On Coherence Maintenance in Human-Machine Dialogue with Contextual Ellipses

Alexander Gelbukh, Grigori Sidorov, and **Igor A. Bolshakov**

Center for Computing Research (CIC) of National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, s/n, esq. Mendizabal, Zacatenco, C.P. *07738*, Mexico D.F., Mexico.
Phone: *(+52) 5729-6000*, ext. *56544*. Fax: *(+52) 5586-2936*
{gelbukh, sidorov, igor}@cic.ipn.mx

Article received on May 1, 1999; accepted on June 21, 2001

Abstract

The paper discusses a method for resolving referential ambiguity related with contextual ellipsis in human-machine dialogue. Necessary conditions for reconstructing elliptical phrases of a special kind—short answers or clarifications—are formulated. The first (most reliable) condition characterizes the lexical combinability; it is tested using a dictionary of collocations. The second one uses general information such as parts of speech, thus characterizing the most general syntactic combinability. The third one is based on the acceptability of the sentence with restored ellipsis. A heuristic algorithm for the restoration of the elliptical links using these three conditions is presented. The experimental results are discussed.

Keywords: human-machine dialogue, coherence, ambiguity resolution, ellipsis, dictionaries.

Resumen

Se presenta un método para la resolución de la ambigüedad referencial relacionada con la elipsis de contexto en el diálogo humano-máquina. Se formulan las condiciones necesarias para la reconstrucción de las frases elípticas de un tipo especial –respuestas cortas o aclaraciones. La primera condición –la más confiable– caracteriza la habilidad de combinación léxica y se comprueba usando un diccionario de colocaciones. La segunda usa la información general tal como las categorías gramaticales, caracterizando así la habilidad de combinación sintáctica más general. La tercera se basa en la gramaticalidad de la oración con la elipsis restaurada. Se presenta un algoritmo heurístico para la restauración de los vínculos elípticos que usa estas tres condiciones. Se discuten los resultados experimentales.

Palabras clave: diálogo humano-máquina, coherencia, resolución de ambigüedad, elipsis, diccionarios.

1 Introduction

One of the difficult problems in the design of a dialogue understanding modules—especially the systems of human-machine dialogue—consists in omission of some parts in an expression that seems obvious for the human but not for the system. Such omission of the words restorable in an "obvious" way from the previous context is called *ellipsis*, for example:

1. System: "*There is a printer in the store*."
   Human: "*Inkjet?*"

In this example, in order to understand the question, the system must restore the omitted word to restore the intended link *printer* ← *inkjet*, since the word *inkjet* here stands for *inkjet printer* (and not *inkjet store*).

In modern natural language analysis systems, there is a tendency to use integrated information (syntactic, semantic, pragmatic) for ellipsis resolution (Hahn *et al.*, 1996). It is worth noting that the same tendency can be observed in anaphora resolution since anaphora is a similar phenomenon connected with reference, which has been investigated to a greater extent; see, for example, (Mitkov, 1997).

The dictionary we use in this paper to solve such syntactic problem as ellipsis can be viewed as a kind of accumulated lexical, semantic, and pragmatic information. This information is accumulated in our dictionary of correlation in the use of the words, supposedly because of lexical, semantic, and pragmatic reasons.

Dialogue also has been recently in the focus of linguistic research; see for example (Carberry and Lambert, 1999). It is well known that dialogic speech has the structure and characteristics that differ in a number of significant aspects from those of written text. One of the characteristic features of dialogue is the presence of elliptical constructions and highly elliptical sentences sometimes reduced to a single word standing for a whole sentence.

On the other hand, dialogue has important commonalities with what is usually understood by language in computational linguistics, i.e., written text. We base our work on the hypothesis that the "broken" dialogic speech has as the "understood" underlying structure the same language as written text, i.e., can be expanded to the "correct," complete sentences. This allows applying the grammar and dictionaries built on the material of written texts, to restore the elliptical constructions of the "broken" dialogic speech.

To explain our method, we will start from some language data exemplifying the problem, discussed in Section 2. In Section 3, we formulate the necessary conditions for the detection of the corresponding phenomenon. In Section 4, we discuss the structure of the dictionary used by the algorithm of ellipsis resolution and then describe the algorithm itself. Experimental results are discussed in Section 5. In Section 6, the limitations of the method and future work directions are discussed.

2 Ellipsis in Dialogue

As we had mentioned above, ellipsis is the phenomenon of structural incompleteness of a sentence that is expected to be restored by the listener as "obvious." There are two most frequent types of ellipsis:[1]

- *Contextual ellipsis*: the omitted part of the sentence is restorable from the immediate previous context, normally being a repetition of a previous word or phrase, see example 1.
- *Pragmatic ellipsis*: the omitted part of the sentence is to be restored by the listener based on the extralinguistic situation. There are other similar types of ellipsis where other kinds of knowledge are used to restore the omitted words. An example of pragmatic ellipsis is:

2. Both participants of communication see a strange man.
   "*Drunk*," says one of them.

Here the elliptical noun phrase stands for "*drunk man*" and the complete sentence for "*This is a drunk man.*"

Since our motivation is human-machine dialogue and current computers typically do not have any means to participate in real-world situations, such type of dialogues is irrelevant for our goals. Note that at least when the referent (or a significant clue for resolving the reference) is of extralinguistic nature, the dialogue viewed out of the situational context—i.e., only as a text—is not coherent. Thus, we do not consider pragmatic ellipsis.[2]

Instead, we will concentrate on contextual ellipsis, namely on its type most frequent in human-machine dialogue—short answers or clarifications about an object or notion recently introduced in the dialogue. Let us consider several examples that we will refer to throughout the article. In each example, the words involved in the ellipsis are underlined. The type of examples is explained in the next section.

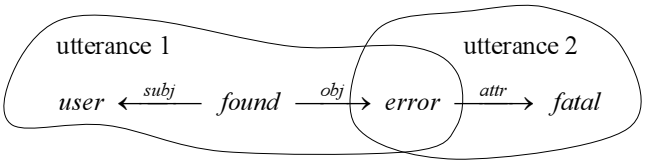
3. Type: (N) → Adj (see also example 1)
   "*The user has found an error in the program.*"
   "*Fatal?*" (Standing for "[*Is this a*] *fatal error?*")
4. Type: (Adj) → Adv
   "*The car uses a dangerous tindery fuel.*"
   "*Ecologically?*"
5. Type: (Adv) → Adv
   "*Does he speak English fluently?*"
   "*Incredibly.*"
6. Type: (V) → Adv
   "*Have you found the software you looked for?*"
   "*Easily.*"
7. Type: (V) → N_{obj}
   "*John was eating.*"
   "*Potatoes, I guess.*"
8. Type: (V) → $N_{indirect\ obj}$
   "*The error was found by the end user.*"
   "*In the program?*"

[1] Some authors include to ellipsis a very frequent in the languages of the world phenomenon of omission of the copula, e.g., Ebonics "*Leslie the boss*," Russian "*Lesli – boss*" or "*Lesli boss*" 'Leslie is the boss'. Other authors classify such phenomena as grammatical, i.e., as normal grammar rules in these languages (Mel'čuk, 1995; Sag & Wasow, 1999). We agree with the latter and thus do not consider such cases as ellipsis.

[2] For a discussion of multimodal dialogs see, for example, (Villaseñor, Massé, and Pineda, 2000). For reference resolution in multimodal dialogues see (Pineda & Garza, 2000).

9. Type: (V) → N_{subj}
   "*Oh damn, again cycling!*"
   "*The program?*"
10. Type: (N) → N
    "*You can find this city on the map.*"
    "*Of Europe?*"
11. Type: (N) ← V
    "*Did John read the book?*"
    "*Only looked through.*"

The problem of interpretation of these examples consists in restoring the omitted structural element. Traditionally ellipsis resolution is viewed as filling in the gap in the syntactic structure by substitution, say, *fatal error* for "[gap] *error*" in the example 3. However, for the purposes of the coherence maintenance in the dialogue it is more appropriate to view the interpretation as establishing a link (similar to a syntactic one) between the words *fatal* and *error* across the utterances. This produces a connected semantic representation of the whole dialogue:



(without duplication of the word *error*). In addition, an approach that treats ellipsis resolution similarly to anaphora resolution has the advantage of allowing to borrow the methods from the area of anaphora resolution, which is a more developed field than ellipsis resolution.

3 Detection and Resolution of Ellipsis

To resolve ellipsis, the potential source of the elliptical link is to be detected and then an appropriate target is to be chosen. Below we discuss these two problems in this order.

3.1 Elliptical relation. Detecting the source

As we have discussed, we suppose the existence of a link between a word in the contextually elliptical phrase and a previous word filling the gap. Let us call such a link *elliptical relation*. In the example 3, such a relation holds between the words *fatal* and *error*, the word *fatal* being the source of the relation and the word *error* being its *target*. We denote this as *error* ← *fatal*.[3]

To simplify our terminology and to emphasize the commonality of elliptical relation with anaphoric one, we will also use the term *an (elliptical) antecedent* for the target of elliptical relation.

As we have seen, restoring these links in a dialogue is crucial to maintain its coherence since this makes the semantic representation of the whole dialogue connected (as discussed in the last paragraph of Section 2). There are three main problems in establishing such links in a specific dialogue:

- Detecting the very presence of an elliptical link,
- Detecting its source,
- Detecting its target (the antecedent).

We will not solve the first of the problems directly. Instead, we will try to find a plausible pair of source and target words, which will thus indicate the presence of the link.

Since by ellipsis we understand structural incompleteness of the sentence, the potential source of the elliptical relation is easy to detect: this is a syntactically obligatory element absent from the surface of the sentence. The most common types of syntactically obligatory elements are the following:

- *Missing governor*: A word that cannot be the root (head) of the sentence is not modifying any other word, i.e., has no governor. Typical examples are an adjective not attached to any noun, as in the example 3; an adverb not attached to any verb, adjective, or adverb, etc.
- *Missing dependent*: A word that requires a modifier is present in the sentence, but there is no required modifier attached to the word. Typical examples are verbs that subcategorize for a component absent from the sentence, as in the example 11; nouns and adjectives also can have subcategorization patterns and thus can indicate this type of ellipsis.

Accordingly, our examples can be classified by the type of the missing syntactic link. In the previous section, the first line of each example indicates its type. The type specifies the parts of speech (noun, verb, adjective, adverb) of the two words involved (which are underlined in the utterances). In each pair, the first word is explicitly present

[3] The direction of this relation is independent from the direction of the implied syntactic relation, which is more often the opposite: *error* $\xrightarrow{attr}$ *fatal* .

in the first utterance but omitted in the second one. E.g., in the example 3 the noun *error* present in the first utterance is omitted in the second one ("*Fatal?*") where *fatal* stands for *fatal error*. In the type, we indicate the omitted element—antecedent—with parentheses: $(N) \rightarrow Adj$ means that N (noun) is the antecedent and the Adj (adjective) is the source.

The arrow indicates the direction of syntactic dependency (possibly mediated by a preposition). Since in all our examples the antecedent appears before the source, the arrow is left to right for missing governor and right to left for missing dependent.

The problem of parsing incomplete sentences is rather difficult technically because of ambiguity. This is why we restricted our considerations to short answers or clarifications, whose syntactical patterns are usually simplified—in particular, they can be elliptical. To build semantic interpretation of the dialogue, in many cases it is enough to restore the elliptical utterance to a complete noun phrase.

One of the possible ways to detect ellipsis in simple short phrases is based on heuristic patterns: say, a standalone adjective indicates ellipsis. This method is very simple, though not reliable for longer phrases. Another, more solid (though more complicated) way to parse incomplete sentences consists in introducing gap features in the grammar (Allen, 1995; Sag & Wasow 1999) and allowing the sentences to contain gaps.

Note that the syntactic rule that reports a potential antecedent unambiguously determined the type of the relation and in particular its syntactic dependency direction—missing governor or missing dependent.

Not any structural incompleteness represents elliptical relation. First, the ellipsis can be of not contextual type. Second, structural incompleteness can be due to some reason other than ellipsis, e.g., a mere error or a prematurely interrupted utterance. Any gap in the sentence, however, can potentially represent a source of elliptical relation. As we have said, we will determine the presence of such relation by looking for a plausible antecedent.

Here we will not develop in more detail the technique for syntactic parsing of elliptical sentences and thus establishing the potential source of the elliptical relation. In the rest of this paper, we will concentrate on finding the antecedent for a given type of gap.

3.2 Properties of elliptical antecedents

Now, given a potential source of a hypothetical elliptical link, we are interested in finding a plausible antecedent. If there is no plausible antecedent, then either the structural incompleteness of the phrase is not due to ellipsis or the ellipsis is not contextual.

On the other hand, since we are interested in finding only one—the most plausible—antecedent, a major problem is, as usually in language processing, ambiguity. For example, how should the program guess that in the example 3 it is the *error* that is *fatal* rather than the *program* or the *user*?

We will call the lexemes 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* $\leftarrow$ *error* is highly probable; however, **fatal* $\triangleleft$ *error* is not true since **fatal* $\rightarrow$ *error* is impossible.[4] Also **program* $\triangleleft$ *fatal* is not true since **fatal program* is not a combination that can be normally expected in the text. Other examples: *error* $\triangleleft$ *program* (*error in the program*), *user* $\triangleleft$ *to find* (*user found an error*), *to find* $\triangleleft$ *error* (*to find an error*). Note that this is a lexical property of the corresponding words taken out of context. It is specified in the dictionary, rather than represents a syntactic relation in a specific context.

We will use the same symbol for whole syntactic categories, such as parts of speech. For example, by $V \triangleleft N$ we will denote the fact that a verb in general can syntactically govern a noun. At the same time the following relations do *not* hold: **Adj* $\triangleleft$ *Adj*, **N* $\triangleleft$ *Adv*, **Adv* $\triangleleft$ *V*, etc. On the other hand, the following relations do hold: $N \triangleleft Adj$, $Adv \triangleleft Adv$, etc. (see the examples above). Similarly, we could write $VP \triangleleft NP$ for a verb group and a noun group. Note that this property for the parts of speech is specified in the grammar.

As we have seen, in the elliptical constructions under consideration, the target and the source of the relation form a syntactic word combination. Thus, denoting x and y the antecedent and the source of the elliptical relation, we get that the corresponding condition $x \triangleleft y$ (example 3) or $x \triangleright y$ (example 11) holds. This gives our first condition:

Condition 1. Elliptical relation $x \rightarrow y$ between the antecedent x and source y is possible only if $x \triangleleft y$ (correspondingly, $x \triangleright y$ for $x \leftarrow y$).

If the lexical information is not available for a specific word, a relaxed condition involving syntactic or semantic categories of the correspondent words can be used. In our implementation, semantic groups are handled internally by the dictionary; see the next subsection.

[4] An asterisk denotes a deliberately incorrect formula or example.

If a simple heuristic-based method is used to detect the source of ellipsis, then a relaxed condition based on syntactic categories can be useful for quickly filtering out impossible candidates. For the parts of speech POS (x) and POS (y):

Condition 2. Elliptical relation $x \rightarrow y$ is possible only if POS (x) $\triangleleft$ POS (y) (correspondingly, POS (x) $\triangleright$ POS (y) for $x \leftarrow y$).

As our examples show, nearly any combination of parts of speech for which Condition 2 holds can be found in an elliptical relation with some appropriate specific lexemes.

Since we treat the type of the ellipsis under consideration as a potential filling of the gap in the utterance, Condition 2 can be generalized as follows:

Condition 3. Elliptical relation between the antecedent x and source y is possible only if the utterance that contains y remains syntactically correct when x is inserted in it and syntactically connected to y.

This condition is equivalent to the solid grammar-based procedure of detecting ellipsis and thus is computationally expensive.

Condition 1 characterizes the lexical combinability and is the most reliable. Condition 2 characterizes the most general syntactic combinability (parts of speech) and is the less reliable. Condition 3 is more specific than Condition 2 and thus more reliable, especially in combination with Condition 1. In the rest of the paper, we concentrate on the discussion of Condition 1.

4 Resolution of Ellipsis Using a Combinatory Dictionary

To restore the elliptical relations in the dialogue we try to find a highly probable antecedent, on the lexical basis. In this section, we describe the structure of the corresponding dictionary and the algorithm that uses it to check Condition 1.

4.1 The dictionary for detecting elliptical antecedents

To test Condition 1, a procedure for checking the combinability of the two lexemes is used. For a given pair of lexemes x and y and the direction of the link, the procedure checks whether $x \triangleleft y$ (correspondingly, $x \triangleright y$) holds. The result is a value between 0 and 1, with 0 indicating that x and y are not combinable, 1 indicating that they are, and other values indicating different degree of certainty in that they might combine.

Internally, the procedure relies on two dictionaries: a dictionary of word combinations and a thesaurus.

The dictionary of word combinations lists, for each headword, the words that can syntactically combine with it, together with the prepositions (or grammatical cases) used in such a combination. In addition to the mere fact of combinability, in some cases the dictionary can specify a quantitative measure of the probability (frequency) of the combination in the texts; such a statistical bigram dictionary is typically trained on large text corpora (Yuret 1998).

The thesaurus lists for each headword its hypernym (more general concept), if there exist any. There might exist more than one hypernym for a headword. One source of such multiple hierarchical relations is the multihierarchical nature of word meanings: e.g., the concept *girl* belongs to both *young* and *female*. Another source of multiple hierarchical relations is ambiguity in the cases when different word meanings are not explicitly distinguished as different headwords.

In our experiments, we use the CrossLexica system (Bolshakov, 1994a, 1994b, Bolshakov and Gelbukh, 2001b) that provides the necessary procedure for checking combinability. A screenshot of the system is shown on Figure 1. The English dictionary used for our experiments contained approximately 120,000 word combinations (collocations) and 150,000 semantic links in its thesaurus and is currently under active augmenting.[5]

The system works in the following way. If for the two input words x and y, the combination $x \rightarrow y$ (correspondingly, $x \leftarrow y$) is explicitly listed in the word combinations dictionary, then $x \triangleleft y$ ($x \triangleright y$). If there is a numeric value (probability) associated in the dictionary with this word combination, it is returned, otherwise 1 is returned.

If the combination is not found in the dictionary, the system tries to infer it using its built-in thesaurus (Bolshakov and Gelbukh, 1999, 2001a). The inferred combinations, if found, are assigned much less probability value than the combinations explicitly listed in the dictionary.

For instance, in example 3 the CrossLexica system finds the correct antecedent because the word combination *fatal*

[5] We also experimented with Russian language texts; currently Russian version of CrossLexica is more complete (700,000 word combinations and 1,000,000 unilateral semantic links).

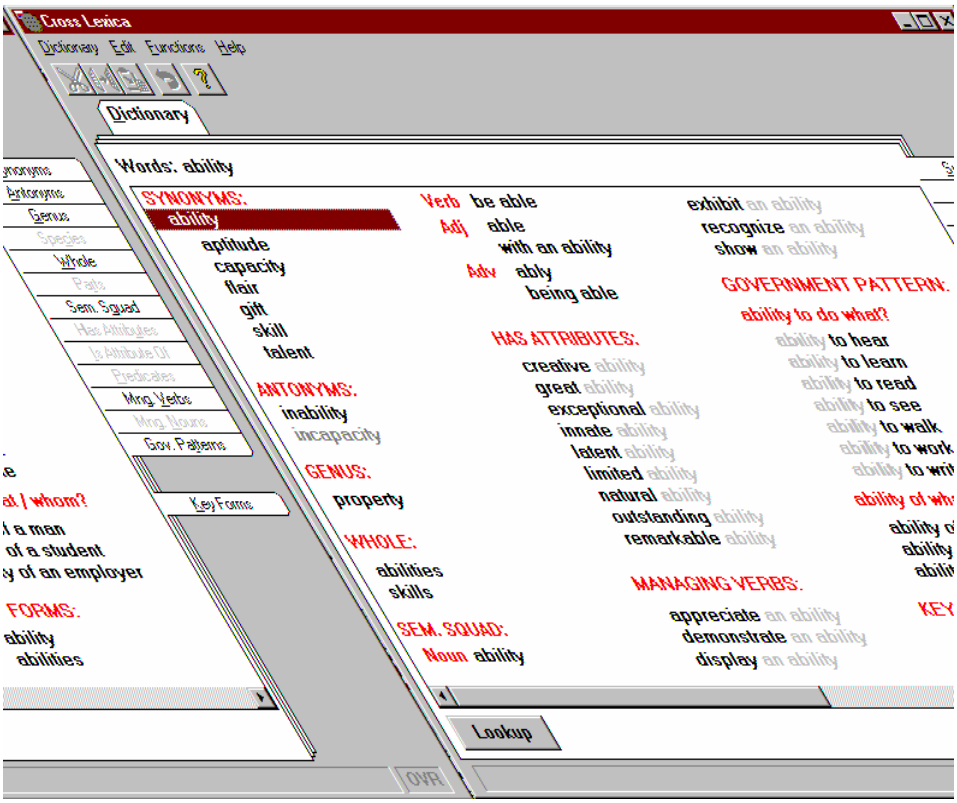


Figure 1. Screenshot of CrossLexica.

error is explicitly listed in the dictionary, while **fatal program* and **fatal user* are not.

The inference mechanism is based on the observation that some lexical relations—such as synonyms, hyponyms, etc.—are transparent for syntactical combinability. Namely, it works in the following way.

Let $a > b$ denote the fact that the concept a is a hypernym of b, e.g., *city* > *London*. Let, for brevity, R stand for any one of the relations $\triangleleft$ or $\triangleright$. Let D be the dictionary of the word combinations. By $(a \; \boldsymbol{R} \; b) \in D$ we will denote the fact that the dictionary explicitly lists the word combination $a \; R \; b$.

The information about combinability of the word w with a word x in a particular direction R is inherited from another word u if any of the following conditions holds:

- u is the nearest superclass (hypernym) of w for which the information on the required part of speech is available. Namely, if $\exists \; u > w$, $(u \; R \; x) \in D$, and there is no a, b such that $u < a < w$, POS (b) = POS (x), and $(a \; R \; b) \in D$, then $w \; R \; x$.

  Example: *nice city* ⇒ *nice London*, since *city* > *London* and (*nice city*) ∈ D.

- u is so-called dominant synonym of w, i.e., the near-synonym that does not have any additional meaning (and thus is similar to a very close hypernym). Namely, if there exists u which is a synonym of w and is marked as the dominant element of its synonym group, and $(u \; R \; x) \in D$, then $w \; R \; x$.

  Example: *happy face* ⇒ *happy mug*, since *face* is a synonym of *mug*, it is marked as the dominant (most neutral) in this synonym group, and (*happy face*) ∈ D.

- w is a noun and u is the other number of the same noun.[6] In other words, if both u and w are nouns, they are marked as number variants of the same lexeme, and $(u \; R \; x) \in D$, then $w \; R \; x$.

  Example: *to open a door* ⇒ *to open doors*, since *door* and *doors* are marked as the number variants of the same noun.

The dictionaries of CrossLexica give various marks associated with individual words or word combinations. These marks help preventing the inference in the cases

[6] Since plural and singular nouns have different combinability, they are handled separately by CrossLexica

when the results are probably incorrect. Specifically, a relation is *not* inherited from u to w if any of the following exceptions apply:

- x is an adjective expressing the selection of a subclass of the modifying noun $u > w$.
  Example: *African city* $\not\Rightarrow$ *African London*, since the word *African* has a special mark telling that it expresses a subclass rather than a property—even though (*African city*) $\in D$ and *city* > *London*.
- The relation u R x is marked as idiomatic or has any stylistic mark.
  Example: *hot dog* $\not\Rightarrow$ *hot poodle*, where the combination (*hot dog*) $\in D$ has a special mark telling that its meaning is idiomatic—even though (*hot dog*) $\in D$ and *dog* > *poodle*.

If for the given pair w and x, $w\,R\,x \notin D$, there are other words of the same part of speech listed in combination of the same type with any one of them—i.e., $\exists\ b$ such that either $(w\,R\,b) \in D$ and POS (b) = POS (x) or $(b\,R\,x) \in D$ and POS (b) = POS (w)—then the probability of any inherited relation is considered very low. In practice, in this case our procedure does not attempt to find any relation and just returns 0, which greatly speeds up the processing. This is based on the hypothesis that if the authors of the dictionary of word combinations listed the high-probable combinations of a given type for a given word, they have listed them all, so that no other combinations of the same type can be inherited.

4.2 The algorithm for detecting elliptical antecedents

The algorithm for detecting potential antecedents works as follows. At the first stage, the cases of structural incompleteness in the phrases are detected, as it was discussed in Section 3.1. In particular, we use heuristic patterns to detect elliptical phrases and the hypothetical sources of elliptical relation.

For each case of structural incompleteness, the hypothesis of the possible elliptical link is checked by trying, word by word, right to left, the candidates in the previous utterance(s) and verifying Conditions 1 to 3. The algorithm stops when the distance becomes too large or when a plausible antecedent is found. The scheme of the algorithm is shown on Figure 2.

As usually with search algorithms, there are variations depending on the desired time to precision tradeoff. If precision is important, then the best candidate must be chosen. For this, the algorithm should find all the possible candidates within some reasonable distance from the source and then choose the best one, as shown in Figure 2. If time is critical, the algorithm stops on the first found candidate that is better than some threshold (not shown on Figure 2).

The algorithm relies on the notion of a distance between two words in the text. There are various possible methods to measure distance. The most obvious one is the linear distance: the number of words between the two locations.

Taking into account another measure—syntactic distance—can improve the precision of the results, though it is computationally more expensive. The idea of the syntactic distance is that the words that form a syntactic unit (a complete constituent) seem to occupy less space in the human temporal memory and thus contribute less to the distance measure than the same number of unrelated words. This is similar to the same effect at the letter level: a word of 10 letters contributes less to the distance than two words of 5 letters each.

However, here we will not discuss the details of syntactic distance, since its precise definition needs more investigation. Thus, the reader can think of the distance as the number of words between the two given words plus one.

There are other ways the syntactic structure can affect the probabilities of the candidates. For instance, in the current version of the algorithm, we assume that the antecedent can be present in the constituent of any level, including subordinate clauses. However, at least in some cases, the fact that the possible antecedent is located in a subordinate clause can affect its plausibility. Here we leave aside the details of handling subordinate clauses; see some additional considerations in (Gelbukh and Sidorov, 1999).

The distance between the source of the elliptical relation and the potential antecedent affects the plausibility weight of the candidate. The weight is calculated as some combination of the distance and the lexical probability (combinability) of the two words out of context. The less the combinability and the greater the distance, the less plausible the candidate. The exact formula for such combination is a topic for further investigation. Currently we use the formula

$$w = \frac{p}{distance + \alpha}, \qquad (1)$$

where p is the lexical probability returned by the CrossLexica system, or, in the case of Condition 2, the small value ε, and α is a constant necessary to smoothen the effect of the distance when it is very short.

This formula affects only precision (but not the speed) of the algorithm. Specifically, the value of the parameter α

```
Apply the heuristic syntactic patterns to detect possible sources of ellipsis (words missing required links)
repeat for each possible source x
    repeat for each word y, from right to left, starting from the last word of the previous utterance
        let distance = the distance between x and y
        if Condition 1 holds for x and y according to CrossLexica system then
            if Condition 3 holds then
                let weight w = combination of distance and the probability returned by CrossLexica;
                the pair (y, w) is remembered as a possible candidate,
        else
        if Condition 2 holds then
            if Condition 3 holds then
                let weight w = combination of distance and a little value ε;
                the pair (y, w) is remembered as a possible candidate,
        if distance < threshold then
            break the loop and go to next x
        else pass to the next word y to the left
    end
end
Out of the set of stored variants, choose and return the one with the highest weight w.
```

Figure 2. The algorithm for finding the elliptical antecedent(s).

should not be too large for too remote candidates to be correctly excluded. However, since we considered only short dialogues, more experiments are necessary to determine its optimal value in case of long dialogs or long utterances. Some intuitive argumentation for its approximate value is as follows. Since the words in one simple phrase (say, preposition + noun + adjective) are ordered according to some fixed laws of language,[7] they should be considered as equally distant from the next sentence. Therefore, the smoothing constant is expected to be of the same order of magnitude as the average size of such a phrase; say, we expect $\alpha = 3$ to be appropriate.

As to the threshold, it affects speed of the algorithm and (given a large enough value) does not affect its precision since too distant candidates are anyway assigned little weights by the formula (1). Thus, the choice of this parameter depends on the desired balance between performance and correctness, provided that it is not too small. Since in most cases the correct candidate is within the last sentence, we consider it enough in practice to choose the threshold equal to the average utterance length, say, *threshold* = 15, though a larger value is desirable whenever it is affordable.

Performance of the algorithm proves to be acceptable in comparison with the speed at which the utterances are produced by the human. Indeed, the algorithm consists of two nested cycles, which for short utterances results in approximately $10 \times 10 = 100$ iterations in order of magnitude.[8] At each iteration, CrossLexica's database is accessed once, and only if the word combination under consideration was found in the dictionary, syntactic analysis is performed. Since the morphological data is cached between such analyses for the same utterance, the syntactic analysis is fast enough. In our experiments, the whole process took less than one second per utterance.

5 Experimental results

Since our goal is to model the natural behavior of a human interlocutor, we used for our tests real human-human dialogues, since they model the desired structure of a natural human-machine dialogue. Namely, we have chosen a corpus of 120 questions and answers of real-world dialogues of the telephone informational service operators with the clients.

In this corpus, we found five non-trivial cases of ellipsis. We did not consider the trivial cases where the second utterance consisted in a mere check-back repetition of some words from the first phrase.

[7] Of syntactic, pragmatic, theme-rhematic, etc. nature.

[8] For very long sentences, the scope for *y* is restricted by the threshold, which is psycholinguistically justified.

After the automatic processing, the results were compared with manual markup. The algorithm resolved correctly all 5 cases. Here are two examples of the trace of the algorithm:

12. "*Hallo, please, the central hospital, the physician-in-chief.*"
    "*For adults, for children*?"

    Source: *adults*.

    Antecedents considered by the algorithm: *hospital, physician-in-chief*.

    Antecedent found: *hospital* ("*Hospital for adults*?").

    Source: *children*.

    Antecedents considered by the algorithm: *hospital, physician-in-chief*.

    Antecedent found: *hospital* ("*Hospital for children*?").

13. "*Hi, the phone of the bus terminal, please.*"
    "*Central*?"

    Source: *central*.

    Antecedents considered by the algorithm: *phone, bus terminal*.

    Antecedent found: *bus terminal*.

6 Discussion and future work

Though the proposed method shows good performance on an average, it has—as any other method—its own limitations (their overcoming would require further investigation). For example:

- The algorithm does not take into account the context, the previous development of the conversation, or real-world knowledge about the situation discussed in the dialogue. For instance, the algorithm would resolve the ellipsis in the dialogue:

  14. "*The user found an error in the program.*"
      "*Is it mine?*"

  as *program of mine* (because the distance from *mine* to *program* is less than to *error*) even if from the previous conversation it is clear that the program is (or is not) of the second interlocutor and thus he or she could not ask whether it is his or her.

  Note that even a human cannot choose the correct antecedent in the example 14 outside of context: the interpretation *error of mine* can be expected in a conversation between a team manager and a programmer while the interpretation *program of mine* might be expected, say, in a conversation between a client and a software provider.
- Some words such as functional words (e.g., *mine*, *the same*), highly combinable words (e.g., *good*), etc. have nearly equal lexical probabilities to combine with almost any other word. In this case, as in the example 14, the algorithm always chooses the nearest candidate.
- The dictionary used by the algorithm gives the lexical probabilities average for the given language, which are not exactly suitable for any specific subject domain (e.g., medical vs. technical), specific company, specific user etc. In theory, for any such case there should be constructed a specific variant of CrossLexica. In practice, this is possible only when a large enough training corpus is available (here we do not discuss the methods of automatic extraction of the CrossLexica dictionary from a corpus).
- The algorithm uses the linear distance between words. However, the "psychological distance" (which we try to roughly estimate using the linear distance) depends on the syntactic structure of the sentence, on the specific words (does a preposition contributes to the distance equally to a verb? does this depend on the length of the word?), etc. A better modeling of the "psychological distance" would be the topic of a future work.

7 Conclusions

We have discussed a dictionary-based method and the corresponding heuristic algorithm for detecting the elliptical antecedents in a special kind of elliptical utterances frequent in human-machine dialogues.

The very fact that the phenomenon in question is present in the utterance at hand is checked by looking for a plausible elliptical antecedent. The algorithm is based on the necessary conditions for ellipsis resolution that we have formulated.

The algorithm uses a large dictionary with a rather simple structure. This dictionary contains information about word collocations. The system that uses this dictionary to measure the lexical combinability of the words is able to infer such information for the collocations absent in the dictionary based on the thesaurus and on the combinability information present in the dictionary.

Acknowledgements

This is a significantly revised version of the paper *Coherence Maintenance in Human-Machine Dialogue with Ellipses*, Proc. of MICAI-2000, Mexican International Conference on Artificial Intelligence, Acapulco, Mexico, 2000. The paper was selected for publication in the journal as one of the best papers of MICAI-2000.

The work was done under partial support of CONACyT, COFAA/CGEPI-IPN, and SNI, Mexico.

References

Allen, James. *Natural Language Understanding*, Benjamin/Cummings, 1995.

Ariel, M., "Referring and accessibility.", *Journal of Linguistics*, 1988, 24: 67–87.

Bolshakov, I.A., "Multifunctional thesaurus for computerized preparation of Russian texts.", *Automatic Documentation and Mathematical Linguistics*. Allerton Press Inc. Vol. 28, No. 1, 1994, p. 13–28.

Bolshakov, I.A., "Multifunction thesaurus for Russian word processing.", In *Proceedings of 4th Conference on Applied Natural language Processing*, Stuttgart, 13–15 October, 1994, p. 200–202.

Bolshakov, I.A., and **A.F. Gelbukh.,** "Enriquecimiento de la base de combinaciones de palabras por medio de un tesauro jerárquico" (in Spanish). In *CIC-99, Simposium Internacional de Computación*, November 15–19, 1999, Mexico D.F.

Bolshakov, I.A. and **A. Gelbukh**. "A Very Large Database of Collocations and Semantic Links", Proc. *NLDB 2000: 5th International Conference on Applications of Natural Language to Information Systems*, *Lecture Notes in Computer Science* N 1959, Springer-Verlag, 2001, pp. 103–114.

Bolshakov, Igor A., and **Alexander F. Gelbukh**. "A Large Database of Collocations and Semantic References: Interlingual Applications", *International Journal of Translation*, Vol. 13, No. 1–2, 2001, pp. 167–193.

Bosch, P., "Representing and accessing focused referents.", *Language and Cognitive Processes*, 1988, 3: 207–231.

Carberry, S., and **Lambert, L.,** "A process model for recognizing communicative acts and modeling negotiation subdialogues.", Computational Linguistics, 1999, 25 (1): 1–54.

Chierchia, G., *Dynamics of Meaning : Anaphora, Presupposition, and the Theory of Grammar*. University of Chicago Press, 1995.

Cowan, R. , "What are Discourse Principles Made of?", In P. Downing and M. Noonan (Eds.), *Word Order in discourse*. Benjamins, Amsterdam/Philadelfia, 1995.

Chafe, W., "Cognitive Constraints in Information Flow.", In R. Tomlin (Ed.), *Coherence and Grounding in Discourse*. Benjamins, Amsterdam, 1987. pp. 21–51.

Chafe, W., *Discourse, Consciousness, and Time*. The University of Chicago Press, Chicago – London, 1994. 327 pp.

Cornish, F., *Anaphora, Discourse, and Understanding : Evidence from English and French*. Oxford Univ Press, 1999.

Downing, P., and **M. Noonan** (Eds.), *Word Order in discourse*. Benjamins, Amsterdam/Philadelfia, 1995. 595 pp.

Fraurud, K., *Processing noun phrases in natural discourse*. Doctoral dissertation, Stockholm University, Stockholm, 1992.

Fraurud, K., "Cognitive ontology and NP form.", In T. Fretheim and J. K. Gundel (Eds.), *Reference and referent accessibility*. John Benjamins, Amsterdam, 1996. pp. 193–212.

Fretheim, T., and **J. K. Gundel** (Eds.), *Reference and referent accessibility*. John Benjamins, Amsterdam, 1996.

Gelbukh, A. and **Sidorov, G.,** "On Indirect Anaphora Resolution.", In *Proceedings PACLING-99, Pacific Association for Computational Linguistics*, University of Waterloo, Waterloo, Ontario, Canada, August 25–28, 1999, pp. 181–190.

Gelbukh, A. F., G. Sidorov, and **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.

Hahn, U., M. Strube, and **K. Markert**, "Bridging textual ellipses.", *Proceedings of the 16th International Conference on Computational Linguistics*, 1996. pp. 496–501.

Hellman, C., "The 'price tag' on knowledge activation in discourse processing.", In T. Fretheim and J. K. Gundel

(Eds.), *Reference and referent accessibility*. John Benjamins, Amsterdam, 1996.

Lambrecht, K., "Information Structure and Sentence Form.", In *Topic, Focus and the Mental Representation of Discourse Referents*. Cambridge University Press, Cambridge, 1994. 388 pp.

Mel'čuk, I. A., *The Russian Language in the Meaning–Text Perspective*. Wiener Slawistischer Almanach. Sonderband 39, Moskau–Wien, 1995.

Mitkov, R., "Pronoun Resolution: the Practical Alternative.", In S. Botley and T. McEmery (eds), *Discourse Anaphora and Anaphor Resolution*, Univ. College London Press, 1997.

Partee, B., and **P. Sgall** (Eds.), *Discourse and Meaning. Papers in Honour of Eva Hajičova.* Benjamins, Amsterdam/Philadelphia, 1996.

Pineda, L., and **G. Garza**. "A model for multimodal reference resolution", *Computational Linguistics*, 26, 2000.

Rose, Carolyn Penstein, B. Di Eugenio, L. Levin, and **C. Van Ess-Dykema**, "Discourse processing of dialogues with multiple threads.", In *Proceedings of the 33th ACL annual meeting*, 1995, pp. 31–38.

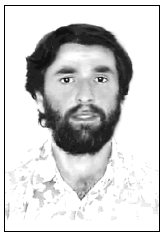
Sag, Ivan A. and **Thomas Wasow**. *Syntactic Theory: A Formal Introduction*, CSLI Publications, 1999.

Tomlin, R. (Ed.), Coherence and Grounding in Discourse. Benjamins, Amsterdam, 1987. 512 pp.

Villaseñor, L., **A. Massé**, and **L. Pineda**. "Towads a Multimodal Dialogue Coding Scheme," A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, Fondo de Cultura Económica, Mexico, to appear in 2001. See abstract in *Proc. of CICLing-2000*, CIC-IPN, Mexico City, 2000.

Ward, G., and **B. Birner,** "Definiteness and the English existential.", *Language*, 1994, 71: 722–742.

Yuret, D., *Discovery of linguistic relations using lexical attraction*, Ph.D. thesis, MIT, 1998. See xxx.lanl.gov / abs / cmp-lg / 9805009.



Alexander Gelbukh *was born in Moscow in 1962. He obtained his Master degree in Mathematics in 1990 from the department of Mechanics and Mathematics of the Moscow State "Lomonossov" University, Russia, and his Ph.D. degree in Computer Science in 1995 from the All-Russian Institute of the Scientific and Technical Information (VINITI), Russia. Since 1997, he is the head of the Natural Language and Text Processing Laboratory of the Computing Research Center, National Polytechnic Institute, Mexico City. He is a member of SNI, Mexico, since 1998. He is the author of about 150 publications on computational linguistics. See http://www.cic.ipn.mx/~gelbukh.*

Grigori Sidorov *was born in Moscow in 1965. He obtained his Master degree in Structural and Applied Linguistics in 1988 from the Phylological faculty of the Moscow State "Lomonossov" University, Russia, and his Ph.D. degree in Structural, Applied and Mathematical Linguistics in 1996 from the same faculty. Since 1998, he works for the Natural Language and Text Processing Laboratory of the Computing Research Center, National Polytechnic Institute, Mexico City. He is a member of SNI, Mexico, since 1999. He is the author of about 40 publications on computational linguistics. See http://www.cic.ipn.mx/~sidorov.*

Igor A. Bolshakov *was born in Moscow in 1934. He obtained his Master degree in Physics in 1956 from the department of Physics of the Moscow State "Lomonossov" University, Russia, his Ph.D. degree in Information Technology in 1961 from the VYMPEL Institute, Moscow, Russia, and his D.Sc. degree in Computer Science in 1966 from the same institute. He received the National Award of the USSR in Science and Technology in 1989. Since 1996, he works for the Natural Language and Text Processing Laboratory of the Computing Research Center, National Polytechnic Institute, Mexico City. He is a member of SNI, Mexico, with level II since 2000. He is the author of about 200 publications on theory of radars, theory of probability, and computational linguistics.*