Efficient Processing of Extra-grammatical Sentences: Comparing and Combining two approaches to Robust Stochastic parsing

Marita Ailomaa[2], Vladimír Kadlec[1], Jean-Cédric Chappelier[2], and Martin Rajman[2]

[1] Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: xkadlec@fi.muni.cz

[2] Artificial Intelligence Laboratory, Computer Science Department
Swiss Federal Institute of Technology (EPFL)
1015 Lausanne, Switzerland
E-mail: {marita.ailomaa,jean-cedric.chappelier,martin.rajman}@epfl.ch

Abstract. This paper compares two techniques for robust parsing of extra-grammatical natural language that might be of interest in large scale Textual Data Analysis applications. The first one returns a "correct" derivation for any extra-grammatical sentence by generating the finest corresponding most probable optimal maximum coverage. The second one extends the initial grammar by adding relaxed grammar rules in a controlled manner. Both techniques use a stochastic parser that selects a "best" solution among multiple analyses. The techniques were tested on the ATIS and Susanne corpora and experimental results, as well as conclusions on performance comparison, are provided.
Keywords: Robust, Parsing, Coverage.

1 Introduction

Formal grammars are traditionally used in NLP applications to describe well-formed sentences. But in large scale Textual Analysis applications it is not practical to rely exclusively on a formal grammar because of the large fraction of sentences that will receive no analysis. This undergeneration problem has lead to a whole field of research called robust parsing, where the goal is to find domain-independent, efficient parsing techniques that return a correct or usefully "close" analysis for almost all of the input sentences [Carroll and Briscoe, 1996]. Such techniques need to handle not only the problems of undergeneration but also the increased ambiguity which is usually a consequence of the robustification of the parser.

In previous works, a variety of approaches have been proposed to robustly handle natural language. Some techniques are based on modifying the input sentence, for example by removing words that disturb the fluency [Bear *et al.*, 1992, Heeman and Allen, 1994]. More recent approaches are based on selecting the right sequence of partial analyses [Worm and Rupp, 1998, van

Noord *et al.*, 1999]. Minimum Distance Parsing is a third approach based on relaxing the formal grammar, allowing rules to be modified by insertions, deletions and substitutions [Hipp, 1992].

Most of these approaches make the distinction between *ungrammaticality* and *extra-grammaticality*. Ungrammatical sentences might contain errors such as wrong agreement in the case of casual written text like mails, or hesitations and other types of disfluencies in the case of spoken language. On the other hand, extra-grammatical sentences are linguistically correct sentences that are not covered by the grammar.

This paper presents two new approaches that focus on extra-grammatical sentences. The first approach described in section 2 is based on the selection of a most optimal coverage with partial analyses, while the second, presented in section 3, uses controlled grammar rule relaxation. Section 4 describes the comparison of these two approaches and shows that they present differences in behavior when given the same grammar and the same test data.

2 Selecting the most probable optimal maximum coverage

2.1 Concepts

For a given sentence a *coverage*, with respect to an input grammar G, is a sequence of non-overlapping, possibly partial, derivation trees, such that the concatenation of the leaves of these trees corresponds to the whole input sentence (see figure 1).

If there are no unknown words in the input sentence, then at least one trivial coverage is obtained, consisting of the trees that all use only lexical rules (i.e. one rule per tree).

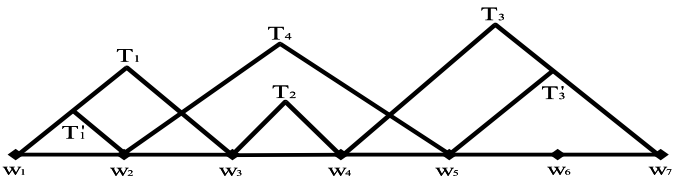

Fig. 1. A coverage $C = (T_1, T_2, T_3)$ consisting of trees T_1, T_2 and T_3. If there are $T_1^{'}$ and $T_3^{'}$, $T_1^{'}$ is a subtree of tree T_1 and $T_3^{'}$ is a subtree of T_3, then we also have coverage $C^{'} = (T_1^{'}, T_4, T_3^{'})$. Conversely $(T_1, T_3^{'})$ and (T_1, T_4, T_3) are not coverages.

A maximum coverage (m-coverage) is a coverage that is maximal with respect to the partial order relation $\leq$, defined as reflexive transitive closure of the subsumed relation $\prec$ (see figure 2). The relation $\prec$ is a relation over coverages such that, for coverages C and $C^{'}$:

$C^{'} \prec C$ iff $\exists i, j, k$, $1 \leq i \leq k, 1 \leq j$ and there exists rule r in the grammar G such that $C = (T_1, ..., T_i, ..., T_k)$, $C^{'} = (T_1, ...T_{i-1}, T_1^{'}, T_2^{'}, ..., T_j^{'}, T_{i+1}, ..., T_k)$ and $T_i = r \circ T_1^{'} \circ T_2^{'} ... \circ T_j^{'}$,

i.e. if there exists a sub-sequence of trees in $C^{'}$ that can be connected by rule r and the resulting tree is element of C, the other trees in $C^{'}$ being the same as in C. Notice that the rule r can be an unary rule.

If there is a successful parse (a single derivation tree that covers the whole input sentence) then there are as many m-coverages as full parse trees and every m-coverage contains only one tree.

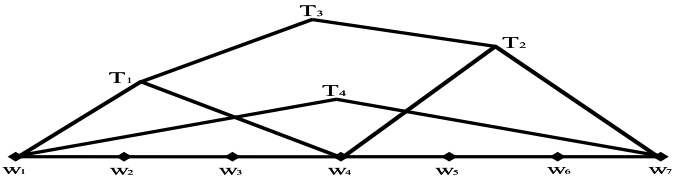

Fig. 2. An example to illustrate a maximum coverage. The coverage $C_1 = (T_3)$ is m-coverage. The coverage $C_2 = (T_1, T_2)$ is not maximum, because $C_2 \leq C_1$. There is also another m-coverage $C_3 = (T_4)$. Notice that C_1 and C_3 are not comparable with relation $\leq$.

In addition to maximality, we focus on *optimal* m-coverage, where optimality could be defined with respect to different measures. In contrast to maximality, the choice of a measure depends on the concrete application. Several optimality measures could be defined. For instance, the optimality measure can look at the intended structure of trees in a coverage, e.g. it can count the number of nodes in trees. In the presented work, we used the following optimality measure which relates to the average width (number of leaves) of the derivation trees in the coverage. For an m-coverage $C = (T_1, T_2, ...T_k)$ of input sentence $w_1, w_2, ..., w_n$, $n > 1$, we define

$$S_1(C) = \frac{1}{n-1}(\frac{n}{k} - 1).$$

Notice that $0 \leq S_1(C) \leq 1$ and $\frac{n}{k}$ is the average width of the derivation trees in the coverage. With this measure, the value of a coverage made exclusively of lexical rules is 0 and the value of a successful parse is 1.

For standard SCFG derivation, the probability of a coverage is defined as the product of the probabilities of the trees it contains. The probability of a coverage could also be viewed as another optimality measure. So the most probable coverages can be found in the same way as optimal m-coverages. But, usually we find all optimal m-coverages (OMC) first (optimal with respect to some other measure than probability) and then the most probable of these is chosen. Notice that both OMC and most probable OMC are not unique.

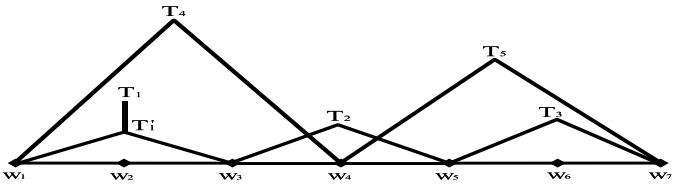

Fig. 3. An example to demonstrate the optimal m-coverage. $C_1 = (T_1, T_2, T_3)$ and $C_2 = (T_4, T_5)$ are m-coverages. The coverage $C_1^{'} = (T_1^{'}, T_2, T_3)$ is not m-coverage. The coverage C_2 is optimal for the measure S_1, $S_1(C_1) < S_1(C_2)$. Notice that the coverages C_1 and C_2 are not comparable with relation $\leq$.

2.2 Algorithm

We use a bottom-up chart parsing algorithm [Chappelier and Rajman, 1998] that produces all possible incomplete parses[1]. The incomplete parses are then combined to find the maximum coverage(s).

The described algorithm finds OMC with respect to the measure S_1 (the average width of the derivation trees in the coverage), but it can be easily adapted to different optimality measures. All operations are applied to a set of Earley's items [Earley, 1970]. In particular, no changes are made during the parsing phase (except some initialization of internal structures for better efficiency of the algorithm).

The Dijkstra's algorithm for shortest path problem in graphs is used to find OMC. The input graph for the Dijkstra's algorithm consists of weighted edges and vertices. The edges are Earley's items and the weight of each edge is 1. The vertices are word positions, thus for n input words we have $n + 1$ vertices. Whenever the Dijkstra's algorithm finds paths with equal length (i.e. identical number of items), we use the probability to select the most probable ones. Notice that, if all the words are known, there exists at least one path from position 0 to n corresponding to the trivial coverage.

The output of the algorithm is a list of Earley's items, which can represent several derivation trees. To get OMC, the most probable tree from each item is selected.

3 Deriving trees with holes

Our second approach to robust parsing is based on the idea that, in the case of a rule-based parser, the parser fails to analyze a given extra-grammatical sentence because one or several rules are missing in the grammar. If a rule-relaxation mechanism is available[2], it can be used to cope with such situ-

[1] Whenever there exists a derivation tree that covers the part of the given input sentence, the algorithm produces that tree

[2] A mechanism that can derive additional rules from the ones present in the grammar

ations. In that case the goal of the robust parser is to derive a full tree where the subtrees corresponding to the used relaxed rules are represented as "holes" (see figure 4).

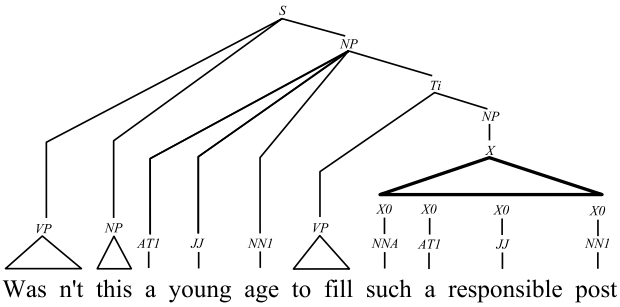

Fig. 4. A tree with a hole representing a missing NP rule $NP \rightarrow NNA \;\; AT1 \;\; JJ \;\; NN1$.

We use the principle called Minimum Distance Parsing which has been introduced in earlier robust parsing applications [Hipp, 1992]. This approach relaxes rules in the grammar by inserting, deleting or substituting elements in their right hand side (RHS). Derivation trees are ranked by the number of modifications that have been applied to the grammar rules to achieve a complete analysis. One important drawback is that, in its unconstrained form, the method produces many incorrect derivations and works well only for small grammars [Rosé and Lavie, 2001].

To prevent such incorrect derivations, we make restrictions on how the rules can be relaxed based on observations and linguistic motivations. One such restriction is to only relax grammar rules for which the LHS is frequently represented in the grammar, e.g. NP. Another restriction is to allow only one type of relaxation, namely insertion. The inserted element is hereafter referred to as a filler. A further refinement of the algorithm is to specify what syntactic category a filler is allowed to have when being inserted into a given position in the RHS. To illustrate the ideas, an example is now provided.

Assume that there is a grammar with two NP rules. (The head is indicated with underlined syntactic categories):

$$R1 : NP \rightarrow ADJ \;\; \underline{N}$$
$$R2 : NP \rightarrow POS \;\; \underline{N}$$

According to this grammar "successful brothers" and "your brother" are syntactically correct NPs while "your successful brother" is not. In order to parse the last one, some NP rule needs to be relaxed. We select the second one, R2 (though both are possible candidates). If the filler that needs to be inserted is ADJ (in this case "successful"), then the relaxed NP rule is expressed as:

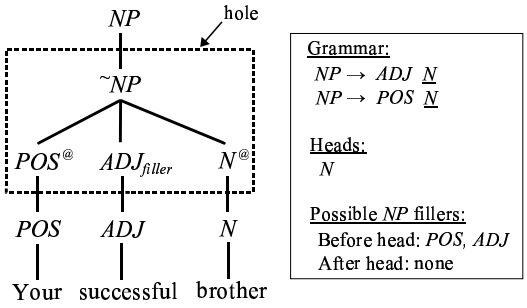

Fig. 5. An example of how a hole is derived by relaxing a rule and inserting a filler.

$$R3 : {\sim}NP \rightarrow POS^{@} \quad ADJ_{filler} \quad N^{@}$$

We use the category ~NP instead of NP to distinguish relaxed rules from initial ones, the "filler" subscripts to identify the fillers in the RHS in the relaxed rule, and the @ to label the original RHS elements. The decision of allowing an insertion of an ADJ as filler is based on whether ADJ is a possible element before the head or not. Since there is a rule in the grammar where an ADJ exists before the head (R1), the insertion is appropriate.

4 Validation

The two robust parsing techniques presented in the previous sections were tested on subsets of two treebanks, ATIS and Susanne. From these treebanks two separate grammars were extracted having different characteristics. Concretely each treebank was divided into a learning set that was used for producing the probabilistic grammar and a test set that was then parsed with the extracted grammar. Around 10% of the sentences in the test set were not covered by the grammar. These sentences represented the real focus of our experiments, as the goal of a robust parser is to process the sentences that the initial grammar fails to describe.

The sentences were first parsed with technique 1 and technique 2 separately and then with a combined approach where the rule-relaxation technique was tried first and only when it failed the most probable OMC was selected. For each sentence the 1-best derivation tree was categorized as good, acceptable or bad, depending on how closely it corresponded to the reference tree in the corpus and how useful the syntactic analysis was for extracting a correct semantic interpretation. The results are presented in table 1. It may be argued that the definition of a "useful" analysis might not be decidable only by observing the syntactic tree. Although we found this to be a quite usable hypothesis during our experiments, some more objective procedure should be defined. In a concrete application, the usefulness might for example be determined by the actions that the system should perform based on the produced syntactic analysis.

	Good (%)	Acceptable (%)	Bad (%)	No analysis (%)
ATIS corpus				
Technique 1	10	60	30	0
Technique 2	24	36	9	31
Technique 1+2	27	58	16	0
Susanne corpus				
Technique 1	16	29	55	0
Technique 2	40	17	33	10
Technique 1+2	41	22	37	0

Table 1. Experimental results. Percentage of good, acceptable and bad analyses with technique 1 (optimal coverage), technique 2 (tree with holes) and with the combined approach.

From the experimental results one can see that, for both grammars, technique 2 is more accurate than technique 1. However, if both good and acceptable results are taken into account, technique 1 behaves better with the ATIS grammar that has relatively few rules, and technique 2 better with Susanne, which is a considerably larger grammar describing a rich variety of syntactic structures.

Regardless of the technique used, the number of bad 1-best analyses that are produced can be explained by the fact that the probabilistically best analysis is not always the linguistically best one. This is a non-trivial problem related to all types of natural language parsing, not only to robust parsers.

An interesting result is that when the sentences are processed sequentially with both techniques, the advantage of each approach is taken into account and the performance is better than when either technique is used alone.

5 Conclusions

In this report we presented and compared two approaches to robust stochastic parsing. First we introduced the optimal maximum coverage framework and associated measures for the optimality of the parser. Then we introduced a rule-relaxation strategy based on the concept of holes, using several linguistically motivated restrictions to control the relaxation of grammar rules.

Experimental results show that a combination of the techniques gives a better performance than each technique alone, because the first one guarantees full coverage while the second has a higher accuracy. The richness of the syntactic structures defined in the initial grammar tends to have some impact on the performance in the second approach but less in the first one. This can be linked to the restrictions that were chosen for the relaxation of the grammar rules. It is possible that different types of restrictions are appropriate for different grammars.

The evaluation of the robust parsing techniques was based on manually checking the derivation trees. An important issue is to integrate the techniques into some target application so that we have more realistic ways of measuring the usefulness of the produced robust analyses.

As a final remark, we would like to point out that this paper has addressed the problem of extra-grammaticality but did not address ungrammaticality, which is also a very important phenomenon in robust parsing, though more relevant in spoken language applications than in textual data analysis.

References

[Bear *et al.*, 1992]John Bear, John Dowding, and Elizabeth Shriberg. Integrating multiple knowledge sources for the detection and correction of repairs in human-computer dialogue. In *Proceedings of the 30th ACL*, pages 56–63, Newark, Delaware, 1992.

[Carroll and Briscoe, 1996]John Carroll and Ted Briscoe. Robust parsing — a brief overview. In John Carroll, editor, *Proceedings of the Workshop on Robust Parsing at the 8th European Summer School in Logic, Language and Information (ESSLLI'96), Report CSRP 435*, pages 1–7, COGS, University of Sussex, 1996.

[Chappelier and Rajman, 1998]J.-C. Chappelier and M. Rajman. A generalized CYK algorithm for parsing stochastic CFG. In *TAPD'98 Workshop*, pages 133–137, Paris, France, 1998.

[Earley, 1970]J. Earley. An efficient context-free parsing algorithm. In *Communications of the ACM*, volume 13, pages 94–102, 1970.

[Heeman and Allen, 1994]Peter A. Heeman and James F. Allen. Detecting and correcting speech repairs. In *Proceedings of the 32th ACL*, pages 295–302, Las Cruces, New Mexico, 1994.

[Hipp, 1992]Dwayne R. Hipp. *Design and development of spoken natural language dialog parsing systems.* PhD thesis, Duke University, 1992.

[Rosé and Lavie, 2001]C. Rosé and A. Lavie. Balancing robustness and efficiency in unification-augmented contextfree parsers for large practical applications. In G. van Noord and J. C. Junqua, editors, *Robustness in Language and Speech Technology.* Kluwer Academic Press, 2001.

[van Noord *et al.*, 1999]Gertjan van Noord, Gosse Bouma, Rob Koeling, and Mark-Jan Nederhof. Robust grammatical analysis for spoken dialogue systems. *Natural Language Engineering*, 5(1):45–93, 1999.

[Worm and Rupp, 1998]Karsten L. Worm and C. J. Rupp. Towards robust understanding of speech by combination of partial analyses. In *Proceedings of the 13th biennial European Conference on Artificial Intelligence (ECAI'98), August 23-28*, pages 190–194, Brighton, UK, 1998.