Part of Speech Tags for Automatic Tagging and Syntactic Structures[1]

Barbora Hladká, Kiril Ribarov

Introduction

During the last years many POS tagging[2] experiments have been performed and the problems of POS tagging for particular languages were discussed. The latter cover the following questions: whether to tag texts manually or automatically, which methods to use to tag texts, which POS tagset is optimal, whether a bigger tagset is better than a smaller one, how the tagging accuracy changes as the size of the POS tagset or the method of tagging changes, where are the differences between tagging inflectional languages and languages with poor flection (see Appendix - Figure 1). In particular, we have tried to solve these problems for Czech, a language with rich inflection (Hajič, Hladká 1997). We believe that the best way how to find a solution is to perform experiments, as we did. Simultaneously, with Czech experiments, we have applied the same program code to tag English texts in order to find out whether the results of the experiments are dependent upon the character of the input language.

Another question which arises immediately is WHY is it so useful to add a part-of-speech information to a text. The adequate and immediate answer says that there are many linguistic applications for which tagged input data are needed, e.g. (Mangu and Brill 1997). It is an obvious fact that these applications presuppose text tagged as thoroughly as possible (see Appendix - Figure 2). We will show below the relation between the tagging accuracy, POS tagset and the results of a parsing procedure (Brill 1993; Ribarov 1996).

1 Methods of POS tagging

A number of approaches have been proposed for automatic tagging. We have used three of them to tag Czech texts: the first one is based on Markov models, i. e. the probabilistic approach (Merialdo 1992), the second one is based on Markov models, too, but it is realised by finite-state automata (Tapanainen 1995; Cutting et al. 1992) and the third one is rule-based (Brill 1992).

The probabilistic approach calculates with lexical probabilities – i.e. the probability of a part of speech given the word, and contextual probabilities – i.e. the probability of a part of speech given the previous part of speech. These probabilities are trained on a manually tagged corpus.

Finite-state tools serve to build finite-state transducers. The tagger lexicon is a finite-state transducer which provides potential POS tags for every input word. These tags correspond to those given by a morphological analysis. In addition to the tagger lexicon the guesser transducer contains word form patterns to determine the potential POS tags for a non-lexicalized word form. Manually tagged corpus, tagger and guesser transducers provide transition probabilities - Hidden Markov Model - for the POS disambiguator.

The rules to predict the most likely tag for unknown words and the contextual rules are trained on a tagged corpus. The rule-based tagger initially tags the words by assigning each of them its most likely tag using the first set of rules and afterwards improves the tagging accuracy by contextual rules.

2 POS tagset

POS tagset is based upon part-of-speech classes and combinations of morphological categories. The following tables illustrate the mapping from the most detailed Czech POS tagset (1171 tags) to POS tagset containing only 34 tags.

Morphological Category	Category Letter	Possible Value	Description
gender	*g*	M	masculine animate
		I	masculine inanimate
		N	neuter
		F	feminine
number	*n*	S	singular
		P	plural
tense	*t*	M	past
		P	present
		F	future
mood	*m*	O	indicative
		R	imperative
case	*c*	1/NOM	nominative
		2/GEN	genitive
		3/DAT	dative
		4/ACC	accusative
		5/VOC	vocative
		6/LOC	locative
		7/INS	instrumental
case	*c'*	NA	nominative or accusative
		GEN	genitive
		DAT	dative
		VOC	vocative
		LOC	locative
		INS	instrumental
voice	*s*	A	active voice
		P	passive voice
polarity	*a*	N	negative
		A	affirmative
degrees of comparison	*d*	1	base form
		2	comparative
		3	superlative
person	*p*	1	1st
		2	2nd
		3	3rd
kind of verb	*k*	PAP	past participle
		PRI	present
		IMP	imperative
		INF	infinitive
		TRA	transgressive

Table 1

Description	POS TAG$_{1171}$	POS TAG$_{206}$	POS TAG$_{47}$	POS TAG$_{43}$	POS TAG$_{34}$
Nouns	N*gnc*	N*n*	NOUN_*c*	NOUN_*c'*	NOUN
Abbreviations	NZ	NZ	NOUN_INV	NOUN_INV	NOUN
Adjectives	A*gncda*	A*nda*	ADJ_*c*	ADJ_*c'*	ADJ
Verbs infinitives	V*Ta*	V*Ta*	VERB_INF	VERB_INF	VERB_INF
Verbs transgressives	VW*ntsga*	VW*ntsa*	VERB_TRA	VERB_TRA	VERB_TRA
Verbs common	V*pnstmga*	V*pnsta*	VERB_*k*	VERB_*k*	VERB_*k*
Pronouns personal	PP[*12*]*nc*, PP*3gnc*	PP*pn*	PRON_*c*	PRON_*c'*	PRON_*c*
Pronouns possessive	PR*gncpgn*	PR*npn*	PRON_*c*	PRON_*c'*	PRON_*c*
Pronouns demonstrative	PD*gnca*	PD*na*	PRON_*c*	PRON_*c'*	PRON_*c*
"Svůj" "his" - subject	PS*gnc*	PS*n*	PRON_*c*	PRON_*c'*	PRON_*c*
Reflexive particle "se"	PE*c*	PE	P_SE	P_SE	P_SE
Adverbs	O*da*	O*da*	ADV	ADV	ADV
Conjunctions	S[S/P]	S[S/S]	CONJ	CONJ	CONJ
Numerals	C*gnc*	C*n*	NUM_*c*	NUM_*c'*	NUM_*c*
Numbers	C*gnc*	C*n*	NUM_INV	NUM_INV'	NUM_INV
Prepositions	R*prep.*	R*prep.*	PREP	PREP	PREP
Interjections	F	F	INTJ	INTJ	INTJ
Particles	K	K	PTCL	PTCL	PTCL
Sentence boundaries	T_SB	T_SB	SENT	SENT	SENT
Punctuation	T_IP	T_IP	PUNCT	PUNCT	PUNCT
Unknown tag	X	X	-	-	-
Proper names	-	-	PROP	PROP	PROP
Comma	-	-	CM	CM	CM
Clitics	-	-	CLIT	CLIT	CLIT
Date	-	-	DATE	DATE	DATE

Table 2

It seems very strange to disregard such important categories for Czech as case and gender (see POS $TAG_{1171} \mapsto$ POS TAG_{206}) and even to assume only part-of-speech information (see POS TAG_{34}). But the experiment results for each POS tagset will be helpful to look for the optimal POS tagset. The Czech corpus that we used in statistical and rule-based experiments was tagged by tags from the tagsets POS TAG_{1171} and POS TAG_{206}.

3 Is tagging of Czech different?

At the first sight we would say YES, the tagging of Czech should be different from the English one. The answer could be found in the cardinality of the two POS tagsets.

The differences between a morphologically ambiguous inflective language and a language with poor inflection are reflected, e.g., in the number of tags for verbs and adjectives as is shown in the following table:

	ENGLISH	CZECH
VERBS	VB,VBD,VBG,VBN,VBP,VBZ	for present tense only
		V[123][SP][AP]P[OR][MIFN][AN]
	6	1 x 3 x 2 x 2 x 1 x2 x 4 x 2 = **192**
ADJECTIVES	JJ, JJR, JJS	A[MIFN][SP][1234567][123][AN]
	3	1 x 4 x 2 x 7 x 3 x 2 **= 336**

To find the answer we took into account the most detailed Czech tagset POS TAG_{1171} and Penn Treebank tagset containing 48 tags for English (Santorini 1990). The numbers 6 vs. 192 and 3 vs. 336 illustrate the differences very clearly.

4 Tagging accuracy

method[3]	BMM	RB	BMM	RB	MMFS	MMFS	MMFS
POS tagset	1 171	1 171	206	206	47	43	34
training data	600K	38K	600K	38K	15K	15K	15K
tag. accuracy	81,5%	79,8%	90,1%	87,2%	91,7%	93,0%	96,2%

Table 3

Should we be satisfied with the levels of performance of our Czecl experiments? Let us look at the numbers in the table, without any links to tagging linguistic background. 81% is still better than our original expectation, 90% is comparable with results for Swedish (Elthworthy 1995) and 96% is the same as for English (Schiller 1996). These numbers show that a smaller tagset achieves better tagging performance than the bigger one does; the statistical approach seems to be a little bit better than a rule-based one[4].

On the one hand these numbers mean that many sentences will contain at least one error; at the same time the subsequent processing requires perfect part of speech analysis, neither 81% neither 96% performance is clearly good enough.

5 Applications: the required input/output

In Elworthy (1995) experiments concerning changing tagsets are presented for three different languages (English, French, Swedish). These experiments show that the relationship between tagset size and accuracy is a weak one and is not consistent even if applied for the same language. The main conclusion which is derived from the results of experiments is **to choose the tagset according to the requirements of a given application rather than to optimise it for tagger**. Figure 3 (see Appendix) supports this conclusion.

What is the state of the art? Three different POS tagsets (POS TAG_{1171}, POS TAG_{206}, POS TAG_{34}), statistical approach to tag text and the tagging accuracy of each statistical tagging with three different tagsets have been developed and tested. Cases of incorrect tag assignment to the words in the input sentence (*Zkrocení zlé ženy mělo úspěch* [lit. *Taming of the shrew had success*]) are in boldface in the tagged input sentence (Zkrocení/**ANS11A**, úspěch/**NIS1**, **ž**eny/**NP**) in Figure 3.

For illustration: how can we grammatically check the input Czech sentence when the only information we know (with precision 96%) is the POS information of the reduced tagset? From the previous simple example it is clear that morphological categories such as gender and number represent information which has to be present in a pretagged text for many applications.

For a user, let us say a linguist, whose aim is to specify morphological categories, we need to perform tagging with a full tagset, more specifically, with carefully selected and detailed morphological classification according to the traditional grammar. The result could be viewed as a process which „added" information to the text, or as a way of classification (clustering) of the word mass. Different tagset sizes result in different classification of the word forms. The members of each cluster are thus given the same tag.

Previous tagset reductions have been connected with specific mutual dependence between each two of the tagsets. The tagsets mappings have to preserve the patterns, the (ir)regularities of the language material.

A special tagset reduction may also be aim-specific: a tagset containing only tags being of interest for the specific user/application. Such a tagset might be of a cardinality lower than any of the previously mentioned ones. In these cases the percentages might, or need not, continue to grow but the extracted information, the new classification, gives a feedback to the process of a future possible improvement. What would then be the required tagset size in order to assign a surface syntactic structure to the sentences, taking the output of the tagging procedure as an input? The answer to this question is not straightforward, but the following results encourage to use non-full (reduced) morphological distinctions.

Let us take the Brill's Rule-Based approach in order to extract, in a rule-formalised way, the syntactic structure of the sentences. The input is a stream of tags (in the sequel: nodes; morphologically disambiguated), while the output would be a bare syntactic structure of the sentence (dependency syntax with unannotated dependencies) (Ribarov 1996; Hajič and Ribarov 1997). The rules can change (add or cancel) a dependency between two nodes and swap the order of nodes. During the process of learning the rules from a manually prepared (both tagged and syntactically annotated) corpus, an ordered list of

rules is constructed. Brill refers to this list as to the grammar of the language the rules are trained on.

If the training corpus is annotated with the POS TAG_{1171} we can observe results geven here in Table 4.

Order	Description of the Rule	Success (%)
1	Swap the dependency between ZIP and NFS4A	33.67
2	Swap the dependency between ANS51A and NFP5A	34.53
3	Swap the dependency between PDFS2 and NFS6A	35.14
4	Swap the dependency between PQFIP1 and VPS3A	35.74
5	Swap the dependency between RV7 and VPS3A	36.27
6	Swap the dependency between ANS51A and NIS4A	36.81
	....	
7	Swap the dependency between ANP71A and NFP7A	44.56
8	Swap the dependency between PQFMP1 and VPP3N	44.76
9	Swap the dependency between ANS53A and NFS6A	44.96
10	Swap the dependency between ANS61A and NNS6A	45.16
	....	
11	Swap the dependency between AFS71A and NMS6A	47.18
12	Swap the dependency between ANS61A and NOMORPH	47.38

Table 4

The situation changes rather dramatically if we train on the reduced tagset POS TAG_{34} as shown in Table 5.

Order	Description of the Rule	Success (%)
1	*Swap the dependency between ADJ and NOUN*	44.38
2	Swap the dependency between CM and CONJ	46.00
3	Swap the dependency between PSE and VERB_PRI	47.48
4	Swap the dependency between ADV and VERB_PRI	48.74
5	Swap the dependency between PROP and VERB_PRI	49.58
6	Swap the dependency between ADJ and NOUN	50.42
7	Swap the dependency between CM ADJ	51.29

8	Delete the dependency between ADJ and CM	52.22
	....	
9	Add a dependency between PUNCT and PROP	64.84
10	Delete the dependency between PRON_INS and PREP	65.02

Table 5

Not all of the rules result in a sentence structure with precise grammatical explanation. The „strange" ones are there to combine with the others in order to correct the structures of the previously applied rules. Although some of the rules are obvious and we believe that a human would derive the same rules, still the grammatical meaning of the rules can be evaluated only when analysing the result of the application of the whole list of the rules. Let us try and examine the relation and dependence of the reduced and non-reduced tagset on the selected rules as given in the above tables. One of the obvious rules in Table 5 is rule 1 (and rule 6; the rules might repeat; during their repetition their influence has a different scope depending on their position in the list). To cope with a more distinct situation in- a more specific POS TAG_{1171}, the algorithm produces more rules in order to capture the relation between an adjective and a noun: rules[5] 2, 6, 7, 9, 10 and 11 in Table 4.

Undoubtedly, the reduced tagset brings better absolute values. We would like to note that the nodes within the syntactic structure could be returned the corresponding values from the full tagset. Thus, no information has been lost.

As for the Rule-Based application for syntactic structure extraction, the reduced tagset leads to a much better start on the learning curve (see Appendix - Figure 4). After the saturation of the process of learning, which comes after several tens of rules have been learnt, the learning might continue after one switches to the more expanded tagset, namely the full tagset[6].

6 Future considerations

So far, a reduced tagset helped us to achieve better results both for tagging tagging and for the input for syntactic analysis procedures. Another question arises: Could the reduced tagset help us for better

achievements in the above stated procedures, involving the full instead of the reduced tagset?

Instead of tag disambiguation, let us consider tag elimination from the set of all possible morphological tags belonging to each word form. The process of tag elimination eliminates those tags which are almost improbable (within HMM each tag must be given a certain probability $\neq 0$). Thus, the text is tagged by sets of more probable tags. On the morphological level, the disambiguation is done only with regard to the neighbouring (predefined context) tags and/or lexemes. Many of the cases would have benefited from a more successful disambiguation, if information about the syntactic structure had been given (phrases; dependencies)[7]. If our Rule-Based approach to syntactic tree structures is modified in a way such that the rules operate on sets of tags rather than on single tags (this would bring more than one possible syntactic structure), then something like „morpho-syntactic bouncing" could lead to a more successful tagging and syntactic algorithms. The bouncing stops when both, the morphological tags and the syntactic tree structures, have been mutually disambiguated. We believe that this process could „eliminate" the differences between the success rate and the tagset size[8]. This neither eliminates nor reduces the importance of what has been said so far concerning the reduced tagset, since the sets of tags could be viewed as a special case of the reduced tagset. Rather, it brings new encouragement and stresses the importance of syntax on the way towards the meaning of the sentence.

Notes

[1]We gratefully recall many friendly and fruitful discussions with Jarmila Panevová on these and several other issues, including especially the one at the occasion of our joint stay at the Colloquium in Leipzig in June 1997.
The research described here is supported by the grant GAČR No. 405/96/K214.

[2]A POS tag describes the part of speech information and the possible combination of morphological categories for each POS class. There are few possible tags for a given word in the sentence. The list of these tags can be found by a morphological analysis of the word. POS tagging is a procedure which assigns a sequence of unique and correct POS tags from the list of possible tags to a sequence of input words within the context of the sentence. For illustration, let us assume the set of POS tags T, T = {ADJECTIVE, ADVERB, CONJUNCTION, INTERJECTION, NOUN, NUMERAL, PARTICLE, PREPOSITION, PRONOUN, VERB}. We want to tag the following input sentence with the POS tags named above. *Mysli na stav mysli!* [lit. *Think of the state of mind*.]

Using morphology analysis each word obtains the list of possible POS tags: *Mysli* {NOUN, VERB} *na* {PREPOSITION} *stav* {NOUN, VERB} *mysli* {NOUN, VERB}.
After the tagging procedure our input sentence will be tagged in the following way: *Mysli*{VERB} *na* {PREPOSITION} *stav* {NOUN} *mysli* {NOUN}. Due to the context information we tagged the first Mysli as VERB and the second one as NOUN.

[3]BMM - Bigram Markov Model, RB - Rule-Based, MMFS - Markov Model realised by finite-state automata

[4] The training of lexically and contextually based rules on Czech corpus was very time-consuming. The process of training took three days and it is important to note that we did not use the complete Czech corpus in the rule-based experiments.

[5] Those rules are not the only ones. The example presented in Table 4 and Table 5 is a selection from an ordered set of rules.

[6] Inclusion of lexical items seems to be necessary in order to improve the success rate of the parsing algorithm. Taking the extreme case, a tagset of an untagged text is the set of all word forms.

[7]There are cases when this is yet not enough, because an assignment of (bare) syntactic structure does not ensure, for all cases, a total morphological disambiguation.

[8]The influence from the tagset (different) structures remains and is projected in the success rate.

References

Eric Brill. A Simple Rule-Based Part of Speech Tagger. Proceedings of The Third Conference on Applied Natural Language Processing, Trento, Italy, 1992.

Eric Brill. Transformation-Based Error-Driven Parsing. Proceedings of the Twelfth National Conference on Artificial Intelligence, 1993.

Doug Cutting, Julian Kupiec, Jan Pedersen and Penelope Sibun. A Practical Part of SpeechTagger. Proceedings of The Third Conference on Applied Natural Language Processing, Trento, Italy, 1992.

David Elworthy. Tagset Design and Inflected Languages. Proceedings of the ACL Sigdat Workshop, Dublin, Ireland, 1995, pp. 1-9.

Jan Hajič and Barbora Hladká. Probabilistic and Rule-Based Tagger of an Inflective Language - a Comparison. Proceedings of The Fifth Conference on Applied Naturale Language Processing, Washington, USA, 1997, pp. 136-143.

Jan Hajič and Kiril Ribarov. Rule-Based Dependencies. Workshop Notes on Empirical Learning of Natural Language Processing Tasks, Prague, 1997, pp. 125-136.

Lidia Mangu and Eric Brill. Automatic Rule Acquisition for Spelling Correction. ICML, 1997.

Bernard Merialdo. Tagging Text with a Probabilistic Model. Computational Linguistics 20(2), 1992, pp. 155-171.

Kiril Ribarov. Automatická tvorba gramatiky přirozeného jazyka. Msc Thesis. Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 1996.

Beatrice Santorini. Part of Speech Tagging Guidelines for The Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990.

Anne Schiller. Multilingual Finite-State Noun Phrase Extraction. ECAI'96, Budapest, Hungary, 1996.

Pasi Tapanainen. RXRC Finite-State Compiler. Technical report MLTT-20, Rank Xerox Research Center, Meylen, France, 1995.

Appendix

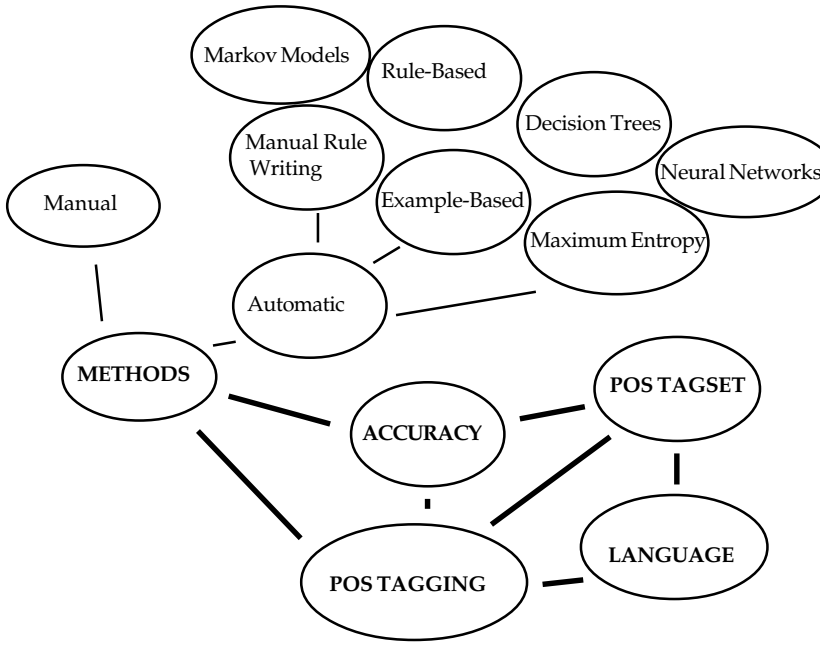


Figure 1

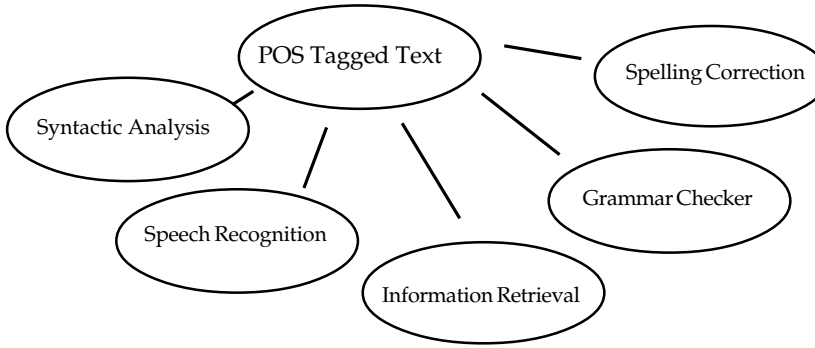
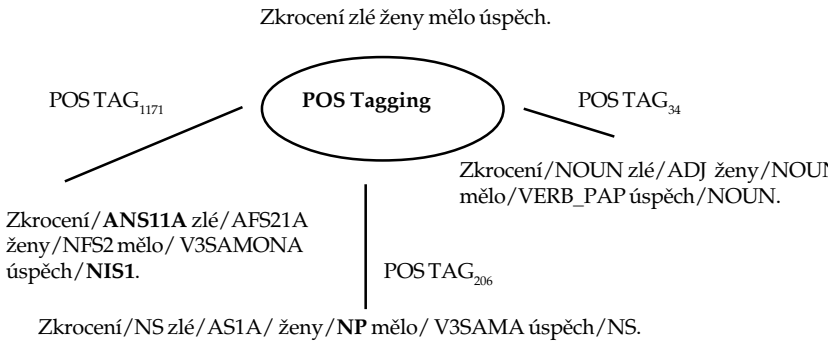


Figure 2



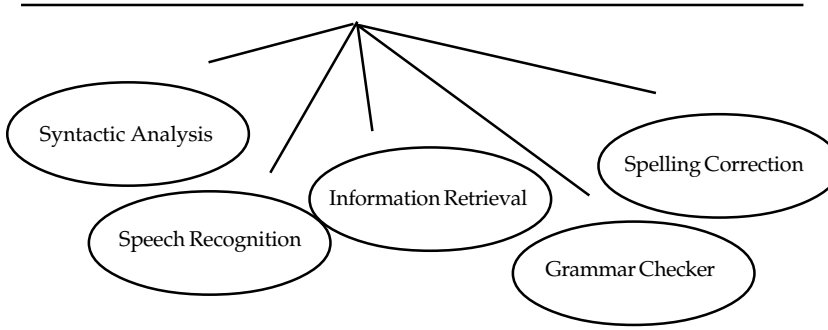


Figure 3

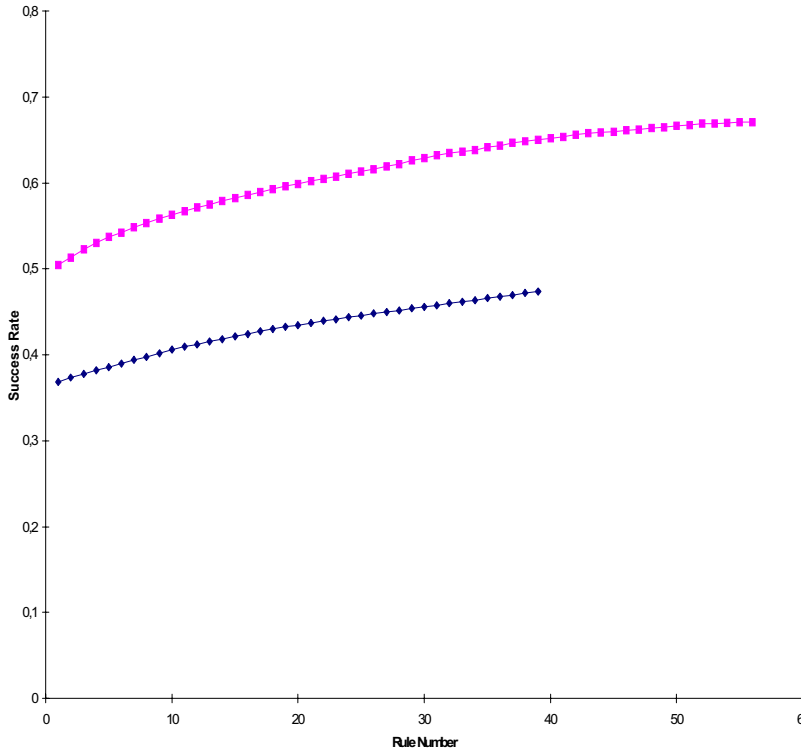

0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Success Rate
0
10
20
30
40
50
60
Rule Number

Figure 4