Language Learning: Beyond Thunderdome

Christopher D. Manning
Computer Science Department
Stanford University
Stanford, CA 94305-9040
manning@cs.stanford.edu

Remember: no matter where you go, there you are.

The eight years from 1988 to 1996 saw the introduction and soon widespread prevalence of probabilistic generative models in NLP. Probabilities were the answer to learning, robustness and disambiguation, and we were all Bayesians, if commonly in a fairly shallow way. The eight years from 1996 to 2004 saw the rise to preeminence of discriminative models. Soon we were all either using SVMs or (in a few cases like myself) arguing that other discriminative techniques were equally as good: the sources of insight were margins and loss functions.

What might the next eight years hold? There will doubtless be many more variants of SVMs deployed, but it seems much less likely to me that major progress will come from new learning methods. NLP pretty much already uses what is known, and commonly the difference between one kernel or prior and another is small indeed. If we are waiting for better two class classifiers to push the performance of NLP systems into new realms, then we may be waiting a very long time. What other opportunities are there?

One answer is to rely on more data, and this answer has been rather fashionable lately. Indeed, it has been known for a while now that "There's no data like more data". One cannot argue with the efficacy of this solution if you are dealing with surface visible properties of a language with ample online text, and dealing with a standard problem over a stationary data set. Or if you have so much money that you can compensate for lacks from any of those directions. But I do not think this approach will work for most of us.

Something that has almost snuck up upon the field is that with modern discriminative approaches and the corresponding widely available software, anyone with modest training can deploy state of the art classification methods. What then determines the better systems? The features that they use. As a result, we need more linguists back in the field (albeit ones with training in empirical, quantitative methods, who are still in short supply, especially in North America). This viewpoint is still somewhat unfashionable, but I think it will increasingly be seen to be correct. If you look through the results of recent competitive evaluations, such as the various CoNLL Shared Task evaluations, many of the groups are using similar or the same machine learning methods. The often substantial differences between the systems is mainly in the features employed. In the context of language, doing "feature engineering" is otherwise known as doing linguistics. A distinctive aspect of language processing problems is that the space of interesting and useful features that one could extract is usually effectively unbounded. All one needs is enough linguistic insight and time to build those features (and enough data to estimate them effectively).

A second direction of the field is a renewed intererst in the deeper problems of NLP: semantics, pragmatic interpretation, and discourse. For both this issue and the previous one, issues of representation become central. At deeper levels of processing, there is less agreement on representations, and less understanding of what are effective representations for language learning. Much of our recent work in NLP has shown the importance and effectiveness of good representations for both unsupervised and supervised natural language learning problems. Working with good representations will be even more important for deeper NLP problems, and will see a revival of rich linguistic representations like in the 1980s.

Finally, a third direction (and perhaps the most productive area for new types of machine learning research) is to build systems that work effectively from *less* data. Whether trying to build a text classifier that can classify email into a folder based on only two examples, porting your work to a different Arabic dialect, or wanting to incorporate context into parsing and semantic interpretation, the challenge is how to build systems that learn from just a little data. This is also the cognitive science challenge of tackling the phenomenon of one-shot learning, and it requires some different thinking from that of relying on large hand-labeled data sets.