronouns "I", "we", "you", likewise selects the object as a preferred center.

5) The repetition of an NP throughout the discourse section increases the probability of its being a center.

6) An NP which occurs in the head of a section, part of which is the current discourse segment, has an increased probability to be a center.

7) A topicalized NP is likely to be center.

8) A definite NP is more likely to be the center than an indefinite NP.

9) NPs in the main clause are more likely to be the center than those in a subordinate clause.

10) In a complex sentence the anaphor often refers to an NP in the previous clause within the same sentence.

There are certain 'symptoms' which determine the subject or the object as a center with very high probability. Cases in point are 3) and 4). Other cases are not so certain, but to some extent quite likely. For example, if a non-concept NP is in subject position and if a repeated concept NP, which is also in a head, is in object position, it is almost certain that the latter is the unambiguous center. Moreover, certain preferences are stronger than others. For example an NP in subject position is preferred over an NP in a section head, but not in subject position.

A sentence which contains more than one clause may have more than one center - each belonging to one clause. Accordingly, we propose the following modification in the description of what is meant by the term "center": the linguistic element expressing a concept or set of concepts that are central to a clause or a sentence. Moreover, it is felt to be natural to distinguish between "clause center", "sentence center", and "discourse segment center".

We have made use of our empirical results (with approximating probability measures) and AI techniques to develop a proposer module which identifies the most likely center. We must point out that even if we do not need one for immediate antecedent disambiguation, a center must still be proposed for each sentence. Or else we will have to go all the way back to track it from the beginning of the segment when one is needed later on.

Tracking the center in a discourse segment is very important since knowing the center of each current sentence helps in many cases to make correct decisions about an antecedent in the event that syntactic and semantic constraints cannot discriminate among the available candidates.

AN ARTIFICIAL INTELLIGENCE APPROACH FOR CALCULATING THE PROBABILITY OF A NOUN (VERB) PHRASE TO BE IN THE CENTER

Based on the results described in the previous section, we use an artificial intelligence approach to determine the probability of a noun (verb) phrase to be the center of a sentence. Note that this approach allows us to calculate this probability in every discourse sentence, including the first one and to propose the most probable center. This approach, combined with the algorithm for tracking the center (in [Brennan et al. 87]), is expected to yield improved results.

Our approach uses an inference engine based on Bayes' theorem which draws an inference in the light of some new piece of evidence. This formula calculates the new probability, given the old probability plus some new piece of evidence.

Consider the following situation. According to our investigation so far, the probability of the subject being a center is 76%. Additional evidence (which we shall refer to as "symptom"), e.g. if the subject represents a domain concept, will increase the initial probability. If this NP is also the head of the section, the probability is increased further.

If the NP occurs more than once in the discourse segment, the probability gets even higher.

An estimation of the probability of a subject, (direct or indirect) object or verb phrase (the only possible centers in our texts) to be centers, can be represented as a predicate with arguments:

center (X, P_I, [symptom1 (weight factor 1_1, weight factor 1_2), ..., symptom N (weight factor N_1, weight factor N_2)])

where center (X, I, list) represents the estimated probability of X to be the center of a sentence (clause), X $\in$ {subject, object1, object2, .., verb phrase} and P_I is the initial probability of X to be the center of the sentence (clause).

Weight factor 1 is the probability of the symptom being observed with a noun (verb) phrase which is the center (we will henceforth refer to this factor as P_Y). Weight factor 2 is the probability of the symptom being observed with a noun (verb) phrase which is not the center (henceforth referred to as P_N).

Following this notation and our preliminary results, we can write for example:

center (object, 25, [symptom (verb_set, 40, 3), symptom (subject_set, 40,2), symptom (domain_concept (95, 80), symptom (repeated, 10, 5), symptom (headline, 10, 9)], symptom (topicalized, 6, 2), symptom (main_clause (85, 30), symptom (definite_form (90, 70)])

center (subject, 73, [symptom (domain_concept (95, 70), symptom (repeated, 10, 4), symptom (headline, 10, 8), symptom (topicalized, 10, 3), symptom (main_clause (85, 30), symptom (definite_form (85, 20)]).

This means that the object is the center in approximately 25% of the cases. In 40% of the cases where the center is the object the verb belongs to the set of verbs {discuss, illustrate, summarize, examine, describe, define...} and it is possible with 3% probability for the verb to be a member of this set while the center of the sentence is not the object.

The above facts, in Prolog notation, are part of a sublanguage knowledge base.

The process of estimating the probability of a given phrase being the center of a sentence (clause), is repetitive, beginning with an initial estimate and gradually working towards a more accurate answer. More systematically, the "diagnostic" process is as follows:

- start with the initial probability
- consider the symptoms one at a time
- for each symptom, update the current probability, taking into account: a) whether the sentence has the symptom and b) the weight factors P_Y and P_N.

The probability for an NP to be the center is calculated by the inference engine represented as a Prolog program (left out here for reasons of space), which operates on the basis of the sublanguage knowledge base and the "local" knowledge base. The latter gives information on the current discourse segment. Initially, our program works with manual inputs. The local knowledge base can be represented as Prolog facts in the following way:

```
observed (headline).
observed (domain_concept).
observed (repeated).
...................
```

The inference engine's task is to match the expected symptoms of the possible syntactic function as center in the knowledge base of the sentence's actual symptoms, and produce a list of (reasonably) possible candidates.

THE PROCEDURE: AN INTEGRATED KNOWLEDGE APPROACH

Our algorithm for assigning (proposing) an antecedent to an anaphor is sublanguage-oriented because it is based on rules resulting from studies of computer science texts. It is also knowledge-based because it uses at least syntactic, semantic and discourse knowledge. Discourse knowledge and knowing how to track the center play a decisive role in proposing the most likely antecedent.

The initial version of our project handles only pronominal anaphors. However, not all pronouns may have specific reference (as in constructions like "it is necessary", "it should be pointed out", "it is clear"...). So before the input is given to the anaphor resolver, the pronoun is checked to ensure that it is not a part of such grammatical construction. This function is carried out by the REFERENTIAL EXPRESSION FILTER.

The procedure for proposing an antecedent to an anaphor operates on discourse segments and can be described informally in the following way:

1) Propose the center of the first sentence of the discourse segment using the method described.

2) Use the algorithm proposed in [Brennan et al. 87], improved by an additional estimation of the correct probability supplied by our method, in order to track the center throughout the discourse segment (in case the anaphor is in a complex sentence, identify clause centers)

3) Use syntactic and semantic constraints to eliminate antecedent candidates.

4) Propose the noun phrase that has been filtered out as the antecedent in case no other candidates have come up; otherwise propose the center as the antecedent

The information obtained in 1) and 2) may not be used; however, it may be vital for proposing an antecedent in case of ambiguity.

To illustrate how the algorithm works, consider the following sample text:

> SYSTEM PROGRAMS
>
> We should note that, unlike user programs, system programs such as the supervisor and the language translator should not have to be translated every time they are used, otherwise this would result in a serious increase in the time spent in processing a user's program. System program are usually written in the assembly version of the machine language and are translated once into the machine code itself. From then on they can be loaded into memory in machine code without the need for any intermediate translation phase. They are written by specialist programmers called system programmers who know a great deal about the computer and the computer system for which their programs are written. They know the exact number of location which each system program will occupy and in consequence can make use of these numbers in the supervisor and translator programs.

The proposed center of the first sentence is "system programs". The center remains the same in the second, third, fourth and fifth sentence. Syntactic constraints are sufficient to establish the antecedent of "they" in the third sentence as "system programs". In the fourth sentence, syntactic constraints only, however, are insufficient. Semantic constraints help here in assigning "system programs" as antecedent to "they". In the fifth sentence neither syntactic nor semantic constraints can resolve the ambiguity. The correct decision comes from proposing the center of the previous sentence, in this case "system programs" (and not "programmers"!), as the most likely antecedent.

CONCLUSION

The model proposed has two main advantages. First, it is an integrated model of different types of knowledge and uses existing techniques for anaphora resolution. Second, it presents a new approach for tracking the center, which proposes centers and subsequently antecedents with maximal likelihood.

ACKNOWLEDGMENT

I would like to express my gratitude to Prof. Pieter Seuren for his useful comments.

REFERENCES

[Aone & McKee 93] Ch. Aone, D. McKee - *Language-independent anaphora resolution system for understanding multilingual texts.* Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, 22-26 June 1993, The Ohio State University, Columbus, Ohio

[Allen87] J. Allen - *Natural language understanding.* The Benjamin/Cummings Publishing Company Inc., 1987

[Brennan et al. 87] S. Brennan, M. Fridman, C. Pollard - *A centering approach to pronouns.* Proceedings of the 25th Annual Meeting of the ACL, Stanford, CA, 1987

[Carbonell & Brown 88] J. Carbonell, R. Brown - *Anaphora resolution: a multi-strategy approach.* Proceedings of the 12. International Conference on Computational Linguistics COLING'88, Budapest, August, 1988

[Dahl & Ball 90] D. Dahl, C. Ball - *Reference resolution in PUNDIT.* Research Report CAIT-SLS-9004, March 1990. Center for Advanced Information Technology, Paoli, PA 9301

[Frederking & Gehrke 87] R. Frederking, M. Gehrke - *Resolving anaphoric references in a DRT-based dialogue system: Part 2: Focus and Taxonomic inference.* Siemens AG, WISBER, Bericht Nr.17, 1987

[Grosz & Sidner 86] B. Grosz, C. Sidner - *Attention, Intention and the Structure of Discourse.* Computational Linguistics, Vol. 12, 1986

[Hayes 81] P.J. Hayes - *Anaphora for limited domain systems.* Proceedings of the 7th IJCAI, Vancouver, Canada, 1981

[Hirst 81] G. Hirst - *Anaphora in natural language understanding.* Berlin Springer Verlag, 1981

[Hobbs 78] J. Hobbs - *Resolving pronoun references.* Lingua, Vol. 44, 1978

[Ingria & Stallard 89] R. Ingria, D. Stallard - *A computational mechanism for pronominal reference.* Proceedings of the 27th Annual Meeting of the ACL, Vancouver, British Columbia, 26-29 June 1989

[Rich & LuperFoy 88] E. Rich, S. LuperFoy - *An architecture for anaphora resolution.* Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas, 9-12 February 1988

[Robert 89] M. Robert - *Résolution de formes pronominales dans l'interface d'interogation d'une base de données.* Thèse de doctorat. Faculté des sciences de Luminy, 1989