Lab Overview

Gideon Maillette de Buy Wenniger
University of Amsterdam
gemdbw - at - gmail - dot -com

1 overview of the Project

The lab is about building a prototype statistical parser. Intuitively sentences have associated syntactic structures. These structures have been manually annotated for a set of training sentences, the so called treebank. The function of a parser is to produce the best (most likely) parses for new sentences. This is a Machine Learning problem, but more than that. In fact there are three parts to the problem, that nicely correspond to general areas of Artificial Intelligence:

1. A modeling problem. How to make a model that describes how threes are build for sentences.
2. An optimization problem: how to optimize/learn/set the parameters of the model to best match the training data
3. A search problem: given a model and its parameters, how to find the best parse under the model.

In step 1 you will extract all depth one sub-trees from the treebank and use these to create a Probabilistic Context Free Grammar (PCFG) whose parameters are estimated by Maximum Likelihood Estimation (in this case simple counting and computation of conditional probabilities based on rule occurrence counts).

In step 2 you will build a bottom up parser, implementing the CYK/CKY Algorithm. Here you produce all trees that are consistent with a sentence under the model (the PCFG you extracted in Step 1). You do not produce probabilities yet (although you may of course do that as well already), this is done in the next step.

In step 3 you add the probabilities. While the parser builds structures it must compute probabilities for these structures. This is done by multiplying the probabilities of the one or two right hand parts with the probability of a rule, whenever a rule is used to add inferences to the chart.

1.1 Step 1

Step 1 of the project corresponds to the modeling and training task. You will extract a Probabilistic Context Free Grammar (PCFG) from the treebank. Such a grammar consists of a set of Context-free rules. Such rules can be seen as depth-one subtree types that are seen in the treebank. The root node of such subtrees are the left hand side of the rule, the children the right hand side of the rules. Basically such rules say that the left hand side may be rewritten or extended into the right hand side. A PCFG is a generative model and describes how to build trees starting from a root/start symbol.

The probability rules of the PCFG have probabilities that are estimated by Maximum Likelihood Estimation. In this case this is as simple as computing conditional probabilities of rules using relative frequency (i.e. counting). This means that the probability of generating children ”B, C” from a node ”A” (rule $A \to B\ C$) is computed by dividing the occurrence count of $A \to B\ C$ in the corpus by the total amount of A’s in the corpus.

Some notes

- Implementation note: extracting rules from the treebank will require you to e.g. count brackets, write recursive functions or do something else. That is up to you, as long as it works and you can explain why it does.
- Regarding the data: the treebank has been binarized. This is a pre-processing step to facilitate parsing. Reason being that the CYK/CKY algorithm only works with binary rules. You can extract rules from this binarized treebank as if it were the original treebank, only in a later step will you have to debinarize the trees your parser produces in order to be compatible with the normal (non-binary)

1.2 Step 2 and 3

These steps correspond to the search problem. You have a certain grammar with parameter values (probabilities). Now you have a sentence, and you want to find the most likely parse for it. This is done by building parses for parts of the sentence of increasing size, using dynamic programing. Crucially, the assumptions of PCFGs with respect to context freeness (only rules of depth one) allows parsing to be done efficiently in $O(n^3)$ time in the first place.

Disclaimer : This is an informal description of the lab, meant to increase your understanding of the general problems and tasks at hand. It is by no means intended to replace or negate the Project description which is available under assignments.