Developing a POS tagger for Magahi: A Comparative Study

Ritesh KUMAR[1] *Bornini LAHIRI*[1] *Deepak ALOK*[1]
(1) CENTRE FOR LINGUISTICS, Jawharlal Nehru University, India
ritesh78_llh@jnu.ac.in, lahiri.bornini@gmail.com,
deepak06alok@gmail.com

ABSTRACT

In this paper, we present a comparative study of the four state-of-the-art sequential taggers applied on Magahi data for part-of-speech (POS) annotation . Magahi is one of the smaller Indo-Aryan languages spoken in Eastern state of Bihar in India. It is an extremely resource-poor language and it is the first attempt to develop some kind of Natural Language Processing (NLP) resource for the language.

The four taggers that we test are – Support Vector Machines (SVM) based SVMTool, Hidden Markov Model (HMM) based TnT tagger, Maximum Entropy based MxPost tagger and Memory based MBT tagger. All these taggers are trained on a miniscule dataset of around 50,000 words using 33 tags from the BIS-tagset for Indian languages and tested on around 13,000 words. The performance of all these taggers are tested against a frequency-based baseline tagger. While all these taggers perform worse than on the English data, the best performance is given by the Maximum Entropy tagger after tuning of certain parameters. The paper discusses the result of the taggers and the ways in which the performance of the taggers could be improved for Magahi.

KEYWORDS : POS TAGGERS, MAGAHI TAGGER, SVM, TNT TAGGER, MXPOST TAGGER, MBT TAGGER

1 Introduction

Historically, Magahi has been classified in different ways by different scholars. While Grierson (1903) puts Magahi under the Eastern group of Outer sub-branch of Indo-Aryan languages, others like Turner have clubbed the 'Bihari' languages with Eastern and Western Hindi (Masica 1991). A classification given by Chatterji (1926) where Magahi is kept together with other languages of Eastern group which is separate from the Western Hindi. Jeffers (1976) gives a classification which is very similar to that of Grierson.

In the present time, Magahi is spoken mainly in Eastern states of India including Bihar and Jharkhand, along with some parts of West Bengal and Orissa. There are three main varieties of Magahi spoken today (Verma, 1991).

- Central Magahi of Patna, Gaya, Hazaribagh (in Bihar)
- South-Eastern Magahi of Ranchi (in Jharkhand) and some parts of Orissa
- Eastern Magahi of Begusarai and Munger (in Bihar)

Some other scholars like Verma (2003) and Grierson (1903) have also classified South-Eastern and Eastern varieties together.

1.1 Magahi: Socio-Political Situation

Socially Magahi is considered a dialect of Hindi even though historically as well as linguistically Magahi distinct enough from Hindi to be called a distinct language. This social attitude towards Magahi where it is considered a dialect (and inferior/distorted form) of Hindi has emanated largely from the political representation of the language as a dialect (or, Mother Tongue in the Census) of Hindi as well as the close lexical affinity of the two languages (Kumar, et.al., 2011). As a result of this socio-political attitude, Magahi has remained a largely ignored language outside linguistic studies despite the presence of quite a large population (counting up to 13.978,565 according to Census of India, 2001) of Magahi speakers.

1.2 Linguistic Features of Magahi

There has been very few linguistic studies on Magahi. However a basic (although not completely accurate) description of Magahi is given by Verma (2003). A basic description of the linguistic features is given here.

An initial analysis of the Magahi sound system shows that it has 35 phonemic sounds – 27 consonants and 8 vowels. Some of the major phonlogical features which distinguish Magahi from Hindi include absence of word-initial consonant cluster, absence of word-initial glides and absence of word-medial and word-final dental laterals.

Morphologically it is a nominative-accusative, inflected language with almost free word order of constituents within phrases and sentences. Both nouns and adjectives have two basic forms. While one form is the basic form (as in *g*^{h}*ora* 'horse', *sona* 'gold', *ʊɟər* ‘white’, etc), the other one is the derived form (as in *g*^{h}*or-ba*, *son-ma, ʊɟər-ka*). The affixes used in the derived forms are the affixal particles which are used for different linguistic function like specificity, definiteness etc (Alok, 2010, 2012).

Unlike Hindi (which is a Noun Class language with two classes, also equated with Masculine and Feminine gender in the language), Magahi is a classifier language. It has three mensural classifiers – *go/t^{h}o, məni*, *sun* (Alok, 2012). These classifiers encode the information about how the referent is measured and are different from the other generally found classifiers which characterise noun in terms of certain inherent properties (Aikhenvald, 2000, 2006). Among these while *go/t^{h}o* measures nouns in terms of length or discrete quantity, *məni* and *sun* are used for measuring nouns in terms of amount (and so are used with the mass nouns) (Alok, 2012). It is to be noted that broadly these are numeral classifiers since they are always attached with the numeral and quantifiers in a noun phrase and never with the noun itself. The presence of classifiers could prove to be a very strong indicator for the Part-of-speech annotation of quantifiers and Noun in Magahi.

Syntactically, Magahi nouns do not have number and gender agreement with verbs. There are only a few nouns in Magahi which could be inflected for number. Moreover adjectives also agree with such nouns in terms of number as well as sex (it should be noted here that sex here refers to the natural sex of the noun in case of animates and not the Noun class as it is used for Hindi since such agreements could occur only with the animates for which males and females are distinctly recognised in the language). Verbs also agree with subjects in person and honorificity. It is to be noted that verbs could also agree with object as well as addressee honorificity of the object or the addressee are honorific.

1.3 Part-of-speech Annotation and Magahi

Part-of-speech annotation is generally considered the most basic step for developing any kind of NLP application. In the recent times several statistical and machine-learning based approaches have been applied to the task of POS annotation. Some of the major and most successful taggers include - Hidden Markov Models (Brants, 2000), Maximum Entropy taggers (Ratnaparkhi, 1996), Transformation–based learning (Brill, 1994, 1995), Memory–based learning (Daelemans, et. al., 2003), Support Vector Machines (Cortes & Vapnik, 2000) besides several others.

All these taggers are trained and evaluated on the WSJ corpus in English. On this corpus all of these have a very comparable accuracy with each giving only slightly different accuracy from the others. In this paper, we have applied Magahi data to four of these POS tagers, viz, HMM tagger, MaxEnt Tagger, Memory-based Tagger and SVM, for the purpose of developing a POS tagger for Magahi. The idea is to test which of these give the best performance on the given dataset with their default settings. The performance is compared against a Maximum-frequency baseline tagger.

2 Experimental Setup

2.1 Dataset

We have used around 50,000 manually POS-tagged data for training each of the tagger and they are tested on around 11,000 words. The corpus consists of data taken from a collection of Magahi folktales. Since Magahi is largely a spoken language and there is very scant availability of written material, the collection of folktales was the most readily available as well as standardised written data available.

2.2 Tagset

For the annotation of Magahi data, we have used a modified version of BIS standard tagset for Indian Languages. The complete tagset, which consists of 33 tags, is given in Table 1 (the corpus tagged with this tagset is same as described in Kumar, et al. (2011) but the tagset is slightly modified).

Sl. No	**Category**		**Label**	**Annotation Convention**	**Examples (in IPA)**
	Top level	Subtype (level 1)			
1	**Noun**		**N**	**N**	c^{h}ɔːɽɑ (boy)
1.1		Common	NN	N__NN	cəcəriː (a small bridge-like st.) ləŋgte (naked)
1.2		Proper	NNP	N__NNP	p^{h}uləva
1.3		Nloc	NST	N__NST	əgaɽiː, pic^{h}aɽiː
2	**Pronoun**		**PR**	**PR**	
2.1		Personal	PRP	PR__PRP	həm, həməniː
2.2		Reflexive	PRF	PR__PRF	əpəne
2.3		Relative	PRL	PR__PRL	ɟe, ɟekər
2.4		Reciprocal	PRC	PR__PRC	əpəne
2.5		Wh-word	PRQ	PR__PRQ	kɑ, ke
2.6		Indefinite	PRI	PR__PRI	koi, kekrɑ

3	**Demonstrative**		**DM**	**DM**	
3.1		Deictic	DMD	DM__DMD	Ĩhã, ʊ̃hã
3.2		Relative	DMR	DM__DMR	ɟe, ɟəun
3.3		Wh-word	DMQ	DM__DMQ	kekrɑ, kəun
3.4		Indefinite	DMI	DM__DMI	i, ʊ
4	**Verb**		**V**	**V**	ləuknɑ (to see)
4.1		Main	VM	V__VM	p^{h}ĩcnɑ (to wash clothes) əɟ^{h}urɑnɑ (to get entangled)
4.2		Auxiliary	VAUX	V__VAUX	həi, həliː, hət̪^{h}iː
5	**Adjective**		**JJ**	**JJ**	cəkəit^{h} (short and well-built) bət̪p^{h}əros (uselessly talkative)
6	**Adverb**		**RB**	**RB**	cəb^{h}ɑk (with splash) cəb^{h}ər-cəb^{h}r (a manner of eating)
7	**Postposition**		**PSP**	**PSP**	ke, me, pər, ɟore
8	**Conjunction**		**CC**	**CC**	
8.1		Co-ordinator	CCD	CC__CCD	ɑʊ, bɑkiː, bəluk
8.2		Subordinator	CCS	CC__CCS	kɑheki, t̪ə, ki
9	**Particles**		**RP**	**RP**	
9.1		Default	RPD	RP__RPD	t̪ə, b^{h}iː

9.2		Classifier	CL	RP__CL	go, t^{h}o
9.3		Interjection	INJ	RP__INJ	əre, he, c^{h}iː, bɑpre
9.4		Intensifier	INTF	RP__INTF	təhtəh, tuhtuh, b^{h}ək-b^{h}ək
9.5		Negation	NEG	RP__NEG	nə, mət̪, binɑ
10	**Quantifiers**		**QT**	**QT**	ek, pəhilɑ, kuc^{h}
10.1		General	QTF	QT__QTF	təniːsun, d^{h}erməniː
10.2		Cardinals	QTC	QT__QTC	ek, du, igɑrəh
10.3		Ordinals	QTO	QT__QTO	pəhilɑ, dʊsrɑ
11	**Residuals**		**RD**	**RD**	
11.1		Foreign word	RDF	RD__RDF	A word in foreign script.
11.2		Symbol	SYM	RD__SYM	For symbols such as $, & etc
11.3		Punctuation	PUNC	RD__PUNC	Only for punctuations
11.4		Unknown	UNK	RD__UNK	
11.5		Echowords	ECH	RD__ECH	(pɑniː-) uniː (k^{h}ɑnɑ-) unɑ

TABLE 1: Magahi Tagset

2.3 Tagger Tools

We have used the following tools to train different taggers on Magahi data -

- MxPost for Maximum-entropy tagger (Ratnaparkhi, 1996). It uses the contextual features like preceeding words and tags as well as morphological feature of the words like the suffixes and prefixes for tagging any given word.
- MBT for Memory-based tagger (Daelemans, et. al., 2003). It creates two separate taggers after the training. One is used exclusively for known words and it uses the

contextual features and the other is used exclusively for unknown words which also uses lexical information alongwith the contextual features. These features are customisable as per the need of the users

- TnT for HMM-based tagger (Brants, 1994). As the name of the tagger itself suggest (Trigrams 'n' Tags), it uses trigram and the tags of the preceding words as features for training
- SVMTool for SVM-based tagger (Gimenez and Marquez, 2004). This tool provides an interface for using SVM-Light (Joachims, 1999). The features that could be used with this tool is similar to the other tools, viz., morphological features of the word and that of the context as well as the tags of the preceding words.

In general we have used these tools with the default/recommended settings for carrying out the experiments. Each of these tools were trained on exactly the same corpus consisting of around 50,000 tokens.

3 Results and Analysis

The results obtained from the four tools are summarised in Table 2 below

	Known Words (86 %)	**Unknown Words (14%)**	**Overall**
TnT	89.75%	67.57%	86.09%
MBT	89.15%	72.97%	86.22%
MxPost	NA	NA	**89.61%**
SVMTool	81.89%	18.11%	**41.46%**
Baseline	NA	NA	71.18%

TABLE 2: Comparison of the taggers

As mentioned above each tagger was also tested on exactly the same dataset which consisted of around 13,000 tokens. Out of these 13,000 tokens around 86% were known tokens (i.e. they were present in the training set also) while 14% were unknown token (they were encountered by the tagger for the first time in the test set itself). As it is shown in the table, MxPost gives the best overall performance; however since the evaluation results are not calculated separately for known and unknown words the break-up is not known. MBT and TnT gives comparable overall results but MBT is significantly more accurate with unknown words. The most dismal performance is given by SVMTool which could be explained only by an extremely small dataset and presence of a large number of classes which needs to be classified.

An error analysis of the data shows that the major source of error in the annotation is the serial verb constructions in all the three taggers. While MxPost performs slightly better, in general, detection of second verb (if it is a compound verb) or the noun of the second verb complex (if it is a conjunct verb) proved to be very problematic. Another source of error was a complete absence of examples of certain closed-class categories in the training set viz., interjections. Besides these two, as expected, lexical ambiguity was also one of the minor sources of annotation error.

Conclusion and Way Ahead

In this paper we have presented a comparison of the four state-of-the-art POS taggers with respect to their performance on the Magahi data. The best overall accuracy as well as accuracy on the known and unknown words individually is given by the maximum-entropy based MxPost tagger. However the accuracy (just below 90%) is much below the general expected accuracy of POS taggers. As the error analysis shows all the taggers perform poorly on very similar kinds of words. So combining different taggers could not solve the problem. The two steps which could be taken to increase the accuracy of the tagger include

- A list of closed-class words will be prepared which would be able to handle the cases where the absence of the word in training set has led to the error by the tagger.
- Some explicit disambiguation rules will also be used in certain cases where sufficiently large number of examples is not present in the training corpus so as to discriminate in between the contexts of occurrence of a particular tag of the word.

A third possible step could be to increase the training set size (which is in any case pretty small by the general standards). However this is very resource-intensive because of the lack of easily available data in the language. Moreover as per our current analysis a hybrid system like this is expected to give a performance at par with most of the other state-of-the-art POS taggers for Indian languages.

Acknowledgments

We would like to thank our supervisors for their constant support and guidance.

References

Alok, D. (2010). Magahi Noun Particles. Paper presented in *4th International Students' Conference of Linguistics in India (SCONLI-4)*, Mumbai, India, February 20-22, 2010.

Alok, D. (2012). *A language without Articles: The Case of Magahi*. Unpublished M.Phil. Dissertation, Jawaharlal Nehru University, New Delhi

Brants , T.. (2000). TnT — A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied NLP Conference(ANLP-2000)*, pages 224–231.

Brill, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. *Proceedings of AAAI*, Vol. 1, pages 722–727.

Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4): 543–565.

Cortes , C. and Vapnik, V. (1995). Support Vector Networks. *Machine Learning*, 20: 273–297.

Daelemans, W., Zavrel, J., van den Bosch, A., van der Sloot, K. (2003). MBT: Memory Based Tagger, version 2.0, *Reference Guide. ILK Research Group Technical Report Series 03-13*, Tilburg.

Jesus Gimenez and Lluis Marquez. (2004). SVMTool: A general POS tagger generator based on support vector machines. In *4th International Conference on Language Resources and Evaluation*, pages 168–176, Lisbon, Portugal.

Joachims, T. (1999). Making Large-scale SVM Learning Practical. In Schölkopf, B., Burges, C., Smola, A., eds.: *Advances in Kernel Methods – Support Vector Learning*, pages 41–56. MIT Press, Boston, MA, USA .

Kumar, R., Lahiri, B. and Alok, D. (2011). Challenges in Developing LRs for Non-Scheduled Languages: A Case of Magahi. In *Proceedings of the 5th Language and Technology Conference Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC '11)*, Adam Mickiewicz University, pages 60-64.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, pages 133–142.