

Pronoun Resolution of “they” and “them”

From: Proceedings of the Eleventh International FLAIRS Conference. Copyright © 1998, AAAI (www.aaai.org). All rights reserved.

Bruce A. Wooley
Box 9637
Mississippi State, MS 39762
bwooley@cs.msstate.edu

Abstract

Pronoun resolution is an uncertain process in machine oriented natural language processing. There are many techniques to determine the referent of the pronoun. This paper addresses the resolution of two pronouns, “they” and “them”, and looks at simplistic rules to improve the process of determining the referent, beyond simply deciding on the basis of most recent plural noun.

Introduction

The majority of text being generated today is available in electronic form. The current advances in speech recognition provide potential access to a vast amount of additional text in electronic form. Automated processing of this text, to obtain information contained within the text and to allow content queries associated with the text, is desirable. The ability to extract information will require resolution of co-references over the entire discourse.

Many of the current techniques identifying co-referents involve identification of phrases within sentences and applying template processes to these in conjunction with rules (Cowie and Lehnert 1996) or hierarchical constructs (McCarthy and Lehnert 1995). The templates are specifically constructed for each domain, and this construction requires some expertise in the overall process in addition to domain knowledge. These templates are then combined with rules or decision trees to evaluate the domain specific text of interest.

Resolving pronominal referents is a subset of the co-reference problem. The ability to identify the person, object, place, time, event, or concept that is a referent of a pronoun is necessary to follow the concepts being communicated and to maintain the general focus of the discourse (Grosz and Sidner 1986).

There are some resolution algorithms outlined by Hobbs (1986) that perform well, but these algorithms require a properly parsed sentence. The process identified by this paper only requires that the words be tagged for part of speech with the single enhancement of identifying plural noun phrases from plural and singular nouns joined with “and” along with “,” delimited tags. The pronouns used in this paper are limited to “they” and “them”, which are general enough to encompass referent types of person, object, place, time, event, and concept. The pronouns “they” and “them” will normally refer to a plural noun, or a

noun phrase containing multiple noun/pronoun elements connected with an “and” or a “,” as described above.

The simplest technique for identifying a referent is to tie the pronoun to the nearest preceding plural noun. This is the basic point of reference for evaluation, where the performance of all other rules is measured against this basis. In keeping with this simplistic approach, rules requiring the parsing of sentences were not considered in this experiment. The rules that are investigated rely only on part-of-speech tagging, and the possible grouping of nouns and plural nouns as a potential referent.

The data used in experimentation for this paper comes from the Association for Computational Linguistics’ (ACL) Data Collection Initiative (ACL/DCI), specifically, the Library of American Texts. The goal is to reduce the error in identifying the referent by 50% over the error obtained by assuming the referent is the most recent plural noun. This involves adding rules that improve a larger percent of the errors than they create.

The results were surprisingly successful. The application of one simple rule provided nearly a 50% improvement. Additionally, the identification of a central focus in a portion of the text, along with one additional rule provided more than a 50% improvement over the nearest preceding pronoun.

Overview

The remainder of this paper is organized in the following manner. Under the section titled DATA, I discuss the source of the data and selection criteria to obtain the sentences containing the pronouns “they/them” along with supporting sentences containing potential referents. Under the section PROCESS, techniques of evaluation and application rules are discussed. The ANALYSIS and OBSERVATION sections present the results and discuss meanings associated with referent resolution. The last two sections, CONCLUSIONS and FUTURE WORK, address the results as applied to the goals, along with potential improvements that can be gained by using more sophisticated techniques (still without complete sentence parsing).

Data

There were three main goals directing the selection of the sentences. First, the data must be tagged for part-of-speech. Second, the data should be as domain independent as possible. Finally, I chose ACL/DCI data files

containing a relatively rich rate of occurrence of "they/them." The ACL/DCI corpus provides sentences that are already tagged. The Penn Treebank text within the treebank/posttext directory was chosen because it is generic (Library of American texts) and is tagged. A search of the files contained within this directory revealed four files (t1.pos, t2.pos, t4.pos, and t20.pos) that contained more occurrences of "they/them" than other files in the same directory.

A program was written to extract sentences from the corpus to evaluate the effectiveness of the selected rules. The criteria used for the extraction processes were as follows. A sentence is included in the study if it contains at least one of the pronouns being inspected ("they" or "them"). Additionally, there are supporting sentences programmatically extracted from the corpus under the following special situations. If a specific selected sentence (containing a pronoun "they/them") does not have any plural nouns preceding this pronoun, then include (as a reference sentence) the sentence immediately prior to the selected sentence. Additionally, if this included sentence does not have any plural nouns, then include the sentence from the corpus that immediately precedes this reference sentence. The scope of the program limits the extraction of these reference sentences to a count of two. If neither of these supporting sentences contain a plural noun, a manual inspection of the corpus was needed to find the nearest prior plural noun. The supporting sentences are not counted as part of the selections from the corpus. There were 500 sentences selected and they contained a total of 611 occurrences of "they" or "them", with the patterns of occurrences as follows: 414 sentences with one (1), 69 sentences with two (2), 13 sentences with three (3), 2 sentences with four (4) and 2 sentences with six (6).

Process

After the 500 sentences containing the desired pronouns along with some reference sentences were selected, the application of rules to the sentences was done manually. Two hand modifications of the data were performed. First, the simple data extraction program did not take text-heading cues into consideration. As a consequence, some titles were included as part of a sentence due to the extraction program logic. These titles were removed from the sentences by hand. Second, the identification of plural and singular nouns connected with "and" or "," were identified with brackets ("[]"). This bracketed phrase was then treated as a simple plural noun.

Analysis

In the following discussion, "nearest prior plural noun" refers to the plural noun that is spatially the closest preceding plural noun (or plural noun phrase). This will cross sentence boundaries if there are no prior plural nouns within the sentence containing the pronoun being

investigated. "The first prior plural noun" refers the plural noun occurring closest to the beginning of the sentence. Identification of "the first prior plural noun" also may cross sentence boundaries. The "nearest prior plural noun" and "first prior plural noun" may refer to the same plural noun.

Application of the ground rule, that is, associating "they" or "them" with the nearest prior plural noun (or plural noun phrase), results in 335 correctly tagged pronouns, or an accuracy of 54.8% (335/611). This requires an accuracy of 77.4% to achieve the goal of 50% improvement in accuracy. Choosing the first plural noun of the sentence (plural noun occurring prior to the "they/them" in question) results in 462 correctly tagged pronouns, raising this accuracy to 75.6% (462/611). The next rules attempted involved the use of other keywords such as "this" and "all" in specific phrase formats. Some examples of these phrases are,

- 1) "all [plural noun]"
- 2) "and they all"
- 3) "all [be-verb]"

I also examined using as the referent the most recent prior pronoun ("they/them") under various conditions. All these rules resulted in marginal improvement or no improvement in identifying the referent, so these rules were discarded.

Upon further analysis of the sentences, I discovered a fictional story contained within the last 44 sentences. In reading these last 44 extracted sentences, I could not identify the referent for most of the "them/they"s. Returning to the original corpus, it was discovered that the major characters of the story were ten young Indian men who were all brothers. Tying all references to "brothers" in these 44 sentences resulted in an 87.1% accuracy within these 44 sentences. Separating these last 44 sentences from the extracted data left the first 456 sentences to process with the rules described above. The results for these 456 sentences are based on 549 occurrences of "they/them." The application of the ground rule, nearest prior plural noun results in 316 correctly tagged pronouns, or an accuracy of 57.6% (316/549). This requires an accuracy of 78.8% to achieve a 50% improvement in accuracy over this new basis. Choosing the first plural noun of the sentence results in 444 correctly tagged pronouns, raising this accuracy to 80.9% (444/549).

The fictional story sets the stage for the introduction of a new rule, wherein the central focus of a story becomes the referent for "they/them" regardless of the spatial locality with regard to the occurrence of "they/them" within the text. For instance, the brothers in the above mentioned story were not mentioned or listed in any of the extracted sentences. Despite this lack of spatial locality of "brothers" to "they/them" this central focus was the correct referent for the pronouns. If the focus of the discourse is constant or re-occurring over a large portion of the

discourse, then this information can be used to re-evaluate the assignments of the referents.

Observations

The final analysis of the experimental results is discussed in the conclusions. It is interesting to note that tying a pronoun to the correct referent for example "brothers" in the text that was analyzed may not always give the exact meaning. The following is a concrete example from the story involving the pronoun "them" that demonstrates the inexact nature of the pronoun resolution process. "The youngest told them..." which references only the remaining nine brothers, as opposed to "it was impossible to track them" which refers to all ten brothers. Although the correct referent to "them" is "brothers", the exact meaning as to the number of brothers cannot be determined using the simple algorithms of this paper.

Conclusions

These conclusions are based on sentences that were taken from a variety of general texts. The rule "choosing the first plural noun in the sentence as the referent" will produce results at least 50% better than "nearest prior plural noun" unless the discourse is lengthy and contains main characters that are the focus throughout the discourse. In the latter case, using the main characters as the referent produced results in the upper 80% range, which is much better than either the first or nearest prior plural noun. These rules did not require the sentences to be parsed. The only addition to the part of speech tagging is the identification of compound singular and plural nouns. For this paper, I used a simplistic algorithm to accomplish this identification. I used the "and" and "," strings within the text, and checked the tags of words on both sides for singular or plural nouns.

Future Work

Future considerations include matching attributes of the pronoun with the attributes of the referent, allowing a pronoun referring to person, to reject a referent that has the attribute "inanimate object", or perhaps even "animal object." Some common attributes include gender, number, mass, and type-of-object.

An additional process that may be found useful is to reverse the normal process that is used in parsing and resolving referents. The normal process is to resolve the referent by using syntactic information along with different foci (i.e. discourse or actor) (Sidner 1986). Assigning the referents first, then computing the syntactical structure and the foci, would give a different view to the problem. I believe humans do this iteratively, attempting to find the concepts that have the highest probabilities in relationship to the individuals knowledge.

Beyond the techniques presented in this paper, a much larger analysis of the concepts in the discourse

(comprehension/understanding) must be built to aid in the resolution of the pronouns. Pronoun resolution is a significant part of automated information extraction.

Acknowledgements

Special thanks go to Dr. Lois Boggess for her comments, suggestions, and support.

References

- Charniak, Eugene. 1993. *Statistical language learning*. Cambridge, MA: The MIT Press.
- Cowie, Jim, and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*. 39 (1): 80-101.
- Grosz, Barbara J. 1986. The representation and use of focus in a system for understanding dialogs, In *Readings in Natural language processing*. Morgan Kaufmann Publishers, Los Altos, CA, pages 352-362.
- Grosz, Barbara J, and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Association for Computational Linguistics*. 12 (3): 175-204.
- Hess, Michael. 1989. Reference and quantification in discourse. <http://www.ifi.unizh.ch/groups/CL/hess/oldpublications.html>.
- Hobbs, Jerry. 1986. Resolving pronoun references, In *Readings in Natural language processing*. Morgan Kaufmann Publishers, Los Altos, CA, pages 339-352.
- Lappin, Shalom, and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Association for Computational Linguistics*. 20 (4): 535-61.
- McCarthy, Joseph F., and Wendy G. Lehnert. 1995. Using Decision trees for coreference resolution. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. IJCAI '95: 1050-5.
- Sidner, Candace L. 1986. Focusing in the comprehension of definite anaphora, In *Readings in Natural language processing*. Morgan Kaufmann Publishers, Los Altos, CA, pages 363-94.
- Wiebe, Janyce, Graeme Hirst, and Diane Horton. 1996. Language use in context. *Communications of the ACM*. 30 (1): 102-11.