A Corpus of Challenging Pronoun Disambiguation Problems, Adapted from Children's Books

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

Xiaoman Pan
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
panx2@rpi.edu

June 8, 2015

Abstract

A corpus has been created containing 403 pronoun disambiguation problems, intended to be much easier for human readers than for artificial intelligence programs in the current state of the art. The corpus has been assembled by extracting short passages out of children's books (narratives and children's histories) in Project Gutenberg. The passages were then manually selected and substantially edited. The pronoun disambiguation problems in the corpus are thus by no means a random or representative sample of any well-defined natural set.

This note briefly describes (a) the form of the corpus; (b) the method by which the corpus was constructed; (c) some statistics over the corpus; (d) the differences between solving disambiguation problems in this artificial setting vs. solving them in reading the full text; (e) issues that arise in doing these disambiguations.

1 Introduction

A corpus has been created containing 403 pronoun disambiguation problems, intended to be much easier for human readers than for artificial intelligence programs in the current state of the art. The immediate purpose is to serve as a "qualification round" for the Winograd Schema Challenge (Levesque, Davis, and Morgenstern, 2012); contestants for the Winograd Schema Challenge would be required first to demonstrate a high degree of accuracy over the problems in this corpus. More generally, it will serve as a resource for research in reference resolution for automated natural language processing.

The corpus has been assembled by extracting short passages out of children's books (narratives and children's histories) in Project Gutenberg.[1] The passages were then manually selected and substantially edited. The pronoun disambiguation problems in the corpus are thus by no means a random or representative sample of any well-defined natural set. For this reason, the statistics enumerated in section 5 are purely characteristics of this particular corpus, possibly useful in evaluating its significance or comparing it to other corpora. These statistics in themselves say nothing about the properties of English text generally or about the task of pronoun disambiguation.

[1] http://www.gutenberg.org

Section 2 reviews related literature. Section 3 presents the form of the corpus. Section 4 describes the method in which the corpus was constructed. Section 5 gives some statistics over the corpus. Section 6 discusses the differences between solving the disambiguation problems in this artificial context versus the same problems when the full text is presented; in some respects they are harder here, in others they are easier. It also discusses how AI programs could "cheat" on the problems i.e. solve the problems without drawing on significant semantic knowledge.

2 Related literature

TO BE WRITTEN.

3 Form of the corpus

The corpus is implemented in XML. The specifics of the XML tags will be self-evident to any reader who looks at the XML source, and is therefore not worth laboriously describing here.

As viewed in a web browser, the corpus is an enumerated list of 403 annotated pronoun disambiguation problems that arise in 334 distinct texts. (Quite a few of the texts contain multiple disambiguation problems). The texts vary in length from one to four sentences.

The presentation of a disambiguation consists of

1. The text, with the ambiguous pronoun in italics.
2. On a new line, the pronoun is repeated. If the same pronoun appears more than once in the text, disambiguating context is placed in brackets.
3. The possible antecedents for the pronoun — i.e. noun phrases that precede the pronoun in the text and agree in gender and number — in backward order; that is, in order of increasing distance from the pronoun.
4. The correct answer is specified, in terms of the index of the antecedent in the list in (3).
5. When appropriate, a note "[Obeys/Violates] 'most recent same case' rule". This is discussed in section 4.4.1

If a text includes multiple ambiguous pronouns, then the entire form is repeated for each pronoun.

For example:

> Bulgaria had to give up much of the territory *it* had conquered to its former allies.
> **Pronoun:** *it.* **Antecedents:** 1: territory. 2: Bulgaria.
> **Answer:** 2
> Obeys "most recent same case" rule
>
> Bulgaria had to give up much of the territory it had conquered to *its* former allies.
> **Pronoun: its. Antecedents:** 1: territory. 2: Bulgaria.
> **Answer:** 2

A noun phrase in the list of antecedents is *never* a pronoun, though it may be a quantified expression (e.g. "Someone", "all"). If two noun phrases that appear before the pronoun are nearly synonymous then only one is included in the list of antecedents, so that the human reader or AI program does not have to make a difficult and arbitrary choice. Generally, we use only the most distinctive word for the antecedent, except when it is necessary for clarity to use multiple words. For instance in the above example, the noun phrases "much of the territory" and "territory" are nearly synonymous, and so the list includes only "territory".

The potential antecedents must agree with the pronoun in gender and number; thus, we exclude texts in which there are not two or more previous noun phrases that agree in gender and number.[2] However, unlike the Winograd Schema Challenge we do not exclude texts in which the disambiguation can be carried out using selectional restrictions or frequency considerations. On the contrary, one of the purposes of this corpus, viewed as a qualification round for the Winograd Schema Challenge, is to make sure that the contestants have mastered the basics of selectional restrictions before they try the Winograd schemas, where such considerations are excluded.

There is one kind of case in which the antecedent is actually not a phrase that appears earlier in the sentence. Rather, the antecedent is the conjunction of two or more people mentioned earlier in the sentence. For example:

> "Why don't you girls form another club" spoke Betty. "We might," said Alice cooly. "Come on, Kittie," she added. "I guess *we're* not wanted here".
> **Pronoun:** *we're*. **Antecedents:** 1: Kittie and Alice. 2: Kittie and Betty. 3: Alice and Betty. 4. Kittie, Alice, and Betty.
> **Answer:** 1

The above example is also unique in our corpus in that the ambiguous pronoun is in the first person. In all the other examples, the ambiguous pronoun is third person.

4 Construction of the corpus

Conceptually, the construction of the corpus consisted of 5 stages:

1. Harvesting single sentences from the online books (section 4.1).
2. Pruning unproductive candidates (section 4.2).
3. Expanding sentences to include previous text. (section 4.3)
4. Further pruning (section 4.4)
5. Editing (section 4.5).

Stage 1 was carried out automatically; the remaining 4 stages were done manually. In practice, states 2-5 were not strictly separated sequentially; however, to a large degree the above sequence was followed.

Broadly speaking, the purposes of the selection and editing was to make sure that the sentences conform to the desired form; that they are easy for human readers; and that they are difficult for AI systems.

4.1 Harvesting

A collection of 178 books, identified as "children's books" were downloaded from Project Gutenberg. This included a mixture of fiction and children's histories. Particularly well represented are R.M. Ballantyne (57 books, adventure stories) "Laura Lee Hope" (author of "The Bobbsey Twins"

[2] In English text, there are occasionally pronouns whose antecedent follows them in the text. We did not encounter any such examples in creating our corpus, and would not have included them.

and similar series; 17 books) Joseph A. Altsheler (16 books, historical fiction), and Horatio Alger (12 books; “rags-to-riches” novels.) A complete list of authors is given in the appendix.

Using the NP-chunker from the Natural Language Toolkit[3] (Bird, Klein, and Loper 2009), a program was written to extract the first 5000 sentences satisfying the following constraints:

1. The sentence had the form `* NP * NP * Pronoun *` (the asterisks here are Kleene stars).
2. The sentence had no more than 20 words.
3. Neither of the first two `NP`’s is a pronoun.
4. The pronoun is in the third person.

4.2 Initial Pruning

Most of the sentences collected in the harvesting phase were eliminated for one of two reasons:

- NP-chunker errors. The NP chunker was trained probabilistically and hence has its percentages. For example, ??. More strangely, it seems to consider any word ending in “our”, such as “colour” or “honour”, to be a pronoun.

  There are also some grammatical forms in which there is perhaps some grammatical justification for the NP-chunker’s claim that these are NP’s, but which certainly can hardly serve as candidate antecedents. A common example is the initial “There” in a sentence; e.g. “There were some natives, but they fled.” For all we know, it is reasonable to consider “There” in this sentence as an NP, but certainly it is not a possible antecedent for “they”.

- Gender/number disagreement. In many cases the pronoun did not agree with both candidate antecedents in gender or number. In some cases this could be fixed by adding previous sentences (see section 4.3), but there were many cases where this was not reasonable.

This stage of pruning eliminated about 4/5 of the sentences, so the corpus now had about 1000 sentences.

4.3 Expanding texts

In many cases, we looked up the sentence in the original book, and brought in some of the preceding text; much more rarely, some of the subsequent text. The cases where we looked up the context were:

- Some pronoun in the sentence (not necessarily the ambiguous pronoun) had its antecedent in an earlier sentence.
- The sentence began with a direct quotation, e.g. “‘Why don’t you girls form another club’ spoke Betty.” In this case, the NP-chunker considered the quotation to be a separate sentence, so t was necessary to bring it in in order to complete the sentence.
- The sentence was hard or impossible to interpret out of context.
- The sentence was particularly intriguing, and we wanted to see what was the context.

[3] http://www.nltk.org/

The new material was brought in if either it was necessary to the interpretation of the original sentence, or if it raised interesting problems of pronoun disambiguation. The new material was often heavily edited.

4.4 Further Pruning

Additional pruning was carried out for a number of reasons.

- The actual antecedent was too far back in the original text.
- The sentence could actually be considered ambiguous to a human reader. This was rare, but there were a few cases.
- In order to make the sentence intelligible, it would be necessary to bring in an unacceptably large amount of context.
- Repetition. Some phrases appeared numerous times, each time disambiguated the same way. For example "shook [his/her/their] head[s]" appeared several times, and in every case, the referent of the pronoun was the subject of "shook" (it never happens that someone shakes someone else's head). In such cases, we included only one sample text.
- The material was distasteful for some reason, in a way that could not easily be corrected by editing. (The original books are often quite sexist, and extremely racist, particularly toward Native Americans.)

4.4.1 Simple syntactic disambiguation

In the corpus of sentences, the following rule, when applicable, was quite reliable: If the ambiguous pronoun is either the subject or object of a verb, and there is a preceding noun phrase that agrees in gender, number, and case, then that noun phrase is the antecedent of the pronoun. For instance, in the sentence "The yeomanry had risen upon the invaders, and they had driven them back," the rule works for both pronouns.

Since this rule can be carried out purely on the basis of syntactic parsing, with no use of semantics or other deeper understanding, we have mostly pruned texts where this rule applies. The exceptions that we have left in are mostly either texts in which these ambiguities appear together with other kinds of ambiguities; or texts that seem particularly interesting.

In the corpus, we have tagged every sentence where the conditions of the rule apply. Sentences where the rule gives the right answer have the tag "Obeys 'most recent same case' rule"; those where the rule gives the wrong answer have the tag "Violates 'most recent same case' rule". In placing these tags, we have followed the following strict form of the rule.

> If the pronoun is either the subject or object of a verb, then find the most recent noun phrase which is not a pronoun and which agrees with the ambiguous pronoun in number, gender, and case. Propose that that is the antecedent of the ambiguous pronoun.

Under that reading of the rule, there are 45 instances in the corpus where the rule is followed and 16 instances where it is violated. Thus, using the rule gives a perceptible edge to an AI program, but not an enormous edge.

No doubt there are other similar syntactic rules that could be useful; but we did not notice any that seemed be both frequent enough and reliable enough to make it worthwhile using them to prune the corpus.

4.5 Editing

Texts were edited to make them either harder for the AI programs or (more frequently) easier for the human reader.

- Genders of either pronouns or of nouns were changed to increase the number of antecedents, to make it harder for the AI program. For instance, pronouns referring to countries and to ships were almost always changed from feminine to neuter.
- Necessary contextual information was incorporated.
- Sentences were simplified; phrases that were extraneous to the disambiguation issue were sometimes pruned, though by no means always.
- Word choices that were either somewhat obsolete or idiosyncratic were replaced. Very unusual names such as "Mun Bun", which might be hard to identify as such, were replaced with something more common.
- Some racist content was altered. Plenty of racist attitudes remain visible, but hopefully not to the point that it is actually unpleasant to look at the corpus. More importantly, hopefully the disambiguations can be carried out without relying on unacceptable stereotypes.
- Sentences were edited for style. Judging from the samples we have looked at, R.M. Ballantyne was not a particularly gifted or careful writer at the sentence level. Horatio Alger, on the other hand, had a more thoroughly tin ear for how the English language actually works than any other writer we've ever run across; some of his sentences are really quite bizarre, especially his dialogue. When the sentences were so bad that they were difficult to understand or that we couldn't stand them, we rewrote them.

5 Statistics

As discussed above, there is no reason to think that the statistics here are anything more than artifacts of the way that the corpus was constructed. That is, they are characteristics of the particular corpus, not characteristics of pronoun disambiguation in English.

Number of texts: 334.
Number of pronoun ambiguities. 403.
Number of sentences: 452.
Average sentence length: 12.0

288 texts with 1 ambiguous pronoun.
34 texts with 2 ambiguous pronouns.
8 texts with 3 ambiguous pronouns.
2 texts with 4 ambiguous pronouns.
1 text with 6 ambiguous pronouns.
1 text with 9 ambiguous pronouns.

235 texts with 1 sentence.
82 texts with 2 sentences.
15 texts with 3 sentences.
2 texts with 4 sentences.

Number of candidate antecedents:
310 ambiguities with 2 possible antecedents.

61 ambiguities with 3 possible antecedents.
17 ambiguities with 4 possible antecedents.
12 ambiguities with 5 possible antecedents.
2 ambiguities with 6 possible antecedents.
1 ambiguity with 7 possible antecedents

The expected accuracy from guessing at random is therefore 0.45.

Location of true antecedent relative to pronoun:
143 antecedents at location 1 (i.e. the candidate immediately preceding the pronoun).
206 antecedents at location 2.
37 antecedents at location 3.
9 antecedents at location 4.
6 antecedents at location 5.
1 antecedent at location 6.
1 antecedent at location 7.

The expected accuracy from always guessing location 2 is therefore 0.51. This value can be therefore taken as a baseline.

It is rather striking that in this corpus, location 2 is the most frequent position for the antecedent. As is well known, in text in general, the recency rule, corresponding to location 1, gives the correct antecedent about 85% of the time. Clearly, the discrepancy is at least in part due to the fact that texts in which there is only one candidate antecedent, and which thus trivially satisfy the recency rule, have been excluded from the corpus. Determining whether this is the entire explanation would require a systematic study.

6 Some observations on the disambiguation task

On the whole, carrying out these disambiguations would probably be easier for both human and artificial readers if the passages occured in context rather than being given in isolation. The human reader finds a number of these passages jarringly incomplete in their detached form; some mental effort at reconstructing the setting is required. For a skilled reader, this should rarely be problematic, but an unskilled reader might sometimes find it difficult to figure out what is going on. In a somewhat analogous way, an AI program might well find the additional information in nearby text useful in carrying out the disambiguation. Of course, existing AI programs are severely limited in the degree to which they can assimilate or use this kind of contextual information, so the degree of gain might not be large.

In a few respects, the process of formulating these texts as multiple-choice problems simplifies the problem of disambiguation.

- In natural texts, the antecedent of a pronoun can be implicit in the preceding text rather than explicit. The antecedent can be a gerund phrase, an infinite phrase, or a subordinate clause. These possibilities do not arise in our corpus.
- In cases where multiple phrases refer to the same thing, we have included only one of these phrases. We have thus carried out a process of identifying the two phrases as coextensional, and thus spared the reader the task of doing this.

AI programs could "cheat" on the test in a couple of ways. Most obviously they can look up the original texts for additional clues. Alternatively, they can look for alternative accounts of the historical events or synopses of the narratives. Neither of these is at all trivial to do effectively, and neither will give decisive information for more than a fraction of the examples in the corpus. Depending on the purposes for which the test is being used and the claims being made, such techniques could be quite legitimate.

Acknowledgements

References

S. Bird, E. Klein, and E. Loper (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media.

H. Levesque, E. Davis, and L. Morgenstern (2012). "The Winograd Schema Challenge." *KR-2012.*

Appendix: List of authors of the books

R.M. Ballantyne	62
Laura Lee Hope	17
Joseph A. Altsheler	16
Horatio Alger	12
Angela Brazil	6
Cyrus Townsend Brady	6
Percy Keese Fitzhugh	5
Roger Thompson Finlay	5
Edward Stratemeyer	4
Howard R. Garis	3
Jacob Abbott	3
H. Irving Hancock	3
George A. Warren	2
Amy E. Blanchard	2
Cyril Burleigh	2
Carolyn Wells	2
Arthur M. Winfield	2
Alice Turner Curtis	1
H. Alger and A.M. Winfield	1
Wilbur F. Gordy	1
Prescott Holmes	1
Ralph Henry Barbour	1
Daniel Defoe	1
Amanda Minnie Douglas	1
Wolfram Eberhard	1
Various	1
D. H. Montgomery	1
Nathaniel Hawthorne	1
Henry Mann	1
George W. Peck	1
Charles W. Whistler	1
Edward S. Ellis	1
A.E. McKinley, C.A. Coulomb, and A.J. Gerson	1
Lady I. A. Gregory	1
Edward Eggleston	1
D.H. Montgomery	1
Joseph Jacobs	1
Ontario Ministry of Education	1
Samuel Griswold Goodrich	1
A. Russell Bond	1
Burton E. Stevenson	1
Mary Roberts Rinehart	1
R.C. Andrews and Y.B. Andrews	1