Weighted Parsing for Grammar-Based Language Models over Multioperator Monoids

Richard Mörbitz Heiko Vogler

We develop a general framework for weighted parsing called *weighted RTG-based language models*, define the *M-monoid parsing problem*, and present an algorithm for its solution.

There is a variety of weighted parsing problems, e.g., probabilistic constituent parsing, where the task is to compute for each English sentence the most probable constituent tree (according to some formal grammar). We view each of these particular problems as a mapping parse: $\mathcal{L} \to \mathbb{K}$, where $\mathcal{L}$ is the set of syntactic objects (e.g., the set of English sentences) and $\mathbb{K}$ is the set of weights (e.g., the set of English constituent trees).

Our approach makes heavy use of universal algebra, in the sense that both $\mathcal{L}$ and $\mathbb{K}$ are identified with the carrier sets of algebras. We employ the initial algebra semantics[1] to model the syntactic structure of $\mathcal{L}$ using regular tree grammars (RTG) and we choose multioperator monoids (M-monoids)[2] to compute the weights in $\mathbb{K}$. The M-monoid parsing problem determines how, given a syntactic object from $\mathcal{L}$, the weight in $\mathbb{K}$ is to be computed using both algebras. Thus particular weighted parsing problems amount to particular choices of $\mathcal{L}$ and $\mathbb{K}$ and are therefore instances of the M-monoid parsing problem.

We show that, in general, the M-monoid parsing problem cannot be solved by a terminating algorithm. We determine a large class of weighted RTG-based language models for which this is still possible. This class is essentially a generalization of the graphs weighted with closed semirings from Mohri's "single source shortest distance" framework.[3] We prove that our algorithm terminates and is correct for the class of *closed weighted RTG-based language models*.

In the end we are also interested in application scenarios of our algorithm, i.e., which weighted parsing problems are covered by the class of closed weighted RTG-based language models. It turns out that our approach subsumes the previous approaches to weighted parsing, semiring parsing[4] and weighted deductive parsing,[5] and also covers weighted parsing problems which are outside the scope of both (e.g., parsing as intersection). Moreover, we can even solve the problems of algebraic dynamic programming.[6]

This work was presented at the 14th International Conference on Finite-State Methods and Natural Language Processing (FSMNLP 2019), September 23–25, 2019, Dresden, Germany.

[1] J. A. Goguen, J. W. Thatcher, E. G. Wagner, and J. B. Wright. "Initial Algebra Semantics and Continuous Algebras". In: *Journal of the Association for Computational Machinery* (1977).

[2] W. Kuich. "Linear systems of equations and automata on distributive multioperator monoids". In: *Contributions to general algebra* (1999).

[3] M. Mohri. "Semiring frameworks and algorithms for shortest-distance problems". In: *Journal of Automata, Languages and Combinatorics* (2002).

[4] J. Goodman. "Semiring parsing". In: *Computational Linguistics* (1999).

[5] M.-J. Nederhof. "Weighted deductive parsing and Knuth's algorithm". In: *Computational Linguistics* (2003).

[6] R. Giegerich, C. Meyer, and P. Steffen. "A discipline of dynamic programming over sequence data". In: *Science of Computer Programming* (2004).