Robust stochastic parsing: comparing and combining two approaches for processing extra-grammatical sentences

Marita Ailomaa[1] and **Vladimír Kadlec**[2]
and **Martin Rajman**[1] and **Jean-Cédric Chappelier**[1]
[1]Artificial Intelligence Laboratory (LIA)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
[2] Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{marita.ailomaa,jean-cedric.chappelier,martin.rajman}@epfl.ch,
xkadlec@fi.muni.cz

Abstract

This paper compares two techniques for robust parsing of extra-grammatical natural language. Both are based on well-known approaches; one selects the optimal combination of partial analyses, the other relaxes grammar rules. Both techniques use a stochastic parser to select the "best" solution among multiple analyses. Experimental results show that regardless of the grammar, the best results are obtained by sequentially combining the two techniques, by first relaxing the rules and only when that fails by then selecting a combination of partial analyses.

1 Introduction

Formal grammars are often used in NLP applications to describe well-formed sentences. But when used in practice, the grammars usually describe only a subset of a NL, and in addition NL sentences are not always well-formed, especially in speech recognition applications. NLP applications that rely exclusively on such grammars cannot be practically used in a large scale because of the large fraction of sentences that will receive no analysis at all. This problem is called undergeneration and has given birth to a field called robust parsing, where the goal is to find domain-independent and practical parsing technique that returns a correct or usefully "close" analysis for almost all (say 90%) of the input sentences (Carroll and Briscoe, 1996). In order to achieve such a high performance, one has 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 work, 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). For most of the approaches it is important to make the distinction between ungrammaticality and extra-grammaticality. Ungrammatical sentences contain speech errors, hesitations etc. while extra-grammatical sentences are linguistically correct sentences that are not covered by the grammar.

This paper presents two approaches that focus on extra-grammatical sentences. The first approach is based on the selection of a most optimal coverage with partial analyses and the second on controlled grammar rule relaxation. The aim of the paper is to compare these two approaches and to investigate if they present differences in behavior when given the same grammar and the same test data.

The rest of the paper is divided into five sections. Section 2 describes the notion of coverage and defines the most probable optimal maximum coverage; section 3 presents the rule-relaxation technique and introduces the con-

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 1–7
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

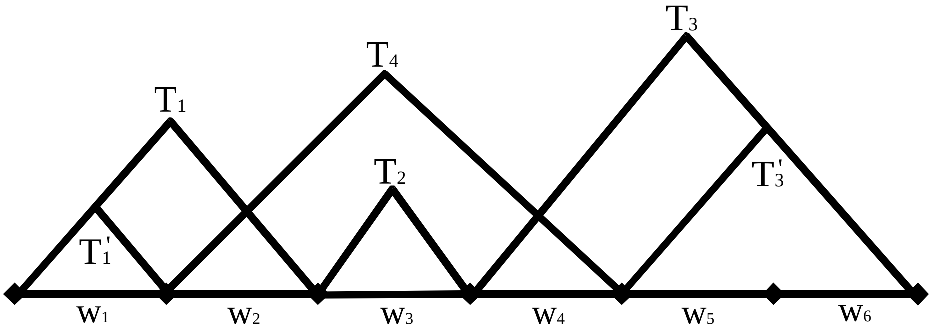

Figure 1: Partial derivation trees. Some of them (e.g. T_1, T_2, T_3 and $T_1^{'}, T_4, T_3^{'}$) can be composed into a coverage.

cept of “holes”. In section 4 the data and methodology for the experiments are presented and section 5 gives a summary of the results. Finally, in section 6 we give a brief conclusion and outline the direction for future work.

2 Selecting the most probable optimal maximum coverage

This section briefly describes the first approach to robust parsing that we tested in our comparative experiments. It is based on selecting partial analyses and gluing them together into an artificial full tree (Kadlec et al., 2005). This technique provides at least one analysis for all input sentences (in the most trivial case, a set of lexical trees).

To describe the technique we will start with the notion of coverage. For any given grammar G and any given input sentence, a coverage is a sequence of non-overlapping, possibly partial derivation trees that cover the whole input. Since the derivation trees can overlap fully or partially, there can be several distinct coverages for the same input sentence.

A maximum coverage is one that consists of maximum derivation trees, i.e. trees that are not subtrees of other ones covering the same subsequence. If a sentence can be analyzed by a single parse tree, the number of maximum coverages is the same as the number of complete parse trees.

Provided that a quality measure is available for coverages, an optimal maximum coverage (OMC) is then a maximum coverage with the highest quality among all maximum coverages derived for the given input sentence. In our experiments, we use a quality measure that favors a coverage with partial analyses of largest average width.

If the used grammar and parsing algorithm can be probabilized, the most probable OMC is the one that is associated with the highest probability. Finding the most probable OMC for stochastic context-free grammars can easily be achieved using a usual bottom-up chart parser (Chappelier and Rajman, 1998).

3 Deriving trees with holes

Our second approach to robust parsing is based on the assumption that extra-grammatical sentences are structurally similar to sentences that the grammar is able to describe. The underlying idea is that, in the case of a rule-based parser, the reason why the parser fails to analyze a given (extra-grammatical) sentence is that one or several rules are missing in the grammar. If a rule relaxation mechanism is available (i.e. a mechanism that can derive additional rules from the ones present in the grammar), it can be used to cope with situations where rules are missing, and in this 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 3).

The main difficulty with this approach is to determine how the hole should be derived. Simple rules allowing holes at any position in the tree are too general and will produce a high number of unacceptable analyses.

In addition we observed that it is more likely that a missing rule corresponds to a syntactic category that frequently appears as rule left hand side in the grammar (Ailomaa, 2004).

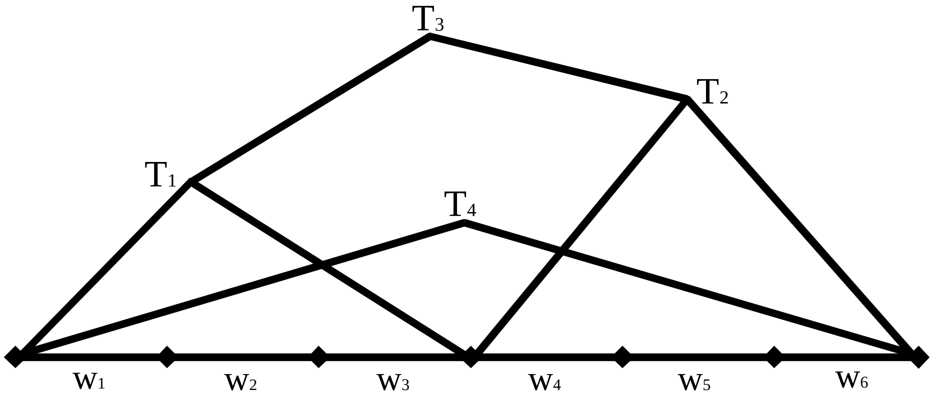

Figure 2: An illustration of maximum coverage. $C_1 = (T_3)$ and $C2 = (T_4)$ are m-coverages but $C_3 = (T_1, T_2)$ is not.

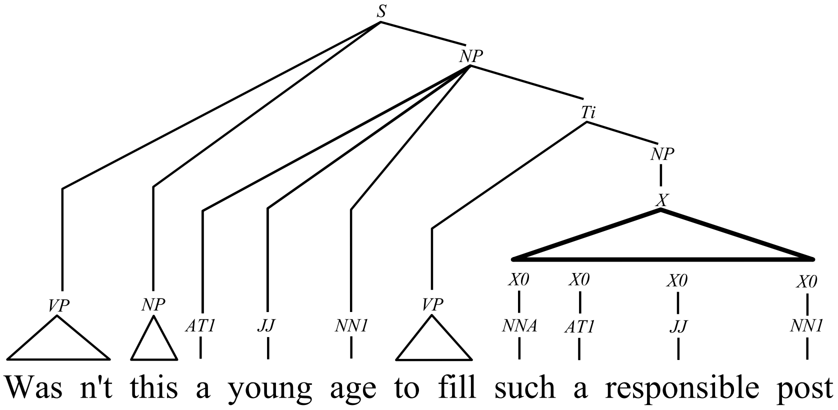

Figure 3: A tree with a hole representing a missing NP rule $NP \rightarrow NNA \;\; AT1 \;\; JJ \;\; NN1$.

Indeed, the higher the number of such rules is, the higher is the variety of possible constructions for the given category and therefore the higher is the probability that some of these constructions are missing in the grammar. NP, VP and S are examples of syntactic categories that are heavily represented in grammars, whereas PPs are not. Consequently, a hole is more likely to have a root of category NP, VP or S than for example PP.

The second issue is how to produce a hole with appropriate leaves. Here we use the principle called Minimum Distance Parsing which has been used in earlier robust parsing applications (Hipp, 1992). The idea is to relax 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 achieve a complete analysis. The problem is that when all the rules are allowed to be relaxed the number of analyses increases dramatically. It becomes difficult to select a good analysis among all the unacceptable ones. Therefore, in its unconstrained form, the method works well only for small grammars (Rosé and Lavie, 2001).

Our solution to the problem is to make restrictions on how the rules can be relaxed. We used two types of restrictions. The first one defines which one of the rules that are allowed to be relaxed. As we previously mentioned, some rules are more likely to be missing in the grammar than others, so we select the ones that correspond to syntactic categories heavily represented in the grammar. It means that we still have a large number of solutions, but they are more likely to be the good ones. The second restriction defines how the rules should be relaxed. In a preliminary test we observed that the most frequent modification that produced the missing rule was insertion. Therefore we limit the relaxation to only this type of modification. The inserted element is hereafter referred to as a filler.

A further refinement of the algorithm is to specify where in the RHS of a rule the filler can appear and what syntactic category that filler is allowed to have. These two restrictions are linguistically motivated. Most phrases are constructed according to a general pattern, e.g. in English NPs are composed of determiners, adjectives and other elements followed by a head, typically a noun, and some complements, e.g. PPs. Finite-state parsing (Ait-Mokhtar and Chanod, 1997) is an example of how such patterns have been successfully used to implement general-purpose robust parsers. In our approach, the elements in the RHS of a rule are considered as belonging to one of three types: (1) elements preceding the head, (2) the head itself, and (3) elements following the head. The reason for this distinction is that, again in English, there are syntactic categories that frequently occur in one part of the phrase but not in another. PPs for instance are often complements of NPs but are less likely to occur before the head. The decision of inserting or not a filler into a RHS is therefore a matter of deciding whether the syntactic category of the filler is appropriate, i.e. whether there is a rule in the grammar in which the category appears in the correct position with respect to the head. As an example, assume that we have a simple grammar with the following 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:

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

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.

	ATIS	Susanne
Sentences	1,381	3,981
CF rules	1,029	8,810
Non-terminals	40	469
Terminals	1,167	10,247
PoS tags	38	122
Average sentence length	12,5	12,9
Av. nb of CF rules/sent	23,3	23,8
Max depth	17	17

Table 1: Some characteristics of the two corpora used for the experiments

4 Data and methodology

The goal of this paper is to compare the two robust parsing approaches described in Section 2 and 3 and to analyze differences in their behavior. In both approaches the sentences are parsed with a probabilistic CFG, which means that the produced analyses have probabilistic scores that allow the selection of the most probable one (or ones).

The probabilistic grammar is extracted and trained from a treebank, which contains parsed sentences representative of the type of sentences that the parser should be able to handle. In our comparative tests we use subsets of two treebanks, ATIS and Susanne, which have very different characteristics (Marcus et al., 1994).

The purpose is to study if the characteristics of the grammar have some impact on the performance of the robust parsing techniques measured in terms of accuracy and coverage. Some of the possible characteristics are given in Table 1.

In addition, we also considered characteristics such as the number and syntactic category of rules that are missing in the grammar to describe the test sentence.

Concretely each treebank is divided into a learning set and a test set. The learning set is used for producing the grammar with which the test set is then parsed. The division is done in such a way that around 10% of the sentences in the test set are not covered by the grammar.

These sentences represent the “true” test set for our experiments, as we only want to process the sentences that the initial grammar fails to describe.

Once the test sentences have been processed with each of the two robust techniques, the most probable analysis is manually characterized as good, acceptable or bad. The categorization is based on the comparison of the produced parse tree with the reference one that can be extracted out of the treebank. A good analysis has a close match to the reference tree; an acceptable one can have some more substantial differences with respect to the phrase segmentation but has to correspond to a useful semantic interpretation. A bad analysis corresponds to an incorrect segmentation that cannot lead to any useful semantic interpretation. Notice that it may be argued that the definition of a “useful” semantic 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 next actions that the system performs based on the received syntactic analysis. Therefore, from this point of view, the results that we obtain have to be considered as preliminary and a future step is to integrate the implemented techniques in some application, e.g. an automatic translator, and to compare the results.

5 Experimental results

This section presents the experimental results of the two robust parsing techniques described in section 2 and 3, using the methodology described in section 4. The number of test sentences is not the same for the two corpora, 89 for ATIS and 250 for Susanne, due to the size and characteristics of each corpus. We parsed the sentences first with technique 1 and technique 2 separately and then with a combined approach where we started with the rule-relaxation technique and only when that failed selected the most probable OMC. All the sentences from ATIS were parsed and analyzed by hand; from Susanne a subset consisting of 134 sentences was chosen. The results of the robust parsing are presented in table 2.

When observing only the number of good analyses, one can see that for both grammars technique 2 performs better than technique 1. But when including all analyses, there is a difference related to the grammars. With ATIS, technique 1 provides acceptable analyses in the majority of cases whereas technique 2 badly suffers from undergeneration. With Susanne, technique 1 produces bad analyses for more than half of the test sentences; on the other hand technique 2 has a good coverage and overall better results than technique 1.

One indicator to why the techniques produce bad analyses, or no analysis at all, is that in more than half of the cases, three or more rules are missing in the grammar to derive the reference tree.

For technique 2 the undergeneration can be explained by the characteristics of the two grammars. ATIS is a very homogeneous corpus, which means that the same rules appear often. A division of 10% learning set and 90% test set was necessary to achieve the desired level of undergeneration ($\simeq 10\%$). With such a small learning corpus, the extracted grammar will have very few rules. There-fore the lists of possible fillers and rules that can be relaxed become very small. (Notice that we refer to two cases of undergeneration; the one mentioned here relates to the preparation of the data for the tests. The other case is when we speak of the results received from the robust parser).

Susanne suffers from the opposite problem. The corpus is very heterogeneous, and the learning set has to be larger in order to keep the undergeneration down to $\simeq 10\%$. The lists of possible fillers and rules to relax are long, but for a sentence that has no analysis, the missing rule is often of some other category than the ones heavily represented in the grammar (NP, VP or S). This is because the number of non-terminals in the corpus is large (469) compared to ATIS (40), and that 77% of the rules appear only once in the corpus.

Here we need to consider if the restrictions we applied on the relaxation of rules are too strong. It is possible that different types of restrictions are appropriate for different types of grammars. Some more experimenting is needed to decide how the change of flexibility in the algorithm affects coverage and accuracy.

Sentences for which the indicators mentioned above do not apply but which nevertheless have bad analyses, the main explanation

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

is that the probabilistically best analysis is not the linguistically best one. This is a non-trivial problem related to all types of natural language parsing, not only for robust parsers.

In short, we conclude that technique 2 is more accurate than technique 1 but cannot stand alone as a robust parser, not being able to provide full coverage. In particular it tends to be less suitable for simple grammars describing small variation of syntactic structures. What we can state 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 stands alone.

6 Conclusions

The aim of this paper was to compare two robust parsing techniques based on different principles. The method we chose was to provide them with the same grammar and test data and to analyze differences in performance. 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 grammar tends to have some impact on the performance in the second approach but less on the first one. This can be linked to the restrictions in technique 2 that were chosen for the relaxation of the grammar rules. These restrictions were based on observations of a relatively small set of extra-grammatical sentences and cannot be considered as final.

A future experiment is to test different levels of flexibility for technique 2 and to see how this affects accuracy and coverage. Another important issue is to integrate the parsing 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 an equally important phenomenon in robust parsing, particularly if the target application deals with spoken language.

References

Marita Ailomaa. 2004. Two approaches to robust stochastic parsing. Master's thesis, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland.

Salah Ait-Mokhtar and Jean-Pierre Chanod. 1997. Incremental finite-state parsing. In *Proceedings of the ANLP97*, pages 72–79, Washington.

John Bear, John Dowding, and Elizabeth Shriberg. 1992. 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.

John Carroll and Ted Briscoe. 1996. 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.

J.-C. Chappelier and M. Rajman. 1998. A generalized cyk algorithm for parsing stochastic cfg. In *TAPD98 Workshop*, pages 133–137, Paris, France.

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

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

Vladimír Kadlec, Marita Ailomaa, Jean-Cedric Chappelier, and Martin Rajman. 2005. Robust stochastic parsing using optimal maximum coverage. In *Proc. of International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 258–263, Borovets, Bulgaria.

Michell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

C. Rosé and A. Lavie. 2001. 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.

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

Karsten L. Worm and C. J. Rupp. 1998. 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.