An Ellipsis Detection Method based on a Clause Parser for Arabic Language

Kais Haddar, Abdelmajid Ben Hamadou

Faculté des Sciences Economiques et de Gestion de Sfax
Laboratoire de recherche LARIS, B.P. 1088 Sfax - TUNISIA
Abdelmajid.Benhamadou@fsegs.rnu.tn

Abstract

Ellipsis has always been a centre of academic interest, be it in linguistic theory or in computational linguistics. In the present paper, we propose an ellipsis detection method for the Arabic language using the ATN formalism. This formalism, which is easy to implement, allows us to construct an efficient clause parser to detect ellipses in complex sentence structures. Then, we give a formal characterisation of ellipsis and a detection method based on an ATN grammar for well-formed clauses. Finally, we present some criteria to classify complex elliptical sentences in order to elaborate specific resolution algorithms.

Introduction

Ellipsis has always been a center of academic interest, be it in linguistic theory or in computational linguistics. A simple example of Arabic ellipsis is given in sentence (1):

(1) ذهب الولد الصغير الى المدرسة و [ذهب] أخوه الى الجامعة.

The young boy went to school and his brother [went] to university.

In this example, the second clause contains the deletion of a verbal phrase, the meaning of which is to be determined from the first clause. According to Dalrymple terminology (Dalrymple, Shieber and Pereira 1991), we call the clause that the verb is copied from the well-formed clause, and the clause which contains the elided verb the elliptical clause.

Syntactic accounts of ellipsis posit the copying of syntactic structure from the well-formed clause representation to the elliptical clause one (Gardent 1992), (Hardt 1993) and (Lappin 1992).

In this paper, we take an interest in the automatic treatment of the phenomenon of ellipsis in discourse written in Arabic language. To identify the well-formed clause in an elliptical sentence and to localize the elliptical ones, we propose to construct an Arabic grammar that recognises only the structure of well-formed clauses. Also, the grammar can be tightly restricted to define the sentence structure in terms of a small number of well-formed clause structures.

The originality of our approach resides in using a clause grammar to detect ellipses of complex sentence structures and in establishing a classification of elliptical sentences which facilitates the construction of efficient resolution algorithms.

The paper is structured as follows: in section 2, we present some related works and compare them to our approach. Section 3 is devoted to the presentation of the clause grammar used for the well-formed clause identification. In section 4, we give an ellipsis detection method based on a formal characterisation of elliptical sentences. In section 5, we propose an elliptical sentence classification based on the elliptical sentence syntactic structure. The last section presents our conclusions and futures works.

Related work

Several methods have been used to treat ellipses in their different forms and in different contexts. We cite in particular:

- The extension of the grammar by adding explicit rules or metarules that prevent ellipses (Gardent 1992), (Huang 1984) and (Sabah 1989).
- The relaxing of constraints: Consistency and ordering constraints can be relaxed on a second pass, allowing the detection and resolution of ellipses (Weischedel and Sondheimer 1982).
- The fitted parsing: It happens when the rules of a conventional syntactic grammar are unable to produce a parse for an elliptical sentence. This technique can be used to produce a reasonable approximate parse that can serve as input to remaining stages of natural language processing (Jensen, Heidorn and Richardson 1993).

Traditional linguistic theory distinguishes ellipsis forms (i.e., Right-node Raising, Gapping, VP-deletion, Coordinate Reduction) according to syntactic structure of elliptical clauses (Gardent 1992). Syntactic accounts of these ellipsis forms posit the copying of syntactic structure from the well-formed clause representation to the elliptical clause one. The used approaches are generally based on a complex sentence grammar that can not easily locate the elliptical parts of the sentence (Boguraev 1983) and (Huang 1984). In this approaches, elliptical clause

occurrence does not play any role. In our approach, we distinguish ellipsis classes according to syntactic structure of elliptical sentences by taking into account the following considerations:

- occurrence of elliptical clauses that a sentence contains,
- alternation of well-formed clause/elliptical clause,
- ellipsis form.

Unlike to the above mentioned approaches, our approach is based on another type of grammar namely an ATN clause grammar used to identify the well-formed clause and to locate the elliptical ones. Writing grammar rules is the first step in designing ATN parser which allows us to have more information about structure and meaning of clauses.

Clause parser

Syntactic analysis can be performed using complex sentence grammar. But, the specification of such a grammar is complex and requires much effort. Besides, it will be difficult to localise the elliptical parts of the sentence. Our proposition is to use a clause grammar. So, we minimize the effort required in writing the sentence grammar, we can easily identify the well-formed clause of an elliptical sentence and then locate the elliptical ones, and we can detect the elliptical fragments of an elliptical clause. In addition, the less structures are allowed, the higher is the potential ambiguity. Note that an Arabic sentence may begin either with a verb phrase (VP) or with a noun phrase (NP). We restrict our study to the sentences that begin with VP. As in (Lappin 1995), we consider an exception phrase fragment not as an instance of ellipsis, but as a displaced NP modifier. In the following, we give a subset of Arabic grammar rules:

$S \rightarrow VP + NP$
$S \rightarrow VP + (NP|PP) + \text{exception prep} + NP$
$S \rightarrow \text{interrogative pro} + VP + NP$
$S \rightarrow S + (NP|PP)$
$NP \rightarrow NP + \text{conj} + NP \mid NP_1 \mid NP_2 \mid NP_3 \mid NP_4 \mid \text{noun} \mid \text{proper noun} \mid \text{pro}$
$PP \rightarrow \text{prep} + (NP_1 \mid NP_2 \mid NP_3 \mid NP_4 \mid \text{noun})$
$NP_1 \rightarrow \text{art} + \text{noun} \mid \text{demonstrative pro} + \text{noun}$
$NP_2 \rightarrow NP_1 + AP$
$NP_3 \rightarrow \text{noun} + (NP_1|NP_2)$
$NP_4 \rightarrow \text{noun} + \text{adj} \mid NP_4 + \text{adj}$
$VP \rightarrow \text{negative adv} + \text{verb} \mid \text{verb}$
$AP \rightarrow \text{art} + \text{adj} \mid \text{art} + \text{adj} + AP.$

The clause grammar that we use for our approach is an augmented context-free clause grammar. To implement it, we construct first a Recursive Transition Network (RTN) (Woods 1970). The main network used in this section is shown in Figure 1. Then, we transform it into an ATN in order to add subcategorization principle and agreement constraints which are useful for the resolution process and to avoid ambiguity. In order to store these information, the ATN uses registers which are responsible for holding a structural element or phrase such as noun, verb, adj, NP, VP, etc.

A sentence like *John and Mary are alike* is accepted by our parser as a well-formed clause but not as a conjuntion as in other parser (Woods 1970). The mechanism also overcomes the inheritance of nonsensical problems.

S:

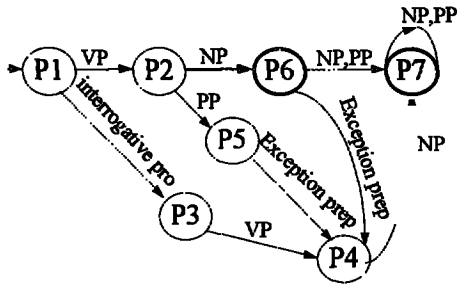

Figure 1: *The main network of a grammar rules fragment.*

The major components of our clause parser are:

- a set of lexical categories for the words of the Arabic language and a lexicon in which each word is assigned the categories that apply to it (Belguith and Ben Hamadou 1996);
- an Arabic clause grammar, using the same categories as the lexicon and whatever constructs are required by the type of grammar used to specify the well-formed clause structures;
- a parsing program, whose inputs are the clause to be processed, the lexicon, and the Arabic clause grammar, and whose output for each clause is a structural description of the clause.

Ellipsis detection method

We begin by recalling that it is possible, from the linguistic viewpoint, to adopt for the Arabic language the same typology of ellipses (i.e., Gapping, Right-node Raising, Coordination Reduction,...) as the one proposed for English and French (Gardent 1992), (Hardt 1993), (Lappin 1992) and (Rau 1985). In the following, we present the main characteristics of ellipses in the Arabic language.

Set forms of ellipses

Certain expressions, and notably letter endings, forms of wishes, certain orders or certain exclamations are elliptical. To understand them, it isn't necessary to construct the complete form. That's why, they are called false ellipses.

Examples:

Happy new year عيدا سعيدا [أتمنى لك] (2)
Fire, fire النار, النار [أحذر] (3)

These set forms of ellipses do not interest us because they can be resolved at the lexical level. But we rather focus our interest on the ellipses that oblige the reader to search in the context the lacking elements without which the message will be incomprehensible. The sentence (1) is an example of this.

The nominal ellipsis

The nominal ellipsis is the omission of the essential part of a noun phrase (i.e., head) that has to be taken from outside of the part of the previous discourse. Then in a sentence that contains a nominal ellipsis, one of the noun phrases is incomplete.

Example:

(2) الأخ الأكبر له من العمر عشرسنين و [الأخ] الأصغر ستة سنوات
The oldest brother is twenty and the youngest [brother] is six years old.

In example (4), the nominal phrase containing ellipsis is represented by an article and an adjective (the youngest). This treated example contains a nominal ellipsis.

The whole phrase ellipsis

The whole phrase ellipsis is distinguished from the nominal one by the fact that the omitted constituent may be a whole phrase (noun phrase or verb phrase). Therefore, in this category, we can distinguish the following ellipsis forms:

- Gapping:

(1) ذهب الولد الصغير الى المدرسة و [ذهب] أخوه الى الجامعة.
The young boy went to school and his brother [went] to university.

- Right-node Raising:

(5) بعت انا دراجة و اعطيت [انا] أختي صنارة.
I sells a bike and [I] gives a fishing rod to my sister.

- Coordination Reduction:

(6) طلبت من أخي دراجته و [طلبت] من أختي صنارتها.
I ask my brother for his bike and my sister for her fishing rod.

Formalism

According to what we have exposed above and the fact that we use an ATN clause grammar, we consider that an elliptical sentence in two distinct parts: an elliptical part preceded by a well-formed part that can be used as reference to resolve the ellipsis. The elliptical part might be composed of a succession of elliptical clauses generally separated with connection words (that we call *connectors*) which are coordinators, correlative conjunctions, subordinators, ...

To formally characterise this phenomena, we must first introduce some notations and definitions.

Elliptical clause. Let V be the finite Arabic alphabet. Let V^* be the set of finite words over V. $V^+ = V^*\setminus\{\varepsilon\}$, where ε is the empty word.

Definition 1. We call *clause* any finite sequence P of words of V^+, $P=w_1w_2...w_m$, $w_i \in V^+$ and $1 \le i \le m$. We note $\mathcal{P}$ the set of all clauses that can be constructed from V^+.

Let G be the Arabic clause grammar over V^+ describing thus a subset $L(G) \subseteq \mathcal{P}$. To verify whether a given clause P is generated by G, we introduce the following function that represents the acceptance function of the clause parser:

$$analyse : \mathcal{P} \rightarrow \mathcal{B} \quad \{true, false\}$$
$$P \mapsto \begin{cases} true \text{ if } P \in L(G) \\ false \text{ otherwise.} \end{cases}$$

Definition 2. A clause P is called *well-formed* if P is in L(G).

$$\boxed{P \text{ well-formed} \Leftrightarrow analyse(P)}$$

Note $\mathcal{S}$ the set of finite sequences of clauses of $\mathcal{P}$.

Definition 3. Let $S=P_0P_1...P_n \in \mathcal{S}$. We define the context of a clause P_i, $1 \le i \le n$ by the subsequence $S_i=P_0P_1...P_{i-1}$ formed from all clauses preceding P_i. We note by W^{Si} the set of all the words composing S_i.

Remark. The context of P_0 is the empty sequence.

Definition 4. Let $S=P_0P_1...P_n$ be a sequence of $\mathcal{S}$ and $P_i = w_{i1}w_{i2}...w_{im}$ a clause of $\mathcal{P}$, $0 \le i \le n$. Let the function *is_recoverable* be defined by:

$$is\ recoverable(P_i,S) = \begin{cases} true \text{ if } \neg analyse(P_i) \wedge W^{Si} \neq \emptyset \wedge \exists w'_{i1},...,w'_{im+1} \in W^{Si} \cup \{\varepsilon\} \\ \quad \wedge analyse(w'_{i1}w_{i1}w'_{i2}..w'_{im}w_{im}w'_{im+1}) \\ false \text{ otherwise.} \end{cases}$$

A clause $P_i, 1 \le i \le n$, is called *elliptical* if *is recoverable* (P_i,S).

Therefore, a clause Pi from a sequence S is elliptical when it isn't well-formed and when it can be completed with words of its context in order to be a well-formed one.

Remarks:

- In the case of a semantic ambiguity the ellipsis resolution will be treated in a semantic level.
- Examples as he is going out and it raining are treated like elliptical clauses.

Elliptical sentence. Let C be a subset of V* that we call set of connectors containing the empty word ε too.

Definition 5. A sentence Ph is a sequence of clauses separated with connectors of C. Ph has the following form:

$$\boxed{Ph = P_0c_1P_1...c_nP_n \text{ where } c_i \in C,\ P_i \in P.}$$

Definition 6. A sentence Ph is called elliptical if

1) the first clause of Ph is well-formed
2) at least one of the remaining clauses is elliptical.

Example 1. Let the sentence Ph be:

يجب ربط السلك الازرق بالنهائي أ و [يجب ربط] السلك الاصفر بالنهائي ب او [يجب ربط] السلك الاسود بالنهائي ج.

The blue cable must be connected to terminal A and either the yellow cable [must be connected] to terminal B or the black cable [must be connected] to terminal C.

P_0: The blue cable must be connected to terminal A.

P_1: the yellow cable to terminal B.

P_2: the black cable to terminal C.

Ph is elliptical because P_0 is well-formed and P_1 and P_2 are elliptical.

Outline of the detection method

Basing on this formalism, we propose an algorithm for elliptical clause location in a sentence consisting of three main stages performed by the clause parser:

–Search of connectors:

We search for all existent connectors in the analysed sentence relying on a connector lexicon. The obtained result is made up of a list containing clauses and of an other one containing connectors.

–Identification of the well-formed clause:

We make a filtering operation of the already obtained lists of clauses and of connectors based on the identification of the longest well-formed clause of them. Once the well-formed clause is identified, the two lists are updated by removing the well-formed clause components from them.

–Labelling of the remaining clauses:

We attribute to each clause of the remaining clause list an etiquette meaning that the clause is either elliptical or not. Labelling allow us to localise the elliptical clauses of the list.

Therefore, in order to have an elliptical sentence, it must begin with a well-formed clause and it must contain at least one clause labelled by the function *etiquette*.

At the end of those three stages, we dispose of enough information to classify elliptical sentences and later to establish resolution algorithms (Haddar and Ben Hamadou 1997).

Classification of elliptical sentences

We can distinguish different classes of elliptical sentences by taking into account the following considerations:

- occurrence of elliptical clauses that a sentence contains,
- alternation of well-formed clause/ elliptical clause,
- ellipsis type.

Formalism

We distinguish the following class definitions:

Definition 7. An elliptical sentence $Ph_e = P_0c_1P_1...c_nP_n$, $n \geq 1$, is called *homogeneous* if all clauses P_i, $1 \leq i \leq n$, are elliptical and have the same type of ellipsis.

Example 2. Take the elliptical sentence of example 1. Since P_1 and P_2 have the same type of ellipsis then Ph is homogeneous.

Definition 8. An elliptical sentence $Ph_e = P_0c_1P_1...c_n P_n$, $n>1$, is called *heterogeneous* if every two successive elliptical clauses P_i et P_{i+1}, $1 \leq i < n$, have not the same type of ellipsis.

Important remark. An elliptical sentence that does not contain two successive elliptical clauses is systematically heterogeneous whatever the type of ellipsis of clauses that the sentence holds.

Example 3. Let the sentence Ph be:

يجب ربط السلك الازرق بالنهائي أ و [يجب ربط] السلك الاصفر بالنهائي ب او يجب عليك ترك النهائيات مفتوحة.

The blue cable must be connected to terminal A and the yellow cable [must be connected] to terminal B or you must keep them free.

P_0: The blue cable must be connected to terminal A.

P_1: the yellow cable to terminal B.

P_2: you must keep them free.

Ph is heterogeneous because P_0 and P_2 are well-formed and P_1 is elliptical.

Definition 9. Let $Ph_e = P_0c_1P_1...c_nP_n$, $n>1$ be an elliptical sentence. Ph_e is called *mixed* if it is neither homogeneous nor heterogeneous.

Example 4. Let the elliptical sentence Ph be:

يجب ربط السلك الازرق بالنهائي أ والسلك الاصفر بالنهائي ب او السلك الاسود بالنهائي ج أو يجب عليك ترك النهائيات مفتوحة.

The blue cable must be connected to terminal A and neither the yellow cable to terminal B or the black cable to terminal C or you must keep them free.

P_0: The blue cable must be connected to terminal A.

P_1: the yellow cable to terminal B.

P_2: the balck cable to terminal C.

P_3: you must keep them free.

Ph is mixed because P_0 and P_3 are well-formed, and P_1 and P_2 are elliptical.

All these definitions, which we have presented, will allow us to establish criteria to apply for the membership class determination of the elliptical sentences with the intention to perform the resolution process.

Class determination

According to definitions 7, 8 and 9, in order to determine the class of an elliptical sentence, we have to know the forms of ellipses of the successive elliptical clauses. Let's recall that an elliptical sentence is systematically heterogeneous when it does not contain successive elliptical clauses. Then, we do not need to know the

ellipsis forms of elliptical clauses of this sentence. The ellipsis form determination of an elliptical clause is not usually easy to implement. That's why, we propose a heuristic, which is clearly easier to implement. This heuristic is based on the behaviour of connectors: when some connectors of the Arabic language are successive in a sentence, the clauses which follow them usually have the same form of ellipsis (if they are elliptical). To represent this heuristic, we propose to use a matrix M[k,k] on the set {0,1} where k = card(C). It is called " the Coherence Matrix of Clause Connectors " and it is defined as follows:

Let c_i , $c_j \in C$.

$M[c_i,c_j]=1$ means that the clause, which follows c_i if it is elliptical, contains the same type as the clause which follows c_j in a given sentence where these two clauses are successive.

$M[c_i,c_j] = 0$ otherwise.

This enables us to define a function that divides the set of connectors of a given elliptical sentence in partitions. The number of this partitions determines the membership class of the elliptical sentence. If the number of partitions is equal to 1 then Ph_e is homogeneous. If the number of partitions is equal to the number of clauses of the remaining clause list then Ph_e is heterogeneous. Otherwise, Ph_e is mixed.

Conclusion

In this study, we have proposed a formal characterisation of ellipsis for the Arabic language and a localisation process of elliptical parts of a sentence. Moreover, we have established classification criteria to define appropriate algorithms for ellipsis resolution. The clause parser, the classification algorithm and the detection algorithm are implemented with C++ within the framework of CORTEXA system (Ben Hamadou 1993).

Finally, our future works concern the improvement of the used clause grammar, the resolution of some ambiguities in particular anaphora cases in ellipsis.

References

Belguith, L.; Ben Hamadou, A. 1996. Marquage morpho-syntaxique robuste de mots ambigus écrit en Arabe non-voyellé. Forum de recherche en Informatique 96. Tunis.

Ben Hamadou, A. 1993. Vérification et correction automatique par analyse affixale des textes écrits en langage naturel: le cas de l'arabe non voyellé. Thèse d'Etat en informatique. Faculté des Sciences de Tunis.

Boguraev, B. K. 1983. Recognizing conjunctions without the ATN framework. Automatic Natural Language Parsing. Ellis Horwood.

Dalrymple, M.; Shieber, M. S. and Pereira, F. 1991. Ellipsis and higher-order unification. *Linguistic and Philosophy* 14: 399-452.

Gardent, C. 1992. A Multi-Level Approach to Gapping. In Proceedings of the Stuttgart Ellipsis Workshop Bericht Nr. 29. Stuttgart: Workshop.

Haddar, K. and Ben Hamadou, A. 1997. Formal Description of Ellipses in Arabic Language and Resolution Process. In Proceedings of IEEE ICIPS'97, 1775-1779. Pekin: IEEE ICIPS.

Hardt, D. 1993. Verb phrase ellipsis: form, meaning, and processing. A dissertation in computer and information science, University of Pennsylvania.

Huang, X. 1984. Dealing with conjunctions in Machine Translation Environment. In Proceedings of 10th International Conference on Computational Linguistics. Standford University, Palo Alto, California: International Conference on Computational Linguistics.

Jensen, K.; Heidorn, G. E. and Richardson, S. D. eds. 1993 *Naturel Language Processing: The PLNLP Approach*. Kulwer academic publishers.

Lappin, S. 1992. The Syntactic Basis of VP Ellipsis Resolution. In Proceedings of the Stuttgart Ellipsis Workshop Bericht Nr. 29. Stuttgart: Workshop.

Lappin, S. 1995. Computational Approaches to Ellipsis Resolution. In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics. Ireland: Conference of the European Chapter of the Association for Computational Linguistics.

Mast, M. et al. 1994. A speech understanding and dialog system with a homogeneous linguistic knowledge base. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16 no. 2, 179-193.

Rau, F. L. 1985. The understanding and generation of ellipses in a natural language system. Technical Report, Cs csd-85-227(dir), 30 pages, University of California Berkeley,.

Sabah, G. eds. 1986. *L'intelligence artificielle et le langage*. Reading, Mass.:Hermès.

Weischedel, R. M. and Sondheimer N. K. 1982. An improved heuristic for ellipsis processing. In Proceedings of the 20th Annual Meeting of the ACL, 85-88.

Woods, W. 1970. Transition networks grammars for natural language analysis, *CACM* 13(10): 591-606.