# Constrained Atomic Term: Widening the Reach of Rule Templates in Transformation Based Learning

Cícero Nogueira dos Santos[1] and Claudia Oliveira[2]

[1] Departamento de Informática,
Pontifícia Universidade Católica, Rio de Janeiro, Brazil
nogueira@inf.puc-rio.br
[2] Departamento de Engenharia de Sistemas,
Instituto Militar de Engenharia, Rio de Janeiro, Brazil
cmaria@de9.ime.eb.br

**Abstract.** Within the framework of Transformation Based Learning (TBL), the rule template is one of the most important elements in the learning process. This paper presents a new model for TBL templates, in which the basic unit, denominated here as an atomic term (AT), encodes a variable sized window and a test that precedes the capture of a feature's value. A case study of Portuguese NP identification is described and the experimental results obtained are presented.

## 1 Introduction

During the last decade Machine Learning (ML) has proven to be a very powerful tool to enable linguistic tasks which would otherwise require an unfeasible amount of time and human resources. ML has been applied to central Natural Language Processing (NLP) problems such as: part-of-speech tagging, word-sense disambiguation, shallow parsing and prepositional phrase attachment ambiguity resolution. The most used ML techniques for these types of learning tasks are Hidden Markov Models, Maximum Entropy Models, Support Vector Machines, Memory Based Learning and Transformation Based Learning (TBL).

Within the TBL framework, the rule template is one of the most important elements in the learning process. This paper presents a new model for TBL templates, in which the basic unit, denominated here as an atomic term (AT), encodes a variable sized window and a test that precedes the capture of a feature's value. The use of this type of AT assumes that an item's feature $X$ should only be captured if another feature $Y$ of the same item complies with a certain test condition.

The research work was initiated as an attempt to build a noun phrase (NP) chunker using TBL for Portuguese. In the initial stages, the framework set up by Ramshaw and Marcus in [1] was used, including the concept of base NP and the rule templates. The results obtained were considerably inferior, which lead

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 622–633, 2005.

to the investigation of the classification errors produced: a large proportion of them were preposition-related mis-taggings.

Indeed, NP identification in Portuguese involves prepositional phrase attachment as a major sub-task, which in turn requires greater flexibility in the manipulation of the context being observed during the learning process. The idea behind the extension in the TBL template model is to provide such flexibility, by enabling long distance dependencies to be learned.

The remainder of this paper is organized as follows: in Sect. 2 we present a brief overview of TBL; in Sect. 3 the concept of constrained atomic term is introduced; in Sect. 4 some details of the implementation of this new type of template component are shown; and Sect. 5 presents a case study of Portuguese NP identification, describing the experiments that were carried out and the results obtained.

## 2 Transformation Based Learning

Transformation Based error-driven Learning (TBL) is a very successful symbolic machine learning method, introduced by Eric Brill in 1992. It has since been used for several important linguistic tasks, such as part-of-speech (POS) tagging [2], parsing, prepositional phrase attachment [3] and phrase chunking [1,4,5], having achieved state-of-the-art performance in many of them.

The central idea of the TBL algorithm is to generate an ordered list of rules, which will correct tagging mistakes in the corpus, which have been produced by an initial guess. The application determines which linguistic feature will be learned, and this feature will be represented by a tag-set (in the case of POS, for instance, the POS tags). The requirements of the algorithm are:

- two instances of a corpus, one that has been correctly annotated with the feature's tag-set, and another that remains un-annotated;
- an initial tagger, the baseline system, which will tag the un-annotated corpus by trying to guess the correct classification of each token, based on the annotated corpus's statistics; and
- a set of rule templates, which are meant to capture the relevant feature combinations, in the neighbourhood of a token, which would determine the tag of that token. Concrete rules are acquired by instantiation of this predefined set of template rules.

The learning algorithm is a mistake-driven greedy procedure that iteratively acquires a set of transformation rules. Figure 1 shows the rule generation process in TBL.

Learning starts with the initial guess classification of the un-annotated training corpus by the baseline system. The resulting classification is compared with the correct one and, whenever a classification error is found, all the rules that can correct it are generated by instantiating the templates with the current token's context. Normally, a new rule will correct tagging errors, but will also generate some other errors by changing correctly tagged tokens. Therefore, after