A DICTIONARY-BASED ALGORITHM FOR INDIRECT ANAPHORA RESOLUTION

Alexander Gelbukh and Grigori Sidorov

Natural Language Processing Laboratory, Computing Research Center, National Polytechnic Institute.
Av. Juan de Dios Batiz s/n, Zacatenco, 07738 México D.F, México.
fax: +52 5-586-2936,
{gelbukh, sidorov}@pollux.cic.ipn.mx

Abstract

In the paper, a dictionary-based method of detecting of implicit links between words in the texts (so-called indirect anaphora) is discussed. The method consists in using of a dictionary of "scenarios" – lists of words semantically related to the given one, and show that detecting the implicit referential relationships can be viewed as intersection of such scenarios. The advantage of the method is in the simplicity of the dictionary being used, since it does not rely on specific semantic relationships between the headword and the words listed in its scenario. Thus, such a dictionary can be derived from some existing semantic dictionaries or even from large corpora.

Keywords: text processing, indirect anaphora, semantic analysis, dictionary.

1. Introduction[*]

Anaphora resolution in general is one of the most challenging tasks of natural language processing. It is necessary in a wide range of NLP tasks, from language understanding to statistics, translation, and abstracting [Aone and McKee 1993, Carter 1987, Hirst 1981, Kameyama 1997, Mitkov 1997]. The resolution of indirect anaphora and even detection of the presence of indirect anaphora are especially difficult [Indirect Anaphora 1996]. Example of indirect anaphora is the discourse *"I had a look at a new house yesterday. The kitchen was extra large"* (*the kitchen* = of the *house*), in which the anaphoric relation holds between two conceptually different words, *kitchen* and *house*; note that there is no coreference between these two words. As we will show, coreference holds between the word *kitchen* in the text and the word *kitchen* implicitly introduced in the discourse by the word *house*. Definite article as in the example above is not the unique way of expression of indirect anaphora. A particular type of indirect anaphora markers is found in expressions with demonstrative pronouns, as in the example *"I sold a house. What can I do with this money?"*.

Two major problems arise with respect to indirect anaphora resolution:

- Detect the presence of the indirect anaphora and
- Resolve the ambiguity of the anaphoric link.

However, we will approach the problem in the opposite order: We will try to plausibly resolve the anaphoric link and, if we succeed, consider that definiteness of the text element has anaphoric nature. Our paper discusses a way of a dictionary-driven resolution of indirect anaphora with a special branch for the demonstrative pronouns in the anaphoric function.

2. Indirect anaphora as references to scenarios

Indirect anaphora can be thought of as coreference between a word and an entity implicitly introduced in the text before. We call such entities implicitly or even potentially introduced by a word, a *prototypic scenario* of this word. Thus, anaphoric relation here holds between a word and

[*] The work done under partial support of REDII-CONACYT, CONACyT grant 26424-A, SNI, and COFAA-IPN, Mexico.

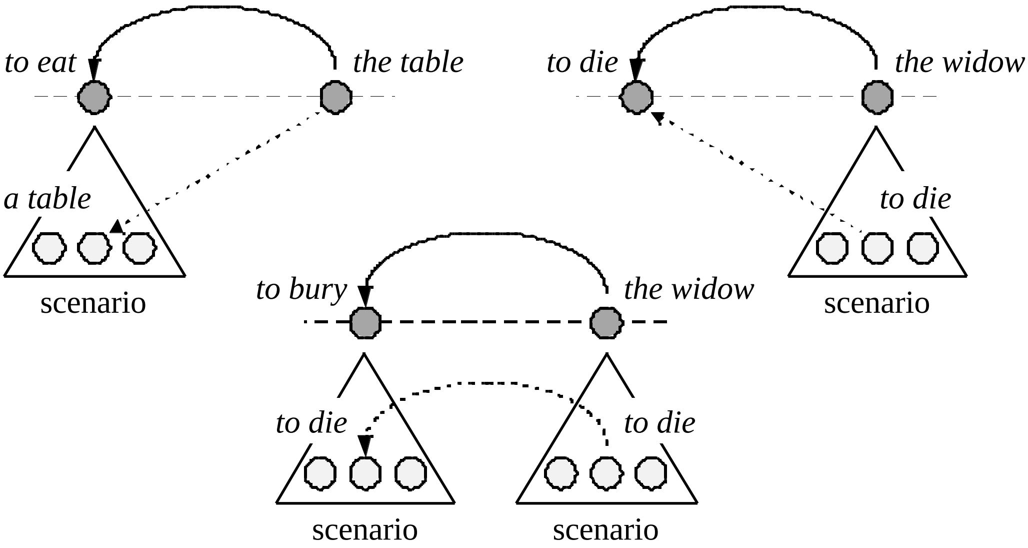

Fig.1 . Three types of indirect anaphoric relationships.

an element of the prototypic scenario of another word in the text; such an element does not have the surface representation in the text.

There are three possible types of the indirect anaphora depending on the relations between the antecedent and the anaphor. (1) The anaphor is a word in the text while the antecedent is an element of a scenario implied by another word; this is the most common case. (2) Vice versa, an implied concept refers to a word in the text (a rather rare case). (3) The reference is made between the implied concepts (an even rarer case). Let us consider the following examples, see Figure 1:

1) *John was eating. The table was dirty.*

2) *John died. The widow was mad with grief.*

3) *John was buried. The widow was mad with grief.*

Here the definite articles are used with the words *table* and *widow*. However, these words (and the corresponding concepts) do not appear literally in the discourse before. What is the reason for their definiteness? It can be explained by the existence of the indirect anaphoric relation: *eat* _ *table*, *die* _ *widow*, *bury* _ *widow*. In the first example the antecedent *to eat* contains in its prototypic scenario a slot for a *place* with a possible value *table*. In the second example, the verb *to die* is included in the lexical meaning of the word *widow*. In the third examples, the concept *to die* is in common in the lexical meanings of *widow* and *to bury*.

Let us consider more examples of indirect anaphora[1]:

4) *I bought a house. The/*This kitchen (walls, roof) was extremely large.*

5) *I bought a house. The/*These dimensions were* 20 _ 20.

6) *I bought a house. The/*This previous owner was happy.*

7) *I was buying a house. I counted the/*this money carefully.*

8) *I sold a house. What can I do with the/this money?*

9) *I bought a house. I liked the/this price.*

10) *John was eating. The/*This table (dish) was dirty.*

[1] The unacceptable variants are marked with an asterisk.

11) *John was eating. It was dark in the/*this forest.*
12) *John was eating. The/This food was delicious.*
13) *John was eating. The/These apples were delicious.*
14) *John was singing. The/This noise disturbed Peter.*
15) *John was singing. Peter disliked the/this noise.*
16) *John was reading. He liked the/this author.*
17) *John died. The/*This widow was mad with grief.*

For example, in the example 4 the indirect anaphoric relation holds between *kitchen* and *house*: *the* kitchen is the kitchen of this house.

In each of these sentences, we consider a purely anaphoric meaning of the definite article or the pronoun; at least these examples *can* have such a meaning. The variants marked with an asterisk are not possible in the anaphoric interpretation. We don't take into account possible non-anaphoric interpretations of examples. One possible interpretation is contraposition: "*this kitchen* is large while the others kitchens are not;" (example 3) in this case a special intonational stress is used which is not reflected in the written text. Another possible non-anaphoric interpretation is deictic function: the speaker is physically in *this kitchen* (example 4) or is showing *this money* (example 7) to the listener.

Yet another example that does not allow the anaphoric relation is:

18) **Peter disliked that John was eating here. The/this table was dirty.*

Thus, a question arises: What are the rules that should be implemented in the algorithm for indirect anaphora resolution?

Indirect anaphora can combine with some phenomena involving substitution of one word for another, such as the use of synonyms, more general (hypernyms) (see example 12) or more specific (hyponyms) (example 10) term, metaphor (example 13), or changing of the surface part of speech (derivation). Such phenomena are transparent for indirect anaphora. We will call the words related with one of these relations *compatible*.

3. Indirect anaphora resolution: general case

As we have seen, to check the possibility of indirect anaphoric link between two words in the discourse, a dictionary can be used that lists the members of the prototypic scenario of a word. In our case, we used a dictionary compiled from several sources, such as Clasitex's dictionary [Guzmán-Arenas 1998], FACTOTUM SemNet dictionary derived from the Roget thesaurus, and some other dictionaries. For example, the dictionary entry for the word *church* includes the words related to this one in the dictionaries mentioned above: *priest*, *candle*, *icon*, *prayer*, etc.

To check compatibility of words (generalization, specification, metaphor) we use a thesaurus compiled on the based of FACTOTUM SemNet dictionary, WordNet, and some other sources.

The algorithm that we use to find the antecedent of a word introduced with a definite article or a demonstrative pronoun first of all uses the heuristics to find the potential antecedents for the current word – for example, it should not be too far in the text. Then the algorithm looks for one of the three cases described in the previous section and checks the following condition:

Condition 1: Indirect anaphora is possible if any of the following conditions holds:

- The word is compatible with an element of the scenario of the potential antecedent, or
- The potential antecedent is compatible with an element of the scenario of the word, or
- Their scenarios intersect (in the meaning of compatibility, see above).

However, as we could see, this condition is necessary but not sufficient for the possibility of an anaphoric link. As the example 18 shows, the following condition is also necessary:

Condition 2: Indirect anaphora is possible only for the uppermost semantic level of the situation.

Really, in the example 18, the uppermost level situation is "*Peter disliked*" and the indirect anaphora to the embedded situation is not possible. For this check, a syntactic parser is used; we use a rather simple context-free parser to quickly reject the incorrect variants.

4. Indirect anaphora resolution: demonstrative pronouns

It can be observed that the anaphors in our examples have different status in the prototypic scenario of the antecedents. Some of them are necessary parts of the lexical meaning of the corresponding antecedent (as in examples 8, 9, 12) and thus are implicitly presented in the situation, while some are not. For example, the Random House dictionary defines the word *sell* as "to transfer (goods) to or render (services) for another in exchange for money; dispose of to a purchaser for a price." Thus, the words "money" (as a concept, but not a physical object) and "price" are parts of the lexical meaning of the word *sell*.

As the analysis of the examples shows, the following condition is also necessary in the case of demonstrative pronouns:

Condition 3: Indirect anaphora can be expressed by a demonstrative pronoun if the both of the following conditions hold:

- The antecedent denotes a process or situation and
- The anaphor is included into the lexical meaning of the antecedent.

Indeed, the examples 4 to 6 have the antecedents denoting objects (*house* _ *kitchen, house* _ *dimensions, house* _ *previous owner*). In the examples 7, 10, 11, 17 the anaphors are not included into the lexical meaning of the antecedents (*buy* _ *money* (as the physical object), *eat* _ *table*, *eat* _ *forest*, *die* _ *widow*).

The other examples (8, 9, 12 to 16) allow the use of the demonstrative pronoun. The examples 8, 9, and 12 are the standard cases; note that in the example 7 *money* is a physical object that is not obligatory in the situation (the buying could be with a credit card, to say), while in the example 8 it is an abstract entity, the price, and is a part of the lexical meaning of the verb, this is why in the example 4 the demonstrative pronoun is forbidden, while in the example 8 it is allowed. Example 15 demonstrates generalization: *sing* _ *noise*, when the prototypic noun would be *singing* or *song*.[2] Example 13 demonstrates specification: *eat* _ *apples* (a kind of *food* which is a part of the lexical meaning of *eat*).

For the algorithm to be able to test the Condition 3, some of the elements of the scenario are marked as "necessary" in our dictionary, while the others are "optional." We took this information mainly from English-English explanatory dictionaries: the words mentioned in the definitions are marked as "obligatory." However, in many cases handwork was necessary to mark additional words.

Additionally, the dictionary contains the basic semantic class of the word: thing versus process or situation (regardless of the surface part of speech). This information was found in the FACTOTUM SemNet dictionary.

5. Conclusions and future work

We have discussed a dictionary-based algorithm of contextual interpretation of definite text expressions by linking them to elements of the prototypic scenario of some another word in the context.

Namely, our algorithm checks the following three conditions: (1) the intersection between the scenarios, (2) the syntactic plausibility of the relation, and (3) in the case of demonstrative

[2] Probably the use of the demonstrative pronoun in case of generalization is preferable.

pronouns, the semantic type of the antecedent and inclusion of the anaphor in the list of the "obligatory participants" of the antecedent.

Note that with our method, the dictionary does not have to specify in what way the element of the scenario is related to the headword. This simplifies the task of compilation of such a dictionary. At the early stages of our experiments, we directly used the "thematic dictionary" of the Clasitex system [Guzmán-Arenas 1998]. In addition, a lexical attraction dictionary [Yuret 1998] automatically extracted from a text corpus can provide useful information.

In the future, we plan to extend the information present in the dictionary. First, the dictionary should include a kind of "weights" of the elements of the scenario. The obligatory elements have the highest weight; however, the "optional" elements can be more closely related to the headword or be rather far from it. For example, the word *table* in the example 10 is not obligatory, but a very probable participant of the situation of *eating*. On the other hand, the word *forest* in the example 11 is a possible, but low-probable participant of this situation. Such weights can be obtained both from some semantic dictionaries as the number of links between the words, and from a large corpus.

References

1. Aone Ch., McKee D. (1993), "Language-independent anaphora resolution system for understanding multilingual texts," *Proc. of the 31st meeting of the ACL*, The Ohio State University, Columbus, Ohio.
2. Carter D. (1987) *Interpreting anaphora in natural language texts* (Chichester: Ellis Horwood).
3. Chafe W. (1976), "Givenness, Contrastiveness, Definiteness, Subject, Topics, and Point of View," *"Subject and Topic,"* Ch.N. Li (ed.), Academic Press, New York, 1976, pp. 27-55.
4. Guzmán-Arenas A. (1998), "Finding the main themes in a Spanish document," *Journal Expert Systems with Applications,* Vol. 14, No. 1/2. Jan/Feb 1998, pp. 139-148.
5. Hirst G. (1981), *Anaphora in Natural Language Understanding* (Berlin, Springer-Verlag).
6. Indirect Anaphora (1996), *Proc. of Indirect Anaphora Workshop.* Lancaster University.
7. Kameyama M. (1997), "Recognizing Referential Links: an Information Extraction Perspective," *Proc. of ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution. Madrid.*
8. Mitkov R. (1997), "Factors in Anaphora Resolution: They are not the Only Things that Matter," *A Case Study Based on Two Different Approaches. Proc. of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution.* Madrid.
9. Shank R. C., Lebowitz M., and Birnbaum L. (1980), "An Integrated Understander," *American Journal of Computational Linguistics*, 1980, Vol. 6, No 1, pp 13-30.
10. Yuret, Deniz. "Discovery of linguistic relations using lexical attraction," Ph.D. thesis, MIT, 1998. See http://xxx.lanl.gov/abs/cmp-lg/9805009.