Parsing using Compositional Vector Grammars

A.U.S.S. Pradeep (deepu@iitk.ac.in), Triya Bhattacharya (triya@iitk.ac.in)
Supervisor: Dr. Amitabh Mukerjee
Dept. of Computer Science and Engineering

Problem Statement:

The aim of the project is to parse natural language sentences using Compositional Vector Grammars (CVG) – which is a combination of Probabilistic Context-Free Grammars (PCFG) and Recursive Neural Networks (RNN) by learning syntactic-semantic representations of linguistic phrases.

Motivation and Related Work:

Syntactic Parsing is an important task which is useful in relations between syntax and semantics and is useful for "relation extraction, semantic role labeling and paraphrase detection" [1]. Earlier, descriptions consisted of coarse categories for identifying phrases but it has been shown that better results are obtained by finer categories, by either manual or automatic learning.

Recent work has included discriminative parsing through careful engineering of features [2]; and lexicalized parsers which associate the discrete categories with lexical phrases [3]. Previous RNN based parsers have used equal, tied weights for computing the vectors (representing a constituent) [4]. Here, the problem of learning features for long, unseen n-grams is tackled by using the CVG for structure prediction. Also, the RNN uses weights dependent on child categories of constituent.

Method Chosen and Dataset:

The method being used here is that phrases are parsed as discrete categories as well as continuous vectors (concatenation of phrases results in a continuous vector, which has its own syntactic representation). PCFGs are used to categorize phrases into discrete, coarse categories, whereas RNNs are used to assign fine-grained information. The CVG model jointly identifies syntactic structure and also the related semantic information.

The vector representation of words is dependent on co-occurrence statistics of a word and its context. The parser combines generative learning and speed of PCFGs with discriminative learning and semantic richness of RNNs.

The dataset being used currently is the Standard Penn Treebank. The benchmark to be used is the standard Stanford parser.

Possible Extensions:

The possible extension could be to parse sentences in a local language, but we would need annotated grammar for that.

References:

[1] Parsing using CVGs, Socher and Manning (2013)

[2] D. Klein and C. D. Manning. 2003a. Accurate unlexicalized parsing.

[2] (Taskar et al., 2004; Finkel et al., 2008).

[3] M. Collins. 2003. Head-driven statistical models for natural language parsing; Charniak 2000, A maximum-entropy-inspired parser.

[4] Socher et al. 2011, Parsing Natural Scenes and Natural Language with Recursive Neural Networks.