

Parsing with an Extended Domain of Locality

John Carroll, Nicolas Nicolov, Olga Shaumyan, Martine Smets & David Weir

School of Cognitive and Computing Sciences

University of Sussex

Brighton, BN1 9QH, UK

Abstract

One of the claimed benefits of Tree Adjoining Grammars is that they have an extended domain of locality (EDOL). We consider how this can be exploited to limit the need for feature structure unification during parsing. We compare two wide-coverage lexicalized grammars of English, LEXSYS and XTAG, finding that the two grammars exploit EDOL in different ways.

1 Introduction

One of the most basic properties of Tree Adjoining Grammars (TAGs) is that they have an **extended domain of locality** (EDOL) (Joshi, 1994). This refers to the fact that the elementary trees that make up the grammar are larger than the corresponding units (the productions) that are used in phrase-structure rule-based frameworks. The claim is that in Lexicalized TAGs (LTAGs) the elementary trees provide a domain of locality large enough to state co-occurrence relationships between a lexical item (the **anchor** of the elementary tree) and the nodes it imposes constraints on. We will call this the **extended domain of locality hypothesis**.

For example, *wh*-movement can be expressed locally in a tree that will be anchored by a verb of which an argument is extracted. Consequently, features which are shared by the extraction site and the *wh*-word, such as case, do not need to be percolated, but are directly identified in the tree. Figure 1 shows a tree in which the case feature at the extraction site and the *wh*-word share the same value.¹

¹The anchor, substitution and foot nodes of trees are marked with the symbols \circ , \downarrow and $*$, respectively. Words in parenthesis are included in trees to provide examples of strings this tree can derive.

Much of the research on TAGs can be seen as illustrating how its EDOL can be exploited in various ways. However, to date, only indirect evidence has been given regarding the beneficial effects of the EDOL on parsing efficiency. The argument, due to Schabes (1990), is that benefits to parsing arise from lexicalization, and that lexicalization is only possible because of the EDOL. A parser dealing with a lexicalized grammar needs to consider only those elementary structures that can be associated with the lexical items appearing in the input. This can substantially reduce the effective grammar size at parse time. The argument that an EDOL is required for lexicalization is based on the observation that not every set of trees that can be generated by a CFG can be generated by a lexicalized CFG. But does the EDOL have any other more direct effects on parsing efficiency?

On the one hand, it is a consequence of the EDOL that wide-coverage LTAGs are larger than their rule-based counterparts. With larger elementary structures, generalizations are lost regarding the internal structure of the elementary trees. Since parse time depends on grammar size, this could have an adverse effect on parsing efficiency. However, the problem of grammar size in TAG has to some extent been addressed both with respect to grammar encoding (Evans et al., 1995; Candito, 1996) and parsing (Joshi and Srinivas, 1994; Evans and Weir, 1998).

On the other hand, if the EDOL hypothesis holds for those dependencies that are being checked by the parser, then the burden of passing feature values around during parsing will be less than in a rule-based framework. If *all* dependencies that the parser is checking can be stated directly within the elementary structures of the grammar, they do not need to be computed dynamically during the parsing process by means of feature percolation. For example, there is no need to use a slash feature to establish filler-gap dependencies over unbounded distances across the tree if the EDOL

Thus, passive sentences such as *The scheme was singled out by a recent Government report* are found difficult³, despite the presence of the syntactic cues *was*, *-ed* and *by*. We therefore replace passive constructions with corresponding active forms. We are currently integrating further rules to split conjoined sentences and extract embedded clauses. Syntactic simplification operates iteratively until a configuration is reached that cannot be simplified. This approach is broadly similar to that proposed by (Chandrasekar et al., 1996).

One of the many challenges in syntactic simplification is the observed effect of the total length of a text being increased when longer sentences are replaced by multiple shorter ones. Also, the removal of cohesive devices such as conjunctions may result in anaphora crossing sentence boundaries. To maintain text coherence and cohesion (Grodzinsky et al., 1993) an anaphor is replaced by its referent if the containing sentence is split.

Lexical Simplifier The lexical simplifier (based on (Devlin, 1999; Devlin and Tait, 1998)) replaces content words with simpler synonyms. It first retrieves a set of synonyms for each word from WordNet (Miller et al., 1993), then, according to the user's desired level of simplification, the original word plus a percentage of the synonym list are looked up in the Oxford Psycholinguistic Database (Quinlan, 1992) for the corresponding Kucera-Francis frequencies. The word with the highest frequency is selected.

Morphological Generator Simplification works on the inflectionally analysed text, so the last stage is morphological generation. The generator is simply an inverted version of the morphological analyser described above. The inversion is performed automatically (Minnen and Carroll, Submitted), so any improvements made to the analyser are reflected in the generator at no extra cost. Finally, inter-word spelling changes (e.g. *a apple* → *an apple*), auxiliary reduction, etc. are performed.

3 Evaluation

We will perform an experimental evaluation of the system with the help of aphasic participants who are matched to the extent that none display visually related reading difficulties, which would confound the results, and all possess a sufficiently high reading ability—determined at the time of the experiment by using an aphasia assessment battery. As the system is a general tool aimed at

all aphasics, the participants will not be screened for aphasia type. The readability of the simplified text and the usability of the system will be assessed by observation and interview; questions will be posed to gauge subjects' comprehension of both explicit and implicit material.

References

- B. Baldwin. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop sponsored by the ACL (Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts)*, Universidad Nacional de Educacion a Distancia, Madrid, Spain.
- E. Briscoe and J. Carroll. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *Proceedings of the 4th ACL/SIGPARSE International Workshop on Parsing Technologies*, pages 48–58.
- J. Carroll and E. Briscoe. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. In *Proceedings of the ACL/SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 92–100.
- R. Chandrasekar, C. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*.
- H. Cunningham, Y. Wilks, and R. Gaizauskas. 1996. GATE—a general architecture for text engineering. In *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*.
- S. Devlin and J. Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. In J. Nerbonne. *Linguistic Databases*. Lecture Notes. Stanford, USA: CSLI Publications.
- S. Devlin. 1999. Simplifying natural language text for aphasic readers. Ph.D. Dissertation, University of Sunderland, UK.
- D. Elworthy. 1994. Does Baum Welch re-estimation help taggers? In *Proceedings of the 4th ACL Conference on Applied Natural Language Processing*, pages 53–58.
- Y. Grodzinsky, K. Wexler, Y. Chien, S. Marakovitz, and J. Solomon. 1993. The breakdown of binding relations. *Brain and Language*, 45(3):396–422.
- G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Teng. 1993. Five papers on WordNet. Technical report, Princeton University, Princeton, N.J.
- G. Minnen and J. Carroll. Submitted. Fast and robust morphological generation in a practical NLP system.
- M. Osborne. Submitted. Minimum description length-based models for practical grammar induction.
- P. Quinlan. 1992. *The Oxford Psycholinguistic Database*. Oxford University Press.

³Semantically reversible sentences such as *The boy was kissed by the girl* are even more difficult, since either noun phrase could be the subject.