How far are we from (semi-) automatic annotation of anaphoric links in corpora?

(From knowledge-poor pronoun resolution to a tentative look at the (semi-) automatic annotation of pronoun-antecedent pairs in corpora)

Ruslan Mitkov
School of Languages and European Studies
University of Wolverhampton
Stafford Street
Wolverhampton WV1 1SB
United Kingdom
R.Mitkov@wlv.ac.uk

Abstract

The paper raises for discussion a proposal for the semi-automatic annotation of pronoun-antecedent pairs in corpora. The proposal is based on robust knowledge-poor pronoun resolution followed by post-editing.

The paper is structured as follows. The introduction comments on the fact that automatic identification of referential links in corpora has lagged behind in comparison with similar lexical, syntactical and even semantic tasks. The second section of the paper outlines the author's practical and robust knowledge-based approach to pronoun resolution which will subsequently be put forward as the core of a larger architecture proposed for the automatic tagging of referential links. Section 3 briefly presents other related knowledge-poor approaches, while section 4 discusses the limitations and advantages of the practical approach. The main argument of the paper is to be found in section 5, where we present the idea of developing a semi-automatic environment for annotating anaphoric links and outline the components of such a program. Finally, the conclusion looks at the anticipated success rate of the approach.

1. Introduction

Annotated anaphoric links in language corpora play an important role in teaching and research. Research roles may include investigation into the distribution of the different types of anaphors, or into the location or distance of the antecedent, and also development of rules or heuristics for anaphora resolution and the testing of anaphora-related hypotheses/theories on the basis of numerous real-life examples.

Annotation of referential links has not yet been able to benefit from the level of automation enjoyed by its lexical, syntactical and semantic "counterparts". Part-of-speech tagging has shown remarkable accuracy (99.2% see [Voutilaen 95]), robust parsing in corpora has delivered very good results and even word sense tagging has reported a considerable improvement. However, "referential tagging" has not been fully explored (and developed) and probably this is due, no doubt, to the complexity of automatic anaphora resolution.

One of the best known tools for anaphoric annotation is XANADU - an X-windows interactive editor written by Roger Garside, which offers the user an easy-to-navigate environment for manually marking pairs of anaphors-antecedents ([Fligelstone 92]). Manual annotation, however, imposes a considerable demand on human time and labour.

In this paper we put forward the idea of incorporating a practical, knowledge-poor approach to anaphora resolution ([Mitkov 97]) within a larger architecture for rough automatic referential annotation of corpora. At this stage our proposal deals with pronominal anaphora only and "rough annotation" implies that a follow-up manual correction would be necessary. Nevertheless, we believe that this partial solution brings us somewhat closer to the automatic annotation of all types of anaphoric links.

2. Outline of our practical pronoun resolution approach

With a view to avoiding complex syntactic, semantic and discourse analysis (vital for real-world applications), we have developed a practical approach to pronoun resolution ([Mitkov

97]) which does not parse and analyse the input in order to identify antecedents of anaphors. It makes use of only a part-of-speech tagger, plus simple noun phrase rules (sentence constituents are identified at the level of noun phrase at most) and operates on the basis of antecedent-tracking preferences (referred to hereafter as "antecedent indicators").

2.1 Antecedent indicators

Our empirical study (restricted to computer and hi-fi technical manuals) enabled us to develop efficient preferences for antecedent tracking in this sublanguage/genre. (We studied more than 400 different documents which had been hand-annotated; referential links were marked by human experts). These antecedent indicators are described in detail in [Mitkov 97]; we shall outline here those which are most frequently used as a supplement to gender and number agreement.

• *Term preference*
NPs representing terms in the field are more likely to be the antecedent than NPs which are not terms (scores 1 if the NP is term and 0 if not).

• *Verb preference*
If the verb is a member of the Verb_set = {discuss, present, illustrate, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal, cover}, then consider the first NP following it as the preferred antecedent (scores 1 and 0). (Empirical evidence suggests that because of their salience, the verbs listed above are particularly likely candidates)

These two preferences can be illustrated by the example:

This table shows a minimal configuration$_i$; it$_i$ does not leave much room for additional applications or other software for which you may require additional swap space.

• *Lexical reiteration*
Lexically reiterated items are likely candidates for antecedent (scores 2 if the NP is repeated within the same paragraph twice or more, 1 if repeated once and 0 if not). Lexically reiterated items include repeated synonymous noun phrases which may often be preceded by definite articles or demonstratives.

• *Section heading preference*
If a noun phrase occurs in the heading of the section, part of which is the current sentence, then consider it as the preferred candidate for the antecedent (1, 0).

• *Collocation pattern preference*
This preference is given to candidates which have an identical collocation pattern with a pronoun. The collocation implemented here is restricted to the pattern "noun/pronoun, verb" or "verb, noun/pronoun" (2, 0). (owing to lack of syntactic information, this preference is somewhat weaker than the collocation preference described in [Dagan & Itai 90] and suggested subsequently in our procedure for semi-automatic annotation)

Press the key$_i$ down and turn the volume up... Press it$_i$ again.

• *Referential distance*
In complex sentences, noun phrases in the previous clause[1] are the best candidate for the antecedent of an anaphor in the subsequent clause, followed by noun phrases in the previous sentence, then by nouns situated 2 sentences further back and finally nouns 3 sentences further back (2, 1, 0, -1).

For anaphors in simple sentences, noun phrases in the previous sentence are the best candidate for antecedent, followed by noun phrases situated 2 sentences further back and finally nouns 3 sentences further back (1, 0, -1)

• *"Non-prepositional" noun phrases*
A "pure", "non-prepositional" noun phrase is given a higher preference than a noun phrase which is part of a prepositional phrase (0, -1)

Insert the cassette$_i$ into the VCR making sure it$_i$ is suitable for the length of recording.

• *Non-candidates*
Constituents introduced by an indefinite article and constituents introduced by determiners such as "another", "other" (-1, 0)

[1] Currently we use simple heuristics for identifying clauses in a complex sentence

Each of the antecedent indicators is assigned a score with a value $\in \{-1, 0, 1, 2\}$. These scores have been determined experimentally on an empirical basis and are constantly being updated. Top symptoms like "lexical reiteration" assign score "2" whereas non-candidates are given a negative score of "-1". We should point out that the antecedent indicators are preferences and not absolute factors. There are cases where an antecedent indicator does not "point" to the correct antecedent. For instance, in the sentence "Insert the cassette into the VCR $_i$ making sure it$_i$ is turned on", the indicator "non-prepositional noun phrases" would give a "wrong" contribution. Within the framework of all preferences (antecedent indicators), however, the right antecedent is still very likely to be tracked down - in the above example, the "non-prepositional noun phrases" heuristics would be overturned by the "collocational preference" one.

2.2 Informal description of the algorithm

The algorithm for pronoun disambiguation can be described informally as follows:

1. Examine the current sentence and the two preceding sentences (if available). Look for noun phrases[2] only to the left of the anaphor[3]
2. Select from the noun phrases identified only those which agree in gender and number[4] with the pronominal anaphor and group them as a set of potential candidates
3. Apply the antecedent indicators to each potential candidate and assign scores; the candidate with the highest score is proposed as antecedent.

For an illustration as to how the approach operates see ([Mitkov 97]).

2.3 Evaluation

For practical reasons, the approach presented does not incorporate syntactic and semantic information (other than a list of domain terms) and it is not realistic to expect its performance to be as good as an approach which makes use of syntactic and semantic knowledge in terms of constraints and preferences. The lack of syntactic information, for instance, means giving up subject preference (or on other occasions object preference, see [Mitkov 94a]) which could be used in center tracking. Syntactic parallelism, useful in discriminating between identical pronouns on the basis of their syntactic function, also has to be forgone. Lack of semantic knowledge rules out the use of verb semantics and semantic parallelism. The preliminary evaluation, however, shows that less is lost than might be feared.

Several documents (user's guides), with an overall length of 40 000 words, served as an initial evaluation corpus. The average success rate was 86%. While the test corpus contained the pronouns "he", "she" and "they", most pronouns were "it" (about 92%). The approach had a very high success rate with sentences which contained one pronoun only (above 90%) but failed in a few paragraphs which contained an abundance of "it"s, 2 (or more) in a sentence, with "it" referring in turn to different antecedents. In these examples, however, a frequent shift of center was observed and, in our view, they were not written in a natural style.

A recent test with Computer Science textbook inputs showed a preliminary accuracy rate of above 80%.

3. Other knowledge-poor approaches

The approaches proposed by Nasukawa ([Nasukawa 94]), Dagan & Itai ([Dagan & Itai 90]) and Kennedy & Boguraev ([Kennedy & Boguraev 96]) address anaphor resolution in a "knowledge-poor" way: the first approach takes into consideration heuristic preferences, the second uses frequency of collocational patterns and the third operates without a parser resorting to salience factors.

3.1 Nasukawa's knowledge-independent approach

T. Nasukawa ([Nasukawa 94]) describes a simple approach which uses intersentential information extracted from a source text in order to improve the accuracy of pronoun resolution. He suggests that collocation patterns (modifier-modifee relationships) can be used to determine whether a candidate for antecedent can modify

[2] A sentence splitter would already have segmented the text into sentences, a POS tagger would already have determined the parts of speech and a simple phrasal grammar would already have detected the noun phrases

[3] In this project we do not treat cataphora; non-anaphoric "it" occurring in constructions such as "It is important", "It is necessary" is eliminated by a "referential filter"

[4] Note that this restriction may not always apply in languages other than English (e.g. German); on the other hand there are certain collective nouns in English which do not agree in number with their antecedents (e.g. "government", "team", "parliament" etc. can be referred to "they"; equally some plural nouns (e.g. "data") can be referred to by "it") and are exempted from the agreement test

the modifee of a pronoun. Nasukawa also finds (similarly to ([Mitkov 93])) that the frequency of preceding noun phrases with the same lemma as the candidate noun phrase may be an indication for preference. Moreover, he suggests a heuristic rule favouring subjects over objects (compare [Mitkov 93] where this preference is sublanguage-based).

Each of the collocational, frequency or syntactic preferences gives its "preference value"; these values are eventually summed up. The candidate with the highest value is picked up as the antecedent.

As an evaluation corpus Nasukawa uses 1904 consecutive sentences (containing altogether 112 third-person pronouns) from eight chapters of two different computer manuals. His algorithm handles the pronoun "it" and has been reported to select a correct antecedent in 93.8% of cases.

3.2 Dagan & Itai's corpus-based approach

I. Dagan and A. Itai ([Dagan & Itai 90]) report on a statistical approach for disambiguating pronouns; this is an alternative solution to the expensive implementation of full-scale selectional constraints knowledge. They perform an experiment to resolve references of the pronoun "it" in sentences randomly selected from the corpus.

In their statistical model, co-occurrence patterns observed in the corpus were used as selectional patterns. Candidates for antecedent were substituted for the anaphor and only those candidates appearing in frequent co-occurrence patterns were approved of.

Dagan and Itai report an accuracy rate of 87% for the sentences with genuine "it" anaphors (sentences in which "it" is not an anaphor have been manually eliminated). It should be pointed out that the success of this experiment depends on the parser used (in this case K. Jennsen's PEG parser).

3.3 Kennedy & Boguraev's approach without a parser

In a recent paper, Kennedy and Boguraev ([Kennedy & Boguraev]) describe an anaphor resolution approach which is a modified and extended version of that developed by Lappin and Leass ([Lappin & Leass 94]). Their system does not require "in-depth, full" syntactic parsing but works from the output of a part of speech tagger, enhanced only by annotations of grammatical function of lexical items in the input text stream.

The basic logic of their algorithm parallels that of Lappin and Leass's algorithm. The determination of disjoint reference, however, represents a significant point of divergence the two. Lappin and Leass's relies on syntactic configurational information, whereas Kennedy and Boguraev's, in the absence of such information, relies on inferences from grammatical function and precedence.

After the morphological and syntactic filters have been applied, the set of discourse referents that remain is subjected to a final salience evaluation. The candidate with highest salience weighting is determined to be the actual antecedent; in the event of a tie, the closest candidate is chosen. The approach works for both lexical anaphors (reflexives and reciprocals) and pronouns.

Evaluation reports 75% accuracy but it should be pointed out that the results were obtained from a wide range of texts/genres: the evaluation was based on a random selection of genres, including press releases, product announcement, news stories, magazine articles, and other World Wide Web documents.

4. Limitations and advantages of the practical approach

We must admit that the practical approach has been tested mainly on a specific genre: computer and hi-fi manuals. It also appears that some of the rules are more genre-specific than others (e.g. "verb preference" and "noun preference"). Therefore, we cannot claim that an equally high level of accuracy would be guaranteed in other genres.

In addition, even though our preliminary results seem to be better than Kennedy and Boguraev's (75%), there is no ground for any real comparison since (i) our evaluation tests are not extensive enough and are of a preliminary nature and (ii) their evaluation is based on a random selection of genres, whereas our method has been applied to a single text genre.

The practical approach presented has been developed recently and is subject to further research and improvements. In particular, we plan to enhance the accuracy of the initial score of each symptom by collecting more empirical evidence and to integrate all the antecedent indicators into a uniform and comprehensive probabilistic model.

On the other hand, the main advantage of the practical approach lies in its independence of syntactic, semantic, domain and real-world knowledge, which makes it not only cheaper to implement but also appropriate for applications

in corpora. Thus we see the pronoun resolution approach as one of the components of a more general methodology aiming to offer a way forward in the automatic annotation of anaphoric links in corpora.

5. Proposed methodology for semi-automatic annotation

Further to our comments in section 4, we would like to propose the development of a semi-automatic procedure for annotating pronominal anaphora in corpora. Such a procedure would speed up the manual marking of pronoun-antecedent pairs. The semi-automatic annotation editor would be practically based on our pronoun resolution approach made more "robust" by a "super" POS tagger and by corpus-based collocation patterns. The process of annotation will consist of the following stages:

a) *sentence splitting*
The first stage will be to segment the input into sentences by identifying their boundaries.

b) *"super" part-of-speech tagging*
We plan to use the so-called super part of speech taggers which (i) determine automatically lexical categories, (ii) provide further lexical information (e.g. gender, number) and (iii) identify the syntactic function of each part-of-speech unit (e.g. subject, object etc.) ([Voutilainen et al. 1992], [Karlsson et al. 95]).

c) *gender and number agreement*
Once the noun phrases (see footnote 2) in a sentence have been identified, agreement constraints for filtering NP candidates in current and preceding sentences will be activated. Certain (e.g. collective) nouns[5] will not be subject to such constraints (see footnote 4).

d) *corpus-based collocation patterns*
Possible antecedents will be substituted for the anaphors and the frequency of the new constructions will be calculated in corpora. Higher weightings will be assigned to NPs which occur more frequently in the same syntactic function as the anaphor (e.g. in combination with a certain verb or subject/object).

e) *antecedent indicators*
The antecedent indicators (as described above) will be used for the final weighting of the candidates and for proposing the antecedent. The candidate with the highest overall score after stages d) and e) will be picked up as the most likely antecedent.

We are aware of the fact that this robust pronoun resolver is unlikely to produce 100% accuracy. Therefore, we envisage the development of a post-editing environment. Anaphors and allocated antecedents will be highlighted with the user accepting or correcting them.

If future comprehensive evaluation suggests that the success of the new approach is restricted to certain genres only, it would be worthwhile to consider using other knowledge-poor approaches (e.g. [Kennedy & Boguraev 96]), which have proved their efficiency, within the framework suggested.

Last but not least, another promising option would be to enhance the framework proposed with a robust parser or even better to select an existing robust platform for pre-processing the input morphologically and syntactically (one such platform which we are currently looking at, is GATE ([Cunningham et al. 96]).

6. Conclusion and expectations

Lancaster University, British Telecom and University of Wolverhampton are interested in setting up a project for the semi-automatic annotation of pronominal anaphora. (In addition, collaboration with the University of Sheffield regarding possible employment of GATE is favoured by both sides). At this stage it is difficult to predict the performance of our robust anaphor resolver, but our expectation is for a level of accuracy in the region of 90%. It should be noted that pronoun disambiguation approaches working exclusively on corpus-based collocation patterns have already reported accuracy levels of 87% ([Dagan & Itai 90]), whereas our practical pronoun resolution approach reports a level of accuracy of 86%. However, we should draw attention to the facts that our evaluation is based on the use of a "standard" part-of-speech tagger and not a "supertagger" (which would additionally give information on syntactic functions) and that we have not benefited from the output of a robust parser or from the highly indicative corpus-based syntactical patterns. If our expectation were right, limited post-editing would mean a considerable gain in speed compared with the existing manual annotation methods.

[5] We are currently preparing an exhaustive list of such nouns

Acknowledgements

I would like to thank Atro Voutilainen and Marco Antonio Da Rocha for their comments on the position paper.

References

[Cunningham et al. 96] H. Cunningham, Y. Wilks, R. Gaizauskas - *Software infrastructure for Language Engineering.* Proceedings of the workshop "Language Engineering for Document Analysis and Recognition", Brighton, 2 April, 1996

[Dagan & Itai 90] I. Dagan, A. Itai - *Automatic processing of large corpora for the resolution of anaphora references.* Proceedings of the 13th International Conference on Computational Linguistics, COLING'90, Helsinki, 1990

[Fligelstone 92] S. Fligelstone - *Developing A Scheme For Annotating Text To Show Anaphoric Relations.* In Leitner G (ed) Proceedings of the 11th International Conference on English Language Research on Computer Corpora. Berlin, Mouton de Gruyter

[Karlsson et al. 95] F. Karlsson, A. Voutilainen, J. Heikkila, A. Antilla - *Constraint grammar: a language-independent system for parsing free text.* Mouton de Gruyter, Berlin/New York, 1995

[Kennedy & Boguraev 96] C. Kennedy, B. Boguraev - *Anaphora for everyone: pronominal anaphora resolutionwithout a parser.* Proceedings of the 16th International Conference on Computational Linguistics COLING'96, Copenhagen, Denmark, 5-9 August 1996

[Lappin & Leass 94] Sh. Lappin, H. Leass - *An algorithm for pronominal anaphora resolution. Computational Linguistics*, 20(4), 1994

[Mitkov 93] R. Mitkov - *Anaphora resolution and natural language understanding*, Unpublished internal report, Universiti Sains Malaysia, Penang, 1993

[Mitkov 94] R. Mitkov- *An integrated model for anaphora resolution.* Proceedings of the 15th International Conference on Computational Linguistics COLING'94, Kyoto, Japan, 5-9 August 1994

[Mitkov 97] Mitkov R. - *Pronoun resolution: the practical alternative.* In S. Botley, T. McEnery (Eds) "Discourse Anaphora and Anaphor Resolution". University College London Press, 1997

[Nasukawa 94] T. Nasukawa - *Robust method of pronoun resolution using full-text information.* Proceedings of the 15th International Conference on Computational Linguistics COLING'94, Kyoto, Japan, 5-9 August 1994

[Stys & Zemke 95] M. Stys, S. Zemke - *Incorporating discourse aspects in English - Polish MT: towards robust implementation.* Proceedings of the international conference "Recent Advances in Natural Language Processing", Tzigov Chark, Bulgaria, 14-16 September 1995

[Voutilainen 95] A. Voutilainen - A *syntax-based part-of-speech tagger.* Proceedings of the 7th conference of the European Chapter of the Association for Computational Linguistics, Dublin, 1995

[Voutilainen et al. 92] A. Voutilainen, J. Heikkila, A. Antilla - *A constraint grammar of English: a performance-oriented approach.* University of HElsinki, Publication No. 21, Helsinki, 1992