Which rules for the robust parsing of spoken utterances with Lexicalized Tree Adjoining Grammars ?

Patrice Lopez
LORIA & University of Nancy 1
lopez@loria.fr

David Roussel
Thomson CSF/LCR
roussel@thomson-lcr.fr

Abstract

In the context of spoken dialogue systems, we investigated a bottom-up robust parsing for LTAG (Lexicalized Tree Adjoining Grammars) that interleaves a syntactic and a semantic structure. When the regular syntactic composition rules fail, the syntactic islands and the corresponding partial semantic structures are combined thanks to additional local rules. We supply some descriptive limits of the grammar with these rules which depend on the immediate syntactic context of the islands. In this paper, we focus on their application to few spoken phenomena.

Introduction

Robust parsing is needed to cope with spontaneous uses of language. In particular, it is needed to deal with out-of-grammar utterances occurring in spoken man-machine interfaces. Because of the restricted application domain of such interfaces, it is expected that a robust architecture can interpret an unexpected utterance. This is illustrated with examples in French like :

(1) *Je voudrais un euh un billet pour Paris*
I would like a hum a ticket for Paris.

(2) *Départ à vers 20h.*
Depart at well at about 8 p.m.

(3) *Départ à huit enfin vingt heures.*
Depart at 8 I mean 8 p.m.

(4) *Je voudrais le premier qui part.*
I would like the first (one) which leaves.

(5) *Je voudrais un billet maintenant pour Paris.*
I would like a ticket now for Paris.

Those utterances represent a typical variety of spoken phenomenon namely a repetition (with hesitation) in (1), a self repair in (2), a correction in (3), a noun ellipsis in (4) and the insertion of an adverbe within a noun phrase (5). Parsing failures are respectively due to the impossible mapping of the parasite determiner into the derived tree (1), to the presence of a self repair (2) and (3), to a non canonical constituent (4) and finally to the prepositional attachment across the adverb barrier (5).

In the LTAG framework, we propose to represent the syntactic (partial) trees as connected routes (section 1.2). Adjunction, substitution but also additional local operation are applied to connected routes to make up the descriptive limits of the TAG formalism. Iu section 2, we expose a small set of rules which handle those routes -instead of the trees- and force operations between the trees. Assuming that local disruptions can be resolved by semautic mechanisms, some robust analyses receive a semantic counterpart in a synchronous TAG framework (section 3). Overgeneration remains a major challenge that we discuss in section 4. We will begin briefly explain the Connection driven parsing principles.

1 Connection driven parsing for lexicalized TAG

1.1 Connected routes

We define a *connected route* as a list of internal and root nodes crossed successively according to a left to right tree transversal (Schabes, 1994) until reaching a substitution or a foot node (included barriers) or an anchor (excluded barrier). Each elementary or derived tree can be represented as a list of connected routes. As the list of connected routes is ordered from left to right, we define the function *next* which gives from a given connected route the next connected route.

In (Lopez, 1998b) we explain how to lead a bottom-up bidirectional parsing focused on connected routes instead of focused on nodes as for other algorithms for TAG. Two data structures are used : the table of connected routes which gathers all the connected routes and a chart of parsing states which stores the sequences of well recognized anchors and their left and right connected routes.

1.2 Island representation with connected route

When no connected parse can span the whole sentence, the result of the parsing consists in representations of islands and its both right and left connected routes. An interesting point of this representation

(a) **Rule for hesitations :**

$$\frac{(i,j,\Gamma_G,\Gamma_D,idf) \qquad (j,k,\Gamma'_G,\Gamma'_D,idf') \qquad (k,l,\Gamma''_G,\Gamma''_D,idf'')}{(i,k,\Gamma_G,\Gamma_D,idf) \qquad (k,l,\Gamma''_G,\Gamma''_D,idf'')} \quad (\Gamma'_G = \Gamma_D = (root, H))$$

(b) **Rule for head ellipsis on the left :**

$$\frac{(i,j,\Gamma_G,\Gamma_D,idf) \qquad (j,k,\Gamma'_G,\Gamma'_D,idf')}{(i,k,\Gamma_G,\Gamma'_D,idf'')} \quad \begin{array}{l}(\exists(foot,X) \in \Gamma_D \wedge \\ (\exists(subs,X) \in \Gamma'_G \quad \vee \quad \exists(foot,X) \in \Gamma'_G))\end{array}$$

(c) **Rule for argument ellipsis on the right :**

$$\frac{(i,j,\Gamma_G,\Gamma_D,idf)}{(i,j,\Gamma_G,\Gamma'_D,idf')} \quad (\exists(subs,X) \in \Gamma_D \quad \wedge \quad \Gamma'_D = next(\Gamma_D))$$

(d) **Rule 1 for self repair :**

$$\frac{(i,j,\Gamma_G,\Gamma_D,idf_p) \qquad (j,k,\Gamma'_G,\Gamma'_D,idf_q)}{(i,k,\Gamma_G,\Gamma'_D,idf_r)} \quad \begin{array}{l}(\exists(v,w,\Gamma''_G,\Gamma''_D,idf) \in \Delta, idf \Rightarrow^* idf_p \quad \wedge \\ (\exists(root|internal,X) \in \Gamma''_D \quad \wedge \quad \exists(foot,X) \in \Gamma'_G) \vee \\ (\exists(subs,X) \in \Gamma''_D \quad \wedge \quad \exists(root,X) \in \Gamma'_G))\end{array}$$

(e) **Rule 2 for self repair :**

$$\frac{(i,j,\Gamma_G,\Gamma_D,idf_p) \qquad (j,k,\Gamma'_G,\Gamma'_D,idf_q) \qquad (k,l,\Gamma''_G,\Gamma''_D,idf_r)}{(i,l,\Gamma_G,\Gamma''_D,idf_s)} \quad \begin{array}{l}((\exists(foot,Y) \in \Gamma'_D \quad \wedge \\ (\exists(root|internal,X) \in \Gamma_D \quad \wedge \\ \exists(foot,X) \in \Gamma''_G) \quad \vee \\ (\exists(subs,X) \in \Gamma_D \quad \wedge \\ \exists(root,X) \in \Gamma''_G))\end{array}$$

Figure 1: Example of repairing rules for connection driven parsing

is that these connected routes correspond to the left and right context of the well recognized islands. A parsing state e is defined as the following 5-tuple :

state : (left index, right index, left connected route, right connected route, idf)

The two indices are the bounds of the input string covered by the island (anchors or the consecutive anchors) corresponding to the parsing state. During the initialization, we build a state for each anchor present in the input string. As each elementary and derived tree is identified, the anchor or the connected anchors belong to the tree *idf*. Those representation allows efficient partial parsing. This is the starting point of our robust strategy.

2 Robust Parsing with rules

2.1 Connected routes as flexible categories

A classical bidirectional TAG parsing (Lavelli and Satta, 1991) (van Noord, 1994) can not directly combine incomplete islands but it is possible to adapt the parser behaviour to the remaining syntactic material. Adaptations can be easily simulated by considering a connected route as a flexible category. The midly context sensitive power of LTAGs and CCGs has already suggested that elementary trees can be considered as flexible structured categories (Doran and Srinivas, 1994). According to the linguistic context, local rules can proceed to local adaptation of the routes. Then, the parser can try again to expand islands in both directions.

2.2 Inference rules system

The new derivation processes can be viewed as inference rules (Shieber et al., 1995) which use the parsing states described in section 1. The inference rules (Schabes, 1994) have the following meaning, if $(item_i)_i$ are present in the chart Δ and if the conditions are verified then add $(item_j)_j$ in Δ :

$$\frac{(item_i)_i}{(item_j)_j} \quad (conditions)$$

We note $\Rightarrow^*$ the reflexive transitive closure of the derivation relation between two elementary or derived trees : if $idf \Rightarrow^* idf'$ then the tree identified with idf' can be obtained from idf after applying to it a set of derivations.

The full system (including adjunction and substitution) increases the worst case complexity to $O(n^8)$ and deals with the following phenomena among others.

2.3 Ellipsis

The TAG formalism presents difficulties to describe these very common spoken productions. For instance, the parsing of utterance (4) does not succeed to find any complete derivation if *premier* does not exist in the lexicon as a noun or without the use of a sophisticated non lexicalized structure.

Two rules and their two symetrical configurations try to detect and recover respectively an empty head (b) and an empty argument (c). For instance, rule (b) attempts to make available an adjunction on a node marked for substitution if adjacent and categorial constraints are respected. When the rule (b) applies during the parsing of the example (4), the N1 node of the structure *N0-vouloir-N1* becomes candidate for an adjunction of the nominal auxiliary trees associated with sequences *le premier* and *qui part*.

2.4 Self repairs

The (Cori et al., 1997) definition of self repairs stipulates that the right side of the interrupted structure (the partial derived tree on the left of the interruption point) and the reparandum (the adjacent syntactic island) must match. Instead of modifing the parsing algorithm as (Cori et al., 1997) do, we consider a connected route matching condition. Rule (d) deals with self repair where the repaired structure has been connected on the target node. Rule (e) applies when the repaired structure has not been connected. In example (2), rule (d) detects the structural matching between the two prepositions à and *vers*. Then the rule reintroduces the target node on which the prepositional phrase *vers Paris* must be adjoined. The corresponding semantic tree of the à preposition is deleted.

Rule (e) remains relevant even if islands are separated by an hesitation (1) or a modification marker (3). Indeed the rule for hesitation (a) absords adjacent elementary trees whose head is a *H* node. Such a tree may correspond to different kind of hesitation forms. Rule (a) deletes an hesitation which can play the role of a barrier and a trace is kept in the chart.

3 Robust parsing with a Synchronous Semantic Tree Grammar

In combinatorial Grammars and lexicalized synchronous Semantic Tree Grammars (Shieber and Schabes, 1990) (Kallmeyer, 1997), predicate argument relations are directly encoded in the lexicon. This provide a syntax/semantic correspondence and additional well-formed criterion to validate an analysis (Abeillé, 1992). Robust parsing can take advantage of this property to only combine the syntactic islands in respect to the combination that the corresponding semantic fragments accept. In the case of robust parsing of an elliptic construction, the mechanism which allows such syntactic and semantic control consists in lambda abstractions.

For instance, the parsing of sentence (4) gives rise in the semantic tree shown Fig 2. Rules (b) and (c) combine islands without considering the empty argument. To control that the missing argument is present at the leftmost side of the partial derivation

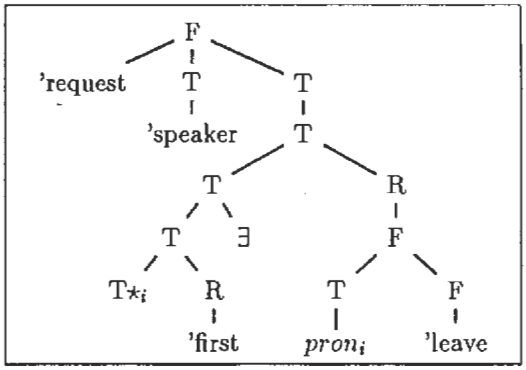

Figure 2: Logical form associated to the robust analyse of sentence (5) by synchronous adjunctions and substitutions

(3) or in the discourse structure (4), the semantic tree (see Fig 2) is translated in a denotational semantic by the introduction of lambda operators. We obtain the following formula for example (3) :

$$\lambda x : term\ (request(speaker(x), \lambda y : term\ \exists y(first(y) \quad \wedge \quad leave(y))))$$

To check if a well- formed formula is obtained, one can test the application of lambda abstraction on the missing predicates and curryfication on independent variables. The resulting semantic tree is correct for the previous example but if we consider a sequence like (6) :

(6) *Je voudrais le.*
I would like the.

the following formula obtained is not correct because the predicate *w* needs to be instanciated :

$$\lambda x : term\ (request(speaker(x), \lambda w : form\ \lambda y : term\ \exists y(w(y)))$$

4 Discussion

4.1 Rules vs specific lexical descriptions

Another way to deal with a sentence like (4) is to adopt a specific elementary tree for the adjective *premier* as explained in (Lopez, 1998a). In that case, the ellipsis resolution is not triggered directly by the parsing failure and a sentence like (7) is rejected.

(7) *Je voudrais le qui part.*
I would like the which leaves.

The same approach could be applied to the description of word order variation. In a Tree Grammar, word order must be determined by dependency relations. While substitution often corresponds to an ordered relation between argument in a syntactic structure, this is not the case for adjoined constituents, especially for adverbs. For instance, the

parsing of utterance (5) needs to consider the adverb *maintenant* as an unusual nominal modifier. The compositionality principle restricts the combination of this syntactic unit to trigger a synchronous combination on the same semantic node that the sentencial adverb does. It is expressed in synchronous TAG by a semantic tree which is synchronously combined at a different node than the syntactic tree.

In this paper, we argue for a rule based approach because we suppose that ambiguous analyses are taken into account at a upper level in a given application domain. By this way, we have to consider more analyses but we avoid inherent restrictions of the "augmented representation".

Indeed, the latter is limited because the semantic derivation can not always be built synchronously with the syntactic derivation. That is the case with the following sentence (8) :

(8) *Un train maintenant pour Paris doit-il partir?*
Does a train now for Paris have to leave?

Moreover, a sentence like (9) triggers redundant analysis because the both elementary trees for the adverb *maintenant* (sentencial and nominal modifier) are valid concurents.

(9) *Je voudrais un train pour Paris maintenant.*
I would like a train for Paris now.

4.2 Constraints vs preferential mechanisms

A previous experiment (Roussel and Halber, 1997) has shown that a robust parsing strategy based on a lexicalized grammar and a set of additional rules can improve the performances of a spoken dialogue system. However, in this experiment, a lot of spurious concurent hypothesis were still hard to eliminate whereas the lexicalized tree grammar was enriched with specific semantic constraints. This result adresses the need of a scoring method to cross-check more knowledge sources. In this framework, the use of semantic control could be use independently among other criteria (hesitation cues, conditions on speech acts, dialogue history, focus, ...) (Roussel and Modave, 1998).

Conclusion

We have shown that connected routes and categorial abstractions gives robustness capacities in a lexicalized tree grammar framework. Many questions are always investigated as the scoring method. A complementary perspective is to extend the rules to more complex discourse representations (Webber and Joshi, 1998).

References

Anne Abeillé. 1992. Synchronous TAGs and French Pronominal Clitics. In *COLING*, Nantes, France.

Marcel Cori, Michel de Fornel, and Jean-Marie Marandin. 1997. Parsing Repairs. In Ruslan Mitkov and Nicolas Nicolov, editors, *Recent advances in natural language processing*. John Benjamins.

Christine Doran and Bangalore Srinivas. 1994. Bootstrapping A Wide-Coverage CCG from FB-LTAG. In *3rd International Workshop on Tree Adjoining Languages (TAG+3)*, Paris, France.

Laura Kallmeyer. 1997. A Syntax-Semantic Interface with Synchronous Tree Description Grammars. In *Formal Grammar Workshop : ESSLLI*, pages 112–124, Aix-en-Provence, France.

Alberto Lavelli and Giorgio Satta. 1991. Bidirectional parsing of lexicalized tree adjoining grammars. In *Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany.

Patrice Lopez. 1998a. A LTAG grammar for parsing incomplete and oral utterances. In *European Conference on Artificial Intelligence (ECAI)*, Brighton, UK.

Patrice Lopez. 1998b. Analyse guidée par la connexité de TAG lexicalisées. In *Conférence sur le Traitement Automatique du Langage Naturel (TALN)*, Paris, France.

David Roussel and Ariane Halber. 1997. Filtering errors and repairing Linguistic Anomalies for Spoken Dialogue Systems. In *Workshop on Interactive Spoken Dialog Systems : ACL/EACL*, pages 74–81, Madrid.

David Roussel and Francois Modave. 1998. A multicriteria scoring method to parse recognition hypotheses. In *the International Workshop on Speech and Computer (SPECOM)*, St.-Petersburg, Russia.

Yves Schabes. 1994. Left to Right Parsing of Lexicalized Tree Adjoining Grammars. *Computational Intelligence*, 10:506–524.

Stuart Shieber and Yves Schabes. 1990. Synchronous Tree Adjoining Grammars. In *COLING*, volume 3, pages 253–260, Helsinki.

Stuart Shieber, Yves Schabes, and Fernando Pereira. 1995. Principles and Implementation of Deductive Parsing. *Journal of Logic Programming*, 24:3–36.

Gertjan van Noord. 1994. Head Corner Parsing for TAG. *Computational Intelligence*, 10:525–534.

Bonnie Lynn Webber and Aravind K. Joshi. 1998. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. In *COLING*, COLING-ACL'98 Workshop on Discourse Relations and Discourse Markers.