Dictionary-based Method for Coherence Maintenance in Man-Machine Dialogue with Indirect Antecedents and Ellipses*

Alexander Gelbukh, Grigori Sidorov, and Igor A. Bolshakov

Natural Language Laboratory, Center for Computing Research (CIC),
National Polytechnic Institute (IPN), Av. Juan de Dios Batiz, s/n, esq. Mendizabal,
Zacatenco, C.P. 07738, Mexico D.F., Mexico.
{gelbukh, sidorov, igor}@cic.ipn.mx

Abstract. Resolution of referential ambiguity is one of the most challenging problems of natural language processing. Especially frequently it is faced with in dialogues. We present a heuristic algorithm for detection of the indirect antecedents for dialogue phrases based on the use of a dictionary of prototypic scenarios associated with each headword as well as of a thesaurus of the standard type. The conditions for filtration of the candidates for the antecedent are presented. We also present a similar algorithm for reconstruction of elliptical phrases of a special kind using a combinatory dictionary.

1 Introduction

The problem of usual (direct) anaphora resolution in a wide range of NLP tasks – from language understanding to statistics, translation, and abstracting – have been the area of active research in recent years [5, 7, 13, 17]. The prevailing approaches to resolution of direct anaphora are nowadays an integrated and an alternative ones [19, 20, 21]. The former approach is based on the integration of different kinds of knowledge like syntactic, semantic, and discourse, see also [1, 9], while the latter uses statistical information, neural networks or the principles of reasoning with uncertainty. For ellipsis resolution there is also a tendency to use integrated information [14]. On the other hand there are attempts to exclude some types of knowledge from algorithms [18]. At the same time the idea to create methods using only dictionaries like WordNet is on the agenda [15]. In this paper we try to exploit the ideas connected with dictionary-based approaches.

Indirect anaphora has drawn less attention than the direct one. The investigation was mainly focused on its various theoretical aspects rather than on practical algorithms of resolution [16, 25]. For example, in [8] and [12], a giveness hierarchy is proposed. A proposal similar to our approach can be found in [23]. The authors note that "this kind of reference (indirect anaphora) has not

* Work done under partial support of CONACyT, REDII-CONACyT, and CGEPI-IPN, Mexico.

been thoroughly studied in natural language processing, but is important for coherence resolution, language understanding, and machine translation." The method they suggest is based on a dictionary of examples of the form "*X* of *Y*" extracted from a text corpus. This dictionary plays the role analogous to that of our scenario dictionary, while the latter is much richer in the semantic and syntactic types of relations; also we build our dictionary on the basis of existing dictionaries of various types rather than extract it from a corpus.

Dialogue is also actively investigated, though not with relation with the anaphoric phenomena. As an example of a recent research in dialogue, see [3].

In this paper, we describe a dictionary-based method of restoration of indirect anaphoric and elliptical links in application to the dialogues where the phenomena in question are much more frequent [27, 10]. In the first part of the paper, we discuss the problem of indirect anaphoric antecedents, and in the second part a modification of our algorithm to be applied to ellipsis. We formulate necessary conditions for the detection of the presence of the phenomenon (anaphora or ellipsis, correspondingly); then we discuss the structure of the dictionary used by the algorithm; finally, the algorithm is described.

For simplicity we consider the microdialogues, containing the phrase of one of the participants (the computer) and the answer of the other (the human), like

1. "*There is a PS/2 computer in the store.*"
   "*Is <u>the</u> motherboard from Intel?*"
2. "*There is a <u>printer</u> in the store.*"
   "*<u>Narrow</u>?*"
3. "*The user found an <u>error</u> in the program.*"
   "*Really <u>fatal</u>?*"

Example 1 demonstrates the case of the indirect anaphora. The problem is in the interpretation of the definite clause in the second utterance of the first dialogue. Usually, the definite clause refers to some entity already introduced in the common view of the participants of the communication (in this case implicitly). Examples 2 and 3 illustrate the cases of the elliptical construction where the omitted word was introduced into the dialogue in the previous utterance(s), and the purpose of the utterance is to clarify or add a detail to a previous one. Thus, we leave aside the cases where the omitted word was introduced, say, by the extralinguistic situation (say, both participants see a strange man; "*Drunk*," says one of them) or by the pragmatics of the communication.

2 Indirect Antecedents in Dialogue

Let us call the entities implicitly or potentially introduced by a word a *prototypic scenario* of this word. For example, the scenario for the word *house* includes *kitchen*, *roof*, *windows*, *district*, *place*, *owner*, etc. We suggest that the indirect anaphoric relation holds between a word and an element of the prototypic scenario of another word previously appeared in the text; such an implicitly introduced element does not have its own surface representation in the text.

We call the words that are equivalent for our algorithms *compatible*. Thus, synonyms, derivatives, hypernyms, hyponyms, and metaphors of a word can be compatible with it.

Let us denote the potential antecedent x, the anaphor y, and the scenario of a word w, $S(w)$. Denoting the compatibility relation between the words w and u as $w \sim u$, we can formulate the following condition:

Condition 1. Indirect anaphoric relation between x and y is possible only if any of the following conditions holds:

$\exists w \in S(x)$ such that $y \sim w$, or
$\exists w \in S(y)$ such that $x \sim w$, or
$\exists w \in S(x)$ and $\exists u \in S(y)$ such that $u \sim w$.

The algorithm for detecting anaphoric links between the utterances in the dialogue works as follows. It analyzes the current utterance word by word. If a word is introduced with a definite article or a demonstrative pronoun, then it is a potential anaphor, and the algorithm tries to find a plausible antecedent for it. It looks for the possible candidates first within the same utterance and then in the previous utterance(s); from right to left. The candidates are a priori scored the lower the greater the distance; the algorithm stops when an antecedent with high score is found or the scores become too little, see Fig. 1.

Namely, let *distance* be the distance between the possible antecedent and the potential anaphor. Currently, we use the linear distance. In a more sophisticated version of the algorithm, the syntactic distance (the number of constituents between the two words) or some combination of the two measures can be used.

Let *probability-of-variant*$_n$ be the a priori probability associated with the corresponding parts 1, 2, and 3 of Condition 1. This probability is a statistical characteristic of the given language (or sublanguage) that characterizes the frequency of the use of the corresponding constructions.

Let *threshold* be the maximum allowed value which depends on *distance* and *probability-of-variant*$_n$. It is used to determine the stop condition of algorithm.

Determining the exact values of these parameters goes beyond the scope of this paper; in our experiments, we used the following approximate values found empirically: *threshold* $= 0.05$, *probability-of-variant*$_1$ $= 1$, *probability-of-variant*$_2$ $= 0.75$, and *probability-of-variant*$_3$ $= 0.4$.

To check the possibility of an indirect antecedent, a prototypic scenario dictionary is used. In our experiments, we used a dictionary compiled from several sources, Currently the dictionary contains about 1,000 entries, and the scenarios contain in total about 40,000 words.

3 Ellipsis in Dialogue

For the purposes of the coherence maintenance in the dialogue with ellipsis it is appropriate to view the interpretation as establishing a link (similar to a syntactic one) between the corresponding words across the utterances. For instance, in example 3 the link holds between the words *fatal* and *error*:

repeat for each definite noun x of the utterance, from left to right
  repeat for each word y, from right to left, from the last word of the previous utterance
  $distance$ = the distance between x and y
  if $\exists$ n such that part n of Condition 1 holds for x and y **then**
    the variant y of the antecedent for x is temporarily stored,
    its weight $w = probability\text{-}of\text{-}variant_n \times (1/distance)$
  if $(max\{probability\text{-}of\text{-}variant_i\} \times (1/distance) < threshold)$ **then**
    break the loop and go to next x
    else pass to the next word y to the left
  end
  The variant y (if any) with the best weight w is chosen and sent to the output.
end

Fig. 1. The algorithm for finding antecedent(s).

$user \xleftarrow{subj} found \xrightarrow{obj} error \xrightarrow{attr} fatal$

Let us call such a link *elliptical relation*, the word *fatal* being the source of the relation and the word *error* being its target.

To restore the elliptical relations in the dialogue, we use the same idea as for the indirect anaphoric relations. Namely, we try to find a highly probable target for such a relation and in this way detect the very presence of the relation.

We call the words that potentially, with high probability can form a word combination *combinable*, and denote this as $u \triangleleft w$. Here u is the governor and w is the dependent: *error* $\triangleleft$ *fatal* is true since the syntactic combination $fatal \longleftarrow error$ is highly probable; however, *fatal* $\triangleleft$ *error* is not true since $fatal \longrightarrow error$ is impossible. Also *program* $\triangleleft$ *fatal* is not true since normally *fatal program* is not a combination that can be normally expected in the text. Note that this is a lexical property of the corresponding words, specified in the dictionary, rather than the syntactic relation in a specific context.

Condition 2. Elliptical relation between x and y is possible only if $x \triangleleft y$.

If the lexical information is not available for a specific word, a relaxed condition using only the parts of speech POS (x) and POS (y) can be used, with less reliability of the positive result:

Condition 3. Elliptical relation between x and y is possible only if POS (x) $\triangleleft$ POS (y).

Since we treat the type of the ellipsis under consideration as a potential insertion of a word into the target utterance, the Condition 3 can be generalized as follows (though such a generalized condition is much more expensive computationally to test):

Condition 4. Elliptical relation between x and y is possible only if the utterance containing the target (x) remains syntactically correct when the source (y) is added to it and syntactically connected to x.

The algorithm for detecting potential antecedents is very similar to the one for detecting the indirect antecedents. Namely, for each case of structural incompleteness, the hypothesis of possible elliptical link is checked by trying, word by word, the candidates in the previous utterance and verifying Conditions 2 to 4. Since Condition 3 is less computationally expensive, it is tried first; then Condition 2. The process stops when a plausible antecedent is found or when the distance from the end of the previous utterance becomes too large.

To check Condition 2, a dictionary of word combinations is used. For each headword, the dictionary lists the words that can syntactically combine with it, together with the necessry prepositions (or grammatical cases). In addition to the mere fact of combinability, the dictionary can specify a quantitative measure of the probability (frequency) of the combination in the texts.

In our experiments, we use the CrossLexica dictionary [2] in its part of word combinations. The Russian version of this dictionary contains more than 500,000 word combinations; the English version is under construction.

As in the case of the indirect anaphora, some lexical relations – such as synonyms, hyponyms, etc. – are transparent for syntactical combinability.

References

1. Azzam, S., K. Humphreys, R. Gaizauskas: Evaluating a Focus-Based Approach to Anaphora Resolution. Proc. of COLING-ACL'98, 1998, pp 74-78.
2. Bolshakov, I.A., A.F. Gelbukh: A Very Large Database of Collocations and Semantic Links. NLDB'2000: 5th International Conference on Applications of Natural Language to Information Systems, Versailles, France, June 28-30, 2000. Lecture Notes in Computer Science, Springer.
3. Carberry, S., Lambert, L.: A process model for recognizing communicative acts and modeling negotiation subdialogues. Comp. Ling., **25** (1), 1999: 1-54.
4. Carter, D.: Interpreting anaphora in natural language texts. Ellis Horwood, Chichester, 1987.
5. Cassidy, P.: An Investigation of the Semantic Relations in the Roget's Thesaurus: Preliminary results, http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/new/FSN_DOC.ASC, 1996.
6. Chierchia, G.: Dynamics of Meaning: Anaphora, Presupposition, and the Theory of Grammar. University of Chicago Press, 1995.
7. Cornish, F.: Anaphora, Discourse, and Understanding: Evidence from English and French. Oxford Univ Press, 1999.
8. Erku, F., J. K. Gundel: The pragmatics of indirect anaphors. In J. Verschueren and M. Bertuccelli-Papi (Eds.), The pragmatic perspective: Selected papers from the 1985 International Pragmatics Conference. John Benjamins, Amsterdam, 1987. pp. 533–545.
9. Ferrandez, A., M. Palomar, L. Moreno: Anaphor resolution in unrestricted text with partial parsing. Coling-98, 1998, pp. 385–391.
10. Gelbukh, A., Sidorov, G.: On Indirect Anaphora Resolution. Proc. PACLING-99, Pacific Association for Computational Linguistics, University of Waterloo, Waterloo, Ontario, Canada, August 25-28, 1999, pp. 181–190.
11. Gelbukh, A., G. Sidorov, A. Guzmán-Arenas: Use of a Weighted Topic Hierarchy for Document Classification. In: Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999.

12. Gundel, J., N. Hedberg, R. Zacharski: Giveness, Implicature and Demonstrative Expressions in English Discourse. Proc. of 25th meeting of Chicago Linguistic Society, Part II (Parasession on Language in Context). Chicago, 1988. pp. 89–103.
13. Fretheim, T, J. K. Gundel (eds.): Reference and referent accessibility. John Benjamins, Amsterdam, 1996.
14. Hahn, U., M. Strube, K. Markert. Bridging textual ellipses. Proc. of the 16th International Conference on Computational Linguistics, 1996. pp. 496–501.
15. Hirst, G., D. St-Onge: Lexical chains as representations of context for the detection and correction of malapropisms. In: Christiane Fellbaum (editor), WordNet: An electronic lexical database and some of its applications, Cambridge, MA: The MIT Press, 1998.
16. Indirect Anaphora Workshop. Lancaster University, Lancaster, 1996.
17. Kameyama, M.: Recognizing Referential Links: an Information Extraction Perspective. Proc. of ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution. Madrid, 1997.
18. Kennedy, C., B. Boguraev. Anaphora for Everyone: Pronominal Anaphor Resolution without a Parser. Coling-96, 1996.
19. Mitkov, R.: Pronoun Resolution: the Practical Alternative. In: S. Botley and T. McEmery (eds), Discourse Anaphora and Anaphor Resolution, Univ. College London Press, 1997.
20. Mitkov, R.: 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, 1997.
21. Mitkov, R.: An Integrated Model for Anaphora Resolution. Coling-94, 1994, pp. 1170–1176.
22. Morris, J., Hirst, G.: Lexical cohesion, the thesaurus, and the structure of text. Computational linguistics, **17(1)**, March 1991, 21–48.
23. Murata, M., H. Isahara, M. Nagao: Resolution of Indirect Anaphora in Japanese Sentences Using Examples "X no Y (Y of X)." ACL'99 Workshop on Coreference and Its Applications, Maryland, USA, June 22, 1999.
24. Partee, B., P. Sgall (Eds.): Discourse and Meaning. Papers in Honour of Eva Hajicova. Benjamins, Amsterdam/Philadelphia, 1996.
25. Sanford, A. J., S. C. Garrod, A. Lucas, R. Henderson: Pronouns without explicit antecedents? Journal of Semantics, 1983, 2: 303–318.
26. Sidorov, G., A. Gelbukh: Demonstrative pronouns as markers of indirect anaphora. Proc. 2nd International Conference on Cognitive Science and 16th Annual Meeting of the Japanese Cognitive Science Society Joint Conference (ICCS/JCSS99), July 27-30, 1999, Tokyo, Japan, pp. 418–423.
27. Sidorov, G., A. Gelbukh: A method of detection and resolution of hidden anaphora (in Russian, abstract in English). Proc. Annual International Conf. on Applied Linguistics Dialogue-99, May 30 - June 5, 1999, Moscow, Russia. pp. 288–297.
28. Ward, G., B. Birner: Definiteness and the English existential. Language, 1994, **71**: 722–742.
29. Coling-ACL'98 Workshop "Usage of WordNet in Natural Language Processing Systems." August 16, 1998, Université de Montréal, Montréal, Canada.

This article was processed using the LaTeX macro package with LLNCS style