Repair Strategies for Lexicalized Tree Grammars

Patrice Lopez
LORIA,
BP239, 54500 Vandœuvre,
FRANCE
lopez@loria.fr

Abstract

This paper presents a framework for the definition of monotonic repair rules on chart items and Lexicalized Tree Grammars. We exploit island representations and a new level of granularity for the linearization of a tree called connected routes. It allows to take into account the topology of the tree in order to trigger additional rules. These local rules cover ellipsis and common extra-grammatical phenomena such as self-repairs. First results with a spoken language corpora are presented.

Introduction

In the context of spoken task-oriented man-machine and question-answering dialogues, one of the most important problem is to deal with spontaneous and unexpected syntactical phenomena. Utterances can be very incomplete and difficult to predict which questions the principle of grammaticality. Moreover large covering grammars are generally dedicated to written text parsing and it is not easy to exploit such a grammar for the analysis of spoken language even if complex syntax does not occur.

For such sentences, robust parsing techniques are necessary to extract a maximum of information from the utterance even if a complete parsing fails (at least all possible constituents). Considering parsing of word-graphs and the large search space of parsing algorithms in order to compute all possible ambiguities, the number of partial parses can be very important. A robust semantic processing on these partial derivations would result in a prohibitive number of hypotheses. We argue in this paper that appropriate syntactical constraints expressed in a Lexicalized Tree Grammar (LTG) can trigger efficient repair rules for specific oral phenomena.

First results of a classical grammatical parsing are presented, they show that robust parsing need to cope with oral phenomena. We argue then that extended domain of locality and lexicalization of LTG can be exploited in order to express repair local rules for these specific spoken phenomena. First results of this approach are presented.

1 LTG parsing and repairing strategy

1.1 Experimental results

Table 1 presents parsing test results of the Gocad corpora. This corpora contains 861 utterances in French of transcribed spontaneous spoken language collected with a Wizard of Oz experiment (Chapelier et al., 1995). We used a bottom-up parser (Lopez, 1998b) for LTAG. The size of the grammar was limited compared with (Candito, 1999) and corresponds to the sublanguage used in the Gocad application. However designing principles of the grammar was close to the large covering French LTAG grammar just including additional elementary trees (for example for unexpected adverbs which can modify predicative nouns) and a notation enrichment for the possible ellipsis occurrences (Lopez, 1998a). The LTAG grammar for the sublanguage corresponds to a syntactical lexicon of 529 entries and a set of 80 non-instancied elementary trees.

A taxonomy of parsing errors occurring in oral dialogue shows that the majority of failures are linked to orality: hesitations, repetitions, self repairs and some head ellipsis. The table 2 gives the occurrence of these oral phenomena in the Gocad corpora. Of course more than one phenomenon can occur in the same utterance.

Prediction of these spoken phenomena would result in a very high parsing cost. However if we can detect these oral phenomena with additional techniques combining partial results, the number of hypotheses at the semantic level will decrease.

Corpus	% complete parses	Average no of parses/utter.	Average no of partial results/utter.
Gocad	78.3	2.0	7.1

Table 1: Global results for the parsing of the Gocad corpora utterances

ill-formed utterances	with hesitations	with repetitions	with self-repairs	agrammatical ellipsis
Occurrences	123	28	22	15

Table 2: Occurrences of error oral phenomena in the Gocad corpora

1.2 Exploiting Lexicalized Tree Grammars

The choice of a LTG (Lexicalized Tree Grammar), more specifically a LTAG (Lexicalized Tree Adjoing Grammar), can be justified by the two main following reasons: first the lexicalization and the extended domain of locality allow to express easily lexical constraints in partial parsing trees (elementary trees), secondly robust bottom-up parsing algorithms, stochastic models and efficient precompilation of the grammar (Evans and Weir, 1998) exist for LTG.

When the parsing of an utterance fails, a robust bottom-up algorithm gives partial derived and derivation trees. With a classical chart parsing, items are obtained from other items and correspond to a well-recognized chunk of the utterance. The chart is an acyclic graph representing all the derivations. A partial result corresponds to the maximal expansion of an island, so to an item which is not the origin of any other item.

The main difference between a Context Free Grammar and a Lexicalized Tree Grammar is that a tree directly encodes for a specific anchor a partial parsing tree. This representation is richer than a set of Context Free rules. We argue that we can exploit this feature by triggering rules not only according to the category of the node N corresponding to an item but considering some nodes near N.

2 Island representation and connected routes in repair local rules

2.1 Finite States Automata representation of an elementary tree

The linearization of a tree can be represented with a Finite State Automaton (FSA) as in figure 2. Every tree traversal (left-to-right, bidirectional from an anchor, ...) can be performed on this automaton. Doted trees used for example in (Schabes, 1994) are equivalent to the states of these automata. It is then possible to share all the FSA of a lexicalized grammar in a single one with techniques presented in (Evans and Weir, 1998).

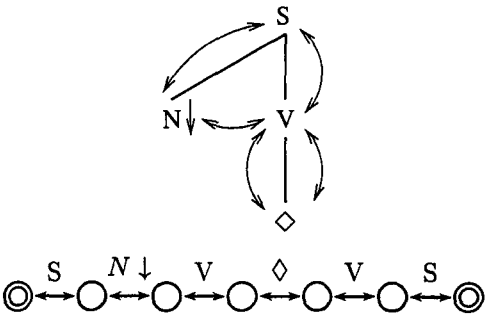

Figure 2: Simple FSA representing an elementary tree for the normal form of French intransive verb.

We consider the following definitions and notations :

- Each automaton transition is annotated with a category of node. Each non-leaf node appears twice in the list of transition framing the nodes which it dominates. In order to simplify our explanation the transition is shown by the annotated category.
- Transitions can be bidirectional in order to be able to start a bidirectional tree walk of a tree starting from any state.
- Considering a direction of transition (left-to-right, right-to-left) the FSA becomes acyclic.

2.2 Parsing invariant and island representation

A set of FSA corresponds to a global representation of the grammar, for the parsing we use a local representation called *item*. An item is defined as a 7-tuple of the following form:

(a) Rule for hesitations :

$$\frac{(i,j,\sigma_L,\sigma_R) \quad (j,k,\sigma'_L,\sigma'_R) \quad (k,l,\sigma''_L,\sigma''_R)}{(i,k,\sigma_L,\sigma_R) \quad (k,l,\sigma''_L,\sigma''_R)} \quad (head(\Gamma'_L) = tail(\Gamma'_R) = H)$$

(b) Rule for head ellipsis on the left :

$$\frac{(i,j,\sigma_L,\sigma_R) \quad (j,k,\sigma'_L,\sigma'_R)}{(i,k,\sigma_L,\sigma'_R)} \quad (tail(\Gamma_R) = X* \ \wedge \ ((head(\Gamma'_L) = X\downarrow \ \vee \ head(\Gamma'_L) = X*) \ \wedge \ tail(\Gamma'_R) = X\uparrow))$$

(c) Rule for argument ellipsis on the right :

$$\frac{(i,j,\sigma_L,\sigma_R)}{(i,j,\sigma_L,next(\Gamma_R))} \quad (tail(\Gamma_R) = X\downarrow)$$

(d) Rule 1 for self repair :

$$\frac{(i,j,\sigma_L,\sigma_R) \quad (j,k,\sigma'_L,\sigma'_R)}{(i,k,\sigma_L,\sigma'_R)} \quad (\exists i = (v,w,\sigma''_L,\sigma''_R) \in \Delta, i \Rightarrow^* (i,j,\sigma_L,\sigma_R) \ \wedge \ (\exists X \in \Gamma''_R \ \wedge \ head(\Gamma'_L) = X*) \vee (tail(\Gamma''_R) = X\downarrow \ \wedge \ head(\Gamma'_L) = X\uparrow))$$

Figure 1: Example of repair rules

```
item: ( left index, right index,
        left state, right state,
        foot left index,
        foot right index, star state)
```

The two first indices are the limits on the input string of the island (an anchor or consecutive anchors) corresponding to the item. During the initialization, we build an item for each anchor present in the input string. An item also stores two states of the same FSA corresponding to the maximal extension of the island on the left and on the right, and *only* if necessary we represent two additional indices for the position of the foot node of a wrapping auxiliary tree and the state *star* corresponding to the node where the current wrapping adjunction have been predicted.

This representation maintains the following invariant: an item of the form $(p, q, \sigma_L, \sigma_R)$ specifies the fact that the linearized tree represented by a FSA Δ is completely parsed between the states σ_L and σ_R of Δ and between the indices p and q. No other attachment on the tree can happen on the nodes located between the anchors p and q-1.

2.3 Connected routes

Considering an automaton representing the linearization of an elementary tree, we can define a connected route as a part of this automaton corresponding to the list of nodes crossed successively until reaching a substitution, a foot node or a root node (included transition) or an anchor (excluded transition). Connected route is an intermediate level of granularity when representing a linearized tree: each elementary (or a derived tree) can be represented as a list of connected routes. Considering connected routes during the parsing permits to take into account the topology of the elementary trees and to locate significative nodes for an attachment (Lopez, 1998b). We use the following additional simplified notations :

- The connected route passing through the state σ_d is noted Γ_d.
- $next(\Gamma)$ (resp. $previous(\Gamma)$) gives the first state of the connected route after (resp. before) Γ according to a left-to-right automaton walk.
- $next(N)$ (resp. $previous(N)$) gives the state after (resp. before) the transition N.
- $head(\Gamma)$ (resp. $tail(\Gamma)$) gives the first right (resp. left) transition of the leftmost (resp. rightmost) state of the connected route Γ.

2.4 Inference rules system

The derivation process can be viewed as inference rules which use and introduce items. The inference rules (Schabes, 1994) have the following meaning, if q items $(item_i)_{0 \leq i < q}$ are present in the chart and if the requirements are fulfilled then add the r items $(item_j)_{0 \leq j < r}$ in the chart *if necessary*:

$$\frac{(item_i)_{0 \leq i < q}}{\text{add} \quad (item_j)_{0 \leq j < r}} \quad (\text{ conditions })$$

We note $\Rightarrow^*$ the reflexive transitive closure of the derivation relation between two items: if $i_1 \Rightarrow^* i_2$ then the item identified with i_2 can be obtained from i_1 after applying to it a set of derivations. We note a root node with $\uparrow$.

Figure 1 presents examples of repair rules. This additional system deals with the following phenomena:

ill-formed utterances	with hesitations	with repetitions	with self-repairs	unexpected ellipsis
% Correctly recovered	79.6	78.5	63.6	46.7

Table 3: Repair results for the Gocad corpora

- Hesitations : Rule (a) for hesitations absorbs adjacent initial trees whose head is a H node. Such a tree can correspond to different kind of hesitation.
- Ellipsis : two rules and their symmetrical configurations try to detect and recover respectively an empty head (b) and an empty argument (c).
- Self-repair : 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 more expressive connected route matching condition. Rule (d) deals with self-repair where the repaired structure has been connected on the target node.

3 First results

The rules has been implemented in Java and are integrated in a grammatical environment system dedicated to design and test the parsing of spoken dialogue system sublangages. We use a two stage strategy (Rosé and Lavie, 1997) corresponding to two sets of rules: the first one is the set for a bottom-up parsing of LTAG using FSA and connected routes (Lopez, 1998b), the second one gathers the repair rules presented in this paper. This strategy separates parsing of grammatical utterances (resulting from substitution and adjunction) from the parsing of admitted utterances (performed by the additional set). This kind of strategy permits to keep a normal parsing complexity when the utterance is grammatical. We present in table 3 statistics for the parsing repairs of the Gocad copora.

Discussion

Connected routes give robustness capacities in a Lexicalized Tree Framework. Note that the results has been obtained for transcribed spoken language. Considering parsing of word-graphs resulting from a state-of-the-art HMM speech recognizer, non-regular phenomena encountered in spoken language might cause a recognition error on a neighbouring word and so could not always be detected.

To prevent overgeneration during the second stage, both semantic additional well-formed criteria and a restrictive scoring method can be used. Future works will focus on a mecanism which allows a syntactic and semantic control in the case of robust parsing based on a LTAG and a synchronous Semantic Tree Grammar.

References

Marie-Hélène Candito. 1999. *Structuration d'une grammaire LTAG : application au fran ais et à l'italien.* Ph.D. thesis, University of Paris 7.

Laurent Chapelier, Christine Fay-Varnier, and Azim Roussanaly. 1995. Modelling an Intelligent Help System from a Wizard of Oz Experiment. In *ESCA Workshop on Spoken Dialogue Systems*, Vigso, Danemark.

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.

Roger Evans and David Weir. 1998. A structure-sharing parser for lexicalized grammars. In *COLING-ALC*, Montréal, Canada.

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

Patrice Lopez. 1998b. Connection driven parsing of Lexicalized TAG. In *Workshop on Text, Speech and Dialog (TSD)*, Brno, Czech Republic.

C.P. Rosé and A. Lavie. 1997. An efficient distribution of Labor in Two Stage Robust Interpretation Process. In *Proceeding of Empirical Methods in Natural Language Processing, EMNLP'97*, Rhode Island, USA.

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