// Tagset developed at IIIT - Hyderabad after consultations with
// several institutions through two workshops.

A Part of Speech Tagger for Indian Languages (POS tagger)

Introduction:

The significance of large annotated corpora in the present day NLP is widely known. Annotated corpora serve as an important tool for investigators of natural language processing, speech recognition and other related areas. It proves to be a basic building block for constructing statistical models for automatic processing of natural languages.

Many such corpora are available for languages across the world and have proved to be a useful step towards natural language processing.

Looking at the scenario for Indian languages, not much work has been carried out in the front of automatic processing of Hindi or any other Indian language. The main bottleneck being unavailability of an annotated corpus large enough to experiment statistical algorithms.

Annotation of corpora can be done at various levels viz, Part of Speech, Phrase/clause level, dependancy level, etc. Part of speech tagging can form a basic step towards building an annotated corpus. This level of tagging can be further extended to higher levels of annotation.

Objectives:

This paper mainly aims at arriving at a standard tagset for part of speech annotation for all Indian languages. It gives a detail description of the various tags used and elaborates the motivations behind the selection of these tags.

This paper also discusses various issues that need to be addressed while preparing a tagset for a POS tagger and how we have tried to solve each one of these in the current tagger.

Later the paper also gives a short analysis of some phenomena which are specific to Indian languages and how this tagger has tried to deal with these.

Organization of the paper:

This paper has been divided in three sections. Each of these sections discusses the three objectives of the paper.

Section I Various issues involved in deciding a tagset for a POS tagger.

Section II Detailed description of various tags designed for Indian languages.

Section III Discussion on some phenomena specific to Indian Languages.

Section I:

This section deals with some of the issues related to any POS tagger and the policy that we have adopted to deal with each of these issues in the current tagger.

In case of a POS tagger, the major issues that need to be dealt with are:

1. Fineness V/s Coarseness in linguistic analysis
2. Syntactic Function V/s lexical category
3. New tags V/s tags from a standard tagger

1. Fineness V/s Coarseness

An issue which always comes up while deciding a tagset for annotating any text is of 'fineness' or 'coarseness' in linguistic analysis. In case of a POS tagger a decision has to be taken whether the tags will account for finer distinctions of the various features of the parts of speech. In other words it has to be decided if plurality, gender and such other information will be marked distinctly or only the lexical category will be marked.

In this tagset we have avoided 'fine' distinction. The motivation behind this is to have less number of tags since less number of tags leads to efficient machine learning. Also accuracy of manual tagging is higher when the number of tags is less.

But we also agree that too coarse an analysis is not of much use. Essentially, we need to strike a balance between fineness and coarseness. The analysis should not be so fine as to hamper machine learning and also should not be so course as to loose important information.

The results of an experiment conducted in the center, shows how more number of tags hampers machine learning.

??In this experiment, initially the noun class was divided into NO (noun oblique) and ND (noun direct). The learning data was of 20K words.

So in case of nouns we have not taken into account the plurality information i.e. we have not marked noun singular and noun plural with different tags. Nouns in Indian languages also inflect for case (direct and oblique). This information too has not been marked distinctly.

In case of adjectives and adverbs we have avoided the distinction between comparative and superlative forms.

Also, we feel that fine distinctions are not relevant for many of the applications (like sentence level parsing, dependancy marking, etc.) for which the tagger may be used in future.

However we agree that plurality and such other information is crucial if the POS tagged corpus is used for any application which needs the agreement information. In case such information is needed at a later stage, the same tagset can be extended and can encompass plurality or case information also. This can be done by providing certain heuristics or linguistic rules.

Thus broadly we have adopted a coarse part of speech analysis.

But finer analysis is made wherever it is found essential.

e.g. VJJ, VRB, VNN

For example, in case this tagger is used for dependancy parsing or any such application, it is essential to preserve the information that the given word is a noun or an adjective or an adverb which is formed from a verb. This helps in understanding the various arguments that this verbal noun/adjective/adverb can take.

In other words the non-finite forms of verbs which are used as nouns or adjectives or adverbs still retain their property of being a verb, for example even in the participle form they can take their own arguments in a sentence. In such a case it is essential that the tag for this participle word indicates that the word is a form of a verb, so that its dependancy structure may be established.

```
e.g. AsamAna/NN meM/PREP uDane/VNN vAlA/PREP ghoDA/NN nIce/NLOC
      "sky"      "in"    "flying"            "horse   "down"
utara/VFM AyA/VAUX.
"climb"   "came"
```

In the above sentence 'uDane' is VNN, it is a noun formed from a verb. And 'AsamAna' is an argument of this verb and not the main verb of the sentence. So in order to preserve such crucial information a finer analysis is essential and so a distinct tag is used for gerunds. Similarly different tags are used for adjectival and adverbial participles too.

2. Syntactic Function V/s lexical category

The part of speech of a word in a sentence can be different from its lexical category.

```
e.g. uttara - noun (lexical category)
    ("North")
uttara bhArata me bhArI varRA HuI. - adjective (syntactic category)
("north" "India" "in" "lots" "rain" "happened")
```

??(jEse)

In such cases a decision has to be taken whether the word should be tagged according to its lexical category or the syntactic category.

In this tagger, the syntactic function of the word is not given as much importance. Pure lexical category of the word is taken into consideration while tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also the machine is able to establish a word-tag relation which leads to efficient machine learning.

3. New tags V/s tags from a standard tagger

While deciding the tags for a tagger one can either come up with a totally new tagset or take any other standard tagger as a reference and make modifications in it according to the objective of the new tagger. The later option is often better to adopt because it is expected that the end users of the tagger are familiar with the standard tagset for one language. They can easily adopt a similar tagset for a new language than a totally new one. It saves time in getting familiar to the new tagset and then work on it.

The Penn tagset is a standard tagset used for English. Many tagsets designed after these have been a variant of this tagset (e.g. Lancaster tagset). So while deciding the tags for this tagger, the Penn tagset has been used as a benchmark. Since the Penn tagset is an established tagset for English, we have extended it to another family of languages. Certain changes have been made in the tags only to make it suitable for Indian languages. Wherever the Penn tags are found to be inadequate for Indian languages, either new tags are introduced or existing tags are modified.

e.g. NVB - new tag for Kriyamuls or light verbs.
QW - modified tag for question words.

Section II:

This section gives the rationale behind each tag used in this tagger.

All the tags used in this tagger are broadly classified into three types. There are some tags which have been adopted with some minor changes in the Penn tagset. They are grouped into one group. The second category of tags is of those which are a modification over the Penn tagset. The last group is of all those tags which are not present in the Penn tagset. They have been designed to cater some phenomena which are specific to Indian languages.

Group 1-

All tags in this group are similar to the Penn tagset. Penn tagset makes finer distinction between singular and plural or comparative and superlative forms, which is not considered in the current tagger. This is in accordance with our policy about fineness and coarseness.

NN Noun

Penn tagset makes a distinction between noun singular and noun plural. As mentioned earlier, this distinction is avoided here. This reduces the number of tags and thus enhances machine learning. Plurality is not crucial information with respect to dependancy level parsing or any other higher level analysis of the sentence. As said before if that information is needed at a later stage it can be incorporated with the help of heuristics and linguistic rules.

NNP   Proper Nouns

This tag is also similar to the Penn tagset. Here too we have not made a distinction between Proper Noun singular and Proper Noun plural as in the Penn tagset.

PRP   Pronoun

Penn tagset makes a distinction between personal pronouns and possessive pronouns. This distinction is avoided here. All pronouns are marked as PRP. In Indian languages all pronouns inflect for all cases (accusative, dative, possessive etc.) Incase we have a separate tag for possessive pronouns; new tags will have to be designed for all the cases. This will increase the number of tags which is unnecessary. So, only one tag is used for all pronouns.

VAUX   Verb Auxiliary

All auxiliary verbs will be marked as VAUX. This tag has been adopted as such from the Penn tagset.

JJ   Adjective

This tag is again same as in Penn tagset. Penn tagset also makes a distinction between comparative and superlative adjectives. This has not been considered here.

So this tag includes the 'tara'(comparative) and the 'tama' (superlative) forms of adjectives in Hindi.
e.g. adhikatara, sarvottama, etc.
   ("more times", "best")

RB   Adverb

This tag is the same RB tag of Penn tagset. Penn tagset also makes a difference between comparative and superlative adverbs, which is not adopted in this tagger. This is in accordance with our philosophy of coarseness in linguistic analysis.

RP   Particle

In Indian languages words like bhI, sA, etc. (Hindi for example) will be marked as RP.

CC   Conjuncts (coordinating and subordinating)

The tag CC will be used for coordinating and subordinating conjuncts both. The Penn tagset has used ‘IN’ tag for prepositions and subordinating conjuncts. Their rationale behind this is that subordinating conjuncts and prepositions can be distinguished because subordinating conjuncts are followed by a clause and a prepositions by a noun phrase.

But in the current tagger all connectors other than prepositions will be marked as CC.

UH   Interjection

Just as in Penn tagset, interjections will be marked as UH. In addition the affirmative word 'HAz'("yes") will also be tagged as UH. This is the only example of such a word so has been clubbed under Interjection.

SYM   Special Symbol

All those words which cannot be classified in any of the other tags will be tagged as SYM. This tag is similar to the Penn 'SYM'. Also special symbols like $, %, etc are treated as SYM. Since the frequency of occurrence of such symbols is very less in Indian languages, no separate tag is used for such symbols.

Group 2-

This group includes those tags which are a modification of some tags in the Penn tagset.

??PREP   Postposition

All Indian languages have the phenomenon of postpositions. Some languages separate the post positions from the noun e.g. Hindi. In such a case, a postposition will be marked as PREP.

For example in Hindi, kheta/NN meM/PREP ("the field"/NN "in"/PREP), here meM is the postposition and is written separately from the noun. So it will be tagged as PREP.

But in Marathi (another Indian language), mulAne/NN("boy by"/NN), here the postposition is written along with the noun. So it will not be tagged separately.

This tag is the same as the IN tag used for prepositions in Penn tagset. But it has been adopted for a parallel concept in this tagger. Postpositions of Indian languages have more or less the same functions as prepositions in English.

The same tag is used by Penn tagset for subordinating conjuncts also. They feel that subordinating conjuncts and prepositions can be distinguished because subordinating conjuncts are followed by a clause and prepositions by a noun phrase. But as pointed out earlier, in this tagger all conjuncts have been clubbed under the tag CC.

QF   Quantifiers

All quantifiers like kama, jyAdA, bahuwa, etc. will be marked as QF. In case these words are used in constructions like 'baHutoM/NN ne/PREP jAne se inkAra kiyA'("many" "by" "to go" "refused") where it is a noun, it will be marked as noun. Quantifiers of number will be marked as below.

QFNUM   Quantifiers Number

No distinction will be made between cardinal and ordinal numbers. Any word denoting numbers will be tagged as QFNUM. Penn tagset has a tag CD for cardinal numbers and they have not talked of ordinals!

VFM   Verb Finite Main

The entire verb category has been dealt with differently in this tagger. The following discussions explain how the verbal category has been dealt with.

The VFM tag is a modification of the VB tag of Penn tagset. Main verb of a finite verb group of a sentence is considered as VFM. Whether the form of the particular word is finite or non-finite it will be tagged as VFM.
E.g. laDZakA seba khA/VFM raHA thA.
    ("boy" "apple" "eating" "was")

VJJ   Verb Non-Finite Adjectival

Unlike Penn tagset all non finite verbs which are used as adjectives will be marked as VJJ. The Penn tagger does not make a distinction between the gerunds and adjectival participles or simple 'ing' type verb forms.

For Hindi, constructions like 'khAte Hue' will be tagged as follows:
khAte/VJJ Hue/VAUX.
("eating")

As explained earlier in the paper, this distinction is made in order to preserve the information that this word is a form of a verb. Every verb is capable of taking its own arguments in a sentence, even if it is not the main verb. In order to be able to show the exact verb-argument structure in the sentence, it is essential that this crucial information is preserved. So this tagger marks all non-finite adjectival participles as VJJ i.e. an adjective which is formed out of a verb.

VRB   Verb Non-finite Adverbial

Again unlike Penn tagset, non-finite forms of verbs which are used as adverbs will be tagged with a different tag VRB.

In Hindi constructions like 'khAte khAte'("while eating"), 'khAkara'("after eating"), etc will be tagged as VRB.
The reason for this distinction between non-finite verbs used as adverbs and other verbs is as explained in VJJ.

VNN   Verb Non-Finite Nominal

In the Penn tagger, VBG is used for gerunds, participles and progressive verb forms. But this tagger will mark gerunds as VNN. This

distinction is being made in order that consructions like 'pIna', etc can be accounted for.
e.g. sharAba pInA/VNN seHata ke liye KAnikAraka HE.
("liquor" "drinking" "heath" "for" "harmful" "is")

QW Question Words

The Penn tagset makes distinction between the wh words which act as questions, as relative pronouns and as determiners. But in this tagger all wh words (ka'kAra's in Hindi) will be tagged as QW. The reason being, in Indian languages the category where 'wh' words act as pronouns or determiners is not present. They all become pronouns like 'jo', 'jisne', etc. in Hindi
e.g. The man who wrote a book ... (vaHa AdamI jisne kItAba likhI ...)
("that" "man" "who" "book" "wrote")

Group 3-

This set is of new tags designed to cater some phenomena which are specific to Indian languages.

NLOC Noun Location

This is an entirely new tag introduced to cover an important phenomenon of Indian Languages. Words like 'Age', 'upara', 'pahele', 'bAda', etc. are used in various ways in Hindi.

1. They act as a postposition along with 'ke'
e.g. ghade ke upara thAlI rakhI HE.
("pot" "on" "plate" "kept" "is")

Here 'ke upara' is a post position which is the direct equivalent of the English preposition 'on'.

2. They also act as adverbs.
e.g. tuma upara jAo.
("You" "up" "go")

Here 'upara' is an adverbial of place.

3. These words also take post positions themselves and so in some sense behave like nouns.
e.g. vaHa upara se AyA.
("He" "above" "from" "came")

4. As pointed out in 3. above, these words take postpositions and act as arguments of the verb in the sentence. And they also take a post position to join with a another noun. So in that sense also they behave like nouns.
e.g. upara kA HissA
("above" "of" "portion")

To tag such words one option is to tag them according to the category to which they belong in the given sentence. For example in 1. above, the word is occurring as a postposition so can be marked as a postposition. In example 2. above, it is an adverb so can be marked as an adverb and so on.

But we feel that these words are more like nouns as is evident from 3. and 4. above, and also if we consider for examples, 'aage', 'upara', etc. as places which are in front, up, etc then we can tag them as nouns.

But these are not pure nouns. They are nouns which indicate a location or time. These also function as adverbs or prepositions in a context. So a new tag NLOC is introduced for such words. This tag will cater to a finite set of such words.
set: (Age, piche, upara, nIce, bAda, pahele)
("front", "behind", "above", "below", "before")

Such words if tagged according to their syntactic function, it will hamper machine learning. So a single tag, NLOC has been devised for such words which indicate location and time.

INTF Intensifier

This tag is not present in Penn tagset. Words like 'baHuta', 'kama', etc. will be covered under this.

NEG Negative

Negatives like 'nahI', 'na', etc. will be marked as NEG. Penn tagset does have a separate tag for this. ??

NNC Compound Nouns

There is no separate tag for Compound nouns in the Penn tagset. But in this tagger, the tag NNC is used for compound nouns. This tag has been introduced in order to indentify unhyphenated compound words as one unit.
e.g. 'keMdra sarakAra' will be tagged as keMdra/NNC sakakAra/NN.
("center" "government")

In this example, 'keMxra' and 'sarakAra' are both nouns which are forming a compound noun. All words except the last one, of compound words will be marked as NNC. Thus any NNC will be always followed by another NNC or an NN. This strategy helps indentify these words as one unit although they are not conjoined by a hyhen.

NNPC Compound Proper Nouns

This tag is also an addition. All words in a compound proper noun will be marked as NNPC excluding the last one.
e.g. aTala/NNPC biHArI/NNPC vAjapeyI/NNP.

Here the first two words are NNPC and the last one will be NNP. Just as the NNC tag this tag too helps identify a compound proper noun as one unit and not confuse it with a list of proper nouns.
e.g. rAma, moHana aur shAma ghara gaye.
("Ram", "Mohan" "and" "Shyam" "home" "went")

Any title like Dr., Col., Lt. etc. which occurs before a proper noun will be tagged as NNC. All such titles are nouns which will always be followed by a Proper Noun. To indicate that these are a part of the proper noun but are nouns they will be tagged as NNC.

e.g. Col./NNC Ranjit/NNPC Deshmukh/NNP

NVB,JVB,RBVB   Kriyamula (light verbs)

This tag has been introduced to account for the concept of kriyamuls of Indian Languages. 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 will be NVB, those formed with an adjective will be JVB and those formed by joining adverbs will be RBVB.

e.g. snAna/NVB karatA/VFM HE/VAUX
("bath" "does")

In the above example 'snAna' is a noun which is joined to the verb 'karanA' to express the sense of the verb 'to bathe'. So here 'snAna' is marked as NVB and the main verb is marked as VFM and 'HE' is its auxilliary.

e.g. lAla/JVB HotA/VFM HE/VAUX
("red" "happens")

In this example the adjective 'lAla' is joined with 'HonA' to express the sense of the verb 'to redden'. So 'lAla' is marked as JVB, 'HotA' as VFM and 'HE' as VAUX.

e.g. yaHa to jarUra/RBVB HE/VFM........

In this example the adverb 'jarUra' is joined with 'HonA' to express the sense 'to be sure'. So 'jarUra' is marked as RBVB and 'HE' is the main verb marked as VFM.

Section III:

This section gives the details of some phenomena in Indian languages which need to be dealt with separately in the tagger. These are issues that one comes across while dealing with Indian languages. And these cannot be handled by just changing or adding tags.

1. Reduplication - This is a phenomenon in Indian languages where the same word is written twice for emphasis.

e.g. choTe choTe -
("small" "small" - very small)
There are two ways in which these words may be written. One is they are separated with a space or sometimes they are separated with a hyphen.

When these words are written with a space in between, the same tag is used for both the words.

e.g. dhIre/RB dhIre/RB, choTe/JJ choTe/JJ, galI/NN galI/NN
("slowly slowly" "small small", lane lane")

But when they are written with a hyphen, they will be tagged as one word.

e.g. galI-galI/NN
   ("lane-lane")
Exceptionally if the hyphen is written with a space then they will be marked with the same tag as in the earlier case.
e.g. galI/NN - galI/NN

2. "vAlA" type constructions - 'vala' is a kind of suffix used in Hindi and some other Indian languages. It conjoins with nouns or verbs to form adjectives. It may also be used rarely as an adverb.

e.g. lAla kamIjZa vAlA AdamI, meHanawa karane vAlA, etc.
    ("red" "shirt" "one with" "man", "efforts" "doing" "one")

This suffix is sometimes is written separately or may also be written together.

Case I- lAThI vAlA
("stick" "one with")
In the case when it is written separately as in 'lAThI vAlA', the word 'lAThI' will be tagged as NN and the word 'vAlA' will be tagged as PREP.

The whole expression 'lATI vAlA' is an adjective, in which 'lATI' is a noun and 'vAlA' is a suffix which makes it an adjective. But since these words are written separately, they will have to be tagged individually. We will treat 'vAlA' like a postposition and tag it as PREP.

In the second case when these words are written together viz. 'lAThIvAlA', then the whole word will be marked as JJ since 'lAThIvAlA' is an adjective.

Case II- karane vAlA
As said earlier 'vAlA' also joins a verb in its nominal form and makes it an adjective. In that case the verb will be marked as VNN and 'vAlA' as PREP.

Again, if the two words are written together then the word will be marked as VJJ. Here we need to mark the information that the word is an adjective which is formed from a verb so that at later stage if we want to annotate the argument structure of the sentence we have the information that the word if a form of a verb and can have its own arguments (discussed earlier in the paper).

Case III- mEM wo karane vAlA HI thA.
(close translation: I was going to do it shortly.)

Here although the word 'karane' has a 'vAlA' suffix, the entire expression is not an adjective but an adverb. In this case too we will mark these words as VNN and PREP. Here again we stand by our policy that the tag will be decided on the basis on the part of speech and not on the basis of the category of the word in the given sentence(syntactic function). This avoids confusion at the level of manual tagging and aids machine learning. So the tags remain same although the function of the words is different in two different places, it is adjective in case I and II and adverb in case III.

3. Honorific particles like 'jI', 'sAHaba' etc. - Hindi and some other Indian languages have a practice of adding a particle 'jI' or 'sAHaba' etc. after proper nouns or personal nouns. They are added when the speaker wants to give respect to the person he is referring to in his speech. Such particles will have the tag RP, the same tag for particles of Penn tagset.

4. As any other language Hindi too has many loan words. Such foreign words will be tagged as per the syntactic function of the word in the sentence.

Future Work

References:

Appendix 1:

List of tags used in this tagger:

NN Noun (laDaZkA, nadI, vicAra, kaThoratA)
("boy", "river", "thought", "hardness")

NNP Proper Noun (rAma, bhAjapA)
(Ram, BJP)

PRP Pronoun (jo, vo, vaHa,"jisa" laDaZke ne, jisane)
("who", "that", "he", "the boy who", "by whom")

VFM Verb Finite Main (vaHa "pItA" HE, vaHa laDaZkA "HE")
("he drinks", "he boy is")

VAUX Verb Auxillary (khA/VFM cukA/VAUX HE/VAUX)
("eat en has")

VJJ Verb NonFinite Adjectival (khAte/VJJ Hue/VAUX)
("eating")
-- negative VJJN :telugu

VRB Verb NonFinite Adverbial (khAkara, pIte/VRB Hue/VAUX)
("after eating", "drinking")
-- negative VRBN :telugu

VNN Verb NonFinite Nominal (pInA)
("drinking")
-- negative VNNN :telugu

JJ Adjective (includes comparative and superlative forms also, adhikatara, sarvottama)

RB Adverb (dhIre/RB dhIre/RB, tejI/RB se/RP)
("slowly slowly", "fast")

NLOC Noun location (upara, Age, pahele, bAda)

PREP Postposition (ne, ke/PREP liye/PREP)
("by", "for")

RP Particle (mIThA sA/RP, taka/RP, HI/RP, to/RP, bhI/RP)

CC Conjunct (Ora, yA, ki)
("and", "or", "that")

QW Question Words (kyA/QW, kEsA/QW)
("what", "how")

QF Quantifier (jyAdA/QF, thoDA/QF, saba/QF, kama/QF, baHuta/QF)
("more", "little", "all", "much")

QFNUM Number Quantifiers (tIsarA, tInoM, tIna)
("third", "three"(oblique), "three")

INTF Intensifier ("baHuta" jyAdA, "Ora" jyAdA) # But note that: [baHutoM/noun ne]
("too much", "much more")

NEG Negative (nA, naHIM)
("no", "not")

Compounds Nouns

NNC Compound Common Nouns (kendra/NNC sarakAra/NN ("center government", rAma/NNC
moHana/NN ("Ram, Mohan"),
laDaZke/NNC laDaZkiyAz/NN ("girls boys"),
laDaZke/NNC laDaZkiyoM/NN ne khAnA khAyA
(girls boys food ate").)

NNPC Compound Proper Nouns (aTala/NNPC biHArI/NNPC vAjapayI/NNP,
shrI/NNC pI./NNPC ke./NNPC mishrA/NNP)

Kriyamula

NVB Noun in kriya mula
(snAna/NVB karatA/VFM HE/VAUX)
(snAna/NVB karate/VJJ Hue/VAUX)
(snAna/NVB karake/VRB)
(snAna/NVB karane/VNN para/PREP)

JVB Adj in kriya mula
(lAla/JVB HotA/VFM HE/VAUX)
(pUrA/JVB HotA/VFM HE/VAUX)
(pUrA/JVB Hote/VRB Hue/VAUX)
(pUrA/JVB Hokara/VRB
(pUrA/JVB Hone/VNM para/PREP

RBVB Adv in kriya mula
In case there is such a usage with xxxx
(xxxx/RBVB HotA/VFM HE/VAUX)

UH Interjection words (HAM and interjections)

SYM Special: Not classified in any of the above