Learning unification-based grammars using the Spoken English Corpus

Miles Osborne and Derek Bridge
Department of Computer Science, University of York, Heslington, York YO1 5DD, U. K.
{miles,dgb}@minster.york.ac.uk
July 12, 1994

Abstract
This paper describes a grammar learning system that combines model-based and data-driven learning within a single framework. Our results from learning grammars using the Spoken English Corpus (SEC) suggest that combined model-based and data-driven learning can produce a more plausible grammar than is the case when using either learning style in isolation.

1 Introduction
In this paper, we present some results of our grammar learning system acquiring unification-based grammars using the Spoken English Corpus (SEC). The SEC is a collection of monologues for public broadcast and is small (circa 50,000 words) in comparison to other corpora, such as the Lancaster-Oslo-Bergen Corpus [JLG78], but sufficiently large to demonstrate the capabilities of the learning system. Furthermore, the SEC is tagged and parsed, thus side-stepping the problems of constructing a suitable lexicon and of creating an evaluation corpus to determine the plausibility of the learnt grammars.

In contrast to other researchers (for example [BMMS92, GLS87, Bak79, LY90, VB87]), we try to learn competence grammars and not performance grammars. We also try to learn grammars that assign linguistically plausible parses to sentences. Learning competence grammars that assign plausible parses is achieved by combining model-based and data-driven learning within a single framework [OB93b, OB93a]. The system is implemented to make use of the Grammar Development Environment (GDE) [CGBB88] and it augments the GDE with 3300 lines of Common Lisp.
Our aim in this paper is to show that combining both learning styles produces a grammar that assigns more plausible parses than is the case for grammars learnt using either learning style in isolation. Plausibility is important in Natural Language Processing as it is very rare that applications need just to determine whether a sentence is grammatical: applications need also to determine the internal structure of sentences (a plausible parse). A grammar that assigns plausible parses is therefore preferable over one that does not assign plausible parses.

The structure of this paper is as follows. Section 2 gives an overview of the combined model-based and data-driven learner. Section 3 then describes the method used to generate the results, which are then presented in section 4. Section 5 discusses these results and points the way forward.

2 System overview

2.1 Architecture

We assume that the system has some initial grammar fragment, G, from the outset. Presented with an input string, W, an attempt is made to parse W using G. If this fails, the learning system is invoked. Learning takes place through the interleaved operation of a parse completion process and a parse rejection process.

In the parse completion process, the learning system tries to generate rules that, had they been members of G, would have enabled a derivation sequence for W to be found. This is done by trying to extend incomplete derivations using what we call super rules. Super rules are the following unification-based grammar rules:

\[
[] \rightarrow [] [] \quad \text{(binary)}
\]
\[
[] \rightarrow [] \quad \text{(unary)}
\]

The binary rule says (roughly) that any category rewrites as any two other categories, and the unary rule says (roughly) that any category rewrites as any other category. The categories in unification grammars are expressed by sets of feature-value pairs; as the three categories in the binary super rule and two categories in the unary super rule specify no values for any of the grammar's features, these rules are the most general (or vacuous) binary and unary rules possible. These rules thus enable constituents found in an incomplete analysis of W to be formed into a larger constituent. In unifying with these constituents, the categories on the right-hand side of the super rules become partially instantiated with feature-value pairs. Hence, these rules ensure that at least one derivation sequence will be found for W.

Many instantiations of the super rules may be produced by the parse completion process described above. Linguistically implausible instantiations must