VARIOUS TAGSETS FOR INDIAN LANGUAGES AND THEIR PERFORMANCE IN PART OF SPEECH TAGGING

[1]NITISH CHANDRA, [2]SUDHAKAR KUMAWAT, [3]VINAYAK SRIVASTAVA

[1,2,3]Department of Computer Science & Engineering, IIT (BHU), Varanasi

Abstract: POS (part of speech) tagging is the processes of anointing each word of a corpus into its appropriate "part of speech" taking into account both its definition and its context. Development of tags which enumerate different part of speech is probably the most important task of pos tagging. For western languages the development of such tags has taken more or less a standard form. The Penn Treebank tagset has emerged as the standard pos tagset for western languages but Indian language scene is much more active. In this paper we do an observation and analysis of the various tagsets developed for pos tagging of Indian languages and suggests features which can make it more reflective of the morphologically rich Indian languages. We also look at the performance of these tagsets in tagging Indian languages. We also look at attempts to promote standardization of tagsets and the continuous improvement in their performance as well as the final standardization and their conformity to EAGLES framework.

General Terms:POS Tagset (Indian Languages IL),POS tagging

Keywords: Part of speech tagset(IL), Part of Speech Tagging.

I. INTRODUCTION

There are four major families of Indian languages – Indo-Aryan (Hindi, Gujarati, Punjabi), Dravidian (Tamil, Kannada), Austro-Asiatic and Tibeto-Burman (Bodo, Manipuri).Of these the Indo-Aryan and Dravidian account for nearly 900 million speakers. While it is true that the two language families (Indo Aryan and Dravidian) have many distinct features, they have a number of similarities as well that allow for a common framework. For example, Dravidian languages are agglutinating in nature while Indo Aryan languages are typologically defined as inflectional. However, we can find some level of agglutination in some of the major Indo Aryan languages like Marathi and Bangla. Both the major language families have pretty much the same structure(subject object verb) and are inflective as well as rich in case morphology. Thus despite strong differences a common framework for Indian Languages (IL) is possible.

Several pos tagsets have been designed by a number of research groups working on Indian Languages viz IIIT (ILMT) tagset, BIS tagset LDC-IL tagset, AU-KBC tagset, Tamil tagset, JNU-Sanskrit tagset (JPOS), Sanskrit consortium tagset (CPOS), MSRI-Sanskrit tagset (IL-POSTS) and CIIL Mysore tagset.

II. ISSUES IN TAGSET DEVELOPMENT

This section deals with various issues regarding Tagset development.

Fineness Vs Coarseness

We need to decide the features which we will consider during the annotation process. We need to decide whether we will consider only the lexical aspect or whether we will mark plurality, gender distinctly. This has an important bearing on the performance of tagger. Though tagsets employing lexical tagsets are much more efficient they sometimes fail to capture some of the basic morph syntactic features of language.

Lexical Vs Syntactic Category

A decision has to be taken whether the word should be tagged according to its lexical category or the syntactic category. Since the lexical category of a word is always the same there is consistency in tagging but a word may belong to different part of speech depending upon its context so even though lexical tagger can be efficiently trained there is loss of accuracy.

New Tags V/S Tags From A Standard Tagger

While deciding the tags for a tagset one can either come up with a totally new tagset or take any standard tagset as reference and make modifications in it according tothe objective of the new tagger.

III. IIIT TAGSET OR ILMT TAGSET

IIIT tagset uses the basic Penn Treebank tagset .It modifies some of the basic tagsets and introduces some other to address the peculiar feature of indian languages. It divides tagsets in three distinct groups.

Group I

This group contains those tagset which are similar to Penn tagset. IIT tagset is a coarse tagset and hence it disregards notion of plurality and gender in its tagset .Some of the tagset of this group like NN(noun), NNP(proper Noun), PRP (pronoun), VAUX (Verb Auxiliary) , JJ (Adjective) RB (adverb) , UH (interjection) are similar to the Penn tagsets. It

introduces two specific tagsets one is CC which is used for conjuncts. The Penn tagsets uses IN for preposition and subordinating tagsets but this tagsets marks all connectors as CC other than prepostions. Another special tagset is SYM which is for foreign symbols in the language like $,% etc. Indian Languages also contain words like bhi.hi these words are tagged as RP.

Group II

This group contains tagsets who are a modified form of Penn Treebank tagset. The Penn tagset has Preposition but in Indian languages we have postposition the tagset for postpositions is called PREP. All Quantifiers in the language are tagged as QF. Any word denoting numbers are tagged as QFNUM. Main verb of a finite verb group of a sentence is considered as VFM. All non-finite verbs which are used as adjectives are tagged as VJJ. Non-finite forms of verbs which are used as adverbs are tagged with VRB. Gerunds will be marked as VNN .All wh words are marked as QW.

Group III

The tags in this group are completely new and address the peculiar features of Indian languages. Indian Languages like Hindi contains words like aage,upar,pahele..These words can behave like adverbs , nouns, and postposition in different contexts .All these words are tagged as NLOC. Words like 'bahuta', 'kama', etc. are tagged as INTF. Negatives like 'nahi', 'na', etc. are marked as NEG. The tag NNC is used for compound nouns. All words except the last one, of compound words are marked as NNC. The tag for compound proper nouns is NNPC and all compound proper nouns are tagged as NNPC excluding the last one.

Kriyamuls are verbs formed by combining a noun or an adjective or an adverb with a (helping) verb. The kriyamuls formed by joining a noun are marked as NVB, those formed with an adjective are tagged as JVB and those formed by joining adverbs are tagged as RBVB.

IV. MSRI TAGSET

MSRI (Microsoft Research India Pvt Ltd) developed this tagset in 2008.This tagset aims at providing a comprehensive tagset that captures as much information as possible from tagging. The guidelines of this tagset contains about 9 categories(Nouns , Pronouns , Verbs , Nominal Modifier , demonstrative , Adverb , Particle , Punctuation and Residual) which branch out in types(such as common , proper , verbal and spatio Temporal) .This tagset has been subdivided into further 14 categories .The various parts of the tagset are connected through dot.

LDC – IL TAGSET

LDC (language Development Consortium) developed tagset is a hierarchical tagset while the previously discussed IIIT tagset is a flat tagset .Flat tagset are a list of mutually exclusive categories and though they are easier to process they cannot capture higher level of granularity .Also if we want to develop such tagsets which are reusable across different linguistic boundaries we have no option but to constantly lengthen the tagset that facilitates addition of new tags. Hierarchical tagsets on the other hand have parameters which can take values for a certain language which can enable it to address the requirements of a group of related languages. The structure for a hierarchical tagset is a top level tagsets which are further split into other bottom level more specific tagsets. The morph syntactic details are encoded in separate layer of hierarchy beginning from the major categories of the top and gradually progressing down to cover morph syntactic features. Decomposability is another feature of a hierarchical tagset design as it allows different features to be encoded in a tag by separate sub-strings. A tag is considered decomposable if the string representing the tag contains one or more shorter sub-strings that are meaningful out of the context of the original tag. Decomposable tags help in better corpus analysis by allowing to search with an underspecified search string.

LDC-IL has 13 top-level categories of tagset and these are Noun (N), Pronoun (P), Demonstrative (D) ,Nominal Modifier(J) ,Verb(V), Adverb (A), Postposition (PP),Particle(C), Numeral (NUM), Reduplication (RDP) , Residual (RD) , Unknown (UNK) , Punctuation (PU).

Table 1. LDC-IL tagset

Category	Types	Attributes
Noun	Common (NC) Proper (NP) Verbal (NV) Spatio-Temporal (NST)	Gender, Number, Case, Distributive, Honorificity, Emphatic dimension
Pronoun	Pronominal (PR) Reflexive (RF) Reciprocal (RC) Relative (RL) Wh (WH)	Gender, Number, Person, Case, Case marker, Distributive, Emphatic, Dimension, Honorificity
Demonstrative (D)	Absolutive (DAB) Relative Demonstrative (DRL) Wh-Demonstrative (DWH)	Number, Case, Dimension, Distributive, Emphatic (not in case of wh)
NominalMod	Adjectives (JJ)	Gender, Number, Case,

ifier (J)	Quantifiers (JQ) Intensifier (JINT)	Numeral, Distributive
Verb(V)	Main verb(VM) Auxiliary verb(VA)	Gender, Number, Person, Tense, Aspect, Mood, Finiteness ,Honorificity
Adverb(A)	Manner(AMN)	Case, Distributive
Post-Position(PP)		Gender, Number, Case marker
Numeral (NUM)	Real (NUMR) Serial (NUMS) Calendric (NUMC) Ordinal (NUMO)	
Residual(RD)	Foreign Word (RDF) Symbol (RDS)	
Unknown		
Punctuation(PU)		

Among other tags are particle(C) which include Co-ordinating (CCD), Subordinating (CSB), Interjection (CIN),Dis) Agreement (CAGR),Emphatic (CEMP),Topic (CTOP) Delimitive (CDLIM), Honorific (CHON) ,Dedative (CDED) ,Exclusive (CEXCL), Interrogative (CINT)Dubitative (CDUB), Similative (CSIM) and Others (CX).

V. BUREAU OF INDIAN STANDARD(BIS) TAGSET

The Bureau of Indian Standards (BIS) Tagset has recommended the use of a common tagset for the part of speech annotation of Indian languages. The tagset, incorporating the advice of the experts and the stakeholders in the area of natural language processing and language technology of Indian languages, has to be followed in the annotation tasks taking place in Indian languages after August, 2010. The annotations taking place under the Indian Languages Corpora Initiative (ILCI) program is following the BIS tagset as proliferated. Many tags in the BIS tagset and the previously discussed LDC-Il tagset are the same .In addition to one type of a category. It also introduces another subtype .Besides it groups together unknown, punctuation and residual in one tagset. Nouns, Pronouns, Demonstratives are the same. All these categories have same subtypes in both BIS standard and LDC-IL standard .Verb(V) has the same two subtypes main verb(VM) and auxiliary verb(VAUX) but here main verb is further divided in finite(VF) ,non-finite(VNF) , gerund(VNG) and infinitive(VINF). Adjective and adverb has no subtype whereas we have two new categories one is called conjunction (CC) which has three subtypes namely coordinator (CCD), subordinator (CCS) and quotative (UT) .These subtypes were grouped under particle (RP) in LDC-IL tagset. Consequently particle(RP) in BIS contains default(RPD) ,classifier(CL),interjection(INJ),Intensifier(INF) and Negation(NEG) as its subtype. There is no category called numerals instead its place is taken by Quantifier(QT) which has Generals(QTF),cardinal(QTC) and ordinal(QTO) as its subtypes.

Except for the three categories of adjective, adverb and postposition, all the categories have some two or more sub-categories. The category of residual is though not part of the language, it is part of the text which is to be annotated. Therefore, this category also has extra-linguistic elements appearing in the text sub-categorized.

The BIS framework is laid out in a hierarchy of two levels:

(i) Categories are the highest level part-of-speech classes. All categories are Obligatory, that is, are generally universal for all languages and hence, must be included in any morph syntactic tagset derived from the framework.

(ii) Types are sub-classes of categories and are Recommended, that is, are recognized to be important sub-classes common to a majority of languages.

The tagset of each word contains two levels which follow the hierarchy of BIS tagset. The tag consist of first level i.e top level followed by each of the two levels whichever is deemed necessary for tagging.

VI. STANDARIZATION

Some of the earlier POS tagsets mentioned previously were designed for English (Greene and Rubin, 1981; Garside, 1987; Santorini, 1990) and remain in popular usage even today. However, even though they were designed for the same language, they differ significantly from each other so that a corpus tagged by one is totally incompatible with the other. Further, as these are English-specific they cannot be reused for any other language without substantial changes .Leech and Wilson (1999) espoused the case of standardization of tagset for their reusability of anointed corpora and interoperability across different languages. The result of their effort was EAGLES guidelines .To achieve the same results BIS has been adopted as the standard for Indian languages.

VII. PERFORMANCE IN TAGGING

IIIT Tagset

Language	No of Words	Words Tagged	Correctly tagged	% Precison(P)	% Recall(R)
Hindi	5677	5677	4515	79.5	79.5
Bengali	5029	5029	3742	74.4	74.4
Telugu	6098	6098	3547	58.2	58.2

Languages	Tagging Accuracy(%)
Hindi	76.34
Bengali	72.17
Telugu	53.17

LDC IL Tagset

Serial No	language	2008-09	2009-10	2010-11	Total words Tagged
1.	Hindi	Tagged set Creation	30,000+	~50,000	84,962
2.	Bengali	Tagged set creation	25,000+	~50,000	75,397
3.	Tamil	Tagged set creation	30,000+	~50,000	88086

VIII. BIS TAGSET

LDC-IL has shown an accuracy of 84.2% .Besides the languages given many more languages has been tagged according to the LDC-IL tagset

Serial No	Languages	Previous Data(before 31.3.2013)	Current Data(1.4.2013 onwards	Total
1	Hindi	233347	410406	643753
2	Bengali	212456	67450	279876
3	Tamil	174488	1202369	1376857

Part of speech tagging using BIS exhibits an accuracy of 88.2%.

It is clear from the result that BIS and LDC-IL tagset performance is better than IIIT (ILMT) tagset. .IIIT (ILMT) tagset represents initial phase in the development of Indian language tagset, the use of flat tagsets and ignoring the morphosyntactic features were the limiting factors of the tagset.

IX. IMPROVEMENT AND ISSUES IN BIS TAGSET

Indian languages have some peculiar features which still cannot be caught properly by tagsets.

2.1 Case of NLOC

The practice of marking some words like upar, niche, aage, piche as NLOC presents some difficulites.Most of the time these words can be followed by ke, ka, se. whereas sometime they occur alone like wo aage se upar gaya This sentence has two words belonging to NLOC category aage and upar .Both these words will be tagged as NLOC in BIS tagset but they are certainly different parts of speech. Moreover there are various similar words like hamesha, aksar.which do not occur by definition in this category but which imply time.

2.2 The Ambigous cases

A lot of ambiguity can occur in Indian languages .Consider the following two sentences Wo laal kila dekhne gaya

Wo laal rang dekh ke daar gaya Kya hua mere laal?

All the three sentences contain the word laal .In the first sentence laal kila is a proper noun, in the second sentence it is an adjective while in the third case it is agin Noun.A POS tagger can catch only one of these multiple cases.

A similar case of ambiguity can occur between other part of speech like verb and adjective Beheta jharna, Pani behata rahta hai

2.3 Case of Sandhi

Indian languages are agglutinative in nature. Two or more words join together to form words which have have alteration of sound at boundaries. Words like sarvottam (sarva+uttam) Here the word sarva is of type Quantifier if it were to exist independently but we rarely use the word sarva independently in Hindi, we almost always use it in conjoined form. Same is the case for words like suputra, kuputra .words like these rarely occur independently. Since these words do not fall under conventional POS categories one has to do a lot of manual work to correctly classify them.

2.4 The case of idioms

Indian language contains a lot of idioms and proverbs .If these words were to be classified according to their part of speech will pose a serious problem in further translation of these phrases. There should be some category of tags which classify all of the phrases as one tag.

2.5 The case of adverbs

The BIS tagset only considers manner adverbs for example Yeh kaam thoda muskil hai Here muskil can be easily tagged as adverb according to BIS tagset, however words like hamesha ,aksar do not fall under this category as defined by BIS.

2.6 IMPROVEMENTS

BIS tagset by any means is not perfect. It is still not comprehensive and unambiguous so its needs constant fine-tune. BIS tagset can be improved in the following way

- Expand the definition of Adverb tagset
- Incorporate tagsets to mark all the words of a proverb or idiom together
- If more than one main verb is found in a sentence let it be distinguished as VM1 and VM2
- List of words under NLOC must be increased.
- Add more tagsets which take into focus the peculiar part of speech of indian language like alankar,etc

ACKNOWLEDGMENTS

We are greatly indebted to Dr Vnayak Srivastava, Assistant Professor at Department of Computer

Engineering, Indian Institute of Technology, (BHU), Varanasi for his able and dedicated guidance and encouragement throughout the project .We would also like to thank Head, Department of Computer Engineering, Indian Institute of Technology, (BHU), Varanasi for allowing us to use the resources of the department in our study and research.

REFERENCES

[1] Sankaran Bhaskar, Kalika Bali. A Common Parts-of-Speech Tagset Framework for Indian Languages..

[2] Pattabhi R K Rao ,Vijay Sundar Ram R,Vijaykrishna R and Sobha LA Text chunker and hybrid pos tagger for Indian languages.

[3] Abbhi A (2001) A manual of Linguistic Fieldwork and Structures for Indian languages.

[4] Bali K Choudhary,M Biswas ,P Choudhary Indian Language Part of Speech Tagset Hindi (LDC catalogue number 2010T24[ISBN: 1-58563-71]

[5] Narayan Choudhary, Pinkey Nainwani,, Ritesh Kumar, Esha Bannerjee ILCI Parts of Speech Annotation Guidelines

[6] htttp://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

[7] Manish Srivastava ,Pushpak Bhattacharya Hindi POS tagger using Naïve Stemming: Harnessing morphological information without extensive linguistic knowledge

[8] Edna Vaz 1, Shantaram V. Walawalikar2, Dr.Jyoti Pawar3, Dr.Madhavi Sardesai BIS annotation standard with refrence to Konkani language

★★★