

An Algorithm for Anaphora Resolution in Aviation Safety Reports

Katia Dilkina and Fred Popowich

Simon Fraser University, Burnaby, BC, CANADA, V5A 1S6

<http://www.sfu.ca/~knd/>

<http://www.cs.sfu.ca/~popowich/>

Abstract. While the general problem of determining the referent of a pronoun like *it*, *he* or *she* is quite difficult, by constraining the domain of discourse, and by constraining the task, it is possible to achieve accurate resolution of the antecedents of pronouns in certain contexts. In this paper, we constrain the domain of discourse by considering only aviation safety reports, and consider the task of pronominal resolution in the context of entity extraction in these aviation texts. In doing so, we are able to provide an algorithm that is better suited to the task than general purpose anaphora resolution algorithms, and which should be applicable to other domains by incorporating similar constraints.

Keywords: Entity extraction, pronouns, reference resolution

1 Introduction

Anaphora resolution is a hard problem; a great deal of linguistic and non-linguistic knowledge must be brought to bear to determine the ‘meaning’ of pronouns (like *it*, *they* and *them*) when they occur in texts. As is often done in natural language understanding tasks, the problem can be simplified by constraining the domain of discourse and by constraining the context in which natural language is used. We constrain the domain of discourse by dealing only with aviation safety text, and we constrain the context by considering only the genre of text found within safety reports, and by considering only the task of entity extraction.

Anaphora can be defined as the process of using pronouns or similar descriptors to refer to entities or expressions found elsewhere in a text, or in the domain of discourse. So, a pronoun, like *she* is an example of an anaphor, while a noun phrase like *The pilot* is an example of a possible antecedent for the anaphor in a sentence like *The pilot thought she might have to abort the landing*. Anaphora resolution can then be defined as the process of finding an appropriate antecedent for each anaphor in the text.

In looking at the techniques used for anaphora resolution, one finds that they differ depending on: 1) the types of anaphors being processed, 2) the degree to which they use syntactic constraints, and 3) their employment of focusing and centering theory techniques. Here, we will restrict our investigation to third person pronouns, specifically: *he*, *his*, *him*, *she*, *her*, *it*, *its*, *they*, *their* and *them*.

2 Aviation Safety Reports

Our ASRS corpus [1] consists of approximately 80,000 reports of incidents involving aircraft. In our study, we manually analyzed 200 of these documents, randomly selected. The average length of a report is 280 words and 16 sentences. The domain is highly specific with a lot of terminology and abbreviations. As noted in [8], about 10% of all the words contained in ASRS reports are abbreviations. For our collection of 200 documents, on average, there are 19 first and third person pronouns in each document, approximately 74% of which are first person pronouns, 22% are third person anaphoric pronouns, and 4% are pleonastic third person pronouns.

In addition to the characteristics outlined above, the ASRS corpus presents the following issues not characteristic for corpora used with anaphora resolution algorithms so far:

First of all, all text is typed in uppercase letters and there is a large number of non-standard abbreviations. Therefore, additional processing is required, with the help of a database of abbreviations and proper nouns, in order to determine the correct POS tags for all lexical items.

Secondly, there is a considerable amount of grammatical and spelling mistakes, which not only hinder basic stages of text processing such as POS tagging and parsing, but also often makes it difficult to resolve seemingly straight-forward co-reference. The seriousness of the mistakes vary from simple misspelling or faulty placement of a period, to gross grammatical errors, which make the text incomprehensible even to an intelligent human reader.

Thirdly, a phenomenon widely observed in the corpus is gender and/or number disagreement between anaphors and antecedents. Often an entity is referred to once by *he* and other times by *they*, or by *he* and *it*, etc., as seen in (1). This peculiarity of the corpus has at least one immediate consequence with respect to the design of an anaphora resolution algorithm, i.e. the algorithm could not include a person/gender/number agreement component, which is present practically in all previous coreference resolution algorithms as an effective filter for antecedent candidates for each anaphor (cf. [4], [5], [6]).

AT TIMES ATC (= air traffic control) WILL GIVE AN INTERSECTION XING ALT, AND WHEN YOU ADVISE THEM THAT IT WOULD BE IMPOSSIBLE TO MAKE THE ALT THEY SAY "DO YOUR BEST." AND THE THING I DON'T UNDERSTAND IS THAT THERE IS A FIX 14 MILES WEST OF MARTINSBURG VOR THAT IS ON THE CHARTS AND WAS ALSO IN THE COMPUTER. WHY DIDN'T HE USE THAT? (1)

Yet another difficulty arises due to the colloquial arbitrary style of this genre of text. Often it is either ambiguous or even impossible, even in manual anaphora resolution, to determine the correct antecedent of an anaphor. This happens because of ambiguity or vagueness of expression, or simply because of incoherent or incomplete thought expressed by a sentence.

In our study, apart from analyzing two hundred aviation safety reports with respect to their style, number and kinds of anaphors, and distance of the antecedent, we also looked in great detail at the nature of the antecedents. Despite their unstructured colloquial style, the ASRS data represent a very constrained domain of knowledge, and this is reflected in the use of anaphors. In particular, we found that 65% of all third person pronouns are used to refer to only four types of entities; namely, aircraft, aircraft crew, air-traffic control, and the person reporting the event (reporter).

3 An Approach to Anaphora Resolution

As we have seen, the ASRS corpus differs greatly from technical manuals corpora, which have been extensively used with previous anaphora resolution algorithms (cf. [5], [6]), and which are characterized by grammaticality and highly specialized language both in terms of lexicon and syntactic structure. The aviation safety reports are also different from journal and newspaper articles, which is the second type of corpora widely used with anaphora resolution algorithms (cf. [7], [9], [4]). Notably, journal and news articles include a range of syntactic structures as well as some amount of quotations, and so they are similar to the ASRS documents. However, they have a considerably larger knowledge domain, they do not use a great amount of abbreviations and jargon, and also they generally obey the grammatical rules of English. So, with respect to lexicon and writing style, the ASRS corpus is somewhere between technical manuals and journal and news articles. This observation leads to the conclusion that a new approach to anaphora resolution is needed if we are to be able to successfully computationally analyze aviation safety reports.

The algorithm consists of four steps, as shown below.

1. when an anaphoric pronoun a_i is identified in sentence s_k , a list of candidate antecedents is created, $\{c_1, c_2, \dots, c_n\}$, which includes all named entities in s_k and s_{k-1} .
2. apply assignment function f to determines the salience value for each c_j ,
3. select c_{max} having largest value for $f(c_j)$
4. if $f(c_{max}) \geq t$, where t is the threshold salience value,
 - a) then return c_{max} as the antecedent of a_i
 - b) else the named entities from s_{k-2} are processed. If none of them yields a candidate value above or at the minimum, entities from s_{k-3} are processed, and so on.

Our approach differs from previous anaphora resolution systems in a number of ways. As discussed in the previous section, there is no matching for gender and number agreement between anaphors and antecedent candidates. Also, there is no morphological or deep syntactic processing – we only need a partial parser to identify the noun phrases (and in particular, we are interested in the head noun). In contrast with previous algorithms, our approach will have no limit on the possible distance between an anaphor and its antecedent. Our approach

also requires a minimum threshold salience value for a candidate antecedent to be determined as an actual antecedent (our initial empirical study of the data suggests setting this number at 10).

Our pilot study shows that entities appearing as first NPs or last NPs in a sentence are preferred as antecedents. Thus, candidate antecedents at those positions will be given extra salience weights. For declarative sentences without a subject (i.e. declarative sentences which do not begin with an NP), the subject (i.e. the first NP) of the previous sentence is to be automatically assumed as the subject of the current sentence. At this preliminary stage, we are not interested in identifying full coreference chains, but only in identifying the antecedent of each anaphor. However, as we shall see below, the named entity will increase its weight by 10 as candidate antecedent for a_2 after having been determined as the correct antecedent of a_1 . This reflects the fact that in ASRS documents we often see a string of sentences with the majority of pronouns referring to one and the same entity, which is named only in the beginning of the string.

Our processing will depend heavily on genre-specific and domain-specific knowledge. The assignment of salience values to candidate antecedents will depend on four factors:

1. **position in the text** — all candidate antecedents which are *in the same sentence* as the anaphor are awarded a weight of 5; all candidate antecedents which are *in the preceding sentence* are awarded a weight of 2; all other candidate antecedents are awarded a weight of 0 for this factor
2. **position in the sentence** — all candidate antecedents which are *the first noun phrase* in the sentence in which they occur are awarded a weight of 5; all candidate antecedents which are *the last noun phrase* in the sentence are awarded a weight of 2; all other candidate antecedents are awarded a weight of 0 for this factor
3. **status of the antecedent** — the candidate antecedent which has already been identified as *the actual antecedent of the previous anaphor* is awarded a weight of 10; all other candidate antecedents are given a weight of 0 for this factor
4. **available semantic information** about the particular named entity which is considered as candidate — special preference will be given to named entities belonging to the four *antecedent categories* mentioned earlier (i.e. *AIRCRAFT*, *REPORTER*, *CREW*, or *CONTROL*). The amount of salience weight given to a candidate will reflect both *the frequency of occurrence of the exact head noun* (e.g. *RPTR*, *WDB*, *CAPT*, etc.), and *the frequency of occurrence of the antecedent category*. We will assume that the frequency of occurrence of the exact head noun ($Freq_n$) is a more important factor than the frequency of occurrence of the category ($Freq_c$). Thus, the salience weight for this factor will be computed using the following formula: $round(\frac{2 \cdot Freq_n + Freq_c}{10}, 0)$.

Let us now consider how our algorithm for anaphoric pronoun resolution would be applied to the ASRS example in (1), which contains three anaphoric pronouns. We have simplified the following discussion to only include potential

antecedents with scores of 2 or higher. According to our algorithm, we first calculate the set of candidate antecedents for the anaphor. So, for *THEM* in (1) we could have the candidate set $\{THE\ INFORMATION, THE\ COMPUTER, THE\ RESTRICTION, ATC, AN\ INTERSECTION\ XING\ ALT\}$. Then, the assignment function will give each candidate the following corresponding salience weights with respect to the four salience factors we defined: $\{2 + 0 + 0 + 0 = 2, 2 + 0 + 0 + 0 = 2, 2 + 2 + 0 + 0 = 4, 5 + 5 + 0 + 3 = 13, 5 + 0 + 0 + 0 = 5\}$. Only one of the values is above 10, therefore *ATC* is chosen as the actual antecedent of the anaphor *THEM*, which according to our analysis is correct. Similarly, the candidate set for *THEY* would be $\{ATC, AN\ INTERSECTION\ XING\ ALT, THE\ ALT\}$, with weights $\{5 + 5 + 10 + 3 = 23, 5 + 0 + 0 + 0 = 5, 5 + 2 + 0 + 0 = 7\}$, resulting in *ATC* also being chosen as the antecedent for *THEY*. Finally, the antecedent set for *HE* would be $\{ATC, AN\ INTERSECTION\ XING\ ALT, THE\ ALT, YOUR\ BEST, THE\ THING, A\ FIX\ 14\ MILES\ WEST\ OF\ MARTINSBURG\ VOR, THE\ CHARTS, THE\ COMPUTER\}$ with weights $\{0 + 5 + 10 + 3 = 18, 0 + 0 + 0 + 0 = 0, 0 + 2 + 0 + 0 = 2, 2 + 0 + 0 + 0 = 2, 2 + 0 + 0 + 0 = 2, 2 + 0 + 0 + 0 = 2, 2 + 0 + 0 + 0 = 2, 2 + 2 + 0 + 0 = 4\}$, again resulting in *ATC* being the antecedent.

4 Analysis and Conclusion

From our ASRS corpus, we chose 10 sentences containing anaphors that were representative of the kind of constructions found in the ASRS corpus. These sentences were then tagged with a part-of-speech tagger based on the Church algorithm [3], and then parsed using the CASS partial parser [2]. A manual analysis of our algorithm, when applied to this very small sample set, showed that it could correctly identify the antecedents for 9 out of 12 anaphors. Based on this analysis, our future research will consist of an implementation of the anaphora resolution algorithm, to be followed by a detailed evaluation of the algorithm, along with a comparison with other algorithms. Our detailed evaluation will also be done in the context of examining the effect of the accuracy of the tagging, and of the partial parsing processes.

Aviation safety reports contain a high degree of domain specific terminology, including abbreviations and acronyms, and as such it is not surprising that additional factors must be taken into consideration when processing instances of anaphora and extracting entities. We have suggested an approach where semantic information about the domain of application is one of the four factors determining the referent of an anaphoric pronoun. This semantic information, together with syntactic information concerning the position of the pronoun and its possible antecedent, and together with the significance of the anaphor (whether it has been previously used as an antecedent), are used in determining salience values for possible antecedents for a pronoun. What is interesting is that some traditional information, like lexical agreement and an explicit window based on the number of sentences separating an anaphor and its antecedent, do not play an explicit role in determining the anaphor/antecedent relationship. The result is

an algorithm that is able to find long distance antecedents, and find antecedents even when the anaphor and antecedent have conflicting agreement values.

Another avenue for further exploration is the automatic construction of the semantic model used by the anaphora resolution algorithm, particularly the tuning of weights for the different semantic classes. Research into this task would also allow the anaphora resolution and entity extraction techniques to be applied to other domains containing a restricted set of entities, which make use of acronyms and other specialized language structures (such as medical reports, insurance claims, patient records, etc.).

Acknowledgements. The authors would like to thank the referees for their observations and suggestions, and Devlan Nicholson, Gordon Tisher and the rest of the development team at Axonwave Software for their assistance in accessing and processing the data found within the ASRS database. This research is partially supported by a Discovery Grant from the Natural Science and Engineering Council of Canada.

References

1. The NASA Aviation Reporting System Incident Database, Windows v1.0.6 Data Current Q2 1999, ARG/US Aeroknowledge, Doylestown, PA. (1999).
2. Abney, S.: Part-of-Speech Tagging and Partial Parsing, In Young S. and Bloothoof G. (eds), *Corpus-Based Methods in Language and Speech Processing*, Chap. 4, pp. 118-136, Kluwer (1997)
3. Church, K.: A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proc. 2nd Conference on Applied Natural Language Processing*, pp. 136-143, ACL, (1988).
4. Dimitrov, M.: A light-weight approach to coreference resolution for named entities in text. MSc thesis, Department of Information Technologies, University of Sofia: Sofia, Bulgaria. (2002)
5. Lappin, S., Leass, H. J.: An algorithm for pronominal anaphora resolution. *Computational Linguistics* **20**(4): 535-561 (1994)
6. Mitkov, R.: Robust pronoun resolution with limited knowledge. In *Proceedings of COLING'98/ACL'98*, Montreal, Quebec, Canada. (1998) 869-875
7. Srinivas, B., Baldwin, B.: Exploiting super tag representation for fast coreference resolution. In *Proceedings of the International Conference on NLP+AI/TAL+AI 96*, Moncton, NB, Canada. (1996)
8. Terada, A., Tokunaga, T.: Corpus Based Method of Transforming Nominalized Phrases into Clauses for Text Mining Application. *IEICE Trans. Inf. & Syst.* (**E86-D**)9, (2003).
9. Williams, S., Harvey, M., Preston, K.: Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing. (1996) In *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC)*, Lancaster, UK. (1996) 441-456