A Methodology for Extending Focusing Frameworks

Linda Z. Suri and Kathleen F. McCoy
University of Delaware
Newark, DE 19716

1 Introduction

The central problem addressed in this work is how to develop and assess local focus tracking algorithms and algorithms for proposing referents of pronouns for use in Natural Language Processing (NLP) systems. By *local focus*, we refer to the person, object, property or concept that a sentence is most centrally about within the discourse context in which it occurs. The appropriate movement and marking of local focus, and the appropriate choice of the form of a Noun Phrase (NP) based on local focus information, are considered to contribute to the local coherence exhibited by discourse ([Sid79], [GJW83], [Car87], and others).

In addition, local focus information is one source of information that is used by readers and hearers for interpreting pronouns. In fact, local focus tracking and pronoun resolution are mutually dependent processes. The local focus information influences pronoun resolution, and pronoun resolution, in turn, influences updating focus information. Therefore, the tracking of local focus is crucial for the interpretation of pronouns.

In this work we explore a robust mechanism for accessing and extending local focusing algorithms. After identifying flaws with some alternative approaches, we introduce a two part methodology which we term the Semantically-Slanted Discourse (SSD) Methodology. The first part of the methodology consists of an exploratory phase in which possible extensions to a focusing algorithm are discovered through the use of carefully constructed discourses which rely on the potential tension between focusing and world knowledge factors in pronoun resolution. We argue that this phase can be followed by a corpus analysis to confirm its findings.

In this paper we briefly introduce the notion of local focusing and what a local focusing algorithm is intended to capture. We note that other work on focusing has not taken the processing of complex (i.e., multi-clausal) sentences into account and describe a number of issues involved in their processing. We motivate and describe our methodology in terms of extending a local focusing algorithm to handle complex sentences. We review the methodology used by other researchers to develop and extend their focusing frameworks, and we identify some difficulties with that methodology. We examine the possibility of using a corpus analysis, and briefly describe potential problems. Finally, we describe our Semantically-Slanted Discourse Methodology, and point out how our methodology overcomes the difficulties of previous approaches.

2 What is local focusing?

We use the term *local focusing framework* to refer to a theory or framework consisting of a (set of) focus tracking algorithm(s) and a (set of) pronoun resolution algorithm(s). A local focusing framework records and makes use of information about focusing factors and indicates how these factors influence pronoun resolution. A local focusing framework is not intended to *independently* interpret pronouns. Rather, a local focusing framework is intended to suggest co-specifications for pronouns in a reasonable order. An inferencing mechanism that makes use of semantic factors (such as semantic case constraints, world knowledge, rhetorical relations, etc.) must be used to confirm or reject a suggested co-specification. Thus, local focusing frameworks are intended to capture a coherence factor in discourse which influences preferences for how to resolve pronouns *independent* of semantic factors.

3 Processing Complex Sentences: A Reason for Extending Focusing Algorithms

Although complex sentences are prevalent in written English, other local focusing research (Focusing: [Sid79], [Car87]; centering: [GJW83], [BFP87], [Wal89], [Kam86], [WIC92]), [Bre93], [Kam93], [KPP93], [Lin93], [HT93], [Wal93]; and PUNDIT: [Dah86], [PDS+86], [DB90]), did not explicitly and/or adequately address how to process complex sentences. Thus, there is a need to extend focusing algorithms.

Notice that there are a number of ways that a given type of complex sentence might be handled. For instance, consider processing a complex sentence of the form "SX because SY," where SX and SY each consist of a single clause. One might imagine processing the SX clause and then the SY clause (i.e., resolving the pronouns in these clauses and updating the focusing data structures) as if the clauses were a sequence of simple (single-clause) sentences. On the other hand, it may be the case that for this type of complex sentence, the sentence should be treated as a single unit of processing with elements of one of the clauses dominating the processing. (For further discussion of these and other possible processing possibilities, see [Sur93].)

The question we address is how one can appropriately extend a focusing mechanism to handle various kinds of complex sentences.[1]

4 A Methodology Used in Other Local Focusing Work

Recall that local focusing theories are attempting to capture patterns of focus movement and patterns of relations between anaphors and their antecedents that are independent of semantics, world knowledge, rhetorical relations, etc. Because of this, the method for determining how to process particular kinds of complex sentences that might seem the most natural is to construct semantically-neutral discourses[2] that involve the type of complex sentence under study, and gather linguistic judgments to determine how people prefer to resolve the pronouns. In fact, in exploring *other aspects* of local focusing frameworks, other literature appears to have tried to make use of semantically-neutral texts in this fashion (e.g., [BFP87], [WIC92]).

However, in trying to construct discourses to determine how to process a particular kind of complex sentence, we realized it is difficult to construct discourses that are truly semantically-neutral *and* sound natural. This task is further complicated by the need to construct a number of semantically-neutral texts in order to control for and isolate each of the factors that might affect how readers prefer to resolve pronouns.[3] These factors include the

[1]Note, it is possible that various types of complex sentence would each need to be handled differently.

[2]I.e., discourses whose pronouns cannot be unambiguously resolved on the basis of semantic/world knowledge factors alone.

[3]This need was not addressed by previous focusing work in an adequate or systematic

influence of the other complex sentence structures in the discourse, and the factors that affect focus computation and pronoun resolution for simple sentences, i.e., focus history, the syntactic roles of pronouns and their potential antecedents, verb aspect and tense, etc.

More importantly, when one is constructing texts *without the benefit of a systematic methodology*, one cannot be sure that the collection of constructed texts are representative of naturally-occurring text in terms of the interactive relationships within and across focusing and semantic factors, and their influence on pronoun interpretation. As a result, there is a danger of tuning a theory to handle a discourse phenomena that is the exception rather then the norm in naturally-occurring situations. (Section 6.1 describes how our methodology avoids this pitfall.)

5 Using a Corpus Analysis to Extend Frameworks

Because of the problems associated with using constructed discourses, it is natural to turn to some kind of corpus analysis to extend a focusing framework. For example, one might measure how well an extension of a framework handles a type of complex sentence by measuring how accurately and efficiently it suggests referents for pronouns in texts which contain the type of complex sentence under consideration. One could count how often the extended framework suggests a wrong referent (which would not be rejected by an ideal inferencing mechanism), and how many referents it suggests (on average) to the inferencing component before the correct referent is selected.

In using this type of approach, one is faced with several potential difficulties. First, a focusing framework is intended to capture the reader's preferences for focus movement and pronoun resolution *independent* of world knowledge, semantics, and other pragmatic factors. Very large amounts of text would have to be analyzed to control the influence of these factors, yet, since there are few tools available for this type of analysis, this task would be formidable.

Second, a corpus analysis may be useful in comparing various extensions, but it is up to the designer to decide *which* extensions to compare. Using

fashion.

this approach, a novel extension cannot emerge by becoming evident as a side effect of the analysis. Generally speaking, before a corpus analysis can be used, the researcher must have made all decisions concerning the processing. That is, all possible extensions of the framework (which are to be tested in the corpus analysis) must be completely identified. Notice, however, that there is no guarantee that the *correct* extension will be specified and tested. One might overlook the appropriate answers to how to segment sentences and how to process a particular kind of complex sentence. In addition, the number of possible extensions is likely quite large, and thus the number of alternative corpora analyses will likely be quite prohibitive.

Perhaps the most significant and problematic obstacle for determining how to extend a focusing framework to handle a particular kind of complex sentence via a corpora analysis is the following: if one does not know how to process many types of complex sentence, it is difficult to perform a corpora analysis to determine how to process a given type of complex sentence, since instances of that type of complex sentence are likely to be preceded and followed by other types of complex sentences. Furthermore, many sentences in the corpus are likely to involve multiple levels of complexity. (This observation points to the fact that one would also need to isolate the influence of one complex sentence structure from the effect of another complex sentence structure.)

6 Our Two-Part Methodology for Determining How to Process Complex Sentences

As we have pointed out, there are several potential problems with analyses using constructed discourses and with corpora analyses. Our methodology combines specific instances of both of these methodologies; these specific instances were designed to overcome the difficulties we identified.

The first part of our methodology involves systematically constructing discourse (of a type to be described) and gathering acceptability judgments on these discourses. The discourses are constructed in such a way so as to help identify a plausible extension of a focusing algorithm which would handle the type of complex sentence in question. The resulting extension must then be confirmed by a corpus analysis to ensure that the constructed discourses uncovered influences actually found in naturally-occurring text.

In the remainder of this paper, we focus on the constructed discourse portion of our methodology. We refer the reader to [Sur93] for a full description of how the methodology was used to extend a particular focusing framework to handle sentences of the form "SX because SY," although we summarize our technique and findings here.

6.1 Semantically-Slanted Discourse (SSD) Methodology

In previous literature on local focusing (e.g., [Sid79], [GJW83], [BFP87]), researchers used a small number of constructed texts to justify aspects of their focusing framework and to assess and compare focusing frameworks. However, they did not explicitly address how one should construct sets of texts in order to draw accurate conclusions about local focusing. The first part of our methodology is intended to help the researcher construct sets of texts (i.e., minimal pairs or minimal quadruples) that allow components of a focusing framework to be systematically isolated and thus allow one to appropriately assess focusing frameworks.

To appreciate the reasoning behind the first part of our methodology, or what we call our *Semantically-Slanted Discourse (SSD) Methodology*, recall that local focusing frameworks (including centering) are intended to capture the preferences for pronoun resolution *independent of* semantics, world knowledge, rhetorical relations and other kinds of pragmatics. Thus, they are intended to capture how one would resolve pronouns and update focusing information in discourse that is neutral in terms of these factors. Presumably in such texts, *only* focusing factors would affect pronoun resolution. (Recall that focusing factors in pronoun resolution include focus history, previous pronominalization, and grammatical roles.) In a semantically-non-neutral discourse, semantic factors (semantics, world knowledge, etc.) can override the preferences of the focusing framework by rejecting potential referents proposed by a focusing framework.

Taking this into account, in order to determine how best to process a particular complex sentence, we decided to construct discourses that are intentionally loaded or slanted for pronoun interpretation based on world knowledge, other pragmatic factors, and semantics.[4] We contend that in a semantically-slanted discourse, if the text seems ambiguous or awkward, or if one needs to re-interpret a pronoun, then the focusing preferences for pronoun resolution are at odds with the preferences based on semantics, other pragmatic factors, or world knowledge. On the other hand, if the text seems acceptable/natural, then we contend the preferences for pronoun resolution based on focusing agree with preferences based on semantic-slanting. Thus, gathering acceptability judgments about semantically-slanted discourses should help us identify what the focusing preferences are, and thus how a focusing framework should be extended to handle a given type of complex sentence. This is the idea at the heart of our methodology.

6.2 Using Semantically-Slanted Discourses

In using semantically-slanted discourses to uncover an extension of a focusing algorithm, discourses must be constructed which:

1. isolate the complexity under study
2. (when taken together) determine an extension of the focusing algorithm by determining how focusing factors interact with each other in the face of the complexity under study.

The first of these issues influences the overall form of the discourses being constructed. The second requires systematically constructing a number of discourses which vary with respect to the various focusing factors in such a way as to isolate their potential influence.

In order to isolate the complexity under study, we construct discourses of the following form, for which the interpretation of the NPs in S3 is fully determined by the semantics of the text and world knowledge:

Example 1

S1) Simple-sentence
S2) Sentence with one level of complexity (i.e., having two clauses), introduced by the syntactic form of interest.
S3) Simple-sentence

[4] Henceforth, *semantically-slanted* discourse.

In examining linguistic judgments about such texts, our goal is to identify preferences imposed by the syntactic form of S2 for:

- resolving pronouns in S2
- updating the focusing data structures after S2 so that the pronouns of S3 can be correctly resolved in a manner that is consistent with resolving pronouns in a sentence following a simple sentence or another kind of complex sentence.

The motivation for having S1 be a simple sentence is to avoid any effect a complex sentence might have on the focusing data structures going into S2. Similarly, the motivation for having S3 be a simple sentence is to avoid any effect that a complex sentence structure in S3 might have on pronoun resolution in S3.

7 Investigating "SX because SY" Sentences"

We illustrate the second issue through an example. To see the types of discourses our methodology calls for constructing, let us consider what would be needed to extend a particular focusing framework, RAFT/RAPR (described in [Sur93]), to handle resolving subject pronouns in sentences of the form "SX because SY" where SX and SY are full sentences.

To address this question, we examined discourses that are "variations" of the form shown in Example 2.

Example 2

S1) Dodge was nearly robbed by an ex-convict the other night.
S2) [Dodge] captured [the ex-con] because [the ex-con] was so stupid and clumsy.
S3) Then [Dodge] called the police.

We needed to construct variations of this text in order to tease out how the various focusing factors interact.

The RAFT/RAPR algorithm prefers to resolve a subject pronoun (in a simple sentence) so that subject co-refers with the subject of the previous sentence if the previous sentence is a simple sentence (if this suggested referent is rejected by world knowledge inferencing, then other elements in the previous discourse are tried in specified order; this ordering is influenced by such things as pronouns used in the previous sentence). Some questions that must be answered in coming up with an extension of RAFT/RAPR in handling these "SX because SY" sentences are:

1. How should Subject(SY) be resolved? I.e., should the algorithm prefer that it co-refer with Subject(S1) or Subject(SX)?
2. How should Subject(S3) be resolved? I.e.,
   - Preferring Subject(SX) always?
   - Preferring Subject(SY) always?
   - Preferring Subject(SX) or Subject(SY) depending on which is pronominalized?
   - Preferring Subject(SX) or Subject(SY) depending on which is co-referential with Subject(S1)?

The answers to these (and all such similar questions) constitute a decision on how and whether the complex sentence should be segmented, and how to weigh the influences of the various focusing factors such as pronominalization and focus history.

In order to answer these questions, we must make up a number of texts which vary with respect to these factors. In particular, the text variations correspond to variations of the following parameters:

1. Whether Subject(S1) is the ex-convict or Dodge. ("An ex-convict nearly robbed Dodge the other night" vs. "Dodge was nearly robbed by an ex-convict the other night.")
2. Whether Subject(SX) of S2 is the ex-convict or Dodge. ("[Dodge] captured [the ex-convict] because [the ex-convict] was so stupid and clumsy" vs. "[The ex-convict] woke [Dodge] up because [the ex-convict] was so stupid and clumsy.")
3. Whether Subject(SY) of S2 is the ex-con or Dodge. ("[The ex-convict] tied [Dodge] up because [the ex-convict] didn't want any trouble" vs. "[The ex-convict] tied [Dodge] up because [Dodge] wasn't co-operating.")

4. Whether Subject(S3) is the ex-con or Dodge.[5] ("Then [the ex-convict] was arrested by the police" or "Then [the ex-convict] took all the money and ran" vs. "Then [Dodge] called the police" or "Then [Dodge] started screaming for help.")

5. Whether Subject(SX) was pronominalized.

6. Whether DirectObject(SX) was pronominalized.

7. Whether Subject(SY) was pronominalized.

By generating texts for all combinations of different values of these parameters, we are able to control for the influence of each focusing factor.

The result of this procedure is a set of texts which can be presented to native speakers for judgments. On the basis of the resulting judgments, a possible extension of the focusing algorithm can be identified. For example, on the basis of the study reported in [Sur93], Suri reports that for the specific case of "SX because SY" sentences, gathered judgments indicate that native speakers of English prefer to resolve subjects in a subsequent (S3) sentence with the Subject(SX), and they prefer to resolve Subject(SY) with the Subject(SX) [Sur93], [SM93]. [Sur93] explains how we extended the RAFT/RAPR framework to process "SX because SY" sentences in light of these findings, and the difficulties encountered when trying to extend centering to account for these findings.

8 Conclusions

The notion of local focusing and its influence on pronoun resolution has been found useful in many aspects of NLP. However, previous work on local focusing has ignored complex sentences even though they are prevalent in naturally-occurring text. The problem that we faced was one of determining a reasonable way to extend a focusing algorithm to handle these sentences. Previous methodology (i.e., using semantically neutral text) was too simplistic and nearly impossible to utilize. A solely corpus-based analysis is impossible because of the variety of *a priori* decisions that needed to be made and

[5] Recall that when Subject(S3) is pronominalized, the referent of Subject(S3) is determined by the semantic-slanting of the text.

because of the complexity of interaction among factors in naturally-occurring discourses. This work presents a methodology that calls for the systematic construction of texts. It relies on the potential tension of semantic factors with focusing factors to identity possible extensions of a focusing framework to account for a particular kind of complex sentence. The methodology has been used to extend a focusing framework to handle one type of complex sentence.

Furthermore, as explained in detail in [Sur93], this methodology can also be used to *compare* local focusing frameworks. Thus, this methodology allows one to study focusing phenomena and algorithms related to focusing phenomena.

References

[BFP87] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, 1987.

[Bre93] Susan E. Brennan. Reference and local structure in interactive discourse. In the proceedings of the Workshop on Centering Theory in Naturally-Occurring Discourse, Institute for Research in Cognitive Science, University of Pennsylvania, May, 1993.

[Car87] David Carter. *Interpreting Anaphors in Natural Language Texts*. John Wiley and Sons, New York, 1987.

[Dah86] Deborah A. Dahl. Focusing and reference resolution in PUNDIT. In *Proceedings of the 1986 National Conference on Artificial Intelligence*, pages 1083–1088, Philadelphia, PA, August 1986.

[DB90] Deborah Dahl and Catherine N. Ball. Reference resolution in PUNDIT. Technical Report CAIT-SLS-9004, UNISYS, Paoli, PA, March 1990.

[GJW83] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Cambridge, Mass, June 1983.

[HT93] Beryl Hoffman and Ümit Deniz Turan. Word order and centering in Turkish. In the proceedings of the Workshop on Centering Theory in Naturally-Occurring Discourse, Institute for Research in Cognitive Science, University of Pennsylvania, May, 1993.

[Kam86] M. Kameyama. A property-sharing constraint in centering. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 200–206, Cambridge, Mass, 1986.

[Kam93] Megumi Kameyama. Intrasentential centering. In the proceedings of the Workshop on Centering Theory in Naturally-Occurring Discourse, Institute for Research in Cognitive Science, University of Pennsylvania, May, 1993.

[KPP93] Megumi Kameyama, Rebecca Passonneau, and Massimo Poesio. Temporal centering. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 70–77, 1993.

[Lin93] Brian Linson. *A Distributional Analysis of Discourse Transitions.* In the proceedings of the Workshop on Centering Theory in Naturally-Occurring Discourse, Institute for Research in Cognitive Science, University of Pennsylvania, May, 1993.

[PDS+86] Martha S. Palmer, Deborah A. Dahl, Rebecca J. Schiffman, Lynette Hirschman, Marcia Linebarger, and John Dowding. Recovering implicit information. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 10–19, June 1986.

[Sid79] Candace L. Sidner. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. PhD thesis, MIT, June 1979.

[SM93] Linda Z. Suri and Kathleen F. McCoy. *Focusing and Pronoun Resolution in Particular Kinds of Complex Sentences.* In the proceedings of the Workshop on Centering Theory in Naturally-Occurring Discourse, Institute for Research in Cognitive Science, University of Pennsylvania, May, 1993.

[Sur93] Linda Z. Suri. *Extending Focusing Frameworks to Process Complex Sentences and to Correct the Written English of Proficient Signers of American Sign Language.* PhD thesis, University of Delaware, 1993. Available as Dept. of CIS Technicial Report TR-94-21.

[Wal89] Marilyn A. Walker. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 251–261, 1989.

[Wal93] Marilyn Walker. *Initial Contexts and Shifting Centers.* In the proceedings of the Workshop on Centering Theory in Naturally-

Occurring Discourse, Institute for Research in Cognitive Science, University of Pennsylvania, May, 1993.

[WIC92] M. Walker, M. Iida, and S. Cote. Japanese discourse and the process of centering. Technical Report 92-14, The Institute for Research in Cognitive Science, The University of Pennsylvania, 1992.