AN ALGORITHM FOR VP ELLIPSIS

Daniel Hardt
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
Internet: hardt@linc.cis.upenn.edu

ABSTRACT

An algorithm is proposed to determine antecedents for VP ellipsis. The algorithm eliminates impossible antecedents, and then imposes a preference ordering on possible antecedents. The algorithm performs with 94% accuracy on a set of 304 examples of VP ellipsis collected from the Brown Corpus. The problem of determining antecedents for VP ellipsis has received little attention in the literature, and it is shown that the current proposal is a significant improvement over alternative approaches.

INTRODUCTION

To understand an elliptical expression it is necessary to recover the missing material from surrounding context. This can be divided into two subproblems: first, it is necessary to determine the antecedent expression. Second, a method of reconstructing the antecedent expression at the ellipsis site is required. Most of the literature on ellipsis has concerned itself with the second problem. In this paper, I propose a solution for the first problem, that of determining the antecedent. I focus on the case of VP ellipsis.

VP ellipsis is defined by the presence of an auxiliary verb, but no VP, as in the following example[1]:

(1) a. It might have rained, any time;
b. only – it did not.

To interpret the elliptical VP "did not", the antecedent must be determined: in this case, "rained" is the only possibility.

The input to the algorithm is an elliptical VP and a list of VP's occurring in proximity to the elliptical VP. The algorithm eliminates certain VP's that are impossible antecedents. Then it assigns preference levels to the remaining VP's, based on syntactic configurations as well as other factors. Any VP's with the same preference level are ordered in terms of proximity to the elliptical VP. The antecedent is the VP with the highest preference level.

In what follows, I begin with the overall structure of the algorithm. Next the subparts of the algorithm are described, consisting of the elimination of impossible antecedents, and the determination of a preference ordering based on clausal relationships and subject coreference. I then present the results of testing the algorithm on 304 examples of VP ellipsis collected from the Brown Corpus. Finally, I examine other approaches to this problem in the literature.

THE ALGORITHM

The input to the algorithm is an elliptical VP(VPE), and VPlist, a list of VP's occurring in the current sentence, and those occurring in the two immediately preceding sentences. In addition, it is assumed that the parse trees of these sentences are available as global variables, and that NP's in these parse trees have been assigned indices to indicate coreference and quantifier binding.

The antecedent selection function is:

```
A-Select(VPlist,VPE)
  VPlist := remove-impossible(VPlist,VPE)
  VPlist := assign-levels(VPlist,VPE)
  antecedent := select-highest(VPlist,VPE)
```

First, impossible antecedents are removed from the VPlist. Then, the remaining items in VPlist are assigned preference levels, and the item with the highest preference level is selected as the antecedent. If there is more than one item with the same preference level, the item closest to the VPE, scanning left from the VPE, is selected.

[1] All examples are taken from the Brown Corpus unless otherwise noted.

The definition of the function **remove-impossible** is as follows:

```
remove-impossible(VPlist,VPE)
 For all v in VPlist
   if ACD(v,VPE) or
      BE-DO-conflict(v,VPE)
   then remove(v, VPlist)
```

There are two types of impossible antecedents: the first involves certain antecedent-containment structures, and the second involves cases in which the antecedent contains a BE-form and the target contains a DO-form. These are described in detail below.

Next, preference levels are assigned to remaining items in VPlist by the **assign-levels** function. (All items on VPlist are initialized with a level of 0.)

```
assign-levels(VPlist,VPE)
For all v in VPlist
  if related-clause(v,VPE) then
    v.level := v.level + 1
  if coref-subj(v,VPE) then
    v.level := v.level + 1
```

An antecedent is preferred if there is a clausal relationship between its clause and the VPE clause, or if the antecedent and the VPE have coreferential subjects. The determination of these preferences is described in detail below.

Finally, the **select-highest** function merely selects the item on VPlist with the highest preference level. If there is more than one item with the highest preference level, the item nearest to the VPE (scanning left) is selected.

IMPOSSIBLE ANTECEDENTS

This section concerns the removal of impossible antecedents from VPlist. There are two cases in which a given VP is not a possible antecedent. The first deals with antecedent-containment, the second, with conflicts between BE-forms and DO-forms.

ANTECEDENT CONTAINMENT

There are cases of VP ellipsis in which the VPE is contained within the antecedent VP:

[V [... VPE ...]]$_{VP}$

Such cases are traditionally termed antecedent-contained deletion (ACD). They are highly constrained, although the proper formulation of the relevant constraint remains controversial. It was claimed by May (1985) and others that ACD is only possible if a quantifier is present. May argues that this explains the following contrast:

(2) a. Dulles suspected everyone who Angelton did.
    b. * Dulles suspected Philby, who Angelton did.

However, it has been subsequently noted (cf. Fiengo and May 1991) that such structures do not require the presence of a quantifier, as shown by the following examples:

(3) a. Dulles suspected Philby, who Angelton did too.
    b. Dulles suspected Philby, who Angelton didn't.

Thus the algorithm will allow cases of ACD in which the target is dominated by an NP which is an argument of the antecedent verb. It will not allow cases in which the target is dominated by a sentential complement of the antecedent verb, such as the following:

(4) That still leaves you a lot of latitude. And I suppose it did.

Here, "suppose" is not a possible antecedent for the elliptical VP. In general, configurations of the following form are ruled out:

[V [... VPE ...]$_S$...]$_{VP}$

BE/DO CONFLICTS

The auxiliary verb contributes various features to the complete verb phrase, including tense, aspect, and polarity. There is no requirement that these features match in antecedent and elliptical VP. However, certain conflicts do not appear to be possible. In general, it is not possible to have a DO-form as the elliptical VP, with an overt BE-form in the antecedent. Consider the following example:

(5) Nor can anyone be certain that Prokofief would have done better, or even as well, under different circumstances. His fellow-countryman, Igor Stravinsky, certainly did not.

In this example, there are two elements on the VP list: "be certain...", and "do better". The target "did not" rules out "be certain" as a possible antecedent, allowing only the reading "Stravinsky did not do better". If the elliptical VP is changed from "did not" to "was not", the situation is reversed; the only possible reading is then "Stravinsky was not certain that Prokofief would have done better...".

A related conflict to be ruled out is that of active/passive conflicts. A passive antecedent is not possible if the VPE is a DO-form. For example:

(6) Jubal did not hear of Digby's disappearance when it was announced, and, when he did, while he had a fleeting suspicion, he dismissed it;

In this example, "was announced" is not a possible antecedent for the VPE "did".

One possible exception to this rule involves progressive antecedents, which, although they contain a BE-form, may be consistent with a DO-form target. The following (constructed) example seems marginally acceptable:

(7) Tom was cleaning his room today. Harry did yesterday.

Thus a BE-form together with a progressive does not conflict with a DO-form.

PREFERENCE LEVELS

If there are several possible antecedents for a given VPE, preferences among those antecedents are determined by looking for other relations between the VPE clause and the clauses containing the possible antecedents.

CLAUSAL RELATIONSHIPS

An antecedent for a given VPE is preferred if there is a configurational relationship between the antecedent clause and the VPE clause. These include comparative structures and adverbial clauses.

Elliptical VP's (VPE) in comparative constructions are of the form

[VP Comparative [NP VPE]]

where Comparatives are expressions such as "as well as", "better than", etc. In constructions of this form there is a strong preference that **VP** is the antecedent for **VPE**. For example:

(8) Now, if Morton's newest product, a corn chip known as Chip-o's, turns out to sell as well as its stock did...

Here, the antecedent of the VPE "did" is the VP "sell".

The next configuration involves VPE's within adverbial clauses. For example,

(9) But if you keep a calendar of events, as we do, you noticed a conflict.

Here the antecedent for the VPE "do" is "keep a calendar of events". In general, in configurations of the form:

[VP ADV [NP VPE]]

VP is preferred over other possible antecedents.

It is important to note that this is a preference rule, rather than an obligatory constraint. Although no examples of this kind were found in the Brown Corpus, violations of this constraint may well be possible. For example:

(10) John can walk faster than Harry can run. Bill can walk faster than Barry can.

If a reading is possible in which the VPE is "Barry can run", this violates the clausal relationship preference rule.

SUBJECT COREFERENCE

Another way in which two clauses are related is subject coreference. An antecedent is preferred if its subject corefers with that of the elliptical VP. An example:

(11) He wondered if the audience would let him finish. They did.

The preferred reading has "they" coreferential with "the audience" and the antecedent for "did" the VP "let him finish".

Subject "coreference" is determined manually, and it is meant to reflect quantifier binding as well as ordinary coreference – that is, standard instances involving coindexing of NP's.

Again, it must be emphasized that the subject coreference rule is a preference rule rather than an obligatory constraint. While no violations were found in the Brown corpus, it is possible to construct such examples.

INTERACTION OF PREFERENCE RULES

There are cases where more than one preference rule applies. The antecedent selected is the item with the highest preference level. If more than one item has the same preference level, the item nearest to the VPE is selected, where nearness is determined by number of words encountered scanning left from the VPE.

In the following example, two preference rules apply:

(12) usually, this is most exasperating to men, who expect every woman to verify their preconceived notions concerning her sex, and when she does not, immediately condemn her as eccentric and unwomanly.

The VPE clause is an adverbial clause modifying the following clause. Thus the VP "condemn

her as eccentric and unwomanly" receives a preference level of 1. The subject "she" of the VPE is coindexed with "every woman". This causes the VP "verify their preconceived notions concerning her sex" to also receive a preference level of 1. Since both of these elements have the same preference level, proximity is determined by scanning left from the VPE. This selects "verify their preconceived notions concerning her sex" as the antecedent.

TESTING THE ALGORITHM

The algorithm has been tested on a set of 304 examples of VP ellipsis collected from the Brown Corpus. These examples were collected using the UNIX grep pattern-matching utility. The version of the Brown Corpus used has each word tagged by part of speech. I defined search patterns for auxiliary verbs that did not have verbs nearby. These patterns did not succeed in locating all the instances of VP ellipsis in the Brown Corpus. However, the 304 examples do cover the full range of types of material in the Brown Corpus, including both "Informative" (e.g., journalistic, scientific, and government texts) and "Imaginative" (e.g., novels, short stories, and humor). I have divided these examples into three categories, based on whether the antecedent is in the same sentence as the VPE, the adjacent (preceding) sentence, or earlier ("Long-Distance"). The definition of sentence is taken from the sentence divisions present in the Brown Corpus.

RESULTS

The algorithm selected the correct antecedent in 285, or 94% of the cases. For comparison purposes, I present results of an alternative strategy; namely, a simple linear scan of preceding text. In this strategy, the first verb that is encountered is taken to be the head of the antecedent VP.

The results of the algorithm and the "Linear Scan" approach are displayed in the following table.

Category		**Algorithm** No. Correct	**Linear Scan** No. Correct
Same-sent	196	193(96%)	172(88%)
Adj-sent	93	85(92%)	72(77%)
Long-Dist	15	7(47%)	2(13%)
Total	304	285(94%)	247(81%)

The algorithm performs considerably better than Linear Scan. Much of the improvement is due to "impossible antecedents" which are selected by the Linear Scan approach because they are closest to the VPE. A frequent case of this is containing antecedents that are ruled out by the algorithm. Another case distinguishing the algorithm from Linear Scan involves coreferential subjects. There were several cases in which the coreferential subject preference rule caused an antecedent to be selected that was not the nearest to the VPE. One example is:

(13) a. But, darn it all, why should we help a couple of spoiled snobs who had looked down their noses at us?
b. But, in the end, we did.

Here, the correct antecedent is the more distant "help a couple of...", rather than "looked down their noses...". There were no cases in which Linear Scan succeeded where the algorithm failed.

SOURCES OF ERROR

I will now look at sources of errors for the algorithm. The performance was worst in the Long Distance category, in which at least one sentence intervenes between antecedent and VPE. In several problem cases in the Long Distance category, it appears that intervening text contains some mechanism that causes the antecedent to remain salient. For example:

(14) a. "...in Underwater Western Eye I'd have a chance to act. I could show what I can do".
b. As far as I was concerned, she had already and had dandily shown what she could do.

In this case, the elliptical VP "had already" means "had already had a chance to act". The algorithm incorrectly selects "show what I can do" as the antecedent. The intervening sentence causes the previous antecedent to remain salient, since it is understood as "(If I had a chance to act then) I could show what I can do." Furthermore, the choice made by the algorithm might perhaps be eliminated on pragmatic grounds, given the oddness of "she had already shown what she could do and had dandily shown what she could do."

Another way in which the algorithm could be generalized is illustrated by the follow example:

(15) a. "I didn't ask you to fight for the ball club", Phil said slowly.
b. "Nobody else did, either".

Here the algorithm incorrectly selects "fight for the ball club" as the antecedent, instead of "ask you to fight for the ball club". The subject coreference rule does not apply, since "Nobody else"

is not coreferential with the subject of any of the possible antecedents. However, its interpretation is dependent on the subject "I" of "ask you to fight for the ball club". Thus, if one generalized the subject coreference rule to include such forms of dependence, the algorithm would succeed on such examples.

Many of the remaining errors involve an antecedent that takes a VP or S as complement, often leading to subtle ambiguities. One example of this is the following:

(16) a. Usually she marked the few who did thank you, you didn't get that kind much in a place like this: and she played a little game with herself, seeing how downright rude she could act to the others, before they'd take offense, threaten to call the manager.

b. Funny how seldom they did: used to it, probably.

Here the algorithm selects "call the manager" as antecedent, instead of "threaten to call the manager", which I determined to be the correct antecedent. It may be that many of these cases involve a genuine ambiguity.

OTHER APPROACHES

The problem addressed here, of determining the antecedent for an elliptical VP, has received little attention in the literature. Most treatments of VP ellipsis (cf. Sag 1976, Williams 1977, Webber 1978, Fiengo and May 1990, Dalrymple, Shieber and Pereira 1991) have focused on the question of determining what readings are possible, given an elliptical VP and a particular antecedent. For a computational system, a method is required to determine the antecedent, after which the possible readings can be determined.

Lappin and McCord (1990) present an algorithm for VP ellipsis which contains a partial treatment of this problem. However, while they define three possible ellipsis-antecedent configurations, they have nothing to say about selecting among alternatives, if there is more than one VP in an allowed configuration. The three configurations given by Lappin and McCord for a VPE-antecedent pair < V,A> are:

1. V is contained in the clausal complement of a subordinate conjunction SC, where the SC-phrase is either (i) an adjunct of A, or (ii) an adjunct of a noun N and N heads an NP argument of A, or N heads the NP argument of an adjunct of A.
2. V is contained in a relative clause that modifies a head noun N, with N contained in A, and, if a verb A′ is contained in A and N is contained in A′, then A′ is an infinitival complement of A or a verb contained in A.
3. V is contained in the right conjunct of a sentential conjunction S, and A is contained in the left conjunct of S.

An examination of the Brown Corpus examples reveals that these configurations are incomplete in important ways. First, there is no configuration that allows a sentence intervening between antecedent and VPE. Thus, none of the Long-Distance examples (about 5% of the sample) would be covered. Configuration (3) deals with antecedent-VPE pairs in adjacent S's. There are many such cases in which there is no sentential conjunction. For example:

(17) a. All the generals who held important commands in World War 2, did not write books.

b. It only seems as if they did.

Perhaps configuration (3) could be interpreted as covering any adjacent S's, whether or not an explicit conjunction is present. [2]

Furthermore, there are cases in which the adjacent categories are something other than S; in the following two examples, the antecedent and VPE are in adjacent VP's.

(18) The experts are thus forced to hypothesize sequences of events that have never occurred, probably never will – but possibly might.

(19) The innocent malfeasant, filled with that supreme sense of honor found in bars, insisted upon replacing the destroyed monacle – and did, over the protests of the former owner – with a square monacle.

In the following example, the adjacent category is S′.

(20) I remember him pointing out of the window and saying that he wished he could live to see another spring but that he wouldn't.

Configurations (1) and (2) deal with antecedent-VPE pairs within the same sentence. In Configuration (1), the VPE is in a subordinate clause, and In (2), the VPE is in a relative clause. In each case, the VPE is c-commanded by the antecedent A. While the configurations cover two

[2] However, a distinction must be maintained between VPE and related phenomena such as gapping and "pseudo-gapping", in which an explicit conjunction is required.

quite common cases, there are other same-sentence configurations in which the antecedent does not c-command the VPE.

(21) In the first place, a good many writers who are said to use folklore, do not, unless one counts an occasional superstition or tale.

(22) In reply to a question of whether they now tax boats, airplanes and other movable property excluding automobiles, nineteen said that they did and twenty that they did not.

In sum, the configurations defined by Lappin and McCord would miss a significant number of cases in the Brown Corpus, and, even where they do apply, there is no method for deciding among alternative possibilities.[3]

CONCLUSIONS

To interpret an elliptical expression it is necessary to determine the antecedent expression, after which a method of reconstructing the antecedent expression at the ellipsis site is required. While the literature on VP ellipsis contains a vast array of proposals concerning the proper method of reconstructing a given antecedent for an elliptical VP, there has been little attention to the question of determining the antecedent.

In this paper, I have proposed a solution to this problem; I have described an algorithm that determines the antecedent for elliptical VP's. It was shown that the algorithm achieves 94% accuracy on 304 examples of VP ellipsis collected from the Brown Corpus. Many of the failure cases appear to be due to the interaction of VPE with other anaphoric phenomena, and others may be cases of genuine ambiguity.

ACKNOWLEDGEMENTS

Thanks to Aravind Joshi and Bonnie Webber. This work was supported by the following grants: ARO DAAL 03-89-C-0031, DARPA N00014-90-J-1863, NSF IRI 90-16592, and Ben Franklin 91S.3078C-1.

REFERENCES

Susan E. Brennan, Marilyn Walker Friedman, and Carl J. Pollard. A Centering Approach to Pronouns, *Proceedings of the 25th Annual Meeting of the ACL*, 1987.

Mary Dalrymple, Stuart Shieber and Fernando Pereira. Ellipsis and Higher-Order Unification. *Linguistics and Philosophy.* Vol. 14, no. 4, August 1991.

Robert Fiengo and Robert May. Ellipsis and Anaphora. Paper presented at GLOW 1990, Cambridge University, Cambridge, England.

Robert Fiengo and Robert May. ndices and Identity. ms. 1991.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. Providing a Unified Account of Definite Noun Phrases in Discourse. In *Proceedings, 21st Annual Meeting of the ACL*, pp. 44-50, Cambridge, MA, 1983.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. Towards a Computational Theory of Discourse Interpretation. ms. 1986.

Isabelle Haik. Bound VP's That Need To Be. *Linguistics and Philosophy* 11: 503-530. 1987.

Jerry Hobbs. Resolving Pronoun References, *Lingua* 44, pp. 311-338. 1978.

Shalom Lappin and Michael McCord. Anaphora Resolution in Slot Grammar, in *Computational Linguistics*, vol 16, no 4. 1990.

Robert May. *Logical Form: Its Structure and Derivation*, MIT Press, Cambridge Mass. 1985.

Ivan A. Sag. *Deletion and Logical Form.* Ph.D. thesis, MIT. 1976.

Marilyn Walker. Evaluating discourse processing algorithms. In *Proceedings, 27th Annual Meeting of the ACL*, Vancouver, Canada. 1989.

Bonnie Lynn Webber. *A Formal Approach to Discourse Anaphora.* Ph.D. thesis, Harvard University. 1978.

Edwin Williams. *Discourse and Logical Form. Linguistic Inquiry*, 8(1):101-139. 1977.

[3]While the problem of antecedent determination for VP ellipsis has been largely neglected, the analogous problem for pronoun resolution has been addressed (cf. Hobbs 1978, Grosz, Joshi, and Weinstein 1983 and 1986, and Brennan, Friedman and Pollard 1987), and two leading proposals have been subjected to empirical testing (Walker 1989).