Automatic analysis of elliptic sentences in the Thetos system[1]

Nina Suszczańska
Institute of Informatics
Silesian Univ. of Technology
44-100 Gliwice, Poland
nsuszcz@polsl.pl

Julia Romaniuk
Institute of Linguistics NASU
Math.&Struct. Linguist. Dept.
01001 Kyiv Ukraine
rdmytro@i.com.ua

Przemysław Szmal
Institute of Informatics
Silesian Univ. of Technology
44-100 Gliwice, Poland
pszmal@polsl.pl

Abstract

The Thetos system translates Polish texts, both monologic and dialogic, into the Polish sign language. The system handles limited ellipsis cases in three main types, specific for parallel and non-parallel structures and for simple dialogues. A rich collection of Polish verbs with their valence schemes is used in this purpose. Our experiments suggest a possibility to reduce the simple-dialogue type ellipses to the remaining two ellipsis types. From another side, it is possible to adopt proposed methods of elliptic sentence processing to different languages.

1 Introduction

Thetos is an experimental system for translating written texts in Polish into the Polish sign language (Szmal and Suszczańska, 2001). It was primarily intended to be the sign language interpreter in the deaf's first contact with the doctor. Then we decided to charge it with interpreting fairy tales to deaf kids. (For this raison examples used in this paper are fragments of tale texts). Due to that dualism, in our research we have – among others – to practically solve problems connected with pronominal anaphora and elliptic structure both in dialogues and in monologic texts. In this paper our focus is the problem of automatic recognition of zero substitution and its reconstruction.

We distinguish three ellipsis types:

Anaphoric ellipsis appears in parallel structures of connected sentences, a complete and incomplete one, for example:

Najstarszy z braci otrzymał młyn, średni (e_1) (e_2) osła, a najmłodszy (e_1), Janek, (e_2) tylko kota. (The eldest of the brothers got the mill, the middle (e_1) (e_2) the donkey, and the youngest (e_1), Johnny, only (e_2) the cat.) (1)

Non-anaphoric (situational) ellipsis appears in non-parallel structures, for example:

Wraca kotek do domu – koguta nie ma (e_4). (The kitty returns home – there is no cock (e_4)). Note: in English, instead of adverbial, it is rather the predicate that would be dropped, giving in effect the sentence: *(e_4') No cock there.* (2)

Dialogic ellipses are specific for simple "question – answer" dialogues, e.g.:

— Kto tam (e_6)? — Słabym głosem zapytała chora babcia. (— Who (is) there? — The granny asked with a faint voice.)

— To ja (e_7) (e_8), kochana babciu, twoja wnuczka — odpowiedział wilk, udając głos Czerwonego Kapturka. (— It (is) me (e_8), dear granny, your granddaughter — the wolf answered imitating the Little Red Riding Hood's voice.) (3)

— Teraz masz własne mieszkanie? (Have you your own flat now?)

— Teraz (e_9) własne (e_{10}). (Now (e_9) own (e_{10})) (4)

In more sophisticated dialogues, all three types of elliptic sentences can be met. That is

[1] This work was supported in part by the Polish Committee for Scientific Research in 2003-2005 under Grant 4 T11C 024 24.

why we consider all of them.

Zero substitutions are governed by a set of rules, which we call "hidden grammar". Those rules allow for dropping components that are actually unessential, well then such ones, without which the whole sentence or its fragment stays fully comprehensible. They also say how to fill up sentences with dropped constructions (Romaniuk, 2001).

2 Ellipsis handling within translation process

In the Polish sign language both anaphoras and ellipses may appear, but rules of using them are a bit different than in phonic language. What we do translate now is a modeled text composed of sentences in so called canonical form (Suszczańska et al., 2004). To transform input sentences to this form we have – among others – to reconstruct the structure of full sentences on the basis of elliptic ones; indeed, it is translation within translation.

First steps involve automatic syntax analysis. The parser (Kulików et al., 2004) produces syntactic representation of the input sentence in the form of a labeled graph. Its nodes represent syntactic groups, and edges – syntactic relations occurring between them. During semantic analysis we transform the syntactic graph and get a predicate–argument structure. In this stage we reconstruct elliptic structures. For each ellipsis type we have to apply a specific algorithm.

Automatic ellipsis type classification is a problem for itself. Many complications for the analysis issue from the fact that the syntax of Polish allows for free sentence word order. It also complicates algorithms for reconstruction of elided components. This is why we haven't till now elaborated algorithms for finding constructions to supplement incomplete sentences but for some cases of ellipses only.

3 Ellipses in parallel structures

Parallel structures are well-known constructions that belong to the good writing-style canon; see e.g. (Nesbitt, 2002). It has been proven that formal structure of sentence where an anaphoric relation appears may be shortened only on condition that structures of connected complete and incomplete sentence are parallel (Gardent, 1993). Such reductions result in anaphoric ellipses, a specific kind of anaphoric connections. Each of them is a zero anaphora meant as a lexical zero.

Surface structure of a sentence mirrors its deep semantic structure. That's why the causes and possibilities of shortening sentences can be sought in their semantic structure; with that, one can refer to the communicative structure of sentences. While analyzing transitions between theme and rheme we may catch the content distributed in the whole text, not only in one sentence. In the case of anaphoric ellipsis of predicative center (PC) or PC's component, the rheme goes to peripheries of the structure of the content of text. E.g. predicate, which typically represents the rheme, in case of anaphoric ellipsis is known from the previous sentence.

Let's take a two-part compound sentence:

Point A lies on the line AB, and point B – (e_1) on the line CD.

In the second component sentence, the predicate (e_1) is in a peripheral position in relation to the rheme. Well now, at angle of the semantics of the sentence (of conveying new information in it), it is not important and – in consequence – it may be dropped.

Let's return to the problem of parallelism of complete and non-complete sentence. With that we will be considering both deep and surface structure. Due to frequently used rule of speaking effort economy, components which are on peripheries of the semantic structure of the sentence may be elided. Since a sentence should be understandable for the receiver even in case of being elliptic, then it should have a readable structure. It should repeat the structure of the previous (parallel) sentence. (Actually, there may be in a sentence more zeroes than PCs; we set up a hypothesis that with anaphoric connection, when structures are parallel, all zeroes may by reconstructed.)

For the case of parallel structures, J. Romaniuk identified some rules for sentence abbreviation in Polish. They say that: 1^o PC or a component which is peripheral in relation to PC may be elided if the structures of the complete and the shortened sentences are parallel, because it is possible to rebuild the structure on the basis of context. 2^o In parallel structures, a sentence component that is dependent on the predicate is elided by stylistic reasons; if the missing component is signaled, it is reconstructed in effect of analysis of non-filled obligatory valence places.

These rules determine the way how to shorten a sentence and to leave it comprehensible in its context as well. On their basis we proposed an algorithm for reconstruction of structural and then lexical composition of elliptic sentences. It is only intended to analyze parallel structures of an incomplete sentence connected with a complete one.

It is easy to recognize two parallel structures in case where syntactic analysis gives an unambiguous parse of both sentences, from which the current one is elliptic, and the preceding one is not. Problems arise in case of ambiguous analysis.

We try to detect parallel structures: 1^o via searching for a dash „–" in the sentence (in such cases as shown in the example, the dash signals the position of ellipsis), 2^o via analyzing the morpho-syntactical traits of words as well as word order in the sentence, in that in parallel sentences the word order is preserved.

At the analysis of the deep structure of the sentence, we assume that in the valence scheme of the verb all obligatory places should be filled, and pretenders to an empty place should be searched in previous sentences. Trying to reconstruct the ellipsis, we limit the scope of searching for anaphora and antecedent by assuming that the anaphora is a PC or a component of PC, where PC should be meant as either subject, or predicate, or subject and predicate.

A similar algorithm works in a more general case, where lacking predicate has been detected in the sentence and the structure of the sentence has been established with using an adequate heuristics. An algorithm applicable for the deep sentence structure instead of the syntactic one is quite similar, too.

4 Ellipses in non-parallel structures

Non-parallel elliptic structures contain information about dropped components in the structure of the sentence and not in the context. Hence the context is useless for their repair. The most often dropped element is predicate that denotes a generalized movement or action. To resolve this type of ellipsis we add a verb (may be synthetic) of such kind to the structure of the sentence. The surrounding scheme for such verb should be fulfilled by any concrete verb of movement. For now we assumed a working variant of the verb and the scheme. It is an urgent task to examine all schemes of movement and action verbs in order to establish the set of common schemes and to define the desired generalizing one.

Evidently, the adopted solution is only the first of many possible steps in solving the problem. For example, there can be more than one dropped element, the verb can have a different meaning, etc.

In case of non-transitive verbs one can assume that the structure of elliptic sentences may be reduced to two subtypes:

$\textit{subject} - 0_{\textit{predicate}} - \textit{adverbial}$

$\textit{adverbial} - 0_{\textit{subject}} - 0_{\textit{predicate}} - \textit{adverbial}$

Having inserted a generalized verb, we can try to reconstruct the dropped subject by using a generalized scheme. Obviously, in this case the analysis becomes unambiguous and imprecise.

5 Ellipses in dialogic texts

As it was mentioned above, in extended dialogues we can meet sentences that contain ellipses of both types discussed in the two preceding sections. Besides that, anaphoric sentences with pronouns and other words intended to replace some elements are used. Our approach to analyzing anaphoric sentences and a method for searching antecedents was discussed in (Kulików et al., 2004).

In dialogues of „question – answer" type, the structure of both sentences is as a rule incomplete. For example:

— *Czy umiesz migać?* (*Can (you) sign?)*
// $0_{subject}$ – predicate *(5.0)*

— *Tak. (Yes.)* // $0_{predicate}$ – $0_{psubject}$ – adverbial *(5.1)*

or other answer variants:

— *Teraz tak. (Now yes.)*
// adverbial – $0_{predicate}$ – $0_{subject}$ *(5.2)*

— *Umiem. ((I) can)* // $0_{subject}$ – predicate *(5.3)*

— *Już umiem. ((I) already can)*
// adverbial – $0_{subject}$ – predicate *(5.4)*

In the sentence (5.0) subject can be easily rebuilt due to the grammatical form of the predicate, which obligatorily requires the subject *"ty" (you)*. The problem consists in detecting the type of shortening, and then – the corresponding reconstruction procedure.

The statement (5.1) is subject to anaphoric sentence analysis with substitutional word *"tak" (yes)*, whose antecedent is the preceding sentence as a total. That means that the structure of the sentence to be reconstructed, (5.1), will be entirely taken from the sentence (5.0), after its completion. There remains a problem with changing the subject expressed with the personal pronoun *"ty" (you)* into the pronoun *"ja" (I)*. The problem no more consists in mechanical change of the form of words, but in preserving both the formal representation of the content of the two utterances and the information that they all are concerned with the same person. In this case, for implementation purposes, we proposed to make use of the pronoun *"ten" (this)*. In consequence, our dialogue takes the following internal form:

— *Czy ten umie migać?* (*Can this sign?)* *(5.0')*

— *Tak. (Yes.)* ⇒ *Ten umie migać. (This can sign.)* *(5.1')*

In this point a new problem arises: transform the input sentence to a form which could be called a standard one.

So that, a possibility appears to reduce the third type of elliptic sentences to the precedent two. This our hypothesis requires additional research. It seems that analysis of the communicative structure of sentence could be helpful in this case.

6 Conclusion

There was no enough place to give a detailed description of algorithms discussed in this paper. The reader can find some additional information in (Kulików et al., 2004; Suszczańska et al., 2004). We have elaborated and implemented a part of exposed ideas. Experiments done in our Thetos translation system seem encouraging. We are intensively working upon accomplishment of remaining thoughts, since we find it necessary for the system to work satisfactorily.

References

Claire Gardent. 1993. A unification-based approach to multiple VP Ellipsis resolution. In: *Proc. of the 6th European Meeting of the ACL, Utrecht, The Netherlands.* Web page ftp://ftp.coli.uni-sb.de/pub/people/claire/multiplevpe.ps

Slawomir Kulików, Julia Romaniuk, and Nina Suszczańska, A syntactical analysis of anaphora in the Polsyn parser. In: *Proc. of the International Conference IIS:IIPWM'04*, Zakopane, Poland, 444–448

Scott Nesbitt. 2002. Parallelism. Web page http://www.unlv.edu/Writing_Center/Parallelism.htm, ed. A. Comeford, updated on 02 June 2002; accessed on 10 March 2005

Julia Romaniuk. 2001. Hidden Grammar of Anaphoric Ellipse. In: *Naukova spadshchyna prof. Semchynskogo i suchasna filologia.* Kyiv, 2:319-325 (in Ukrainian)

Nina Suszczańska, Przemysław Szmal, and Sławomir Kulików. 2004. Continuous Text Translation using Text Modeling in the Thetos System. *Int. J. of Computational Intelligence*, 1(4):338-341. Web page http://www.ijci.org/volumes/1304-2386-1.pdf

Przemysław Szmal and Nina Suszczańska. 2001. Selected Problems of Translation from the Polish Written Language to the Sign Language. A*rchiwum Informatyki Teoretycznej i Stosowanej*, 13(1):37-51