PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 89 JUNE 2008 23–40

Combining Statistical and Rule-Based Approaches to Morphological Tagging of Czech Texts

Drahomíra “johanka” Spoustová

Abstract

This article is an extract of the PhD thesis (Spoustová, 2007) and it extends the article (Spoustová et al., 2007). Several hybrid disambiguation methods are described which combine the strength of hand-written disambiguation rules and statistical taggers. Three different statistical taggers (HMM, Maximum-Entropy and Averaged Perceptron) and a large set of hand-written rules are used in a tagging experiment using Prague Dependency Treebank. The results of the hybrid system are better than any other method tried for Czech tagging so far.

1. Introduction

Inflective languages pose a specific problem for tagging due to two phenomena: highly inflective nature (causing sparse data problem in any statistically based system), and free word order (causing fixed-context systems, such as n-gram HMMs, to be even less adequate than for English).

The average tagset contains about 1,000–2,000 distinct tags; the size of the set of possible and plausible tags can reach several thousands. There have been attempts at solving this problem for some of the highly inflective European languages, such as (Daelemans, 1996), (Erjavec, 1999) for Slovenian and (Hajič, 2000) for five Central and Eastern European languages.

Several taggers already exist for Czech, e.g. (Hajič et al., 2001b), (Smith, 2005), (Hajič et al., 2006) and (Votrubec, 2006). The last one reaches the best accuracy for Czech so far (95.12%). Hence no system has reached – in the absolute terms – a performance comparable to English tagging (such as (Ratnaparkhi, 1996)), which stands above 97%.

We are using the Prague Dependency Treebank (Hajič et al., 2006) (PDT) with about 1.8 million hand annotated tokens of Czech for training and testing. The tagging experiments in this paper all use the Czech morphological (pre)processor, which includes a guesser for “unknown” tokens and which is available from the PDT website (PDT Guide, 2006), to disam-

Please cite this article as: Drahomíra “johanka” Spoustová, Combining Statistical and Rule-Based Approaches to Morphological Tagging of Czech Texts. The Prague Bulletin of Mathematical Linguistics No. 89, 2008, 23–40.

	Name	Description
1	POS	Part of Speech
2	SUBPOS	Detailed POS
3	GENDER	Gender
4	NUMBER	Number
5	CASE	Case
6	POSSGENDER	Possessor's Gender
7	POSSNUMBER	Possessor's Number
8	PERSON	Person
9	TENSE	Tense
10	GRADE	Degree of comparison
11	NEGATION	Negation
12	VOICE	Voice
13	RESERVE1	Unused
14	RESERVE2	Unused
15	VAR	Variant

Table 1. Czech Morphology and the Positional Tags

biguate only among those tags which are morphologically plausible.

The meaning of the Czech tags (each tag has 15 positions) we are using is explained in Table 1. A detailed linguistic description of the individual positions can be found in the documentation for the PDT (Hajič et al., 2006).

2. Components of the hybrid system

2.1. The HMM tagger

The HMM tagger is based on the well known formula of HMM tagging:

$$\hat{T} = \arg\max_T P(T)P(W \mid T) \tag{1}$$

where

$$\begin{aligned} P(W|T) &\approx \textstyle\prod_{i=1}^{n} P(w_i \mid t_i, t_{i-1}) \\ P(T) &\approx \textstyle\prod_{i=1}^{n} P(t_i \mid t_{i-1}, t_{i-2}). \end{aligned} \tag{2}$$

The trigram probability $P(W \mid T)$ in formula 2 replaces (Hajič et al., 2001b) the common (and less accurate) bigram approach. We will use this tagger as a baseline system for further improvements.

Initially, we change the formula 1 by introducing a scaling mechanism[1]: $\hat{T} = \arg\max_T(\lambda_T * logP(T) + logP(W \mid T))$.

[1] The optimum value of the scaling parameter λ_T can be tuned using held-out data.

We tag the word sequence from right to left, i.e. we change the trigram probability $P(W \mid T)$ from formula 2 to $P(w_i \mid t_i, t_{i+1})$.

Both the output probability $P(w_i \mid t_i, t_{i+1})$ and the transition probability $P(T)$ suffer a lot due to the data sparseness problem. We introduce a component $P(ending_i \mid t_i, t_{i+1})$, where $ending$ consists of the last three characters of w_i. Also, we introduce another component $P(t^*_i \mid t^*_{i+1}, t^*_{i+2})$ based on a reduced tagset T^* that contains positions POS, GENDER, NUMBER and CASE only (chosen on linguistic grounds).

We upgrade all trigrams to fourgrams; the smoothing mechanism for fourgrams is history-based bucketing (Krbec, 2005).

The final fine-tuned HMM tagger thus uses all the enhancements and every component contains its scaling factor which has been computed using held-out data. The total error rate reduction is 13.98% relative on development data, measured against the baseline HMM tagger.

2.2. Morče

The Morče[2] tagger assumes some of the HMM properties at runtime, namely those that allow the Viterbi algorithm to be used to find the best tag sequence for a given text. However, the transition weights are not probabilities. They are estimated by an Averaged Perceptron described in (Collins, 2002). Averaged Perceptron works with features which describe the current tag and its context.

Features can be derived from any information we already have about the text. Every feature can be true or false in a given context, so we can regard current true features as a description of the current tag context.

For every feature, the Averaged Perceptron stores its weight coefficient, which is typically an integer number. The whole task of Averaged Perceptron is to sum all the coefficients of true features in a given context. The result is passed to the Viterbi algorithm as a transition weight for a given tag. Mathematically, we can rewrite it as:

$$w(C,T) = \sum_{i=1}^{n} \alpha_i . \phi_i(C,T) \tag{3}$$

where $w(C,T)$ is the transition weight for tag T in context C, n is number of features, α_i is the weight coefficient of i^{th} feature and $\phi(C,T)_i$ is evaluation of i^{th} feature for context C and tag T.

Weight coefficients (α) are estimated on training data, cf. (Votrubec, 2006). The training algorithm is very simple, therefore it can be quickly retrained and it gives a possibility to test many different sets of features (Votrubec, 2005). As a result, Morče gives the best accuracy from the standalone taggers.

[2] The name Morče stands for “MORfologie ČEštiny” (“morphology of Czech”).

2.3. The Feature-Based Tagger

The Feature-based tagger, taken also from the PDT (Hajič et al., 2006) distribution used in our experiments uses a general log-linear model in its basic formulation:

$$p_{AC}(y \mid x) = \frac{\exp(\sum_{i=1}^{n} \lambda_i f_i(y,x))}{Z(x)} \tag{4}$$

where $f_i(y,x)$ is a binary-valued feature of the event value being predicted and its context, λ_i is a weight of the feature f_i, and the $Z(x)$ is the natural normalization factor.

The weights λ_i are approximated by Maximum Likelihood (using the feature counts relative to all feature contexts found), reducing the model essentially to Naive Bayes. The approximation is necessary due to the millions of the possible features which make the usual entropy maximization infeasible. The model makes heavy use of single-category Ambiguity Classes (AC)[3], which (being independent on the tagger's intermediate decisions) can be included in both left and right contexts of the features.

2.4. The rule-based component

The approach to tagging (understood as a stand-alone task) using hand-written disambiguation rules has been proposed and implemented for the first time in the form of Constraint-Based Grammars (Karlsson, 1995). On a larger scale, this aproach was applied to English (Karlsson, 1995) and (Samuelsson, 1997), and French (Chanod, 1995). Also (Bick, 2000) uses manually written disambiguation rules for tagging Brazilian Portuguese, (Karlsson, 1985) and (Koskenniemi, 1990) for Finish and (Oflazer, 1997) reports the same for Turkish.

2.4.1. Overview

In the hybrid tagging system presented in this paper, the rule-based component is used to further reduce the ambiguity (the number of tags) of tokens in an input sentence, as output by the morphological processor (see Sect. 1). The core of the component is a hand-written *grammar* (set of rules).

Each rule represents a piece of knowledge of the language system (in particular, of Czech). The knowledge encoded in each rule is formally defined in two parts: a sequence of tokens that is searched for in the input sentence and the tags that can be deleted if the sequence of tokens is found.

The overall strategy of this "negative" grammar is to keep the highest recall possible (i.e. 100%) and to gradually improve precision. In other words, whenever a rule deletes a tag, it is (almost) 100% safe that the deleted tag is "incorrect" in the sentence, i.e. the tag cannot be present in any correct tagging of the sentence.

Such an (virtually) "error-free" grammar can partially disambiguate any input and prevent the subsequent taggers (stochastic, in our case) to assign tags that are "safely incorrect".

[3] If a token can be a N(oun), V(erb) or A(djective), its (major POS) Ambiguity Class is the value "ANV".

2.4.2. The rules

Formally, each rule consists of the description of the *context* (sequence of tokens with some special property), and the *action* to be performed given the context (which tags are to be discarded). The length of context is not limited by any constant; however, for practical purposes, the context cannot cross over sentence boundaries.

For example: in Czech, two finite verbs cannot appear within one clause. This fact can be used to define the following disambiguation rule:

- context: unambiguous finite verb, followed/preceded by a sequence of tokens containing neither a comma nor a coordinating conjunction, at either side of a word x ambiguous between a finite verb and another reading;
- action: delete the finite verb reading(s) at the word x.

It is obvious that no rule can contain knowledge of the whole language system. In particular, each rule is focused on at most a few special phenomena of the language. But whenever a rule deletes a tag from a sentence, the information about the sentence structure "increases". This can help other rules to be applied and to delete more and more tags.

For example, let's have an input sentence with two finite verbs within one clause, both of them ambiguous with some other (non-finite-verbal) tags. In this situation, the sample rule above cannot be applied. On the other hand, if some other rule exists in the grammar that can delete non-finite-verbal tags from one of the tokens, then the way for application of the sample rule is opened.

The rules operate in a loop in which (theoretically) all rules are applied again whenever a rule deletes a tag in the partially disambiguated sentence. Since deletion is a monotonic operation, the algorithm is guaranteed to terminate; effective implementation has also been found in (Květoň, 2006).

2.4.3. Grammar used in tests

The grammar is being developed since 2000 as a standalone module that performs Czech morphological disambiguation. There are two ways of rule development:

- the rules developed by syntactic introspection: such rules are subsequently verified on the corpus material, then implemented and the implemented rules are tested on a testing corpus;
- the rules are derived from the corpus by introspection and subsequently implemented.

In particular, the rules are not based on examination of errors of stochastic taggers.

The set of rules is (manually) divided into two (disjoint) reliability classes — *safe* rules (100% reliable rules) and *heuristics* (highly reliable rules, but obscure exceptions can be found). The safe rules reflect general syntactic regularities of Czech; for instance, no word form in the nominative case can follow an unambiguous preposition. The less reliable heuristic rules can be exemplified by those accounting for some special intricate relations of grammatical agreement in Czech.

The grammar consists of 1,727 safe rules and 504 heuristic rules. The system has been used in two ways:

- *safe rules only*: in this mode, safe rules are executed in the loop until some tags are being deleted. The system terminates as soon as no rule can delete any tag.
- *all rules*: safe rules are executed first (see *safe rules only* mode). Then heuristic rules start to operate in the loop (similarly to the safe rules). Any time a heuristic rule deletes a tag, the *safe rules only* mode is entered as a sub-procedure. When safe rules' execution terminates, the loop of heuristic rules continues. The disambiguation is finished when no heuristic rule can delete any tag.

The rules are written in the *fast LanGR* formalism (Květoň, 2006) which is a subset of a more general LanGR formalism (Květoň, 2005). The LanGR formalism has been developed specially for writing and implementing disambiguation rules.

3. Methods of combination

The motivation for the combination experiments is following: if we have several different methods solving the same problem with similar error rate, it is probable that they do not make exactly the same mistakes. If we identify the strong and weak aspects of each method and find the optimal way to combine them, the resulting method's performance should be better than the performance of all of its components.

In our experiments we use the components described above – three statistical taggers (Feature-based – „a", HMM – „b", Morče – „m") and two sets of hand-written rules („safe", safe + heuristics – „all"). Most of the ideas for the experiments were original, except the serial combination rules – tagger, which was already published in (Hajič et al., 2001b) and we only performed the same experiment with new versions of the components.

All the methods presented in this paper have been trained and tested on the PDT version 2.0[4]. Taggers were trained on PDT 2.0 training data set (1,539,241 tokens), the results were achieved on PDT 2.0 development-test data set (201,651 tokens), and for the best methods also the PDT 2.0 evaluation-test data set (219,765 tokens) was used. The morphological analysis processor and all the taggers were used in versions from April 2006 (Hajič et al., 2006), the rule-based component is from September 2006.

For evaluation, we use both precision and recall (and the corresponding F-measure) and accuracy, since we also want to evaluate the partial disambiguation achieved by the hand-written rules alone. Let t denote the number of tokens in the test data, let c denote the number of tags assigned to all tokens by a disambiguation process and let h denote the number of tokens where the manually assigned tag is present in the output of the process.

- In case of the morphological analysis processor and the standalone rule-based component, the output can contain more than one tag for every token. Then *precision* (p), *recall*

[4] The results cannot be simply (number-to-number) compared to previous results on Czech tagging, because different training and testing data (PDT 2.0 instead of PDT 1.0) are used since 2006.

Tagger	accuracy
Feature-based (a)	94.27%
HMM (b)	95.13%
Morče (m)	**95.43%**

Table 2. Evaluation of the taggers alone

	precision	recall	f-measure
morphology	25.72%	99.40%	40.87%
safe rules	58.76%	98.90%	73.72%
all rules	67.36%	98.24%	**79.92%**

Table 3. Evaluation of the rules alone

(r) and *F-measure* (f) characteristics are defined as follows:

$$p = h/c \qquad r = h/t \qquad f = 2pr/(p + r).$$

- The output of the stochastic taggers contains always exactly one tag for every token — then $p = r = f = h/t$ holds and this ratio is denoted as *accuracy*.

The initial performance of the components is presented in table Table 2 and Table 3

3.1. Serial combination rules – tagger

The simplest way of combining a hand-written disambiguation grammar with a stochastic tagger is to let the grammar reduce the ambiguity of the tagger's input. Formally, an input text is processed as follows:

1. morphological analysis (every input token gets all tags that are plausible without looking at context);
2. rule-based component (partially disambiguates the input, i.e. deletes some tags);
3. the stochastic tagger (gets partially disambiguated text on its input).

This algorithm was already used in (Hajič et al., 2001b), only components were changed — the ruled-based component was significantly improved and two different sets of rules were tried, as well as three different statistical taggers. The results (compared to the results of the standalone taggers) are presented in Table 4.

The best result was (not surprisingly) achieved with the set of safe rules followed by the Morče tagger.

An identical approach was used in (Tapanainen, 1994) for English.

	–	safe rules	all rules
tagger a	94.27%	92.51%	92.55%
tagger b	95.13%	95.48%	95.30%
tagger m	95.43%	**95.64%**	95.44%

Table 4. Evaluation of the serial combination rules – tagger

Tagger	accuracy
tagger a	99.31%
tagger b	99.22%
tagger m	**99.25%**

Table 5. Accuracy of the taggers in SUBPOS disambiguation

3.2. Serial combination with SUBPOS pre-processing

Manual inspection of the output of the application of the hand-written rules on the development data (as used in the serial combination described in the previous section) discovered that certain types of deadlocked ("cross-dependent") rules prevent successful disambiguation.

Cross-dependence means that a rule A cannot apply because of some remaining ambiguity, which could be resolved by a rule B, but the operation of B is still dependent on the application of A. In particular, ambiguity in the Part-of-Speech category is very problematic. For example, only a few safe rules can apply to a three-word sentence where all three words are ambiguous between finite verbs and something else.

If the Part-of-Speech ambiguity of the input is already resolved, precision of the rule-based component and also of the final result after applying any of the statistical taggers improves. Full Part-of-Speech information is represented by the first two categories of the Czech morphology tagset — POS and SUBPOS, which deals with different types of pronouns, adverbs etc. As POS is uniquely determined by SUBPOS (Hajič et al., 2006), it is sufficient to resolve the SUBPOS ambiguity only.

All three taggers achieve more than 99% accuracy in SUBPOS disambiguation (see Table 5). For SUBPOS disambiguation, we use the taggers in usual way (i.e. they determine the whole tag) and then we put back all tags having the same SUBPOS as the tag chosen by the tagger.

Thus, the method with SUBPOS pre-processing operates in four steps:

1. (morphological analysis)
2. SUBPOS disambiguation (any tagger)
3. rule-based component
4. final disambiguation (any tagger)

Results after performing the first, the second and the third step are presented in Tables 6, 7,

	precision	recall	f-measure
tagger a	30.05%	98.92%	46.10%
tagger b	30.10%	98.83%	46.15%
tagger m	30.10%	98.87%	46.15%

Table 6. Combination with SUBPOS pre-processing: results of the first step

	precision	recall	f-measure
tagger a + safe rules	64.81%	98.68%	78.24%
tagger a + all rules	70.53%	98.36%	82.15%
tagger b + safe rules	65.07%	98.59%	78.40%
tagger b + all rules	70.81%	98.27%	82.31%
tagger m + safe rules	65.07%	98.62%	78.41%
tagger m + all rules	70.81%	98.30%	**82.32%**

Table 7. Combination with SUBPOS pre-processing: results of the second step

	tagger a	tagger b	tagger m
tagger a + safe rules	92.81%	95.68%	**95.78%**
tagger a + all rules	93.08%	95.69%	95.77%
tagger b + safe rules	92.76%	95.63%	95.72%
tagger b + all rules	93.02%	95.64%	95.71%
tagger m + safe rules	92.79%	95.63%	95.75%
tagger m + all rules	93.05%	95.64%	95.73%

Table 8. Combination with SUBPOS pre-processing: final accuracy (lines – tagger and rules used in the first two steps, columns – tagger used in the third step)

8, respectively.

The best result was achieved with tagger *a* in the first step, the set of safe rules in the second step and the tagger *m* in the third step. If we want to use only one tagger (i.e. the same in the first and the third step), the result with tagger *m* and the set of safe rules is nearly as good as the best result.

We performed also experiments with the second step (rules) omitted, because we wanted to check, whether the rules really have some significant impact on the final result, or if the only important step is the SUBPOS pre-processing.

The results in Table 9 show that rules are really important, because the method without

	tagger a	tagger b	tagger m
tagger a	92.96%	95.18%	**95.42%**
tagger b	92.90%	95.13%	95.37%
tagger m	92.92%	95.15%	95.40%

Table 9. Combination with SUBPOS pre-processing: check of the rules efficiency (lines – tagger used in the first step, columns – tagger used in the last step)

rules does not even reach the accuracy of the best of the standalone taggers.

3.3. Combining more taggers in parallel

This method is quite different from previous ones, because it essentially needs more than one tagger. It consists of the following steps:

1. (morphological analysis;)
2. running N taggers independently;
3. merging the results from the previous step — each token ends up with between 1 and N tags, a union of the taggers' outputs;
4. the rule-based component;
5. final disambiguation (single tagger).

This method is based on the assumption that different stochastic taggers make complementary mistakes, so that the recall of the "union" of taggers is almost 100%. Several existing language models are based on this assumption — (Brill, 1998) for tagging English, (Borin, 2000) for tagging German and (Vidová-Hladká, 2000) for tagging inflective languages. All these models perform some kind of "voting" — for every token, one tagger is selected as the most appropriate to supply the correct tag. The model presented in this paper, however, entrusts the selection of the correct tag to another tagger that already operates on the partially disambiguated input.

Results after performing the first two steps, the third and the final step are presented in Tables 10, 11, 12, respectively.

The best results were achieved with two taggers in Step 1 (*a* and *m*), the set of all rules in Step 3 and the tagger *b* in Step 4.

We also measured the accuracy of this method with the rules step omitted. The results of this experiment presented in Table 13 lead to two important conclusions: 1) the rules significantly improve the result (but) 2) the paralell combination without rules performs better than any other purely statistical method or combination.

4. Results

Table 14 shows overall results of the best methods in each category (depending on number of components) measured on the dev-test and on the eval-test data.

	precision	recall	f-measure
$a \cup b$	92.18%	96.90%	94.48%
$a \cup m$	92.30%	97.04%	94.61%
$b \cup m$	93.19%	97.05%	**95.08%**
$a \cup b \cup m$	90.81%	97.66%	94.11%

Table 10. Paralell combination: results of the first two steps (union of the tagger’s outputs)

	precision	recall	f-measure
$(a \cup b)$ + safe rules	93.56%	96.74%	95.12%
$(a \cup b)$ + all rules	93.99%	96.63%	95.29%
$(a \cup m)$ + safe rules	93.71%	96.86%	95.26%
$(a \cup m)$ + all rules	94.15%	96.77%	95.44%
$(b \cup m)$ + safe rules	94.11%	96.90%	95.48%
$(b \cup m)$ + all rules	94.46%	96.81%	**95.62%**
$(a \cup b \cup m)$ + safe rules	92.67%	97.46%	95.00%
$(a \cup b \cup m)$ + all rules	93.32%	97.32%	95.28%

Table 11. Paralell combination: results of the third step (union + rules)

	tagger a	tagger b	tagger m
$(a \cup b)$ + safe rules	95.43%	95.49%	95.96%
$(a \cup b)$ + all rules	95.54%	95.58%	95.96%
$(a \cup m)$ + safe rules	95.56%	96.03%	95.73%
$(a \cup m)$ + all rules	95.68%	**96.09%**	95.82%
$(b \cup m)$ + safe rules	95.81%	95.58%	95.77%
$(b \cup m)$ + all rules	95.89%	95.71%	95.86%
$(a \cup b \cup m)$ + safe rules	95.52%	95.66%	95.84%
$(a \cup b \cup m)$ + all rules	95.69%	95.80%	95.95%

Table 12. Paralell combination: final accuracy (lines – taggers and rules used in the first three steps, columns – the tagger used in the last step)

Table 15 shows the relative error rate reduction. The best method presented by this paper (parallel combination of taggers with all rules) reaches the relative error rate decrease of 11.48% in comparison with the tagger Morče (which achieves the best results for Czech so far).

	tagger a	tagger b	tagger m
$a \cup b$	94.94%	95.13%	**95.87%**
$a \cup m$	95.05%	**95.87%**	95.46%
$b \cup m$	95.56%	95.13%	95.48%
$a \cup b \cup m$	94.85%	95.14%	95.47%

Table 13. Paralell combination: check of the rules efficiency (lines – taggers used in the first step, columns – the tagger used in the last step)

Components available	The best method	dev-test	eval-test
one tagger	m	95.43%	95.12%
two taggers	–	–	–
three taggers	$(a \cup m) + b$ or $(a \cup b) + m$	95.87%	95.52%
one tagger + rules	SUBPOS m + safe rules + m	95.75%	95.44%
two taggers + rules	$(b \cup m)$ + disheu1 + m	95.86%	95.49%
thee taggers + rules	$(a \cup m)$ + disheu1 + b	**96.09%**	**95.68%**

Table 14. Overall results

Method	Morče	Parallel without rules
Parallel without rules	8.20%	–
Parallel with all rules	11.48%	3.57%

Table 15. Relative error rate reduction

4.1. Error analysis

Table 16 shows error rate (100% – *accuracy*) of various methods[5] on particular positions of the tags (13 and 14 are omitted). The most problematic position is CASE (5), whose error rate was significantly reduced.

The CASE confusion matrices 18 and 17 show the final situation in more detail. Ambiguity between nominative and accusative remains to be the most problematic even for the hybrid tagging methods.

[5] *Par* stands for parallel combination without rules, *Par+Rul* for parallel combination with rules.

	a	b	m	Par	Par+Rul
1	0.61	0.70	0.66	0.57	0.57
2	0.69	0.78	0.75	0.64	0.64
3	1.82	1.49	1.66	1.39	1.37
4	1.56	1.30	1.38	1.18	1.15
5	4.03	3.53	3.08	2.85	2.62
6	0.02	0.03	0.03	0.02	0.02
7	0.01	0.01	0.01	0.01	0.01
8	0.06	0.07	0.08	0.06	0.05
9	0.05	0.08	0.07	0.05	0.04
10	0.29	0.28	0.30	0.26	0.27
11	0.29	0.31	0.33	0.28	0.28
12	0.05	0.08	0.06	0.05	0.04
15	0.31	0.31	0.31	0.28	0.29

Table 16. Error rate [%] on particular positions of tags

tg/an	-	1	2	3	4	5	6	7	X
-	82753	37	41	0	18	3	4	7	21
1	53	26027	286	11	939	21	8	5	81
2	9	205	29363	21	146	0	25	14	24
3	1	41	70	5265	54	0	50	23	1
4	50	1835	404	12	21302	1	155	44	15
5	0	8	0	3	2	36	0	1	0
6	3	18	54	15	128	0	17914	3	3
7	29	26	19	8	73	0	0	9010	3
X	115	312	90	7	44	21	14	5	4242

Table 17. CASE confusion matrix: paralell combination without rules (rows – output of the combination, columns – annotation)

5. Conclusion

We have presented several variations of a novel method for combining statistical and hand-written rule-based tagging. The best variation improved the accuracy of the best-performing standalone statistical tagger by over 11% (in terms of relative error rate reduction), and the inclusion of the rule-component itself improved the best statistical-only combination by over 3.5% relative.

Our experiments produced a software suite which gives the all-time best results in Czech

tg/an	-	1	2	3	4	5	6	7	X
-	82747	39	43	2	18	3	2	7	23
1	50	26063	290	13	883	22	6	7	97
2	8	188	29397	23	128	0	18	16	29
3	0	37	71	5310	48	0	14	24	1
4	37	1561	406	13	21597	1	145	41	17
5	0	10	0	8	2	29	0	1	0
6	3	17	56	18	120	0	17917	3	4
7	31	22	20	8	62	0	0	9022	3
X	109	285	86	6	48	21	11	6	4278

Table 18. CASE confusion matrix: paralell combination with rules

tagging and which was used to re-tag the existing 200 mil. word Czech National Corpus. It should significantly improve the user experience (for searching the corpus) and allow for more precise experiments with parsing and other NLP applications that use that corpus.

Different variants of the method are available for different tasks – without the rule-bassed component, the accuracy is not much lower and the system runs ten times faster, which makes this variant suitable for large data processing.

6. Recent Advances and Outlook

The goal of this paper was to present the main results of the PhD thesis (Spoustová, 2007). There are also some new, unpublished results, which immediately follow the work described in the thesis and in this paper. We would like to present them here (very briefly) before they will be published in a definite form.

We have developed a method of a semi-supervised training of the Morče tagger. The main idea consists in the preparation of the training data: for every iteration, the training data set is unique. Each of the training sets begins with the PDT 2.0 train data set, which is followed by a (unique) part of the Czech National Corpus processed by the parallel combination with rules (the results of this combination are passed to the tagger instead of the human morphological annotation, which is not available for such a large corpus). Thus, every training set contains the same supervised part as the other sets and a unique unsupervised part.

We have experimented with various sizes of the unsupervised parts (from 500k tokens to 5M) and also with various numbers of iterations. During the last year also the supervised Morcče tagger, so we used the newest version (”gangrena”).

The preliminary results (PDT 2.0 devel-test) are presented in Table 19. The table contains results of the standalone Morče tagger, results of the two versions of parallel combination, and finally, results of the semi-supervised taggers trained on the parallel combinations.

This preliminary results show that our method of semi-supervised training allows Morče

Method	accuracy
Morče gangrena alone	95.99%
Parallel combination without rules (P1)	96.03%
Parallel combination with rules (P2)	96.22%
Semi-supervised Morče trained on P1	96.22%
Semi-supervised Morče trained on P2	96.23%

Table 19. Accuracy of the semi-supervised Morče compared to other methods (devel-test)

tagger to perform at least as good as the corresponding parallel combination. The output of the parallel combination is needed in the training stage of the tagger, but the tagging process is as fast and simple as when running the supervised tagger.

This method is in development for various languages (Czech, English, Slovak) and final results will be published soon in more detail.

Acknowledgements

The research described here was supported by the projects *LC536* of *Ministry of Eduation, Youth and Sports* of the Czech Republic.

Bibliography

Bick, Eckhard. 2000. The parsing system "Palavras" — automatic grammatical analysis of Portuguese in a constraint grammar framework. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation, TELRI.* Athens.

Borin, Lars. 2000. Something borrowed, something blue: Rule-based combination of POS taggers. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Vol. 1, pp. 21–26. Athens.

Brill, Eric and Jun Wu. 1998. Classifier combination for improved lexical disambiguation. In: *Proceedings of the 17th international conference on Computational linguistics*, Vol. 1, pp. 191–195. Montreal, Quebec.

Chanod, Jean-Pierre and Pasi Tapanainen. 1995. Tagging French — comparing a statistical and a constraint-based method. In: *Proceedings of EACL-95*, pp. 149–157. Dublin.

Collins, Michael. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In: *Proceedings of EMNLP'02*, July 2002, pp. 1–8. Philadelphia.

Daelemans, W., Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In: *Proceedings of the 4th WVLC*, pp. 14–27. Copenhagen.

Erjavec, Tomaz, Saso Dzeroski, and Jakub Zavrel. 1999. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. *Technical Report*, Dept. for Intelligent Systems, Jozef Stefan Institute. Ljubljana.

Hajič, Jan and Barbora Hladká. 1997. Tagging of inflective languages: a comparison. In: *Proceedings of ANLP '97*, pp. 136–143. Washington, DC.

Hajič, Jan. 2000. Morphological tagging: Data vs. dictionaries. In: *Proceedings of the 6th ANLP / 1st NAACL'00*, pp. 94–101. Seattle, WA.

Hajič, Jan, Pavel Krbec, Pavel Květoň, Karel Oliva, and Vladimír Petkevič. 2001. Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics.* CNRS – Institut de Recherche en Informatique de Toulouse and Université des Sciences Sociales, pp. 260–267. Toulouse.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank v2.0. CDROM. Linguistic Data Consortium, Cat. LDC2006T01. Philadelphia. ISBN 1-58563-370-4. Documentation also at `http://ufal.ms.mff.cuni.cz/pdt2.0`.

Karlsson, Fred. 1985. Parsing Finnish in terms of a process grammar. In: Fred Karlsson (ed.): *Computational Morphosyntax: Report on Research 1981-84*, University of Helsinki, Department of General Linguistics Publications No. 13, pp. 137–176.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila (eds.). 1995. Constraint Grammar: a language-independent system for parsing unrestricted text. *Natural Language Processing.* Vol. 4, Mouton de Gruyter, Berlin and New York.

Koskenniemi, Kimmo. 1990. Finite-State Parsing and Disambiguation. In: *Proceedings of Coling-90*, University of Helsinki, 1990, pp. 229–232. Helsinki.

Krbec, Pavel. 2005. *Language Modelling for Speech Recognition of Czech.* PhD Thesis, MFF, Charles University Prague.

Květoň, Pavel. 2005. *Rule-based Morphological Disambiguation.* PhD Thesis, MFF, Charles University Prague.

Květoň, Pavel. 2006. Rule-based morphological disambiguation: On computational complexity of the LanGR formalism. In: *The Prague Bulletin of Mathematical Linguistics*, Vol. 85, pp. 57–72. Prague.

Oflazer, Kemal and Gökhan Tür. 1997. Morphological disambiguation by voting constraints. In: *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pp. 222–229. Madrid.

Oliva, Karel, Milena Hnátková, Vladimír Petkevič, and Pavel Květoň. 2000. The Linguistic Basis of a Rule-Based Tagger of Czech. In: Sojka P., Kopeček I., Pala K. (eds.): *Proceedings of the Conference ”Text, Speech and Dialogue 2000”, Lecture Notes in Artificial Intelligence*, Vol. 1902. Springer-Verlag, pp. 3–8. Berlin-Heidelberg.

PDT Guide. `http://ufal.ms.mff.cuni.cz/pdt2.0`

Ratnaparkhi, A.. 1996. A maximum entropy model for part-of-speech tagging. In: *Proceedings of the 1st EMNLP*, May 1996, pp. 133–142. Philadelphia.

Samuelsson, Christer and Atro Voluntainen. 1997. Comparing a linguistic and a stochastic tagger. In: *Proceedings of ACL/EACL Joint Converence*, pp. 246–252. Madrid.

Smith, Noah A., David A. Smith, and Roy W. Tromble. 2005. Context-Based Morphological Disambiguation with Random Fields. In: *Proceedings of HLT/EMNLP*, pp. 475–482. Vancouver.

Spoustová, Drahomíra “johanka”. 2007. *Kombinované statisticko-pravidlové metody značkování češtiny. (Combining Statistical and Rule-Based Approaches to Morphological Tagging of Czech Texts).* PhD Thesis, MFF UK.

Spoustová, Drahomíra “johanka”, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. Cooperation of Statistical and Rule-Based Taggers for Czech. In: Proceedings of Balto-Slavonic Natural Language Processing Workshop, ACL, Prague 2007. pp. 67–74. Prague.

Tapanainen, Pasi and Atro Voutilainen. 1994. Tagging accurately: don’t guess if you know. In: *Proceedings of the 4th conference on Applied Natural Language Processing*, pp. 47–52. Stuttgart.

Vidová-Hladká, Barbora. 2000. *Czech Language Tagging.* PhD thesis, MFF UK. Prague.

Votrubec, Jan. 2005. *Volba vhodných rysů pro morfologické značkování češtiny. (Feature Selection for Morphological Tagging of Czech.)* Master thesis, MFF, Charles University, Prague.

Votrubec, Jan. 2006. Morphological Tagging Based on Averaged Perceptron. In: *WDS’06 Proceedings of Contributed Papers*, MFF UK, pp. 191–195. Prague.