Part of Speech Tagging from a Logical Point of View

Torbjörn Lager¹ and Joakim Nivre²

¹ Uppsala University, Department of Linguistics,
Torbjorn.Lager@ling.uu.se
² Växjö University, School of Mathematics and Systems Engineering,
Joakim.Nivre@msi.vxu.se

Abstract. This paper presents logical reconstructions of four different methods for part of speech tagging: Finite State Intersection Grammar, HMM tagging, Brill tagging, and Constraint Grammar. Each reconstruction consists of a first-order logical theory and an inference relation that can be applied to the theory, in conjunction with a description of data, in order to solve the tagging problem. The reconstructed methods are compared along a number of dimensions including ontology, expressive power, mode of reasoning, uncertainty, underspecification, and robustness. It is argued that logical reconstruction of NLP methods in general can lead to a deeper understanding of the knowledge and reasoning involved, and of the ways in which different methods are related.

1 Introduction

Comparing different methods for solving a particular problem is a necessary and natural part of the development of any field of science or engineering. So also in the field of NLP, where, for example, newly proposed methods for part of speech tagging are compared to already existing ones.

Unfortunately, within the field of NLP, the use of different mathematical frameworks, or idiosyncratic formalisms and notations, tends to make such activities hard. It has led to a situation where different solutions to a problem are usually compared only on the level of performance, rather than on the level of knowledge representation and inference, and where attempts to combine different methods often fail to take advantage of their respective strengths.

In an attempt to find a remedy for this situation, we want to explore ways to reconstruct different NLP methods within a single framework. We choose to work with first order predicate logic (FOPL), often considered the lingua franca of knowledge representation. In this paper, we reconstruct four methods for part of speech tagging: Finite State Intersection Grammar, HMM tagging, Brill tagging, and Constraint Grammar.

The common framework will allow us to compare these four tagging methods with respect to issues such as the following: What kind of inference engines are these part of speech taggers? What kind of reasoning do they perform? What
kind of knowledge do they exploit and how can the required knowledge be represented in first-order predicate logic? In this way, we hope to contribute to a better understanding of these particular methods and, thereby, also to demonstrate the usefulness of the methodology of logical reconstruction in general. (Further motivation for this kind of study can be found in section 2.3 below.)

2 Background

2.1 Reconstruction of Part of Speech Tagging

What does it mean to reconstruct a tagging method as a logical theory? For each method $M$, the reconstruction consists of two parts:

1. Specifying a FOPL theory $T_M$, representing the knowledge that the method uses in order to tag a sequence of words.
2. Specifying an inference relation $I_M$, such that the use of the inference relation for the representation will yield a solution to the tagging problem.

In addition, we need some way of representing sequences of words and their analyses. Let $\text{yield}(s)$ denote the description of a sequence of words $s$, in the form of a set (or, equivalently, a conjunction) of ground, atomic sentences. For example, $\text{yield}('the can smells')$ is:

\[
\text{word}(1, \text{the}). \text{word}(2, \text{can}). \text{word}(3, \text{smells}).
\]

Let $\text{analysis}(s)$ denote the assignment of parts of speech to positions in $s$, again in the form of a set of ground, atomic sentences. For example, we assume that $\text{analysis}('the can smells')$ is:

\[
\text{pos}(1, \text{dt}). \text{pos}(2, \text{nn}). \text{pos}(3, \text{vb}).
\]

Throughout the paper, we will use the simple example of tagging the string the can smells, which has the advantage of not requiring too much space. We will assume that the correct analysis is the one given above.

Finally, we note that, even if we adopt FOPL as our common framework, some of our reconstructions will only make use of a subset of the full system (e.g. Horn clause logic). We will also apply different inference strategies for different theories. However, this does not alter the fact that the methods are reconstructed within the same framework, although it may say something about the expressive power or inference mechanisms required for reconstructing different methods.

2.2 Ontology

A commitment to FOPL does not carry with it a commitment to a particular ontology. Before we continue, we therefore need to say something about our conceptualization of the objects, properties and relations in our domain. For each of the methods we will talk about three kinds of things: