Choosing a Parser Evaluation Metric

Woodley Packard

Introduction

Evaluation metrics are handy tools for guiding parser design decisions. However, the innate differences between commonly used evaluation metrics prompt the question:

What effect will my choice of guiding metric have on the parser I build?

I examined the influence of a variety of parser evaluation metrics on two common parser design tasks: **setting a regularization parameter** and **feature selection** for machine learning.

In this work, the parser being designed is a maximum entropy disambiguation model with quadratic regularization coupled to the English Resource Grammar. The resulting analyses consist of a syntax tree and an MRS logical form, each of which can also be interpreted as types of dependencies. With around 60 different feature sets and around 40 different regularization levels being considered, an evaluation metric is essential for selecting the optimal model out of the roughly 2400 combinations available.

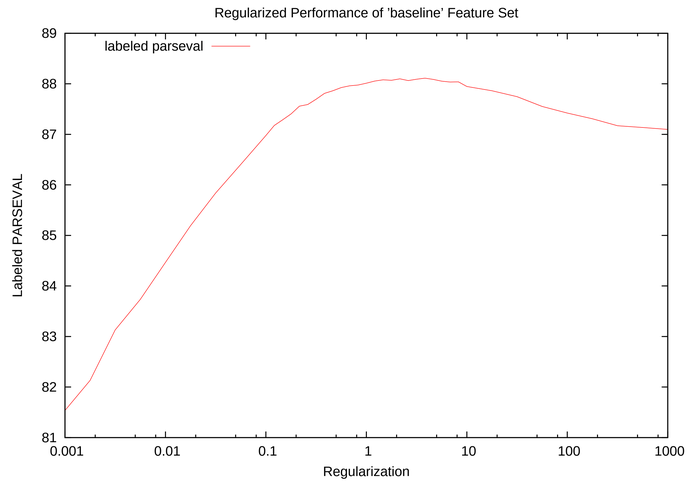
The following metrics are considered:

- Labeled PARSEVAL
- Unlabeled PARSEVAL
- Average Crossing Brackets
- Zero Crossing Brackets
- Leaf Ancestor
- Exact Tree Match
- Labeled Syntactic Dependencies
- Unlabeled Syntactic Dependencies
- Exact MRS Match
- Labeled MRS Dependencies
- Unlabeled MRS Dependencies
- Percentage of Correct Lexemes*
- Percentage of Correct POS*
- Exact Tree Node Count Match*

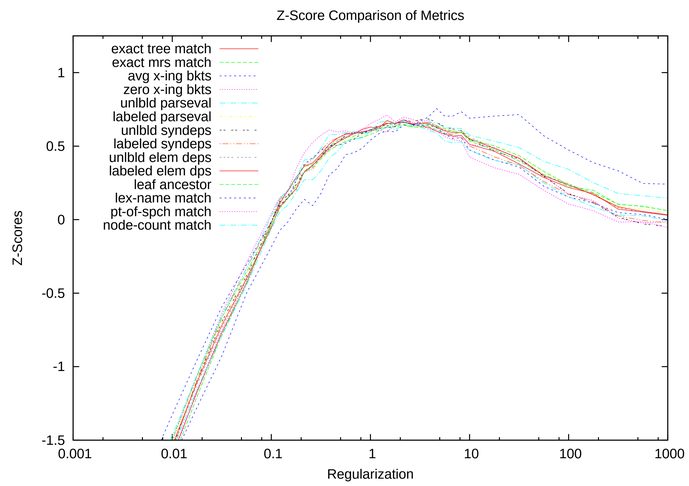
*These last three metrics are not commonly used in the literature; rather, they are purposefully blind to most of the linguistic information present in the analyses, with a view to seeing how much of a difference this task can detect between such "stupid" metrics and commonly accepted ones.

Regularization

The first graph below shows the values of the Labeled PARSEVAL metric as a function of the regularization parameter for the "baseline" feature set (which consists of PCFG-like depth-1 subtree configuration features). Using this metric as our parser design heuristic, we would most likely choose roughly $\rho = 2$ as the value of our regularization parameter.

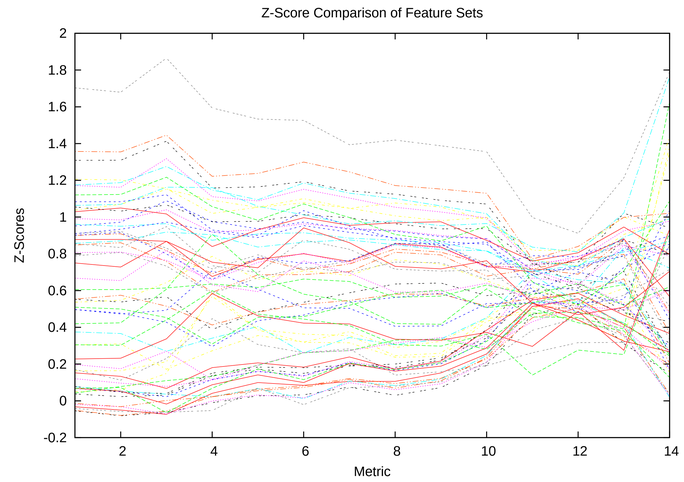


The next graph shows, for the same "baseline" feature set, the values of all 13 metrics as a function of the regularization parameter. For the sake of comparability, their values have been linearly rescaled on the vertical axis. With the notable exception of the "Average Crossing Brackets" metric, we can see that they all take their optimal values at approximately the same $\rho = 2$.



Feature Selection

In this graph, each colored line is an (optimally regularized) feature set, and the 13 horizontal positions are different metrics. As before we plot the rescaled values of the metrics. To the extent that the lines do not cross, the metrics agree with one another about the relative ranking of the different feature sets. The outlier metric "Average Crossing Brackets" is on the far right.



In the final graph, the colored lines are the metrics and the horizontal positions are the different feature sets, sorted by PARSEVAL score. The high correlation between most of the different metrics is evident. "Average Crossing Brackets" and the three "stupid" metrics are apparent outliers in this view.