Research Paper Summarization

Max Entropy Inspired Parser

This paper describes a method to choose an appropriate parser when parsing text. The paper begins by explaining a generative model. In the generative model, a sentence to be parsed is broken into constituents. These constituents can be labelled as noun-phrases, verb-phrases, punctuation etc. A single constituent can be labelled/parsed as many types. The right label must be found since this can determine the head of a sentence (the most important term) and is pivotal in interpreting the sentence's meaning. The right label is considered the label with the highest probability. The generative model has a preset (empirically obtained set) of probability which may not always be accurate, especially with low empirical data. The Maximum-Entropy Inspired parsing tries to reconcile this dilemma. More specifically, the maximum-entropy method employs log-linear models to smooth out the gap between abundant and sparse empirical data.

The new parser outperforms previous state-of-the-art parsers in all measures of testing. The average precision/recall measure is a good representation of improvement. The parser of this paper has 91.1% precision/recall for sentences less than 40 characters and 89.5 precision/recall for sentences less than 100 characters long. This leads to a 13% less parsing error when compared to previous parsers.

The max-entropy model isn't the only reason Charniak's et al. parser works so well. The max-entropy model uses features to relate a constituent with historical constituents. In order to change the model, parser builders need only to change the features. This flexibility allowed the authors to try different 'tweaks' to the model. One tweak is the parser finding the preterminal before the head of a sentence. Finding the header given the preterminal is easier as the probability is conditioned with the preterminal and thereby reducing the set of possible header candidates. Another important tweak is using Markov grammar. Markov grammar uses surrounding words for context whereas the alternative, tree-bank grammar, uses a preset dictionary to deduce context. Markov grammar has degrees/orders. First order Markov grammar uses one constituent from the left and one from the right of the constituent in question. Second order will use two from the left and two from the right of a constituent. This pattern goes on for higher orders. The flexibility from the max-entropy parsers allowed the authors to try many Markov orders with ease. A third order Markov grammar performed well above tree-bank grammar. All these innovations to existing parsers helped the Maximum-Entropy Inspired parser significantly improve performance standards.