Knowledge Representation and Extraction at Scale

Christos Christodoulopoulos
Amazon Research, Cambridge, UK
chrchrs@amazon.co.uk

Bio

Christos Christodoulopoulos is a Research Scientist at Amazon Research Cambridge (UK), working on knowledge extraction and verification. He got his PhD at the University of Edinburgh, where he studied the underlying structure of syntactic categories across languages. Before joining Amazon, he was a post-doctoral researcher at the University of Illinois working on semantic role labeling and psycholinguistic models of language acquisition. He has experience in science communication including giving public talks and producing a science podcast.

Abstract

These days, most general knowledge question-answering systems rely on large-scale knowledge bases comprising billions of facts about millions of entities. Having a structured source of semantic knowledge means that we can answer questions involving single static facts (e.g. "Who was the 8th president of the US?") or dynamically generated ones (e.g. "How old is Donald Trump?"). More importantly, we can answer questions involving multiple inference steps ("Is the queen older than the president of the US?").

In this talk, I'm going to be discussing some of the unique challenges that are involved with building and maintaining a consistent knowledge base for Alexa, extending it with new facts and using it to serve answers in multiple languages. I will focus on three recent projects from our group. First, a way of measuring the completeness of a knowledge base, that is based on usage patterns. The definition of the usage of the KB is done in terms of the relation distribution of entities seen in question-answer logs. Instead of directly estimating the relation distribution of individual entities, it is generalized to the "class signature" of each entity. For example, users ask for baseball players' height, age, and batting average, so a knowledge base is complete (with respect to baseball players) if every entity has facts for those three relations.

Second, an investigation into fact extraction from unstructured text. I will present a method for creating distant (weak) supervision labels for training a large-scale relation extraction system. I will also discuss the effectiveness of neural network approaches by decoupling the model architecture from the feature design of a state-of-the-art neural network system. Surprisingly, a much simpler classifier trained on similar features performs on par with the highly complex neural network system (at 75x reduction to the training time), suggesting that the features are a bigger contributor to the final performance.

Finally, I will present the Fact Extraction and VERification (FEVER) dataset and challenge. The dataset comprises more than 185,000 human-generated claims extracted from Wikipedia pages. False claims were generated by mutating true claims in a variety of ways, some of which were meaning-altering. During the verification step, annotators were required to label a claim for its validity and also supply full-sentence textual evidence from (potentially multiple) Wikipedia articles for the label. With FEVER, we aim to help create a new generation of transparent and interprable knowledge extraction systems.

Proceedings of SemDeep-3, the 3rd Workshop on Semantic Deep Learning, page 46
Santa Fe, New Mexico, USA, August 20, 2018.