Encoding and Decoding Graph Representations of Natural Language

Samuel Thomson

Abstract

When analyzing natural language text, it is often helpful to process the text into structured representations of its linguistic content. Labeled directed graphs are a natural and flexible representation, especially for semantics. Their generality over trees, for instance, allows them to represent relational semantics while handling phenomena like coreference and coordination.

This thesis focuses on such graphical representations of text. We develop novel algorithms and models for decoding—automatically extracting graphs from raw text; and encoding—reducing arbitrarily sized graphs to tensors so that they can be used as input to a downstream system. We primarily focus on semantic graphs—graphs that represent sentence meanings. Semantic graphs are not necessarily trees, so they require methods that can handle arbitrary graphs. We make three main contributions. First, we introduce neural turbo parsing. We show that it is an effective approach for semantic dependency parsing, and in particular multitask semantic dependency parsing, where we get state-of-the-art results for three formalisms. Next, we show that the problem of decoding connected semantic graphs is an extension of a classic, well-studied problem in network optimization called the prize-collecting Steiner tree (PCST) problem. This allows us to prove theoretical properties about decoding under certain graph constraints, and to adapt and extend existing PCST solving techniques to decoding semantic graphs. Finally, we introduce two new lawful neural encoders whose constrained forms allow for efficient exact inference. Soft Patterns (SoPa) encode sequential text using an array of small neurally-weighted finite state automata; and Kleene Graph Encoder Networks (KGEN) extend SoPa to work on graphs. KGEN is able to encode all paths in a graph, grouped by source and destination, using dynamic programming.