Chapter 6

Conclusion

> "It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it disagrees with experiment, it's wrong."
>
> —Richard P. Feynman

> "If your intuitions are good, you should follow them and you'll eventually be successful. If your intuitions are not good, it doesn't matter what you do."
>
> —Geoffrey E. Hinton

Statistical shift-reduce parsing poses a number of challenges. It is a representation learning challenge, usually boiled down to learning richer representations for various aspects of parser states. It is a structured learning challenge, demanding models to take into account the structural properties of the output. These two interrelated and orthogonal challenges are further compounded by inexact search, and together they constitute three essential elements in any shift-reduce model. While improving each element independently has shed light on their individual significance, approaches that treat them holistically have shown their merits.

The holistic approach is the one that has been considered in this study, which began with the question: Should we model the derivations, the dependencies, or both, for shift-reduce CCG parsing? The resulting algorithmic and structured learning issues that arose in the context of the linear structured perceptron dependency model

were addressed by a novel dependency oracle and by extending and generalizing provably correct theories marrying the structured perceptron and inexact search (Huang et al., 2012), while at the same time preserving and capitalizing on, rather than interfering with, the strengths of the normal-form shift-reduce CCG model (Zhang and Clark, 2011a), such as global structured learning and its rich feature sets (§3).

Drawing on recent work on using feed-forward neural networks for learning feature representations for parser states (Chen and Manning, 2014), I then described a framework for training RNN shift-reduce parsing models optimized for a task-specific loss based on expected F-measure (§4). Being agnostic to the underlying neural network architecture, this framework was also applied to an LSTM parser inspired by the stack-LSTM of Dyer et al. (2015) (§5). In both cases, the models were also globally normalized and the three aforementioned essential elements were tightly integrated.

Empirically, extensive experiments were performed throughout, and results were state-of-the-art. Clearly, however, the present study is far from complete.

First, all shift-reduce parsing models introduced still required a separate supertagging model, which has a large impact on the final parsing accuracy. Further conjoined with the POS tagging model, this pipelined approach has been dominating in CCG parsing for over a decade. But with the flexibility provided by neural networks, and as suggested by some recent works (Zhang and Weiss, 2016; Søgaard and Goldberg, 2016), it is reasonable to expect that an end-to-end neural model for CCG parsing can be derived, ideally even without being confined to a specific parsing paradigm, either chart-based or shift-reduce.

Second, I have just barely scratched the surface of investigating structured learning for neural shift-reduce models. Combining structured learning with deep learning is an emerging theme, and the same intuition applies to the models presented above, especially in devising both empirical and formal methods that faithfully take into account the respective neural models, instead of relying upon techniques originally developed for other models (Watanabe and Sumita, 2015; Weiss et al., 2015; Andor et al., 2016). Moreover, how to integrate such methods with a framework like expected F-measure training is of particular interest.

Aside from possible extensions, however, it is debatable under what guiding principles should the present study be further extended.

Statistical parsing is a well-defined research area that has attracted a lot of attention in computational linguistics, and it has generated a set of techniques many of which have been used for or adapted to other tasks (Wu, 1997; Chelba and Jelinek, 1998; Goodman, 1999; Ramshaw et al., 2001; Collins and Roark, 2004; Huang and Chiang, 2005; Zettlemoyer and Collins, 2005; Chiang, 2007; Dyer, 2010). Statistical parsers (Collins, 1997; Klein and Manning, 2003; Briscoe et al., 2006; McDonald, 2006; Nivre et al., 2006; Curran et al., 2007), on the other hand, have mainly played the role of feature generators, producing annotations shown to be useful in various settings (Yamada and Knight, 2001; Chiang et al., 2008; Marton and Resnik, 2008; Mi et al., 2008; Chiang et al., 2009; Dyer and Resnik, 2010). In addition to being used in such capacities, however, it is apparent that their future role is likely to be limited in end-to-end language processing approaches, and in the endeavour to achieve automated human-level language understanding—beyond formalism-dependent parsing and natural language text processing—which remains elusive with currently available language technologies.

As a related issue, the practical implications for formalisms like CCG also calls for revisiting, with one notable reason being that alternatives such as dependency grammars have usually demonstrated their superior simplicity and cross-lingual scalability that are preferred in production systems (McDonald et al., 2013; Andor et al., 2016), which put more emphasis on empirical results rather than the linguistic properties of a grammar.

But more importantly on a higher level, it might be illuminating to ask: How useful is syntax for language understanding, is it relevant at all?

Overall, this thesis does not intend to serve as a proponent for parsing nor CCG. Instead, it purposes itself as an exploration of structured learning with inexact search—in the context of shift-reduce CCG parsing. It is hoped that future work will continue in this spirit.