Towards Surface Realization with CCGs Induced from Dependencies

Michael White
Department of Linguistics
The Ohio State University
Columbus, OH 43210, USA
mwhite@ling.osu.edu

Abstract
We present a novel algorithm for inducing Combinatory Categorial Grammars from dependency treebanks, along with initial experiments showing that it can be used to achieve competitive realization results using an enhanced version of the surface realization shared task data.

1 Introduction
In the first surface realization shared task (Belz et al., 2011), no grammar-based systems achieved competitive results, as input conversion turned out to be more difficult than anticipated. Since then, Narayan & Gardent (2012) have shown that grammar-based systems can be substantially improved with error mining techniques. In this paper, inspired by recent work on converting dependency treebanks (Ambati et al., 2013) and semantic parsing (Kwiatkowski et al., 2010; Artzi and Zettlemoyer, 2013) with Combinatory Categorial Grammar (CCG), we pursue the alternative strategy of inducing a CCG from an enhanced version of the shared task dependencies, with initial experiments showing even better results.

A silver lining of the failure of grammar-based systems in the shared task is that it revealed some problems with the data. In particular, it became evident that in cases where a constituent is annotated with multiple roles in the Penn Treebank (PTB), the partial nature of Propbank annotation and the restriction to syntactic dependency trees meant that information was lost between the surface and deep representations, leading grammar-based systems to fail for good reason. For example, Figure 1 shows that with free object relatives, only one of the two roles played by how much manufacturing strength is captured in the deep representation, making it difficult to linearize this phrase correctly. By contrast, Figure 2 (top)

Figure 1: Shared Task Input for Economists are divided as to [how much manufacturing strength], they expect to see t in September reports on industrial production and capacity utilization, also due tomorrow (wsj.2400.6, “deep” representation)

shows an experimental version of the shallow representation intended to capture all the syntactic dependencies in the PTB, including the additional object role played by this phrase here. Including all PTB syntactic dependencies in the shallow representation makes it feasible to define a compatible CCG; at the bottom of the figure, a corresponding CCG derivation for these dependencies is shown. In the next section, we present an algorithm for inducing such derivations. In contrast to Ambati et al.’s (2013) approach, the algorithm integrates the proposal of candidate lexical categories with the derivational process, making it possible to derive categories involving unsaturated arguments, such as $s_e, del / np_x / (s_e', to / np_y)$; it also makes greater use of unary type-changing rules, as with Artzi & Zettlemoyer’s (2013) approach.

---

1 Kudos to Richard Johansson for making these enhancements available.
Unlike their approach though, it works in a broad coverage setting, and makes use of all the combinator s standardly used with CCG, including ones for type-raising.

2 Inducing CCGs from Dependencies

Pseudocode for the induction algorithm is given in Figure 3. The algorithm takes as input a set of training sentences with their gold standard dependencies. We pre-processed the dependencies to make coordinating conjunctions the head, and to include features for zero-determiners. The algorithm also makes use of a seed lexicon that specifies category projection by part of speech as well as a handful of categories for function words. For example, (1) shows how a tensed verb projects to a finite clause category, while (2) shows the usual CCG category for a determiner, which here introduces a (NMOD) dependency.  

1) \[ \text{expect} \vdash s_{w,d} : \emptyset_s(\text{expect} \wedge \langle \text{TENSE},\text{pres} \rangle) \]

2) \[ \text{the} \vdash \text{n}_{w_2}/\text{n}_{w_3} : \emptyset_s((\text{NMOD} \langle d \wedge \text{the} \rangle)) \]

The algorithm begins by instantiating the lexical categories and type-changing rules that match the input dependency graph, tracking the categories in a map (edges) from nodes to edges (i.e., signs with a coverage vector). It then recursively visits each node in the primary dependency tree bottom up (combineEdges), using a local chart (doCombos) at each step to combine categories for adjacent phrases in all possible ways. Along the way, it creates new categories (extendCats and coordCats) and unary rules (applyNewUnary). For example, when processing the node for expect in Figure 2, the nodes for they and to are recursively processed first, deriving the categories \( np_{w_9} \) and \( s_{w_{11},to} \) \( np_{w_5} / np_{w_8} \) for they and to see . . . , respectively. The initial category for expect is then extended as shown in (3), which allows for composition with \( to \) see . . . (as well as with a category for simple application). When there are coordination relations for a coordinating conjunction (or coordinating punctuation mark), the appropriate category for combining like types is instead constructed, as in (4). Additionally, for modifiers, unary rules are instantiated and applied, e.g. the rule for noun-noun compounds in (5).

\[ a_{w_6} / a_{w_7} : \emptyset_s(d \wedge \langle \text{NMOD},\text{gen} \rangle) \]

Figure 3: CCG Induction Algorithm
economists are divided as to how much manufacturing strength they expect to see in ... .

... September reports on industrial production and capacity utilization, also due tomorrow

Figure 2: Augmented Syntactic Dependencies with Corresponding CCG Derivation (dashed dependencies indicate relations from additional parents beyond those in the primary tree structure)
3 Experiments and Future Work

We ran the induction algorithm over the standard PTB training sections (02–21), recovering complete derivations more than 90% of the time for most sections. Robust treatment of coordination, including argument cluster coordination and gapping, remains a known issue; other causes of derivation failures remain to be investigated. To select preferred derivations, we used a complexity metric that simply counts the number of steps and the number of slashes in the categories. We then trained a generative syntactic model (Hockenmaier and Steedman, 2002) and used it along with a composite language model to generate n-best realizations for reranking (White and Rajkumar, 2012), additionally using a large-scale (giga-word) language model. Development and test results appear in Table 1. Perhaps because of the expanded use of type-changing rules with simple lexical categories, the generative model and hyperpatter (Espinosa et al., 2008) performed worse than expected. Combining the generative syntactic model and composite language model (GEN) with equal weight yielded a devtest BLEU score of only 0.4513, while discriminatively training the generative component models (GLOBAL) increased the score to 0.7679. Using all features increased the score to 0.8083, while doubling the beam size (ALL+) pushed the score to 0.8210, indicating that search errors may be an issue. Ablation results show that leaving out the large-scale language model (NO-BIGLM) and dependency-ordering features (NO-DEPORD) substantially drops the score.3 Focusing only on the 80.5% of the sentences for which a complete derivation was found (COMPLETE) yielded a score of 0.8668. By comparison, realization with the native OpenCCG inputs (and the large-scale LM) on all sentences (NATIVE) yields a score more than five BLEU points higher, despite using inputs with more semantically-oriented relations and leaving out many function words, indicating that there is likely substantial room for improvement in the pre-processing and grammar induction process. Towards that end, we tried selecting the best derivations using several rounds of Viterbi EM with the generative syntactic model, but doing so did not improve realization quality.

A similar pattern is seen in the Section 23 results, with a competitive BLEU score of 0.8260 with the expanded beam, much higher than Narayan & Gardent’s (2012) score of 0.675 with 38.8% coverage, the best previous score with a grammar-based system. This score still trails the shared task scores of the top statistical dependency realizers by several points (STUMABA-S at 0.8911 and DCU at 0.8575), though it exceeds the score of a purpose-built system using no external resources (ATT at 0.6701). In future work, we hope to close the gap with the top systems by integrating an improved ranking model into the induction process and resolving the remaining representational issues with problematic constructions.

Acknowledgments

Thanks to the anonymous reviewers, Richard Johansson and the University of Sydney Schwa Lab for helpful comments and discussion. This work was supported in part by NSF grants IIS-1143635 and IIS-1319318.

Table 1: Development set (Section 00) & test set (Section 23) results, including exact match and complete derivation percentages and BLEU scores

<table>
<thead>
<tr>
<th>Model</th>
<th>Exact</th>
<th>Complete</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sect 00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEN</td>
<td>2.4</td>
<td>79.5</td>
<td>0.4513</td>
</tr>
<tr>
<td>GLOBAL</td>
<td>29.7</td>
<td>79.0</td>
<td>0.7679</td>
</tr>
<tr>
<td>NO-BIGLM</td>
<td>29.1</td>
<td>78.2</td>
<td>0.7757</td>
</tr>
<tr>
<td>NO-DEPORD</td>
<td>34.3</td>
<td>77.9</td>
<td>0.7956</td>
</tr>
<tr>
<td>ALL</td>
<td>35.8</td>
<td>78.4</td>
<td>0.8083</td>
</tr>
<tr>
<td>ALL+</td>
<td>36.4</td>
<td>80.5</td>
<td>0.8210</td>
</tr>
<tr>
<td>COMPLETE</td>
<td>44.4</td>
<td></td>
<td>0.8668</td>
</tr>
<tr>
<td>NATIVE</td>
<td>48.0</td>
<td>88.7</td>
<td>0.8993</td>
</tr>
</tbody>
</table>

| Sect 23        |       |          |        |
| GEN            | 2.8   | 80.3     | 0.4560 |
| GLOBAL         | 31.3  | 78.5     | 0.7675 |
| ALL            | 37.6  | 77.2     | 0.8083 |
| ALL+           | 38.1  | 80.4     | 0.8260 |
| COMPLETE       | 47.0  |          | 0.8743 |
| NATIVE         | 46.4  | 86.4     | 0.8594 |

4 All differences were statistically significant at \( p < 0.01 \) with paired bootstrap resampling (Koehn, 2004).
References


