

Extending CCGbank with quotes and multi-modal CCG

Daniel Tse and James R. Curran

School of Information Technologies

University of Sydney

NSW 2006, Australia

{dtse6695, james}@it.usyd.edu.au

Abstract

CCGbank is an automatic conversion of the Penn Treebank to Combinatory Categorical Grammar (CCG). We present two extensions to CCGbank which involve manipulating its derivation and category structure. We discuss approaches for the automatic re-insertion of removed quote symbols and evaluate their impact on the performance of the C&C CCG parser. We also analyse CCGbank to extract a multi-modal CCG lexicon, which will allow the removal of hard-coded language-specific constraints from the C&C parser, granting benefits to parsing speed and accuracy.

1 Introduction

Combinatory Categorical Grammar (Steedman, 2000) is a powerful lexicalised grammar formalism. In CCG, the combination of *categories* using a small set of *combinatory rules* allows a parser to simultaneously build up syntax and semantics.

CCGbank, a corpus automatically converted from the Wall Street Journal section of the Penn Treebank (Marcus et al., 1994), forms the cornerstone of research in wide-coverage CCG parsing. Its coverage enables the construction of practical, efficient and robust CCG parsers (Clark and Curran, 2007).

We describe corpus transformations on CCGbank which improve its linguistic fidelity and discriminative power. We restore the quote symbols to CCGbank derivations removed by the CCGbank derivation procedure (Hockenmaier, 2003). Quotes yield useful local information for many corpus applications, such as speaker or topic segmentation,

and the *supertagging* phase in a CCG parser (the assignment of categories to lexical items).

Multi-modal Combinatory Categorical Grammar (Baldrige, 2002) is an extension to CCG for finer-grained control over the applicability of combinatory rules. We develop a method for the semi-automatic extraction of a MMCCG corpus from CCGbank. These corpus transformations increase the fidelity and precision of CCGbank and hence its usefulness in a range of applications.

2 Multi-modal CCG

In CCG, lexical items are mapped to categories, which are combined through a set of combinatory rules. Typically, permitting all of the combinators to be considered by the parser leads to the undesired acceptance of ungrammatical examples, known as *overgeneration*. To prevent this in pure CCG, restrictions on the set of applicable rules must be specified as hard-coded parser constraints.

$$\begin{array}{l} X/\star Y \quad Y \rightarrow X \\ Y \quad X \backslash \star Y \rightarrow X \end{array} \quad \begin{array}{l} Y/\circ Z \quad X \backslash \circ Y \rightarrow X/\circ Z \end{array}$$

$$\begin{array}{l} X/\circ Y \quad Y/\circ Z \rightarrow X/\circ Z \\ Y \backslash \circ Z \quad X \backslash \circ Y \rightarrow X \backslash \circ Z \end{array} \quad \begin{array}{l} X \rightarrow T/i(T \backslash_i X) \\ X \rightarrow T \backslash_i(T/i X) \end{array}$$

To restore lexicality, Baldrige (2002) devised *multi-modal* CCG, in which each slash of a category encodes an indication (or *mode*) defining the rules in which it may participate. Modes simultaneously yield efficiency benefits and a reduction in derivational ambiguity by limiting the set of combinatory rules a MMCCG parser has to consider.

3 Re-quoting CCGbank

We restore quotes to CCGbank by consulting the Penn Treebank to determine the original location of

the opening and/or closing quote. Having found the corresponding leaf in the CCGbank derivation, we ascend to its parent as long as the leaves of its subtree contain strictly a sequence of tokens which were originally between quotes in the text. We splice in the quote below the node at which this condition is no longer true, so that the re-inserted quote leaf dominates the text that it quotes.

We correctly reinstate quotes in 8477 of the 8677 derivations (97.7%) originally containing quotes. We evaluate the re-quoted CCGbank on the C&C CCG parser (Clark and Curran, 2007).

<i>Corpus</i>	Labelled F	Supertagger acc
C&C orig	85.12%	93.05%
Re-quoted	85.03%	93.18%

Figure 1: C&C evaluation

As per our hypothesis, the extra local information provided by the presence of quotes slightly increases supertagging accuracy. However, there is a small tradeoff against parser accuracy, due to the additional complexity the re-added quotes entail.

4 Multi-modal CCG

A simple approach to mode annotation is to examine each slash of each category occurring in CCGbank. If a given slash is *consumed* by a given rule with a proportion α of total cases, then we assign that slash a mode (\star, \bowtie, \circ) compatible with that rule. The *null mode* (\bowtie) permits no rules, allowing us to mark a category such as $S[pt] \backslash NP$, whose slash is never directly consumed.

The problem with this automated approach is sparseness: rarely attested rules can nevertheless contribute to legitimate analyses. We address this by performing manual annotation on those slashes consumed often by the more powerful composition rules, while relying on our frequency cutoff criterion to assign the application-only and null modes.

There are two outcomes in annotating a given CCG category with modes: for each slash, we either assign the least permissive mode that preserves the vast majority of CCGbank derivations, or else we discover that a given CCG category corresponds to two or more MMCCG categories differing by mode.

We give an example of the additional fine-grained control provided by MMCCG, allowing us to make a grammaticality distinction previously impossible to specify lexically in CCG. The adverbs *freely* and *evidently* belong to the classes *VP adverb* and *sentential adverb*, respectively. The former are characterised by their ability to undergo a degree of shifting unavailable to the latter.

- (1) Adverbs permute within their phrase *freely*.
- (2) Adverbs permute *freely* within their phrase.
- (3) He knows some judo *evidently*.
- (4) *He knows *evidently* some judo.

This distinction cannot be made in CCG, because both *freely* and *evidently* share the structural category $(S \backslash NP) \backslash (S \backslash NP)$. However, the additional derivational control of MMCCG allows us to partition the VP and sentential adverbs. In particular, the VP adverbs would carry a category $(S \backslash NP) \backslash_{\circ} (S \backslash NP)$, the mode \circ permitting the use of the combinatory rules of composition which enable a CCG analysis of movement, while the sentential adverbs would receive the category $(S \backslash NP) \backslash_{\star} (S \backslash NP)$, the mode \star only permitting the non-associative and non-permutative rules of application. A further benefit of moving these distinctions into the lexicon is that we can make these grammaticality distinctions with the granularity of lexical items.

5 Conclusion

We have described two transformations on CCGbank, which enhance and extend its utility as a CCG corpus. We have produced a CCGbank of greater fidelity through an algorithm for re-instating quote symbols removed during its corpus conversion process, demonstrated the role of multi-modal CCG in addressing overgeneration inherent in pure CCG, and described a strategy for the generation of a MMCCG corpus. We have considered automatic and manual strategies for the annotation of a MMCCG corpus, and justify our chosen solution of a compromise between them. The focus of our work is now to refine the corpus obtained, and implement a full MMCCG parser.

Through the generation of a multi-modal version of CCGbank, we have the potential for more accurate, and at the same time more efficient wide-coverage CCG parsing.

References

- Jason Baldridge. 2002. *Lexically Specified Derivational Control in Combinatory Categorical Grammar*. Ph.D. thesis, University of Edinburgh.
- Stephen Clark and James R. Curran. 2007. Wide-Coverage Statistical Parsing with CCG and Log-Linear Models. *To appear in Computational Linguistics 2007. Unpublished draft manuscript.*
- Julia Hockenmaier. 2003. *Data and Models for Statistical Parsing with Combinatory Categorical Grammar*. Ph.D. thesis, University of Edinburgh.
- Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.
- Mitchell 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.
- Mark Steedman. 2000. *The Syntactic Process*. MIT Press. Cambridge, MA, USA.