

Pronominal Anaphora Resolution for Unrestricted Text

Anna Kupś^{*}, Teruko Mitamura[†], Benjamin Van Durme[†], Eric Nyberg[†]

^{*}Polish Academy of Sciences, Institute of Computer Science
Ordonia 21, 01-237 Warsaw, Poland
aniak@ipipan.waw.pl

[†]Carnegie Mellon University, Language Technologies Institute
5000, Forbes Ave., Pittsburgh, PA, 15213, USA
{teruko, vandurme, ehnl}@cs.cmu.edu

Abstract

The paper presents an anaphora resolution algorithm for unrestricted text. In particular, we examine portability of a knowledge-based approach of (Mitamura et al., 2002), proposed for a domain-specific task. We obtain up to 70% accuracy on unrestricted text, which is a significant improvement (almost 20%) over a baseline we set for general text. As the overall results leave much room for improvement, we provide a detailed error analysis and investigate possible enhancements.

1. Introduction

Pronominal anaphora resolution is one of the crucial problems in text understanding. It deals with identifying elements which are coreferent with a pronoun in the text and has been applied to NLP tasks such as information extraction, question answering, text summarization or machine translation. Multiple approaches to anaphora resolution include: knowledge- and linguistic-based (Carbonell and Brown, 1988; Lappin and McCord, 1990), machine learning (Aone and Bennett, 1995; Soon et al., 2001) or statistical techniques (Ge et al., 1998), applied both to restricted and unrestricted text.

In this paper, we present an anaphora resolution algorithm for an open-domain, unrestricted text. As a starting point, we take a knowledge-based approach to anaphora resolution (Mitamura et al., 2002) adopted for domain-specific machine translation. In order to accommodate this technique to our current task, several modifications are necessary. First, we incorporate various robust processing tools to compensate for the lack of an exhaustive lexicon in the open-domain task. Second, as we deal with unrestricted text, we extend the basic algorithm to include all types of pronouns ((Mitamura et al., 2002) consider only the pronouns *it*, *they* and *them*) and we employ general linguistic principles in addition to heuristics. We conclude with an evaluation of the algorithm and remarks on the obtained results.

2. Anaphora Resolution Process

The general coreference resolution process presented in this paper consists of the following steps: text processing (tokenization, stemming, incorporating lexical information, parsing), assignment of agreement features to all noun phrases (NPs), identification of candidate antecedents and, finally, pruning of candidates. Details of this process are described below.

2.1. Text Processing

For text processing, we combine several publicly available tools and resources in order to obtain a single robust tool. We use the RASP toolkit (Briscoe and Carroll, 2002)

for text segmentation and tokenization, as well as for getting the part of speech (POS) and the stem. We employ the Link grammar parser (Grinberg et al., 1995) for the assignment of grammatical functions as it performed better than the RASP parser in our initial tests. We use WordNet (Fellbaum, 1998) for lexical lookup. We do not employ WordNet for stemming and getting POS as the database covers four syntactic categories only. Initially, we incorporated also the BBN named entity tagger (BBN, 2000) which recognizes about 20 categories. However, its performance on our texts was rather poor and we could not retrain it, so we decided not to use it here. We partly compensate for the lack of a named entity tagger by using the rich CLAWS2 POS tagset (used in RASP) and lexical lookup to recognize people names, organizations and locations.

2.2. Feature Assignment

In order to identify possible antecedents in unrestricted text, we use 4 agreement features as in (Siddharthan, 2003): person, number, gender and animacy. The agreement features and their possible values are presented in (1).

(1)

FEATURE	POSSIBLE VALUES
PERSON	1, 2, 3
NUMBER	singular, plural
GENDER	male, female, neuter
ANIMATE	true, false

Additionally, we use the ‘all’ value if more than one value can be used, e.g., to mark unisex gender of first names, such as *Chris*, or last names.

Resources and a general strategy we employ for feature assignment are presented below:

- **PERSON:** 1st and 2nd person pronouns are assigned 1st and 2nd person, respectively; all other nouns are 3rd person;
- **NUMBER:** number is determined based on POS tags and WordNet; if a noun is tagged plural, its number is plural; if the noun is tagged singular, its NUMBER is singular unless it is a hypernym of the *group* synset in WordNet, which assigns plural;

- GENDER: we used three main techniques to assign gender: 1) heuristic rules for titles, e.g., *Mr* or *Mrs*, unambiguously specify gender¹, and for acronyms (sequences of two or more capital letters, possibly with dots) which are assigned neuter; 2) name and location lists (about 12000 female and 13000 male first names, 134000 last names and 163000 locations): nouns POS-tagged as proper names are verified in the lexicons and assigned gender; the name lexicons are checked first, then the location lexicon; 3) WordNet: for common nouns, if the noun is a hypernym of *male* (*female*) synset, it is assigned male (female) gender;²

If none of these techniques can assign gender, the noun is considered neuter.

- ANIMATE: nouns assigned male (female) gender, people names and nouns which are hypernyms of *animate thing*, *biological group* or *social group* are animate, locations are inanimate; other nouns are considered inanimate.

Once the agreement features are assigned, we create a list of possible antecedents by checking agreement with the pronoun. We specified the following rules of agreement:

- strict agreement: all agreement values of the pronoun and a candidate antecedent have to be identical;
- relaxed gender agreement: if PERSON, NUMBER and ANIMATE have the same values, a unisex candidate antecedent can agree with either male or female pronoun; this enables a coreference between a last name and a personal pronoun, e.g., *Jones* vs. *(s)he*;
- relaxed number and animacy agreement: if the candidate and the pronoun have the same GENDER and PERSON values, either NUMBER or ANIMATE values of the candidate and the pronoun have to agree but agreement of the two values is not obligatory; this relaxation allows us to refer to *police* by either a singular (*it*) or a plural (*they*) pronoun, cf. (Siddharthan, 2003).

2.3. Candidate Selection

After creating a list of candidate antecedents, we have to trim this list so that the unique antecedent is selected. As an initial disambiguation filter, we employ general linguistic principles:

- reflexives must be coreferent with an NP argument of the same verb;
- possessive pronouns have to be coreferent with a preceding NP;

¹We used the following titles: *Mrs*, *Miss*, *Madame*, *Madam*, *Lady*, *Mlle*, *Medemoiselle*, *Ms* and *Mr*, *Mister*, *Monsiuer*, *Sir*, *Lord*.

²We have also run an experiment to learn gender-specific nouns from lexical resources by using a technique presented in (Thelen and Riloff, 2002). We acquired, however, only proper names which were already comprised in the lexicons mentioned above.

- personal pronouns cannot be coreferent with an NP in the same clause.

If there is more than one candidate left, we apply anaphora resolution heuristics of (Mitamura et al., 2002). The heuristics are applied sequentially, i.e., if one heuristic fails, another one is tried until a single candidate is left. The heuristics are tried in the order specified below³:

- prefer an antecedent that is also an anaphor;
- prefer an antecedent that is not a proper noun;
- if two antecedents *NP1* and *NP2* occur in the form *NP1 of NP2*, prefer *NP1* unless *NP2* is one of the nouns *type*, *part*, *length*, *size*; in this case prefer *NP2*;
- prefer antecedents that are arguments of a word which has the same stem as the word the pronoun is an argument of;
- prefer antecedents that are arguments of a word which has the same part of speech as the word the pronoun is an argument of;
- prefer antecedents that fill the same grammatical function as the pronoun;
- prefer antecedents that have a quantifier, determiner, possessor, or are a named entity;
- prefer antecedents that have a definite determiner;
- prefer the most recent antecedent.

In our implementation, antecedents can be searched back up to the beginning of the paragraph: if the current sentence does not contain an appropriate candidate, the previous sentence is checked, and so on.

3. Experiments

In order to test the algorithm on unrestricted text, we made some experiments using a set of documents from the Center for Nonproliferation Studies (CNS) made available under the AQUAINT program. We have selected a sub-corpus associated with al-Qa'ida activities and then we extracted sentences containing male pronouns, i.e., *he*, *him*, *his*, *himself*. These sentences contained also other pronouns, e.g., *her* or *our*, the total of 361 pronouns.

In order to set a baseline, we ran a basic version of the algorithm using the last heuristic only, i.e., the most recent agreeing candidate was always selected. We distinguish two types of antecedents: a relative antecedent (in case the candidate antecedent is a pronoun, the reference of the pronoun is not resolved) and an absolute antecedent (if the candidate antecedent is a pronoun, its reference has to be resolved). The baseline test gave us 50.1% accuracy for a relative antecedent and 55.9% for an absolute antecedent.

³Since the analysis of conjunctions is problematic in Link, we did not apply the heuristic which takes conjunctions as antecedents. As indicated in (3) below, this simplification was not crucial in the final evaluation.

We tested the full algorithm in two main experiments: with and without WordNet. In the latter test, we used lexicons and the default feature assignments (see sec. 2.2.) and obtained 67.3% accuracy. In this test, only the reference to the relative antecedent was checked.

In tests with WordNet, we checked accuracy of the algorithm both for relative and absolute antecedents. In these tests, we obtained 70.0% (relative antecedent) and 63.9% (absolute antecedent) accuracy. The general test results are summarized in (2).

TEST	ACCURACY
baseline:	
relative antecedent	50.1%
absolute antecedent	55.9%
without WordNet:	
relative antecedent	67.3%
with WordNet:	
relative antecedent	70.0%
absolute antecedent	63.9%

As the above results indicate, the full algorithm performs significantly better than the baseline: 17.2% without WordNet and up to 19.9% with WordNet. On the other hand, the difference between versions with and without WordNet is minimal: there is only 2.7% improvement. In order to better understand these results, we further examined them and made a detailed error analysis. The table in (3) presents data for the WordNet test with a relative antecedent.

PROBLEM	#	%
lexicon	28	25.9%
heuristics	28	25.9%
world knowledge	9	8.3%
expletives	7	6.5%
conjunction	7	6.5%
wrong POS	7	6.5%
distance	6	5.5%
feature assignment	5	4.6%
backreference	4	3.7%
quotations	4	3.7%
ellipsis	1	0.9%
parsing	1	0.9%
agreement	1	0.9%
total	108	100%

As (3) shows, the majority of problems were related to the lexicon (unrecognized last names, people names misinterpreted as locations or vice versa) and heuristics, i.e., the correct antecedent was among considered candidates but the heuristics selected a different one. Another important issue (overall 28 errors) was processing (unrecognized expletives, conjunctions, wrong POS, incorrect feature assignment and/or agreement, parsing). A discourse analysis would have to be incorporated to deal with another 9 errors: backreferences, quotations and NP-ellipsis. A similar problem is lack of world knowledge, e.g., required to assign gender to professions which are not gender specific, e.g., *biologist*, or nationalities. Finally, more of a technical

PRONOUN CLASS	ACCURACY	
	relative antecedent	absolute referent
HE	76.7% (181/236)	69.1% (163/236)
SHE	100% (2/2)	100% (2/2)
IT	38.6% (17/44)	29.5% (13/44)
THEY	42.8% (18/42)	40.5% (17/42)
I	90.9% (10/11)	90.9% (10/11)
WE	95.4% (21/22)	95.4% (21/22)
YOU	100% (4/4)	100% (4/4)

Figure 1: Accuracy for different pronoun classes.

issue is the distance between the antecedent and the pronoun. These errors might have been corrected should we restricted the search, e.g., to 2 previous sentences.

As mentioned above, the original algorithm of (Mitamura et al., 2002) dealt with the pronouns *it*, *they* and *them* only. Hence, we examined the accuracy of the current algorithm for specific pronoun classes. The results of the WordNet test are presented in Fig. 1 (each class comprises all forms of the pronoun given in the table). The results in Fig. 1 show that the accuracy for the pronouns IT and THEY is very low in comparison to those in (Mitamura et al., 2002): 88.5% (IT) and 92.8%⁴ (THEY)). Given that in unrestricted text gender and animacy, required by the algorithm, are assigned robustly and we use heuristics proposed for a specific domain and application, the low numbers should be less surprising. On the other hand, due to sparse data only the results for HE can be considered reliable.

4. Conclusions

The paper explores a possibility of reusing an anaphora resolution algorithm, originally proposed for a domain-specific task, and extending it to unrestricted text. In our final tests we obtain 70% accuracy on general texts. Although these results are not very impressive with respect to the original algorithm, which achieves almost 90% accuracy, they show a significant improvement (almost 20%) over the baseline we set for this paper. However, as the current approach heavily relies on heuristics proposed for a specific application and requires rich knowledge (lexicons, parsing) which cannot be fully reliably provided for general texts, many problems of the present algorithm will remain unresolved. An alternative would be to use the current algorithm as a basis to assign features and employ machine learning techniques to learn resolution rules. This topic is left for future study.

5. Acknowledgements

This work was supported in part by the Advanced Research and Development Activity (ARDA) under AQUAINT contract MDA904-01-C-0988. We would like to thank Curtis Huttenhower, for his work on integrating the tools we used for text analysis, as well as three anonymous reviewers and Adam Przepiórkowski for useful comments on earlier versions of this paper.

⁴This is a joint accuracy for *they* and *them* pronouns: (24+41)/(24+46) which is not reported in their paper.

6. References

- Aone, Ghinatsu and Scott William Bennett, 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of 33rd ACL Annual Meeting*.
- BBN, 2000. *IdentiFinder User Manual*. BBN Technologies.
- Briscoe, Ted and John Carroll, 2002. Robust accurate statistical annotation of general text. In *3rd LREC Proceedings*.
- Carbonell, Jaime and Ralph Brown, 1988. Anaphora resolution: A multi-strategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics*.
- Fellbaum, Christiane, 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Ge, N., J. Hale, and E. Charniak, 1998. A statistical approach to anaphora resolution. In *Proceedings of 6th Workshop on Very Large Corpora*.
- Grinberg, D., J. Lafferty, and D. Sleator, 1995. A robust parsing algorithm for link grammars. In *Proceedings of the Fourth International Workshop on Parsing Technologies*.
- Lappin, Shalom and Mike McCord, 1990. Anaphora Resolution in Slot Grammar. *Computational Linguistics*, 16(4):197–212.
- Mitamura, Teruko, Eric Nyberg, Enrique Torrejon, David Svoboda, Annelen Brunner, and Kathryn Baker, 2002. Pronominal anaphora resolution in the KANTOO multilingual machine translation system. In *Proceedings of TMI 2002*.
- Siddharthan, A., 2003. Resolving pronouns robustly: Plumbing depths of shallowness. In *Proceedings of EACL 2003*.
- Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim, 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.
- Thelen, Michael and Ellen Riloff, 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of EMNLP 2002*.