High Performance Part-of-Speech Tagging of Bulgarian

Veselka Doychinova and Stoyan Mihov

Institute for Parallel Processing
Bulgarian Academy of Sciences
stoyan@lml.bas.bg

Abstract. This paper presents an accurate and highly efficient rule-based part-of-speech tagger for Bulgarian. All four stages – tokenization, dictionary application, unknown words guessing and contextual part-of-speech disambiguation – are implemented as a pipeline of a couple deterministic finite state machines and transducers. We present a description of the Bulgarian ambiguity classes and a detailed evaluation and error analysis of our tagger. The overall precision of the tagger is over 98.4% for full disambiguation and the processing speed is over 34K words/sec on a personal computer. The same methodology has been applied for English as well. The presented realization conforms to the specific demands of the semantic web.¹

1 Introduction

Part-of-speech tagging has important applications to many areas of computational linguistics including syntax analysis, corpus linguistics, grammar checking, text-to-speech generation etc. The part-of-speech tagger is an essential resource for many semantic web applications like information extraction and knowledge acquisition. Recent efforts for providing a semantic-based multilingual infrastructure of the world wide web require new formal models and methods for language engineering. For the specific needs of the semantic web, in addition to the preciseness, a part-of-speech tagger has to provide the following features:

– High performance – crucial in respect to the vast amount of information presented on the web;
– XML and Unicode compliance;
– Technology applicable to other languages.

Our solution addresses all of the above mentioned problems.

For English and some other languages there are various accurate part-of-speech taggers based on different methodologies [3–5, 15]. Steven Abney gives a survey on the main methods in [1]. Most of the known approaches do not provide high performance.

¹ This work was funded by a grant from VolkswagenStiftung.
Emmanuel Roche and Yves Schabes introduce in [11] a deterministic tagger which has significantly better performance than the other systems [2, 4, 5]. The authors compose the contextual rules from the Brill tagger into a sequential transducer. Tokenizer, dictionary and guesser are implemented using other techniques. We extend this approach by:

- providing an uniform technique for the representation and utilization of the tokenizer, dictionary, guesser and the contextual rules;
- application of bimachines for supporting rewriting rules with no restrictions on the length of the left and right context;
- support of lexical and morphological constraints in the contextual rules;
- processing the text without modifying it – by inserting XML tags in the original.

Reports for Bulgarian part-of-speech taggers are given by Hristo Tanev and Ruslan Mitkov in [14] and by Kiril Simov and Petya Osenova in [13]. The development of the first tagger was performed without relying on a large tagged corpus. The reported resulting precision of it is 95% for 95% recall. For the development of the second tagger a corpus consisting of 2500 sentences was used. The achieved precision is 95.17%.

Our part-of-speech tagger is developed using a large manually tagged corpus kindly provided by Svetla Koeva from the Institute for Bulgarian Language. The corpus consists of 197K tokens (over 150K words) randomly extracted from an 1M words Bulgarian corpus, structured along the standards of the Brown University corpus. This is a running text of edited Bulgarian prose divided into 500 samples of over 2000 words each, representing a wide range of styles and varieties of prose.

In our research initially we tried to train the Brill tagger [2] for Bulgarian. We used a 160K tokens tagged corpus for training. The results were disappointing – although the tagger performed very well on the training corpus – 98.7%, on a unseen 40K tokens corpus it performed poorly – 95.5%. We suppose that the reason for the low accuracy on Bulgarian texts is a consequence of the inflectional nature of the Bulgarian morphology leading to a large amount of wordforms and the free word order in the Bulgarian sentence.

We present a rule-based approach [3, 15] leading to 98.4% precision implemented by finite state devices [8, 11, 12]. The first step to solving the ambiguity is tokenizing the text. For the second step we use a 75K base forms grammatical dictionary [9], which assigns to each known word its most-probable tag and the set of all possible tags (its ambiguity class). If a word is not in the dictionary, a guesser is consulted in the third step. Finally, 148 manually constructed contextual rules are applied on the text to handle the ambiguities.

We present the ambiguity classes and the tagset in the next section. Afterwards we proceed with the tokenizer, lexicon and guesser description. Section 4 describes the contextual rules. The evaluation results are presented in Section 5. Implementation details are given in Section 6. Finally the conclusion presents some general comments and directions for further work.