On Statistical Methods in Natural Language Processing

Joakim Nivre

1 Introduction

What is a statistical method and how can it be used in natural language processing (NLP)? In this paper, we will try to throw some light on this question by examining the different ways in which NLP methods deserve to be called "statistical", an exercise that will hopefully throw some light also on methods that do not deserve to be so called.

2 NLP: Problems, Models and Methods

NLP problems have to do with natural language input and output. Here are a few typical and uncontroversial examples of NLP problems:

- Part-of-speech tagging: Annotating natural language sentences or texts for parts-of-speech.
- Natural language generation: Producing natural language sentences or texts from non-linguistic representations.
- Machine translation: Translating sentences or texts in a source language to sentences or texts in a target language.

In part-of-speech tagging we have natural language input, in generation we have natural language output, and in translation we have both input and output in natural language.

If our aim is to build effective components for computational systems, then we must develop *algorithms* for solving these problems. However, this is not always possible, simply because the problems are not well-defined enough. The way out of this dilemma is the same as in most other branches of science. Instead of attacking real world problems directly with all their messy details, we build mathematical models of reality and solve abstract problems within the models instead. Provided that the models are worth their salt, these solutions will provide adequate approximations for the real problems. An abstract problem Q is a binary relation on a set I of problem *instances* and a set S of problem *solutions*. The abstract problems that are relevant to NLP are those where either I or S (or both) are linguistic entities or representations of linguistic entities. More precisely, an NLP problem P can be modeled by an abstract problem Q if the instance set I is a subset of the set of permissible inputs to P and the solution set S is a subset of the set of possible solutions to P.

Online Proceedings of NODALIDA 2001

2.1 Application Methods

A method for solving an NLP problem P typically consists of two elements:

1. A mathematical model M defining an abstract problem Q that can be used to model P.
2. An algorithm A that effectively computes Q.

We will say that M and A together constitutes an *application method* for problem P with Q as the *model problem*. For example, let G be a context-free grammar intended to model the syntax of a natural language NL and let Q be the parsing problem for G. Then G together with, say, Earley's algorithm is an application method for syntactic analysis of NL with Q as the model problem.

2.2 Acquisition Methods

The term *acquisition method* will be used to refer to any procedure for constructing a mathematical model that can be used in an application method. For example, any procedure for developing a context-free grammar modeling a natural language or a hidden Markov model for part-of-speech tagging is an acquisition method in this sense. In the following, we will concentrate almost exclusively on acquisition methods that make use of machine learning techniques in order to induce models (or model parameters) from empirical data, specifically *corpus* data. An empirical and algorithmic acquisition method typically consists of two elements:

1. A parameterized mathematical model M_θ such that providing values for the parameters θ will yield a mathematical model M that can be used in an application method for some NLP problem P.
2. An algorithm A that effectively computes values for the parameters θ when given a sample of data from P.

If the data sample must contain both inputs and (correct) outputs from P, then A is said to be a *supervised* learning algorithm. If it is sufficient with a sample of inputs, we have an *unsupervised* learning algorithm.

2.3 Evaluation Methods

If acquisition and application methods were infallible, no other methods would be needed. In practice, however, there are many factors which may cause an NLP system to perform less than optimally and we therefore need methods for evaluating NLP systems. We will use the term *evaluation method* to refer to any procedure for evaluating NLP systems. However, the discussion will focus on extrinsic evaluation of systems in terms of their accuracy. For example, let P be an NLP problem, and let (M_1, A_1) and (M_2, A_2) be two different application methods for P. A common way of evaluating and comparing the accuracy of these two methods is to apply them to a representative sample of inputs from P and measure the accuracy of the outputs produced by the respective methods. A special case of this evaluation scheme is where $A_1 = A_2$ and the models M_1 and M_2 are the results of applying two different acquisition methods to the same parameterized model M_θ and training corpus C. In this case, it is primarily the acquisition methods that are being evaluated.

3 Statistical Models and Methods

For the purpose of this paper, we will say that a model or method is *statistical* (or *stochastic*) if it involves the concept of probability (or related notions such as entropy and mutual information) or if it uses concepts of statistical theory (such as statistical estimation and hypothesis testing).

4 Statistical Methods in NLP

4.1 Application Methods

Most examples of statistical application methods in the literature are methods that make use of a stochastic model, but where the algorithm applied to this model is entirely deterministic. Typically, the abstract model problem computed by the algorithm is an *optimization problem* which consists in maximizing the probability of the output given the input. Here are some examples:

- Language modeling for automatic speech recognition using smoothed n-grams to find the most probable string of words $w_1, \ldots, w_n$ out of a set of candidate strings compatible with the acoustic data [Jelinek 1990].
- Part-of-speech tagging using hidden Markov models to find the most probable tag sequence $t_1, \ldots t_n$ given a word sequence $w_1, \ldots w_n$ [Merialdo 1994].
- Syntactic parsing using probabilistic grammars to find the most probable parse tree T given a word sequence $w_1, \ldots, w_n$ (or tag sequence $t_1, \ldots, t_n$) [Stolcke 1995].
- Word sense disambiguation using Bayesian classifiers to find the most probable sense s for word w in context C [Gale *et al* 1992].
- Machine translation using probabilistic models to find the most probable target language sentence T for a given source language sentence S [Brown *et al* 1990a].

4.2 Acquisition Methods

Statistical acquisition methods are methods that rely on *statistical inference* to induce models (or model parameters) from empirical data, in particular corpus data. The model induced may or may not be a stochastic model, which means that there are as many variations in this area as there are different NLP models. We will therefore limit ourselves to a few representative examples:

- Supervised learning of HMM taggers [Merialdo 1994].
- Unsupervised learning of HMM taggers [Cutting *et al* 1992]
- Transformation-based learning [Brill 1995].
- Decision tree parsing [Magerman 1995].

4.3 Evaluation Methods

Evaluation of NLP systems can have different purposes and consider many different dimensions of a system. Consequently, there are a wide variety of methods that can be used for evaluation. Many of these methods involve empirical experiments or quasi-experiments in which the system is applied to a representative sample of data in order to provide quantitative measures of aspects such as efficiency, accuracy and robustness. These evaluation methods can make use of statistics in at least three different ways:

- Descriptive statistics is often used to derive measures such as accuracy rate, recall and precision.
- Statistical estimation may be used to derive confidence intervals for descriptive measures.
- Hypothesis testing may be used to test the significance of any differences found when comparing alternative methods.

5 Conclusion

In this paper, we have discussed three different kinds of methods that are relevant in natural language processing:

- An *application method* is used to solve an NLP problem P, usually by applying an algorithm A to a mathematical model M in order to solve an abstract problem Q approximating P.
- An *acquisition method* for an NLP problem P is used to construct a model M that can be used in an application method for P. Of special interest here are empirical and algorithmic acquisition methods that allow us to construct M from a parameterized model M_θ by applying an algorithm A to a representative sample of P.
- An *evaluation method* for an NLP problem P is used to evaluate application methods for P. Of special interest here are experimental (or empirical) evaluation methods that allow us to evaluate application methods by applying them to a representative sample of P.

We have argued that statistics, in the wide sense including both stochastic models and statistical theory, can play a role in all three kinds of methods and we have supplied numerous examples to substantiate this claim.[1] We have also tried to show that there are many ways in which statistical methods can be combined with traditional linguistic rules and representation, both in application methods and in acquisition methods.

[1] More examples will be given in the full paper.

References

[Brill 1995] Brill, E. (1995) Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4), 543–566.

[Brown *et al* 1990a] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R. and Rossin, P. (1990) A Statistical Approach to Machine Translation. *Computational Linguistics* 16(2), 79–85.

[Cutting *et al* 1992] Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. (1992). A Practical Part-of-speech Tagger. In *Third Conference on Applied Natural Language Processing*, ACL, 133–140.

[Gale *et al* 1992] Gale, W. A., Church, K. W. and Yarowsky, D. (1992) A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities* 26, 415–439.

[Jelinek 1990] Jelinek, F. (1990) Self-Organized Language Modeling for Speech Recognition. In Waibel, A. and Lee, K.-F. (eds) *Readings in Speech Recognition*, pp. 450–506. Los Altos, CA: Morgan Kaufman.

[Magerman 1995] Magerman, D. (1995) Statistical Decision-Tree Models for Parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 276–283.

[Merialdo 1994] Merialdo, B. (1994) Tagging English Text with a Probabilistic Model. *Computational Linguistics* 20(2), 155–172.

[Stolcke 1995] Stolcke, A. (1995) An Efficient Probabilistic Context-Free Parsing Algorithm That Computes Prefix Probabilities. *Computational Linguistics* 21(2), 165–202.