Analysis of Syntax-Based Pronoun Resolution Methods

Joel R. Tetreault
University of Rochester
Department of Computer Science
Rochester, NY, 14627
tetreaul@cs.rochester.edu

Abstract

This paper presents a pronoun resolution algorithm that adheres to the constraints and rules of Centering Theory (Grosz et al., 1995) and is an alternative to Brennan et al.'s 1987 algorithm. The advantages of this new model, the Left-Right Centering Algorithm (LRC), lie in its incremental processing of utterances and in its low computational overhead. The algorithm is compared with three other pronoun resolution methods: Hobbs' syntax-based algorithm, Strube's S-list approach, and the BFP Centering algorithm. All four methods were implemented in a system and tested on an annotated subset of the Treebank corpus consisting of 2026 pronouns. The noteworthy results were that Hobbs and LRC performed the best.

1 Introduction

The aim of this project is to develop a pronoun resolution algorithm which performs better than the Brennan et al. 1987 algorithm[1] as a cognitive model while also performing well empirically.

A revised algorithm (Left-Right Centering) was motivated by the fact that the BFP algorithm did not allow for incremental processing of an utterance and hence of its pronouns, and also by the fact that it occasionally imposes a high computational load, detracting from its psycholinguistic plausibility. A second motivation for the project is to remedy the dearth of empirical results on pronoun resolution methods. Many small comparisons of methods have been made, such as by Strube (1998) and Walker (1989), but those usually consist of statistics based on a small hand-tested corpus. The problem with evaluating algorithms by hand is that it is time consuming and difficult to process corpora that are large enough to provide reliable, broadly based statistics. By creating a system that can run algorithms, one can easily and quickly analyze large amounts of data and generate more reliable results. In this project, the new algorithm is tested against three leading syntax-based pronoun resolution methods: Hobbs' naive algorithm (1977), S-list (Strube 1998), and BFP.

Section 2 presents the motivation and algorithm for Left-Right Centering. In Section 3, the results of the algorithms are presented and then discussed in Section 4.

2 Left-Right Centering Algorithm

Left-Right Centering (LRC) is a formalized algorithm built upon centering theory's constraints and rules as detailed in Grosz et. al (1995). The creation of the LRC Algorithm is motivated by two drawbacks found in the BFP method. The first is BFP's limitation as a cognitive model since it makes no provision for incremental resolution of pronouns (Kehler 1997). Psycholinguistic research support the claim that listeners process utterances one word at a time, so when they hear a pronoun they will try to resolve it immediately. If new information comes into play which makes the resolution incorrect (such as a violation of binding constraints), the listener will go back and find a correct antecedent. This incremental resolution problem also motivates Strube's S-list approach.

The second drawback to the BFP algorithm is the computational explosion of generating and filtering anchors. In utterances with two or more pronouns and a Cf-list with several candidate antecedents for each pronoun, thousands of anchors can easily be generated making for a time consuming filtering phase. An exam-

[1] Henceforth BFP

ple from the evaluation corpus illustrates this problem (the italics in U_{n-1} represent possible antecedents for the pronouns (in italics) of U_n):

U_{n-1}: Separately, the *Federal Energy Regulatory Commission* turned down for now a *request* by *Northeast* seeking *approval* of *its* possible *purchase* of *PS* of *New Hampshire*.

U_n: Northeast said *it* would refile *its* request and still hopes for an expedited review by the FERC so that *it* could complete the purchase by next summer if *its* bid is the one approved by the bankruptcy court.

With four pronouns in U_n, and eight possible antecedents for each in U_{n-1}, 4096 unique Cf-lists are generated. In the cross-product phase, 9 possible *Cb's* are crossed with the 4096 *Cf's*, generating 36864 anchors.

Given these drawbacks, we propose a revised resolution algorithm that adheres to centering constraints. It works by first searching for an antecedent in the current utterance[2], if one is not found, then the previous Cf-lists (starting with the previous utterance) are searched left-to-right for an antecedent:

1. **Preprocessing** - from previous utterance: $Cb(U_{n-1})$ and $Cf(U_{n-1})$ are available.
2. **Process Utterance** - parse and extract incrementally from U_n all references to discourse entities. For each pronoun do:
   (a) Search for an antecedent intrasententially in Cf-partial(U_n)[3] that meets feature and binding constraints.
   If one is found proceed to the next pronoun within utterance. Else go to (b).
   (b) Search for an antecedent intersententially in $Cf(U_{n-1})$ that meets feature and binding constraints.
3. **Create Cf** - create Cf-list of U_n by ranking discourse entities of U_n according to grammatical function. Our implementation used a left-right breadth-first walk of the parse tree to approximate sorting by grammatical function.
4. **Identify Cb** - the backward-looking center is the most highly ranked entity from $Cf(U_{n-1})$ realized in $Cf(U_n)$.
5. **Identify Transition** - with the Cb and Cf resolved, use the criteria from (Brennan et al., 1987) to assign the transition.

[2]In this project, a sentence is considered an utterance

[3]Cf-partial is a list of all processed discourse entities in U_n

It should be noted that BFP makes use of Centering Rule 2 (Grosz et al., 1995), LRC does not use the transition generated or Rule 2 in steps 4 and 5 since Rule 2's role in pronoun resolution is not yet known (see Kehler 1997 for a critique of its use by BFP).

Computational overhead is avoided since no anchors or auxiliary data structures need to be produced and filtered.

3 Evaluation of Algorithms

All four algorithms were run on a 3900 utterance subset of the Penn Treebank annotated corpus (Marcus et al., 1993) provided by Charniak and Ge (1998). The corpus consists of 195 different newspaper articles. Sentences are fully bracketed and have labels that indicate word-class and features. Because the S-list and BFP algorithms do not allow resolution of quoted text, all quoted expressions were removed from the corpus, leaving 1696 pronouns (out of 2026) to be resolved.

For analysis, the algorithms were broken up into two classes. The "N" group consists of algorithms that search intersententially through all Cf-lists for an antecedent. The "1" group consists of algorithms that can only search for an antecedent in $Cf(U_{n-1})$. The results for the "N" algorithms and "1" algorithms are depicted in Figures 1 and 2 respectively.

For comparison, a baseline algorithm was created which simply took the most recent NP (by surface order) that met binding and feature constraints. This naive approach resolved 28.6 percent of pronouns correctly. Clearly, all four perform better than the naive approach. The following section discusses the performance of each algorithm.

4 Discussion

The surprising result from this evaluation is that the Hobbs algorithm, which uses the least amount of information, actually performs the best. The difference of six more pronouns right

Algorithm	Right	% Right	% Right Intra	% Right Inter
Hobbs	1234	72.8	68.4	85.0
LRC-N	1228	72.4	67.8	85.2
Strube-N	1166	68.8	62.9	85.2

Figure 1: "N" algorithms: search all previous Cf lists

Algorithm	Right	% Right	% Right Intra	% Right Inter
LRC-1	1208	71.2	68.4	80.7
Strube-1	1120	66.0	60.3	71.1
BFP	962	56.7	40.7	78.8

Figure 2: "1" algorithms: search $Cf(U_{n-1})$ only

between LRC-N and Hobbs is statistically insignificant so one may conclude that the new centering algorithm is also a viable method. Why do these algorithms perform better than the others? First, both search for referents intrasententially and then intersentially. In this corpus, over 71 % of all pronouns have intrasentential referents, so clearly an algorithm that favors the current utterance will perform better. Second, both search their respective data structures in a salience-first manner. Intersententially, both examine previous utterances in the same manner. LRC-N sorts the Cf-list by grammatical function using a breadth-first search and by moving prepended phrases to a less salient position. While Hobbs' algorithm does not do the movement it still searches its parse tree in a breadth-first manner thus emulating the Cf-list search. Intrasententially, Hobbs gets slightly more correct since it first favors antecedents close to the pronoun before searching the rest of the tree. LRC favors entities near the head of the sentence under the assumption they are more salient. The similarities in intra- and intersentential evaluation are reflected in the similarities in their percent right for the respective categories.

Because the S-list approach incorporates both semantics and syntax in its familiarity ranking scheme, a shallow version which only uses syntax is implemented in this study. Even though several entities were incorrectly labeled, the shallow S-list approach still performed quite well, only 4 percent lower than Hobbs and LRC-N.

The standing of the BFP algorithm should not be too surprising given past studies. For example, Strube (1997) had the S-list algorithm performing at 91 percent correct on three New York Times articles while the best version of BFP performed at 81 percent. This ten percent difference is reflected in the present evaluation as well. The main drawback for BFP was its preference for intersentential resolution. Also, BFP as formally defined does not have an intrasentential processing mechanism. For the purposes of the project, the LRC intrasentential technique was used to resolve pronouns that were unable to be resolved by the BFP (intersentential) algorithm.

In additional experiments, Hobbs and LRC-N were tested with quoted expressions included. LRC used an approach similar to the one proposed by Kamayema (1998) for analyzing quoted expressions. Given this new approach, 70.4% of the 2026 pronouns were resolved correctly by LRC while Hobbs performed at 69.8%, a difference of only 13 pronouns right.

5 Conclusions

This paper first presented a revised pronoun resolution algorithm that adheres to the constraints of centering theory. It is inspired by the need to remedy a lack of incremental processing and computational issues with the BFP algorithm. Second, the performance of LRC was compared against three other leading pronoun resolution algorithms based solely on syntax. The comparison of these algorithms is significant in its own right because they have not been previously compared, in computer-encoded form, on a common corpus. Coding all the algorithms allows one to quickly test them all on a large corpus and eliminates human error, both shortcomings of hand evaluation.

Most noteworthy is the performance of Hobbs and LRC. The Hobbs approach reveals that a walk of the parse tree performs just as well as salience based approaches. LRC performs just as well as Hobbs, but the important point is that it can be considered as a replacement for the BFP algorithm not only in terms of performance but in terms of modeling. In terms of implementation, Hobbs is dependent on a precise parse tree for its analysis. If no parse tree is available, Strube's S-list algorithm and LRC prove more useful since grammatical function can be approximated by using surface order.

6 Future Work

The next step is to test all four algorithms on a novel or short stories. Statistics from the Walker and Strube studies suggest that BFP will perform better in these cases. Other future work includes constructing a hybrid algorithm of LRC and S-list in which entities are ranked both by the familiarity scale and by grammatical function. Research into how transitions and the Cb can be used in a pronoun resolution algorithm should also be examined. Strube and Hahn (1996) developed a heuristic of ranking transition pairs by cost to evaluate different Cf-ranking schemes. Perhaps this heuristic could be used to constrain the search for antecedents.

It is quite possible that hybrid algorithms (i.e. using Hobbs for intrasentential resolution, LRC for intersentential) may not produce any significant improvement over the current systems. If so, this might indicate that purely syntactic methods cannot be pushed much farther, and the upper limit reached can serve as a base line for approaches that combine syntax and semantics.

7 Acknowledgments

I am grateful to Barbara Grosz for aiding me in the development of the LRC algorithm and discussing centering issues. I am also grateful to Donna Byron who was responsible for much brainstorming, cross-checking of results, and coding of the Hobbs algorithm. Special thanks goes to Michael Strube, James Allen, and Lenhart Schubert for their advice and brainstorming. We would also like to thank Charniak and Ge for the annotated, parsed Treebank corpus which proved invaluable.

Partial support for the research reported in this paper was provided by the National Science Foundation under Grants No. IRI-90-09018, IRI-94-04756 and CDA-94-01024 to Harvard University and also by the DARPA research grant no. F30602-98-2-0133 to the University of Rochester.

References

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings, 25th Annual Meeting of the ACL*, pages 155–162.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. *Proceedings of the Sixth Workshop on Very Large Corpora.*

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Jerry R. Hobbs. 1977. Resolving pronoun references. *Lingua*, 44:311–338.

Megumi Kameyama. 1986. Intrasentential centering: A case study. In *Centering Theory in Discourse.*

Andrew Kehler. 1997. Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, 23(3):467–475.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Lingusitics*, 19(2):313–330.

Michael Strube and Udo Hahn. 1996. Functional centering. In *Association for Computational Lingusitics*, pages 270–277.

Michael Strube. 1998. Never look back: An alternative to centering. In *Association for Computational Lingusitics*, pages 1251–1257.

Marilyn A. Walker. 1989. Evaluating discourse processing algorithms. In *Proceedings, 27th Annual Meeting of the Association for Compuational Linguisitcs*, pages 251–261.