Cross-lingual CCG Induction: Learning Categorial Grammars via Parallel Corpora

Anonymous ACL submission

Abstract

Combinatory categorial grammars are linguistically motivated and useful for semantic parsing, but costly to acquire in a supervised way and difficult to acquire in an unsupervised way. We propose an alternative making use of cross-lingual learning: an existing source-language parser is used together with a parallel corpus to induce a grammar and parsing model for a target language. On the PASCAL benchmark, cross-lingual CCG induction outperforms CCG induction from gold-standard POS tags on 3 out of 8 languages, and unsupervised CCG induction on 6 out of 8 languages. We also show that cross-lingually induced CCGs reflect known syntactic properties of the target languages.[1]

1 Motivation

Combinatory Categorial Grammar (CCG; Steedman, 2001; Steedman and Baldridge, 2011) is a grammar formalism known for its elegant handling of coordination, its transparent syntax-semantic interface, its computational efficiency, and its strongly lexicalized approach, which uses few and very general rules, and puts all that is language-specific into the lexicon. CCG has been successfully used for statistical syntactic parsing and has emerged as a leading grammar formalism in semantic parsing. CCGs used in most work to date are either hand-crafted (Zettlemoyer and Collins, 2007; Kwiatkowski et al., 2013; Artzi et al., 2015) or extracted from large syntactically annotated corpora (Curran et al., 2007; Reddy et al., 2014). In either case language-specific human effort is required. Acquiring CCGs in an unsupervised way is difficult and does not reach the performance of supervised methods (Bisk and Hockenmaier, 2013). As a result, most research focuses on English and other languages are neglected, meaning that speakers of other languages have delayed or no access to CCG-based semantic parsing technology.

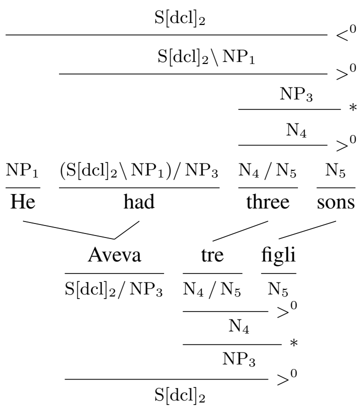

Figure 1: Projection of an English CCG derivation to an Italian translation. The indices distinguish different instances of categories.

We propose to overcome this bottleneck by inducing CCGs cross-lingually, i.e., transferring an existing grammar from English to other languages via unannotated parallel data. The process is illustrated for one English-Italian sentence pair in Figure 1: the English sentence is parsed by an existing CCG parser and word-aligned to the Italian sentence. Italian words receive categories equivalent to those of the aligned English words, and a semantically equivalent derivation is built for the Italian sentence. With enough derivations projected in this way, they can be used to extract a CCG lexicon and to estimate parameter weights for parsing the target language.

Unlike previous competitive methods for CCG induction such as Bisk and Hockenmaier (2013), our method does not require the training data to be POS-tagged. It also induces more fine-grained labels. In this work, we compare the performance

[1] This submission is an extended abstract of Evang (2019).

of parsers trained using our method to previous induced CCG parsers. We also investigate whether the cross-lingually induced CCG lexicons correspond with linguistic insights about the target languages.

2 Summary of Method

We start by extracting sentence pairs from a parallel corpus, parsing the source-language side with the existing parser, and do unsupervised word alignment using standard methods from machine learning. The result is exemplified in the upper half of Figure 1: a source-language derivation where the lexical categories are aligned to target-language words. We now induce the target-language derivation by transfering lexical source-language categories to the aligned target-language words and applying CCG's combinatory rules to obtain a derivation. There are some complications to this:

1. **Many-to-one alignments.** In our example, *He had* is aligned to *Aveva*. We deal with such cases by applying CCG's combinatory rules to the source-language phrase, thereby obtaining a single category to assign to the target-language word.

2. **Unaligned source-language words.** A common occurrence is that the source language uses an article where the target language does not, leaving it unaligned. We ensure successful projection by converting NP / N words to type-changing rules in such cases.

3. **Word order.** Word order may differ between source and target sentence. This means we need to change the directionality of some slashes. We present an algorithm to do this fully automatically.

4. **Ambiguity.** Since we construct the target derivation by applying CCG's combinatory rules, we need a way to deal with parsing ambiguity. We eliminate most ambiguity by indexing categories for this purpose (e.g., NP_1 and NP_3 in our example are treated as distinct) and by enforcing normal-form constraints. The remaining ambiguity mostly stems from ambiguous word alignments, which we use to improve recall. We describe a heuristic that we use to determine a single winning parse per sentence pair.

Once we have constructed target-language derivations in this way, we use them to train a supervised CCG parser for the target language in the standard way.

3 Summary of Results

The PASCAL benchmark for multilingual grammar induction includes Arabic, Czech, Danish, Basque, Dutch, Portuguese, Slovenian, and Swedish. Our approach outperforms the best unsupervised CCG induction system (Bisk and Hockenmaier, 2013) on Danish, Dutch, and Portuguese, and additionally on Arabic, Czech, and Danish when no use of gold-standard POS tags is allowed (Bisk et al., 2015).

We show that the CCG lexicons induced by our system reflect known typological properties of some target languages, such as 1) SVO and SOV word order in German and Dutch, 2) absence of *do*-support of German, Dutch, and Italian, 3) particle verbs in German and Dutch, 4) order of noun and attributive adjective in German and Dutch vs. Italian, 5) pro-drop in Italian.

4 Typological Significance

Our approach is an interesting testbed for CCG's approach to multilinguality, where combinatory rules are thought of as universal, and everything that is language-specific is in the lexicon. This obviously concerns the specific vocabulary of each language, but also features that are not normally thought of as lexical, such as word order. For example, a CCG lexicon for Italian would contain a high-frequency category $\mathrm{N} \backslash \mathrm{N}$ for attributive adjectives, whereas for languages such as English N / N would be the more frequent one due to the differing word order. Similarly, in a lexicons for languages such as German and Dutch, both $(\mathrm{S} \backslash \mathrm{NP}) / \mathrm{NP}$ and $(\mathrm{S} \backslash \mathrm{NP}) \backslash \mathrm{NP}$ would be frequent transitive verb categories due to the mixed SVO/SOV order. We have shown that this view is very promising for multilingual NLP and in particular for cross-lingual grammar induction, because it reduces grammar induction to lexicon induction. This let us obtain promising results with a conceptually simple method.

Future work will delve more deeply into typological features other than word order, such as analytic vs. synthetic encoding of tense, and into a typologically more diverse range of languages than we have so far explored.

References

Yoav Artzi, Kenton Lee, and Luke Zettlemoyer. 2015. Broad-coverage CCG semantic parsing with AMR. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1699–1710.

Yonatan Bisk, Christos Christodoulopoulos, and Julia Hockenmaier. 2015. Labeled grammar induction with minimal supervision. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–876. Association for Computational Linguistics.

Yonatan Bisk and Julia Hockenmaier. 2013. An HDP model for inducing combinatory categorial grammars. *Transactions of the Association for Computational Linguistics*, 1:75–88.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume: Proceedings of the Demo and Poster Sessions*, pages 33–36.

Kilian Evang. 2019. Cross-lingual CCG induction. In *Proceedings of NAACL-HLT*. To appear.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1556. Association for Computational Linguistics.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association of Computational Linguistics*, 2:377–392.

Mark Steedman. 2001. *The Syntactic Process*. The MIT Press.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. In Robert Borsley and Kersti Borjars, editors, *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell.

Luke S. Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687.