Using word similarity lists for resolving indirect anaphora

Caroline Gasperin and **Renata Vieira**
PIPCA - Unisinos
São Leopoldo, Brazil
{caroline,renata}@exatas.unisinos.br

Abstract

In this work we test the use of word similarity lists for anaphora resolution in Portuguese corpora. We applied an automatic lexical acquisition technique over parsed texts to identify semantically similar words. After that, we made use of this lexical knowledge to resolve coreferent definite descriptions where the head-noun of the anaphor is different from the head-noun of its antecedent, which we call indirect anaphora.

1 Introduction

In this work we investigate the use of word similarity list for treating coreference, especially the cases where the coreferent expressions have semantically related head nouns (instead of same head nouns), which we call indirect anaphora.

We applied a lexical acquisition technique (Gasperin, 2001) over Portuguese parsed corpora to automatically identify semantically similar words. After that, we made use of this lexical knowledge to resolve the coreferent definite descriptions where the head-noun of the anaphor is different from the head-noun of its antecedent.

Previous work on anaphoric resolution of English texts has used acquired lexical knowledge in different ways, examples are (Poesio et al., 2002; Schulte im Walde, 1997; Bunescu, 2003).

This paper is organised as follows. The next section explain our notion of indirect anaphora. Section 3 details the tools and techniques used to the construction of our lexical resource. Section 4 presents our heuristic for solving the indirect anaphors on the basis of such resource. Section 5 details the corpus we are using for evaluating the proposed heuristics. Section 6 reports the implementation of the heuristic and in Section 7 we present our experiments over Portuguese annotated corpora. In Section 8 we discuss our results and compare them to previous works. Finally, Section 9 presents our concluding comments.

2 Indirect anaphora

Coreference has been defined by (van Deemter and Kibble, 2000) as the relation holding between linguistic expressions that refer to the same extralinguistic entity. A slightly different discourse relation is anaphora. In an anaphoric relation the interpretation of an expression is dependent on previous expressions within the same discourse (in various ways). Therefore, an anaphoric relation may be coreferent or not. An expression may be anaphoric in the strict sense that its interpretation is only possible on the basis of the antecedent, as it is in general the case of pronouns in written discourse. On the other hand, it might be just coreferent, in the sense that the entity has been mentioned before in the text.

In this work, we focus on the expressions that are anaphoric and coreferent, and restricting even more, just the indirect cases, when the antecedent head-noun and the anaphor head-noun are not same but semantically related.

To clarify what we mean by indirect anaphora, we detail the classification we adopted in our previous work (Vieira et al., 2002; Vieira et al., 2003). Our classes of analyses were based on the analyses of English texts presented in (Poesio and Vieira, 1998), with the difference that we divided the Bridging class of their analyses into two different classes, separating coreferent (Indirect Anaphora) and non-coreferent (Other Anaphora) cases. Each definite description (d) is classified into one of the following four classes:

1. Direct anaphora: d corefers with a previous expression a; d and a have the same nominal head:

   a. ***A Comissão*** *tem conhecimento do livro...* (**the Commission** knows the book)

   d. ***a Comissão*** *constata ainda que o livro não se debruça sobre a actividade das várias...* (**the Commission** remarks that the book ignores the activity of various)

2. Indirect anaphora: d corefers with a previous

expression a; d and a have different nominal heads:

a. *a circulação* ***dos cidadões que dirigem-se...*** (the flow of **the citizens heading to...**)

d. *do controle* ***das pessoas*** *nas fronteiras* (the control of **the people** in the borders)

3. Other Anaphora: d does not corefer with a previous expression a, but depends for its interpretation on a:

   a. ***o recrutamento de pessoal científico e técnico***... (**the recruitment of scientific and technical employees**)

   d. ***as condições de acesso à carreira científica*** (**the conditions of employment for scientific jobs**)

4. Discourse New: the interpretation of d does not depend on any previous expression:

   d. *o livro não se debruça sobre* ***a actividade das várias organizações internacionais***... (the book ignores **the activity of various international organisation**...)

In (Schulte im Walde, 1997) acquired lexical knowledge is used for solving bridging descriptions, a broader class of anaphoric relations that includes our class, indirect anaphora. (Poesio et al., 2002) presents alternative techniques, based on syntactic patterns, focusing on meronymy relations. Finally, (Bunescu, 2003) deals with another class of anaphoric descriptions, which is also included in the bridging class, called as associative anaphora, following (Hawkins, 1978), where associative anaphora is an anaphoric relation between non-coreferent entities.

3 Lexical resource

Our lexical resource consists on lists of semantically related words. These lists are constructed automatically by a syntax-based knowledge-poor technique. The technique used is described in (Gasperin et al., 2001; Gasperin, 2001), and it is an extension of the technique presented in (Grefenstette, 1994).

Briefly, this technique consists on extracting specific syntactic contexts for every noun in the parsed whole corpus and then applying a similarity measure (the weighted Jaccard measure) to compare the nouns by the contexts they have in common (more contexts they share, more similar they are). As syntactic context, we understand any word that establishes a syntactic relation with a given noun in the corpus. An example of one kind of syntactic context considered is subject/verb, meaning that two nouns that occur as subject of the same verb share this context. Other examples of syntactic contexts are verb/object, modifier/noun, etc. To each context it is assigned a global and a local weight: the first related to the context frequency in the corpus, and the second related to its frequency as a context of the noun in focus. As output, we have a list of the most similar nouns to each noun in the corpus, ordered by the similarity value. We present the similarity list for the noun `acusação` (accusation) in Table 1 as an example.

Table 1: Similarity list for the noun `acusação`

acusação (*accusation*)	denúncia (*denunciation*) escândalo (*scandal*) crime (*crime*) pedido (*demand*) declaração (*declaration*) proposta (*proposal*) notícia (*news*) carta (*letter*) lista (*list*) cargo (*post*) ataque (*attack*) arma (*gun*) caso (*case*) impressão (*impression*) reclamação (*complain*)

The similarity lists can contain any kind of semantic relation (e.g. synonymy, hyponymy, etc.) between the words, but they are not classified. In general, the similarity lists for the less frequent words in the corpus contain some non-semantically related words (noise), since the relations were based on few syntactic contexts they shared along the corpus.

The main advantage of this technique is the possibility of having a corpus-tunned lexical resource built completely automatically. This resource reflects closely the semantic relations present in the corpus used to create the lists. So, we believe the similarity lists are more suitable for being used as lexical knowledge for resolving the anaphoras than a generic lexical base (e.g. Wordnet), since it focus on the semantic relations between the terms that appear in the corpus, without considering extra meanings that some words could have. New lists could be generated from each corpus that one aims to resolve the anaphoras.

To generate the similarity lists for Portuguese we utilised a 1,400,000-words corpus from the Brazilian newspaper 'Folha de São Paulo', containing news about different subjects (sports, economics, computers, culture, etc.). This corpus includes the

set of texts that was hand-annotated with coreference information in previous work (Vieira et al., 2002; Salmon-Alt and Vieira, 2002). The corpus was parsed by the Portuguese parser PALAVRAS (Bick, 2000), provided by VISL project[1].

We created two different sets of similarity lists: one considering just nouns and the other considering nouns and proper names. So, the first set of lists includes one list for each noun in the corpus and each list is composed by other common nouns. The second set of lists has one list for each noun and proper name in the corpus, and each list is composed by other nouns and proper names. The first set contains 8019 lists and the second 12275, corresponding to the different nouns (and proper names) appearing in the corpus. Each similarity list contains the 15 words that are more similar to the word in focus, according to the calculated similarity values.

Having lexical information about the proper names in the corpus is important, since we have many coreference cases whose anaphor or antecedent is a proper name. But when generating the similarity lists, proper names bring noise (in general they are less frequent then common nouns) and the lists became more heterogeneous (includes more non semantically related words).

4 Using similar words lists to solve indirect anaphora

From the manual annotation and classification of 680 definite descriptions we selected those cases classified as indirect anaphora (95). For each of them there is a list of candidate antecedents. This list is formed by all NPs that occur in the text. We consider as candidates all the NPs that occur in the text before the anaphor being mentioned.

Our heuristic for solving indirect anaphoras using lists of similar words is the following. Consider:

- H_{ana} is the head-noun of the anaphor
- H_{can_i} is the head-noun of the antecedent candidate i
- L_{ana} is the anaphor's list of similar nouns
- L_{can_i} is the list of similar nouns for the candidate i
- So, H_{can_i} is considered the antecedent of H_{ana} if

$$(1) H_{can_i} \in L_{ana}$$

or

$$(2) H_{ana} \in L_{can_i}$$

or

$$(3) L_{ana} \ni H_j \in L_{can_i}$$

We call (1) 'right direction', (2) 'opposite direction', and (3) 'indirect way'.

We consider (1) > (2) > (3) when regarding the reliability of the semantic relatedness between H_{ana} and H_{can_i}.

If the application of the heuristic resulted in more than one possible antecedent, we adopted a weighting scheme to choose only one among them. The candidate with the lowest weight wins. For ranking the possible antecedents, we considered two parameters:

- reliability: how the possible antecedent was select, according to (1), (2) or (3). A penalising value is added to its weight: 0, 40, 200, respectively. The higher penalty for the 'indirect way' is because we expected it could cause many false positives;
- recency: we consider the distance in words between the anaphor and the possible antecedent.

The penalty values for the reliability parameter were chosen in such a way they could be in the same magnitude as the recency parameter values, that are measured in words. For example, if candidate A is 250 words far from the anaphor and was selected by (1) (getting weight=250) and a candidate B is 10 words far from the anaphor and was selected by (3) (getting weight=210), candidate B will be selected as the correct antecedent.

5 Our evaluation corpus

As result of previous work (Vieira et al., 2002; Vieira et al., 2003), we have a Portuguese corpus manually annotated with coreference information. This corpus is considered our gold-standard to evaluate the performance of the heuristic presented in the previous section. The study aimed to verify if we could get a similar distribution of types of definite descriptions for Portuguese and English, which would serve as an indication that the same heuristics tested for English (Vieira et al., 2000) could apply for Portuguese. The main annotation task in this experiment was identifying antecedents and classifying each definite description according to the four classes presented in section 2.

For the annotation task, we adopted the MMAX annotation tool (Müller and Strube, 2001), that requires all data to be encoded in XML format. The corpus is encoded by <word> elements with sequential identifiers, and the output - the anaphors and

[1] See http://visl.hum.sdu.dk/visl/pt/

its antecedents - are enconded as <markable> elements, with the anaphor markable pointing to the antecedent markable by a 'pointer' attribute.

The annotation process was split in 4 steps: selecting coreferent terms; identifying the antecedent of coreferent terms; classifying coreferent terms (direct or indirect); classifying non-coreferent terms (discourse new or other anaphora). About half of the anaphoras were classified as discourse new descriptions, which account for about 70% of non-coreferent cases. Among the coreferent cases the number of direct coreference is twice the number of indirect coreference. This confirms previous work done for English.

For the present work, we took then the 95 cases classified as indirect coreference to serve as our evaluation set. In 14 of this cases, the relation between anaphor and antecedent is synonymy, in 43 of the cases the relation is hyponymy, and in 38, the antecedent or the anaphor are a proper name.

6 Implementing heuristics for indirect anaphora in ART

Our heuristics were implemented as an XSL stylesheet on the basis of the Anaphora Resolution Tool (ART) (Vieira et al., 2003).

The tool integrates a set of heuristics corresponding to one or more stylesheets to resolve different sorts of anaphora. The heuristics may be applied in a sequence defined by the user. As resolving direct anaphoric descriptions (the ones where anaphor and antecedent have the same head noun) is a much simpler problem with high performance rates as shown in previous results (Vieira et al., 2000; Bean and Riloff, 1999), these heuristics should be applied first in a system that resolves definite descriptions. In this work, however, we decided to consider for the experiments just the anaphoras that were previously annotated as indirect and check if the proposed heuristic is able to find the correct antecedent.

ART allows the user to define the set of anaphors to be resolved, in our case they are selected from previously classified definite descriptions. The stylesheet for indirect anaphora takes as input this list of indirect anaphors, a list of the candidates and the similarity lists. We consider all NPs in the text as candidates, and for each anaphor we consider just the candidates that appear before it in the text (we are ignoring cataphora at moment).

All the input and output data is in XML format, based on the data format used by MMAX. Our stylesheet for solving indirect anaphora takes the <markable> elements with empty 'pointer' attribute (coming unsolved from passing by the previously applied stylesheets/heuristics) and create and intermediate file with <anaphor> elements to be resolved. The resolved <anaphor>s are again encoded as <markable>s, with the 'pointer' filled. A detailed description of our data encoding is presented in (Gasperin et al., 2003).

Table 2: Results considering just nouns

Description		Numbers
Total indirect anaphors		57
Correctly resolved anaphors	Right direction	8
	Opposite direction	5
	Indirect way	6
	TOTAL	19 (33.3%)
Unsolved anaphors		21

7 Experiments

We run two experiments: one using the similarity lists with proper names and another with the lists containing just common nouns.

With these experiments we verify the values for precision, recall and false positives on the task of choosing an semantically similar antecedent for each indirect anaphor. Our annotated corpus has 95 indirect anaphors with nominal antecedents, where 57 of them do not include proper names (as anaphor or as antecedent). We use a non annotated version of this corpus for the experiments. It contains around 6000 words, from 24 news texts of 6 different newspaper sections.

Firstly, we reduced both sets of similarity lists to contain just the list for the words present in this portion of the corpus (660 lists without proper names and 742 including proper names).

7.1 Experiment 1

Considering the 57 indirect anaphoras to be solved (the ones that do not include any proper name), we could solve 19 of them. It leads to a precision of 52.7% and a a recall of 33.3%. Table 2 shows the result of our study considering the set of common noun lists.

Most of the cases could be resolved by 'right direction', that represents the more intuitive way. 21 of the cases didn't get any antecedent. We got 17 false positives, with different causes:

1. the right antecedent was not in the lists, therefore it could not be found but other wrong antecedents were retrieved. For example, in *meu **amigo** Ives Gandra da Silva Martins escreveu para esse jornal ... o **conselheiro** Ives* (my friend Ives_Gandra_da_Silva_Martins wrote

to this newspaper ... the councillor Ives), two more candidates head-nouns are similar words to "conselheiro" (councillor): "arquiteto" (architect) and "consultor" (consultant), but not "amigo" (friend);

2. the right antecedent was in the lists but another wrong antecedent was given higher weights, because of proximity to the anaphora, as in the example *a **rodovia** Comandante João Ribeiro de Barros ... próximo a **ponte** ... ao tentar atravessar a **estrada*** (the highway Comandante Joao Ribeiro de Barros ... near to the bridge ... while trying to cross the road). Here, the correct antecedent to "a estrada" (the road) is "rodovia" (the highway) and it is present in "estrada"'s similarity list (right direction), but also is "ponte" (the bridge) and it is closer to the anaphor in the text.

As expected, most of the false positives (11 cases) were 'resolved' by "indirect way".

Considering all similar words found among the candidates, not just the one with highest weight, we could find the correct antecedent in 24 cases (42%). The average number of similar words among the candidates was 2.8, taking into account again the positive and false positive cases. These numbers report how much the similarity lists encode the semantic relations present in the corpus. 64% of the synonymy cases and 28% of the hyponymy cases could be resolved. 35% of the hyponymy cases resulted in false positives, the same happened with just 14% of the synonymy cases.

7.2 Experiment 2

We replicated the previous experiment now using the similarity lists that include proper names. Table 3 shows the results considering the set of lists for nouns and proper names. Considering the 95 indirect anaphoras to be solved, we could solve 21 of them. It leads to a precision of 36.8% and a a recall of 22.1%. There was no antecedent found for 38 anaphors, and 36 anaphors got wrong antecedents (half of them by "inderect way"). We observed the same causes for false positives as the two presented for experiment 1.

Considering all cases resolved (correct and false ones), we could find the correct antecedent among the similar words of the anaphor in 31 cases (32.6%). The average number of similar words among the candidates was 2.75. The numbers for synonymy and hyponymy cases were the same as in experiment 1 - 64% and 28% respectively. The numbers for proper names were 50% of false positives and 50% of unresolved cases. It means none of the cases that include proper names could be resolved, but do not means they hadn't any influence in other nouns similarity lists. In 26% of the false positive cases, the correct antecedent (a proper name) was in the anaphor similarity list (but was not selected due to the weighting strategy).

Table 3: Results considering nouns and proper names

Description		Numbers
Total indirect anaphors		95
Correctly resolved anaphors	Right direction	13
	Opposite direction	3
	Indirect way	5
	TOTAL	21 (22.1%)
Unsolved anaphors		38

The experiment with the similarity lists that include proper names was able to solve more cases, but experiment 1 got better precision and recall values.

8 Related work

An evaluation of the use of WordNet for treating bridging descriptions is presented in (Poesio et al., 1997). This evaluation considers 204 bridging descriptions, distributed as follows, where NPj is the anaphora and NPi is antecedent.

- synonymy relation between NPj and NPi: 12 cases;
- hypernymy relation between NPj and NPi: 14 cases;
- meronymy between NPj and NPi: 12;
- NPj related with NPi being a proper name: 49;
- NPj sharing a same noun in NPi other than head (compound nouns): 25;
- NPj with antecedent being an event 40;
- NPj with antecedents being an implicit discourse topic: 15;
- other types of inferences holding between NPj and antecedent: 37.

Due to the nature of the relations, only some of them were expected to be found in WordNet. For Synonymy, hypernymy and meronymy, 39% of the 38 cases could be solved on the basis of WordNet. From this related work we can see the large variety of cases one can found in a class such as bridging. In our work we concentrated on coreference relations, these can be related to synonymy, hypernymy, and

proper name sub-classes evaluated in (Poesio et al., 1997).

The technique presented in (Schulte im Walde, 1997) based on lexical acquisition from the British National Corpus was evaluated against the same cases in (Poesio et al., 1997). For synonymy, hypernymy and meronymy, it was reported that 22% of the 38 cases were resolved. In (Poesio et al., 2002) the inclusion of syntactic patterns improved the resolution of meronymy in particular, resulting in 66% of the meronymy cases being resolved. Bunescu (Bunescu, 2003) reports for his method on resolving associative anaphora (anaphoric relation between non-coreferent entities) a precision of 53% when his recall is 22.7%.

9 Concluding remarks

We tested the use of word similarity lists on resolving indirect anaphoras on Portuguese newspaper texts. We presented our heuristic for searching word similarity lists to be able to find the relation between an anaphor and its antecedent. We considered similarity lists containing proper names and lists containing just common nouns. Our heuristic was able to resolve 33.3% of the cases, with precision of 52.7% when considering just common nouns, and we got 22.1%recall with precision of 36.8% when including proper names. Even though considering proper names give us the possibility of treating more anaphora cases, we got lower precision than using the lists with only nouns, since such lists are more homogeneous. These results are comparable to previous work dealing with such complex anaphora.

As future work, we intend to integrate our heuristic for indirect anaphora with other heuristics for anaphora resolution into ART and investigate the best combination of application of these. Concerning refining the proposed heuristic, we intend to run more experiments aiming to tune the penalising weights when choosing an antecedent among the candidates already selected by the search on the similarity lists.

Acknowledgements

We would like to thank CNPq (Brazil) / INRIA (France) for their financial support, and Susanne Salmon-Alt, for her collaboration in this work.

References

D. Bean and E. Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*.

Eckhard Bick. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Århus University, Århus.

Razvan Bunescu. 2003. Associative anaphora resolution: A web-based approach. In *Proceedings of EACL 2003 - orkshop on The Computational Treatment of Anaphora*, Budapest.

Caroline Gasperin, Pablo Gamallo, Alexandre Agustini, Gabriel Lopes, and Vera Lima. 2001. Using syntactic contexts for measuring word similarity. In *Proceedings of the Workshop on Semantic Knowledge Acquisition and Categorisation*, Helsink, Finland.

Caroline Gasperin, Renata Vieira, Rodrigo Goulart, and Paulo Quaresma. 2003. Extracting xml syntactic chunks from portuguese corpora. In *Traitement automatique des langues minoritaires - TALN 2003*, Btaz-sur-mer, France.

Caroline Varaschin Gasperin. 2001. Extração automática de relações semânticas a partir de relações sintáticas. Master's thesis, PUCRS, Porto Alegre.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

John A. Hawkins. 1978. *Definiteness and Indefiniteness*. Humanities Press, Atlantic Highland, NJ.

Christoph Müller and Michael Strube. 2001. MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the IJCAI 2001*, pages 45–50, Seattle.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging descriptions in unrestricted texts. In *Proceedings of the Practical, Robust, Anaphora Resolution for Unrestricted Texts, Workshop on Operational Factors*, Madrid.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte Im Walde, and Renata Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of LREC 2002*, Las Palmas De Gran Canaria.

Susanne Salmon-Alt and Renata Vieira. 2002. Nominal expressions in multilingual corpora: Definites and demonstratives. In *Proceedings of the LREC 2002*, Las Palmas de Gran Canaria.

Sabine Schulte im Walde. 1997. Resolving Bridging Descriptions in High-Dimensional Space. Master's thesis, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, and

Center for Cognitive Science, University of Edinburgh.

K. van Deemter and R. Kibble. 2000. On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4).

Renata Vieira, Susanne Salmon-Alt, and Emmanuel Schang. 2002. Multilingual corpora annotation for processing definite descriptions. In *Proceedings of the PorTAL 2002*, Faro.

Renata Vieira, Caroline Gasperin, and Rodrigo Goulart. 2003. From manual to automatic annotation of coreference. In *Proceedings of the International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization*, Venice.

Vieira et al. 2000. Extração de sintagmas nominais para o processamento de co-referência. In *Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada - PROPOR*, Atibaia.