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 bimachines 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.

2 Restricted Tagset and Ambiguity Classes

The problem of correct part-of-speech tagging is a consequence of the implicit ambiguity of the words in respect of their morphology, syntactic function and meaning. For example, in the sentence: *Напразно тичат инкасаторите в бели престилки - навсякъде всичко е наред.*, the word *бели* could be tagged as five different forms: a plural adjective, a plural noun, a present singular verb, a past singular verb and an imperative verb. And whereas in this case the solution is quite straightforward, for the correct tagging of the verbs *завърши* and *стане* in the sentence *Нямаше търпение да завърши висше образование, да стане козметичен хирург.* a more complex analysis is required.

It is hard to use all 946 different morphological tags from the dictionary in our rule-based method. Moreover, some distinctions in the given text cannot be determined reliably without full syntactic and semantic analysis. For example the dative and accusative cases of pronouns in short form are hardly distinguishable. Hence, the tagset should be quite restricted but it should still provide the deserved functionality. In our tagset only morphological characteristics, which are considered essential for part-of-speech tagging are taken into account. For example, past singular verbs have the same tag, although only second and third person forms are ambiguous. All forms, which differ in number are divided. Verbs are also classified according to whether they are in present or past tense. Pronouns are grouped according to gender and short forms. The final tagset has 40 tags given in Table 1.

Nmfs	feminine and masculine singular nouns, singular neutral nouns w. definite article	PLs	singular personal pronoun
		PLp	plural personal pronouns
Nns0	singular neutral nouns without definite article	PL	personal reflexive pronoun
		PLz	short forms of personal reflexive pronoun
Np	plural nouns	PLsz	short forms of singular personal pronouns
Nv	vocative nouns	PLpz	short forms of plural personal pronouns
Nc	countable masculine nouns	PPmfs	singular feminine and masculine possessive pronoun
Amfs	singular feminine and masculine adjectives, singular neutral adjectives with definite article	PPns	singular neutral possessive pronouns
		PPp	plural possessive pronouns
		PPz	short forms of possessive pronouns
Ans0	singular neutral adjectives without definite article	POmfs	demonstrative, relative, indefinite, collective, negative and interrogative singular masculine and feminine pronouns
Ap	plural adjectives		
VRs	singular present tense verbs		
VRp	plural present tense verbs	POns	demonstrative, relative, indefinite, collective, negative and interrogative singular neutral pronouns
VPs	singular past tense verbs		
VPp	plural past tense verbs		
VXs	singular present tense active voice participle	POp	demonstrative, relative, indefinite, collective, negative, interrogative plural pronouns
VXp	plural present tense active voice participle		
		ADV	adverbs
VYs	singular past tense active voice participle	NUMO	ordinal numerals
VYp	plural past tense active voice participle	NUMC	cardinal numerals
VQs	singular past tense passive voice participle	CONJ	conjunctions
		PREP	prepositions
VQp	plural past tense passive voice participle	PUNCT	punctuation
VZ	adverbial participle	MISC	particles, interjection, other miscellaneous tokens
VI	imperative verbs		

Table 1. Bulgarian part-of-speech tagset.

In respect of the tagset the words in our dictionary are divided into 259 ambiguity classes. 47 of them consist of more than 100 words. As shown on Table 2 the largest class consists of present and past verbs, followed by the one of masculine and countable nouns. The first column presents the number of entries in the class and the second one shows the occurrence frequency on 1000 words derived from a 20M words untagged corpus. Statistics on the right shows how many times words from the class have appeared with a corresponding tag in our manually tagged representative corpus.

Entries	Freq.	Class	Examples	Realization
12967	32.30	VPs/VRs	абдикира|абонира|абортира|абсолютизира	689/4162
5376	17.10	Nc/Nmfs	абажура|абзаца|аблатива|абонамента	274/1754
4575	4.69	Amfs/VQs	адаптираната|адаптирана|адаптираният|адаптирания	571/148
3192	2.79	Amfs/Ap	абисински|абитуриентски|аборигенски|абсолвентски	222/228
3165	10.05	VI/VPs/VRs	агни|бави|безбожничи|безделничи	15/430/636
2182	18.38	ADV/Ans0	абсолютно|абстрактно|абсурдно|авантюристично	2122/1011
2141	1.37	Np/VQs	абортирания|авансирания|анонсирания|аранжирания	260/16
1881	0.53	VI/VRp	ахнете|барнете|бафнете|белнете	17/45
1521	2.97	Ap/VQp	адаптираните|адаптирани|активираните|активирани	343/87
594	0.69	Ans0/VQs	адаптирано|активирано|арестувано|асфалтирано	57/50
554	0.56	VQs/VRp	барнат|близнат|блъвнат|блъснат	20/61
408	1.11	Amfs/VYs	аглутиниращата|аглутиниращa|аглутиниращият	142/1
394	0.46	VI/VRs	бди|бележи|блести|бръмчи	1/109
387	0.69	Amfs/VXs	буренясалата|буренясала|буренясалият|буренясалия	46/31
350	2.52	Amfs/Nmfs	абаджийска|абхазката|абхазка|авантюристката	115/292
273	1.71	Np/VRp	багрите|барабаните|бедите|благословите	196/24
260	6.42	Np/VI/VPs/VRs	багри|барабани|беди|благослови	475/4/117/276
219	0.41	Ap/Np	абхазките|авантюристките|адвентистките	26/44
168	0.60	ADV/Ans0/VQs	вдъхновено|вперено|втренчено|вцепенено	50/43/8
136	0.43	Ap/VYp	аглутиниращите|аглутиниращи|благоденстващите	69/0
129	0.68	Np/VXs	белилата|белила|бесилата|бесила	35/47
126	0.50	Ap/VXp	буренясалите|буренясали|велите|вели	30/10
111	1.24	Nmfs/VXs	белилото|бесилото|билото|бил	43/72
106	0.51	Amfs/Np/VQs	активирания|вдъхновения|възвишения|въздържания	4/88/0

Table 2. Main Bulgarian ambiguity classes.

3 Tokenizer, Dictionary and Guesser

3.1 Tokenizer

We built a sophisticated tokenizer, which marks and categorizes tokens as numeral expressions, dates, hours, abbreviations, URLs, items, punctuation, words, abbreviations etc. Words in Latin and Cyrillic are differently marked when capitalized or upper case. Punctuation is classified according to the type and the function of the sign. Sentence boundaries are recognized as well.

The tokenizer is implemented as a composition of 53 contextual rewriting rules. Some of the rules require that the length of the right context is unlimited. All tokenization rules are composed into 4 bimachines.

3.2 Dictionary

The dictionary assigns to each dictionary word in the text its ambiguity class and its initial tag. Usually the initial tag is the most probable tag for the ambiguity class. For example, the class which consist of adverbs and neutral adjectives gets adverb as a most probable tag, because in the corpus these words are adverbs 2122 times and adjectives 1011 times. There are exceptions from this principle in some cases. For example in the class singular present tense verb / singular past tense verb, the perfective verbs are initially tagged as past verbs and the imperfective and dual aspect verbs are tagged as present verbs, because usually the perfective verbs are used in present tense with *да* or *ще* and in this way there is a more reliable context information to transform the past verb tag to present. In the sentences:
Тръгна пеша и стигна до стадиона за около час. – *тръгна* and *стигна* are past verbs.
Кога да тръгна, за да стигна навреме? – *тръгна* and *стигна* are present verbs.

	tokens	share
All tokens:	101207	
Dictionary coverage	97760	96.59%
Correct initial tag from dictionary	91128	90.04%
Wrong initial tag from dictionary	6632	6.55%

Table 3. Result of dictionary application.

The dictionary contains about 1M wordforms. Since each wordform can occur in lower case, upper case or capitalized, the dictionary contains 3M strings. It is implemented as one (big) rewriting rule and represented by a sequential transducer using the construction method presented in [10].

The overall dictionary performance is given in Table 3.

3.3 Guesser

The words that are not in the dictionary are handled by the guesser. The constructed rules are analysing the suffix for guessing the word's morphology and for assigning the initial part-of-speech tag and the ambiguity class. For example, the word *суопнахме*, which was found in a computer text is an english word with a Bulgarian suffix for plural third person verb. Hence, it will be correctly tagged by the guesser. Words with the suffix *ирания*, receive a tag for ambiguity class, which allows them to be tagged as a masculine singular adjective, a plural noun or a singular passive participle. For some words there is only one option. Such are the words with suffix *аемо*, *тел* and others.

The capitalized words in the middle of the sentence, which in most cases are unknown proper names are tagged as singular nouns. The same applies to Latin text.

After the guesser is applied to the text the precision reaches 93.28%. The total number of the words in our 100K tokens corpus that have not been found in the dictionary is 3447. For all of them the guesser has suggested its initial tag, which is wrong in only 4.72% of the cases treated by the guesser. Table 4 presents the exact numbers.

	tokens	share
All tokens:	101207	
Not in Dictionary	3447	3.41%
Correct initial tag from dictionary	91128	90.04%
Correct initial tag from Dictionary and Guesser	94412	93.29%
Wrong initial tag from guesser	163	0.16%

Table 4. Guesser performance.

The guesser is implemented as a composition of 73 rewrite rules all compiled into a single bimachine.

4 Contextual Disambiguation Rules

The part-of-speech ambiguity ratio for our corpus is 1.51 tags/word, which means that in average every second word is ambiguous. For solving the ambiguities we apply 148 contextual rules, which can utilize part-of-speech, lexical or dictionary information on the context. Below some exemplary rules are given:

```
change ADV to Amfs if next word is Nmfs
change Np to VRs if previous 1 or 2 word is "ще/MISC"
change Nmfs to VPs if word is perfective verb and previous word is
"се/MISC"
```

A rule is applied if and only if the context is fulfilled and the ambiguity class of the targeted word allows the new tag.

Experiments have shown that the rule order significantly influences the results. Because of that the rules are applied in a certain order to handle specific ambiguities. For example the rules that change a noun into past verb precede the rules that change a past verb into present verb.

All 148 contextual rules are composed into 2 bimachines which we apply in a pipeline.

5 Evaluation Results and Error Analysis

5.1 Preciseness Results

After all the rules are applied to the training corpus the overall accuracy reached 98.44%. Totally 5610 tags were changed by the context rules. For the unseen test corpus the result is slightly worse - the accuracy is 98.29% and the context rules changed 5297 tags. Table 5 presents the details.

	Training corpus		Test corpus	
	tokens	share	tokens	share
All tokens:	101207		96146	
Correct initial tags from dictionary & guesser	94412	93.29%	89599	93.19%
Correct tags by Tagger	99630	98.44%	94505	98.29%
Tags changed by context rules	5610	5.54%	5297	5.51%

Table 5. Tagger results on training and test corpus.

At that point a few new rules were developed to handle specific cases that were found mainly in the test corpus. We have got a slight increase of the preciseness after applying the additional rules:

	Additional rules	Overall result
Training corpus	+0.15%	98.59%
Test corpus	+0.16%	98.46%

5.2 Error Analysis

After the application of the tagger on both corpora 2900 words received a wrong tag. 2117 of them were not changed by the contextual rules and 783 cases handled by the context rules were not tagged correctly. From the 783 cases 452 times the tagger changed a correct initial tag given by the dictionary or guesser and 331 times both – initial and context rules tags were wrong.

The largest group of wrong tags is adverb / neutral adjective. Most of those errors were made because in many cases the syntactic context of both adverbs and adjectives is similar. For example when the following word is a neutral noun we have no reliable context for distinguishing the two cases.
*Закъснението на самолета причинява * само/Ans0 известно забавяне на програмата.*
*Като статистическа величина БВП отчита стойността на крайните стоки и услуги за определен период, * обикновено/Ans0 тримесечие или година.*

The same applies to dual aspect verbs. Our statistics showed that in most cases the dual aspect verbs are in present so the dictionary initially gives them

the relevant tag. Certainly, some cases were incorrectly tagged using this statistic approach.
*Така лидерът на СДС Екатерина Михайлова * коментира/VRs изявлението на Първанов по повод вота на недоверие на правителството.*

The ambiguity, which remains in the singular adjective / singular noun class is difficult in the case of proper nouns. Our context rules cannot completely solve this ambiguity at that point. For solving it a thorough semantic analysis is necessary.
Пристигнахме късно от Света/Amfs Гора.
** Дарън/Amfs Ралфс от Щатите поднесе сензацията на Световното първенство по ски алпийски дисциплини.*

Short forms of pronouns are known to be very movable in the Bulgarian sentence. Generally, they are likely to be personal pronouns in contiguity with a verb and possessive pronouns in contiguity with a noun. But there are many difficult cases where the short form is between a noun and a verb.
Изображението се е запазило в душата ми/PPz.
*Мама * ми/PPz е обещала да отидем в ботаническата градина.*

Most errors were made in cases where semantic analysis is required for disambiguation.
*Целта беше да * й/PLsz грабне вниманието.*
*Следователно с Волкерс и с романа му "Нещо сладко" осъществяваме едно закъсняло запознанство с една практически неизвестна * ни/PLpz литература.*

The main ambiguity classes, which are wrongly tagged are given in Table 6.

325 words	Ans0/ADV	(бързо, обикновено)
199 words	PLz/MISC	(се, си)
156 words	VRs/VPs	(води, каза)
143 words	Amfs/Nmfs	(бос, свят, лек, имена)
134 words	PPz/PLsz	(му,ми,й,ти)
130 words	Np/Nmfs	(листа, крака, господа)
101 words	PPz/MISC	(си)
80 words	PPz/PLpz	(им, ни)
69 words	Np/VRs	(работи, води, мисли)
50 words	Np/VPs	(обяви, отговори, уреди)

Table 6. Main errors by the tagger.

The 10 classes in Table 6 represent 43.10% of all errors. The rest of the wrongly tagged words are distributed into 110 ambiguity classes.

6 Implementation Details

As already mentioned the whole tagger is implemented as a composition of 275 rewriting rules. For the construction of the rules we used the methods presented

in [8, 7, 6]. The rules are then composed and represented by bimachines by the techniques presented in [12]. For the construction of the sequential transducer for the rewriting dictionary rule we developed a new method presented in [10]. In this way the tagger consists of 7 bimachines and 1 sequential transducer applied in a pipeline.

The text is transduced by each of the finite state devices deterministically. After each step the text is enriched with additional information which is used by the following steps. At the end only part-of-speech tags are left. The tagger is capable of processing of Unicode and the markup is down by inserting XML tags without modifying the input text.

The size of the tagger is 82002KB including all data and the executable. We measured the processing speed on a Pentium III 1GHz computer running Linux. The performance is measured for the whole process of tagging, including the disk operations needed for reading in and writing out the text file. The performance is **34723 words/sec**.

The processing speed is proportional to the number of devices in the pipeline. Theoretically we could compose all the rules into one single bimachine. In that case we would have about 8 times faster processing. The problem is that the size of this device would be unacceptable for the current technology. Nevertheless we could trade off speed for size if needed. For example by representing our tagger by 208 bimachines the size is 21123KB by a processing speed of 1550 words/sec.

The same technique was tested by realizing an English tagger. We constructed the bimachine pipeline by composing the rules from the Brill system [2] trained on the Penn Treebank Tagged Wall Street Journal Corpus. In addition to the technique presented in [11] we implemented the extended version of Brill's system, which is supplemented by a guesser and the contextual rules are able to make use of lexical relationships. All 148 guesser rules were composed into 3 bimachines. The 70K words dictionary was implemented as a sequential transducer. The 288 contextual rules were composed into 13 bimachines. The size of our implementation is 56340KB. As expected our system delivered identical result compared to the result of Brill's system and performed with a processing speed of 16223 words/sec.

7 Conclusion

We presented a Bulgarian part-of-speech tagger with 98.4% precision. The analysis of the ambiguity classes showed that the problem of part-of-speech disambiguation for Bulgarian is as complex as the one for English, but the ambiguities are consequences of other language phenomena. We showed that the cases where the tagger assigns a wrong tag require a more complex syntactic or semantic analysis. Hence, further improvement of the tagger preciseness would require much more efforts.

The methodology presented in the paper provides a very efficient, language independent implementation of all stages of the process. We successfully applied the same technique with the rules of the Brill tagger for English.

Further we plan to exploit the presented methodology for other languages and other applications like information extraction, grammar checking and prosody and phonetic description generation.

Acknowledgments: We are grateful to Svetla Koeva for the valuable discussions during the beginning of our research.

References

1. Steven P. Abney, Part-of-Speech Tagging and Partial Parsing, In Ken Church and Steve Young and Gerrit Bloothooft, editor, Corpus-Based Methods in Language and Speech Kluwer Academic Publishers, Dordrecht, 1996.
2. Eric Brill, Some advances in rule-based part of speech tagging, Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa., 1994.
3. Jean-Pierre Chanod, Pasi Tapanainen, Tagging French - comparing a statistical and a constraint-based method, Proceedings of Seventh Conference of the European Chapter of the Association for Computational Linguistics, 1995.
4. Kenneth Church, A stochastic parts program and noun phrase parser for unrestricted texts, Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas, 1988.
5. Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun, A practical part-of-speech tagger, Proceedings of Third Conference on Applied Natural Language Processing (ANLP-92), pages 133–140, 1992.
6. Hristo Ganchev, Stoyan Mihov and Klaus U. Schulz, One-Letter Automata: How to Reduce k Tapes to One. CIS-Bericht, Centrum für Informations- und Sprachverarbeitung, Universität Munchen, 2003.
7. Dale Gerdemann and Gertjan van Noord, Transducers from Rewrite Rules with Backreferences, Proceedings of EACL 99, Bergen Norway, 1999.
8. Ronald Kaplan and Martin Kay, Regular Models of Phonological Rule Systems, Computational Linguistics 20(3), 331-378, 1994
9. Svetla Koeva, Grammar Dictionary of the Bulgarian Language Description of the principles of organization of the linguistic data, Bulgarian language magazine, book 6, 1998.
10. Stoyan Mihov and Klaus U. Schulz, Efficient Dictionary-Based Text Rewriting using Sequential Transducers, CIS-Bericht, Centrum für Informations- und Sprachverarbeitung, Universität Munchen, 2004. To appear.
11. Emmanuel Roche, Yves Schabes, Deterministic Part-of-Speech Tagging with Finite-State Transducers, Computational Linguistics, Volume 21, Number 2, June 1995.
12. Emmanuel Roche and Yves Schabes, Introduction, Finite-State language processing, E Roche, Y Schabes (Eds), MIT Press, 1997.
13. Kiril Simov, Petya Osenova, A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian, Proceedings of the RANLP 2001 Conference, Tzigov Chark, Bulgaria, 5-7 September 2001.
14. Hristo Tanev; Ruslan Mitkov, Shallow Language Processing Architecture for Bulgarian, Proceedings of COLING 2002: The 17th International Conference on Computational Linguistics, 2002.
15. Atro Voutilainen, A syntax-based part-of-speech analyser , Proceedings of Seventh Conference of the European Chapter of the Association for Computational Linguistics, 1995.