CORRECTING ILLEGAL NP OMISSIONS USING LOCAL FOCUS

Linda Z. Suri[1]
Department of Computer and Information Sciences
University of Delaware
Newark DE 19716
Internet: suri@udel.edu

1 INTRODUCTION

The work described here is in the context of developing a system that will correct the written English of native users of American Sign Language (ASL) who are learning English as a second language. In this paper we focus on one error class that we have found to be particularly prevalent: the illegal omission of NP's.

Our previous analysis of the written English of ASL natives has led us to conclude that language transfer (LT) can explain many errors, and should thus be taken advantage of by an instructional system (Suri, 1991; Suri and McCoy, 1991). We believe that many of the omission errors we have found are among the errors explainable by LT.

Lillo-Martin (1991) investigates null argument structures in ASL. She identifies two classes of ASL verbs that allow different types of null argument structures. *Plain verbs* do not carry morphological markings for subject or object agreement and yet allow null argument structures in some contexts. These structures, she claims, are analogous to the null argument structures found in languages (like Chinese) that allow a null argument if the argument co-specifies the topic of a previous sentence (Huang, 1984). Such languages are said to be discourse-oriented languages.

As it turns out, our writing samples collected from deaf writers contain many instances of omitted NP's where those NP's are the topic of a previous sentence and where the verb involved would be a plain verb in ASL. We believe these errors can be explained as a result of the ASL native carrying over conventions of (discourse-oriented) ASL to (sentence-oriented) English.

If this is the case, then these omissions can be corrected if we track the topic, or, in computational linguistics terms, *the local focus*, and the *actor focus*.[2] We propose to do this by developing a modified version of Sidner's focus tracking algorithm (1979, 1983) that includes mechanisms for handling complex sentence types and illegally omitted NP's.

[1]This research was supported in part by NSF Grant #IRI-9010112. Support was also provided by the Nemours Foundation. We thank Gallaudet University, the National Technical Institute for the Deaf, the Pennsylvania School for the Deaf, the Margaret S. Sterck School, and the Bicultural Center for providing us with writing samples.

[2]Grosz, Joshi and Weinstein (1983) use the notion of centering to track something similar to local focus and argue against the use of a separate actor focus. However, we think that the example they use does not argue against a separate actor focus, but illustrates the need for extensions to Sidner's algorithm to specify how complex sentences should be processed.

2 FOCUS TRACKING

Our focusing algorithm is based on Sidner's focusing algorithm for tracking local and actor foci (Sidner 1979; Sidner 1983).[3] In each sentence, the actor focus (AF) is identified with the (thematic) agent of the sentence. The Potential Actor Focus List (PAFL) contains all NP's that specify an animate element of the database but are not the agent of the sentence.

Tracking local focus is more complex. The first sentence in a text can be said to be about something. That something is called the current focus (CF) of the sentence and can generally be identified via syntactic means, taking into consideration the thematic roles of the elements in the sentence. In addition to the CF, an initial sentence introduces a number of other items (any of which can become the focus of the next sentence). Thus, these items are recorded in a potential focus list (PFL).

At any given point in a well-formed text, after the first sentence, the writer has a number of options:

- Continue talking about the same thing; in this case, the CF doesn't change.
- Talk about something just introduced; in this case, the CF is selected from the previous sentence's PFL.
- Return to a topic of previous discussion; in this case, that topic must have been the CF of a previous sentence.
- Discuss an item previously introduced, but which was not the topic of previous discussion; in this case, that item must have been on the PFL of a previous sentence.

The decision (by the reader/hearer/algorithm) as to which of these alternatives was chosen by the speaker is based on the thematic roles (with particular attention to the agent role) held by the anaphora of the current sentence, and whether their co-specification is the CF, a previous CF, or a member of the current PFL or a previous PFL. Confirmation of co-specifications requires inferencing based on general knowledge and semantics.

At each sentence in the discourse, the CF and PFL of the previous sentence are stacked for the possibility of subsequent return.[4] When one of these items is returned to, the stacked CF's and PFL's above it are popped, and are thus no longer available for return.

[3]Carter (1987) extended Sidner's work to handle intrasentential anaphora, but for space reasons we do not discuss these extensions.

[4]Sidner did not stack PFL's. Our reasons for stacking PFL's are discussed in section 4.

2.1 FILLING IN A MISSING NP

We propose extending this algorithm to identify an illegally omitted NP. To do this, we treat the omitted NP as an anaphor which, like Sidner's treatment of full definite NP's and personal pronouns, co-specifies an element recorded by the focusing algorithm. This approach is based on the belief that an omitted NP is likely to be the topic of a previous sentence. We define preferences among the focus data structures which are similar to Sidner's preferences.

More specifically, when we encounter an omitted NP that is not the agent, we first try to fill the deleted NP with the CF of the immediately preceding sentence. If syntax, semantics or inferencing based on general knowledge cause this co-specification to be rejected, we then consider members of the PFL of the previous sentence as fillers for the deleted NP. If these too are rejected, we consider stacked CF's and elements of stacked PFL's, taking into account preferences (yet to be determined) among these elements.

When we encounter an omitted agent NP, in a simple sentence or a sentence-initial clause, we first test the AF of the previous sentence as co-specifier, then members of the PAFL, the previous CF, and finally stacked AF's, CF's and PAFL's. To identify a missing agent NP in a non-sentence-initial clause, our algorithm will first test the AF of the previous clause, and then follow the same preferences just given. Further preferences are yet to be determined, including those between the stacked AF, stacked PAFL, and stacked CF.

2.2 COMPUTING THE CF

To compute the CF of a sentence without any illegally omitted NP's, we prefer the CF of the last sentence over members of the PFL, and PFL members over members of the focus stacks. Exceptions to these preferences involve picking a non-agent anaphor co-specifying a PFL member over an agent co-specifying the CF, and preferring a PFL member co-specified by a pronoun to the CF co-specified by a full definite description.

To compute the CF of a sentence with an illegally omitted NP, our algorithm treats illegally omitted NP's as anaphora since they (implicitly) co-specify something in the preceding discourse. However, it is important to remember that discourse-oriented languages allow deletions of NP's *that are the topic of the discourse*. Thus, we prefer a deleted non-agent as the focus, as long as it closely ties to the previous sentence. Therefore, we prefer the co-specifier of the omitted non-agent NP as the (new) CF if it co-specifies either the last CF or a member of the last PFL. If the omitted NP is the thematic agent, we prefer for the new CF to be a pronominal (or, as a second choice, full definite description) non-agent anaphor co-specifying either the last CF or a member of the last PFL (allowing the deleted agent NP to be the AF and keeping the AF and CF different).[5] If no anaphor meets these criteria, then the members of the CF and PFL focus stacks will be considered, testing a co-specifier of the omitted NP before co-specifiers of pronouns and definite descriptions at each stack level.

[5] As future work, we will explore how to resolve more than one non-agent anaphor in a sentence co-specifying PFL elements.

3 EXAMPLE

Below, we describe the behavior of the extended algorithm on an example from our collected texts containing both a deleted non-agent and agent.

Example: "*(S1) First, in summer I live at home with my parents. (S2) I can budget money easily. (S3) I did not spend lot of money at home because at home we have lot of good foods, I ate lot of foods. (S4) While living at college I spend lot of money because _ go out to eat almost everyday. (S5) At home, sometimes my parents gave me some money right away when I need _.* "

After S1, the AF is I, the CF is I, and the PFL contains SUMMER, HOME, and the LIVE VP. For S2, I is the only anaphor, so it becomes the CF, the PFL contains MONEY and the BUDGET VP, and the focus stack contains I and the previous PFL.

S3 is a complex sentence using the conjunction "because." Such sentences are not explicitly handled by Sidner's algorithm. Our analysis so far suggests that we should not split this sentence into two[6], and should prefer elements of the main clause as focus candidates. Thus, we take the CF from the first clause, and rank other elements in that clause before elements in the second clause on the PFL.[7] In this case, we have several anaphora: I, money, at home.... The AF remains I. The CF becomes MONEY since it co-specifies a member of the PFL and since the co-specifier of the last CF is the agent. Ordering the elements of the first clause before the elements in the second results in the PFL containing HOME, the NOT SPEND VP, GOOD FOOD, and the HAVE VP. We stack the CF and the PFL of S2.

Note that S4 has a missing agent in the second clause. To identify the missing agent in a non-sentence-initial clause, our algorithm will first test the AF of the preceding clause for possible co-specification. Because this co-specification would cause no contradiction, the omitted NP is filled with "I", which is eventually taken as the AF of S4. The CF is computed by first considering the first clause of S4, since the X clause is the preferred clause of an X BECAUSE Y construct. Since "money" co-specifies the CF of S3, and nothing else in the preferred clause co-specifies a member of the PFL, MONEY remains the CF. The PFL contains COLLEGE, the SPEND VP, EVERY DAY, the TO EAT VP, and the GO OUT TO EAT VP. We stack the CF and PFL of S3.

S5 contains a subordinate clause with a missing non-agent. Our algorithm first considers the

[6] If we were to split the sentence up, then the focus would shift away from MONEY when we process the second clause (which contradicts our intuition of what the focus is in this paragraph).

[7] The appropriateness of placing elements from both clauses in one PFL and ranking them according to clause membership will be further investigated. This construct ("X BECAUSE Y") is further discussed in section 4.

CF, MONEY, as the co-specifier of the omitted NP; syntax, semantics and general knowledge inferencing do not prevent this co-specification, so it is adopted. MONEY is also chosen as the CF since it is the co-specifier of the omitted NP occurring in the verb complement clause which is the preferred clause in this type of construct.

4 DISCUSSION OF EXTENSIONS

One of the major extensions needed in Sidner's algorithm is a mechanism for handling complex sentences. Based on a limited analysis of sample texts, we propose computing the CF and PFL of a complex sentence based on a classification of sentence types. For instance, for a sentence of the form "X BECAUSE Y" or "BECAUSE Y, X", we prefer the expected focus of the effect clause as CF, and order elements of the X clause on the PFL before elements of the Y clause. Analogous PFL orderings apply to other sentence types described here. For a sentence of the form "X CONJ Y", where X and Y are sentences, and CONJ is "and", "or", or "but", we prefer the expected focus of the Y clause. For a sentence of the form "IF X (THEN) Y", we prefer the expected focus of the THEN clause, while for "X, IF Y", we prefer the expected focus of the X clause. Further study is needed to determine other preferences and actions (including how to further order elements on the PFL) for these and other sentence types. These preferences will likely depend on thematic roles and syntactic criteria (e.g., whether an element occurs in the clause containing the expected CF).

The decisions about how these and other extensions should proceed have been or will be based on analysis of both standard written English and the written English of deaf students. The algorithm will be developed to match the intuitions of native English speakers as to how focus shifts.

A second difference between our algorithm and Sidner's is that we stack the PFL's as well as the CF's. We think that stacking the PFL's may be needed for processing standard English (and not just for our purposes) since focus sometimes revolves around the theme of one of the clauses of a complex sentence, and later returns to revolve around items of another clause. Further investigation may indicate that we need to add new data structures or enhance existing ones to handle focus shifts related to these and other complex discourse patterns.

We should note that while we prefer the CF as the co-specifier of an omitted NP, Sidner's recency rule suggests that perhaps we should prefer a member of the PFL if it is the last constituent of the previous sentence (since a null argument seems similar to pronominal reference). However, our studies show that a rule analogous to the recency rule does not seem to be needed for resolving the co-specifier of an omitted NP. In addition, Carter (1987) feels the recency rule leads to unreliable predictions for co-specifiers of pronouns. Thus, we do not expect to change our algorithm to reflect the recency rule. (We also believe we will abandon the recency rule for resolving pronouns.)

Another task is to specify focus preferences among stacked PFL's and stacked CF's, perhaps using thematic and syntactic information.

An important question raised by our analysis is how to handle a paragraph-initial, but not discourse-initial, sentence. Do we want to treat it as discourse-initial, or as any other non-discourse-initial sentence? We suggest (based on analysis of samples) that we should treat the sentence as any non-discourse-initial sentence, unless its sentence type matches one of a set of sentence types (which often mark focus movement from one element to a new one). In this latter case, we will treat the sentence as discourse-initial by calculating the CF and PFL in the same manner as a discourse-initial sentence, but we will retain the focus stacks. We have identified a number of sentence types that should be included in the set of types which trigger the latter treatment; we will explore whether other sentence types should be included in this set.

5 CONCLUSIONS

We have discussed proposed extensions to Sidner's algorithm to track local focus in the presence of illegally omitted NP's, and to use the extended focusing algorithm to identify the intended co-specifiers of omitted NP's. This strategy is reasonable since LT may lead a native signer of ASL to use discourse-oriented strategies that allow the omission of an NP that is the topic of a preceding sentence when writing English.

REFERENCES

David Carter (1987). *Interpreting Anaphors in Natural Language Texts.* John Wiley and Sons, New York.

Barbara J. Grosz, Aravind K. Joshi and Scott Weinstein (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, 44–50.

C.-T. James Huang (1984). On the distribution and reference of empty pronouns. *Linguistic Inquiry*, 15(4):531–574.

Diane C. Lillo-Martin (1991). *Universal Grammar and American Sign Language.* Kluwer Academic Publishers, Boston.

Candace L. Sidner (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse.* Ph.D. thesis, M.I.T., Cambridge, MA.

Candace L. Sidner (1983). Focusing in the comprehension of definite anaphora. In Robert C. Berwick and Michael Brady, eds., *Computational Models of Discourse*, chapter 5, 267–330. M.I.T. Press, Cambridge, MA.

Linda Z. Suri and Kathleen F. McCoy (1991). Language transfer in deaf writing: A correction methodology for an instructional system. TR-91-20, Dept. of CIS, University of Delaware.

Linda Z. Suri (1991). Language transfer: A foundation for correcting the written English of ASL signers. TR-91-19, Dept. of CIS, University of Delaware.