ASSESSMENT AND DEVELOPMENT OF POS TAG SET FOR TELUGU

Dr.Rama Sree R.J
Rashtriya Sanskrit Vidyapeetha,
Tirupati
rjramasree@yahoo.com

Dr.Uma Maheswara Rao G
Central University
Hyderabad
guraohyd@yahoo.com

Dr. Madhu Murthy K.V
S.V.U.College of Engineering
Tirupati
kvmmurthy@yahoo.com

ABSTRACT

In this paper, we first had a overall study of existing POS tag sets for European and Indian languages. Till now, most of the research done on POS tagging is for English. We observed that even though the research on POS tagging for English is done exhaustively, part-of-speech annotation in various research applications is incomparable which is variously due to the variations in tag set definitions. We understand that the morpho-syntactic features of the language and the degree of desire to represent the granularity of these morpho-syntactic features, domain etc., decide the tags in the tag set. We then examined how POS tagset design has to be handled for Indian languages, taking Telugu language into consideration.

1. Introduction

Annotation is the process of adding some additional information (grammatical features like word category, case indicator, other morph features) about the word to each word of the text. This additional information is called a tag. The set of all these tags is called a tag set. When words are considered in isolation, they can have one or more number of tags for each word. But when these words are used in a certain context, the tags representing morphological and syntactic feature reduce to one tag. The information to be captured as a tag is an application specific issue (Anne,1997, David, 1994 and David, 1995). A number of tag sets have been evolved for a number of languages. These tag sets not only differ with each other from language to language, but vary within the language itself. The reasons for the variation of tags in the tag sets are as follows. As taggers give additional information like grammatical features such as number, gender, person, case markers for noun inflections; tense markers for verbal inflections, the number of tags used by different systems varies depending on the information encoded in the tag. However the tag set design plays a vital role when data is tagged according to it and hence it affects the development of NLP application tools within and across that language. Language independent representation of a tag set help to find out the hidden information like context, structure, syntactic and semantic aspect of the word. It also gives an overview of language modeling features.

2. Desirable Features of a Tag Set

Unfortunately, there does not seem to be much literature on standard tag set design. There is a need to have standard tag set labels for the words to encode the same linguistic information across the languages. The tag set labels of a given language should satisfy the following characteristics.

(1) The words carrying same syntactic, categorical information should be grouped under the same tag. For example, all adjectives should be tagged as JJ.

(2) The words which have same syntax and come under different categories should clearly be distinguished depending on the categorical sense in which it is used in the given context. For example, the word **book** can be tagged both as noun (NN) and Verb (VB).

(3) The tag set should also help us to classify and predict the sense, category of the unknown and foreign words. For example, consider the sentence, "Give it to xyxxy", POS tagger should be in a position to predict **xyxxy** (or any non-sensical string) could be a noun.

3. Sources of Variations among POS Tag Sets for English

In order to identify the reasons for tag set variations for English, the tag sets viz., the Penn Treebank (Mitchel, 1993) tag set(PT), UCREL CLAWS7 tag set (UCREL_C7), the International Corpus of English (ICE) tag set (Greenbaum, 1992) and the Brown Corpus (BC) tag set (Green, 1997) for English are examined; the POS tag labels are extracted for some important morpho-syntactic features and studied to demonstrate the present study.

After a careful study, the following points were observed with regard to the differences in POS tag sets.

(i) Desire to capture more semantic content: BC, ICE, URCEL tag set are making more subtle distinctions within one category than PT. For example, POS tags for adjectives- PT is not making any clear distinction for adjectives other than JJ, JJS, JJR, whereas other tag sets are maintaining fine granularity. Such differences can be observed for several morpho-syntactic features.

(ii) Corpus Coverage: Depending on the syntactic distribution of the test corpus under consideration, there may be variations. For example, BC tag set made a wide provision for foreign words (not shown in the above table). In British corpus, there may be a possibility of the presence of the test corpus where more number of words are borrowed from other languages into English.

(iii) Desire for precision: The reason for more number of tags in a tag set is to precisely capture all linguistic criteria which describe morpho-syntactic features in detail. However, there should be a balance between theoretical and actual distribution of these syntactic features.

4. Tag Sets for Indian Languages

The two POS tag sets developed for Hindi (revised on Nov 15, 2003) and Telugu by IIIT, Hyderabad and CALTS, Hyderabad respectively are examined and **t**he following points are observed.

Telugu POS tag set contains more number of POS tag labels. This difference is due to the reason that Telugu is more inflective than Hindi. In Hindi nouns are non-inflectional. *Karaka* roles are not encoded in Hindi noun word forms as in Telugu. Similarly main verbal roots appear as non-inflective in Hindi. The verbs co-occur with tense, aspect and modality as separate words whereas aspect and modality are packed into a single verbal inflection word in Telugu. For example, consider the following sentences.

English: **Ram killed Ravana**.

Hindi : **RAm ne mArA Ravana ko**.

Telugu: **RAmudu caMpAdu RAvanunni**.

For convenience, the word order is maintained as it is in all the three languages. In case of English language, position gives the roles played by Rama (subject) and Ravana (object). In case of Hindi, case markers *ne* and *ko* exist, but they do not inflect Ram and Ravana. But Telugu noun inflections give the information of case markers also. Hence there are differences in the tag labels of Hindi and

Telugu language tag sets. In order to capture these syntactic (more over they are also semantic) information, Telugu has more number of POS tags (nn1,nn2,nn3, nn4,nn5,nn6,nn7) in the place of a single tag (nn) of Hindi.

The POS tags of Telugu are described below in detail.

(i) Nouns (nAma vAcakAlu- nn) :These tags capture the nouns and their roles played in the sentence. The different tags in the subclass are *nn1,nn2,nn3,nn4,nn5, nn6* and *nn7*. Depending on the *vibhakti*, the nouns get the number label to main class, i.e., *nn* based on the *karaka* relations. The tag *nni* stands for noun oblique form indicating that the noun is in a position to get attached with the succeeding noun inflection.

(ii) Locative affixes (swAna vAcakAlu – nl) : Here some locative prepositions combined with the six *vibhaktis* are listed as nl1 (pEna-పైన), nl4 (pEki-పైకి), nl5 (pEnuMdi-పైనుండి), nl6 (pEni-పైని) etc.

(iii) Prepositions (Vibhakti – pp): Sometimes prepositions can occur independently. For example, varaku (వరకు). Hence all vibhaktis are labelled as *pp1,pp2* etc.

(iv) Pronouns (sarva nAmAlu - pr) : Like nouns, all pronouns form inflections with *vibhaktis*. Accordingly they are named as *pr1*, *pr*2 etc.

(v) Adjectives (Visheshana) : Special type of adjectives like Verbal adjectives (*kriya visheshana*) as vjj, Nominal adejctives (*saMjna viseshana*) as jj and noninfinitive verbal adjectives (*sahAyaka asamapaka kriya*) as ajj etc.

(vi) Other syntactic categories : The tags for other syntactic categories like quantifiers as *qf*, negative meanings as *ng* etc., are given.

5. Improvement of Telugu Tag Set

In addition to the above mentioned tags, some new tags are introduced to capture and provide finer discrimination of the semantic content of some of the linguistic expressions a corpus of 12,000 words. They are explained briefly in the succeeding paragraphs.

(a) Verbal finite negative : Some words like *kAxu* (కాదు), *lexu* (లేదు) are verbal finites but they give the negative meaning of the verbal action. If they are tagged simply as vf, it is undestood that some action has taken place. But these words are used in negative sense. In order to capture this feature, we have labelled them as **vng**.

(b) Verbal nouns wih vibhakti: Verbal nouns behave in the same way as nouns do, in forming their inflections with vibhaktis like *AdataM*-(ఆడటం), *Adatanni*-(ఆడటాన్ని), *Adatamcewa (*ఆడటం చేత) etc. At present they are labelled as *nn1, nn2, nn3* etc. depening on the affix. In doing so, the semantic content of verb is lost. This would lead to difficulties in disambiguating words at the semantic level. Hence the introduction of POS tags like *vnn1, vnn2* etc is proposed.

(c) Words expressing doubts: There are linguistic expressions that express the doubtfulness as explained below.

Doubtfulness of :

(i) **Verbal finites**:Words like *uMxo* (ఉందో) *vunnavo* (ఉన్నవో) etc., express the doubtfullness of the occurrence of action. To capture this semantic discrimination, POS tag *vfw* is introduced. Previously they are labelled as *vf*.

(ii) **Nouns:** Words which express the doubtfulness a noun participation in the action like *rAmudo* (రాముడో), *axo* (అదో)

etc. Instead of labelling them *nn1*, they are labelled them with the tag *nnw*.

The above mentioned improvements made to the existing POS tag sets and the advantages thereof are as follows.

(i) A finer discrimination is made. For example consider *vfw*. In the absence of this tag, the verbal inflections which end with *lexu* (లేదు) could be tagged as *vf*.

Due to this, the verbal inflections which are completed can be clearly distinguished from theose verbal inflections where action has not been completed.

(ii) *vnn* tag captures more information that the noun present in the verbal inflection is just a simple common noun. In the absence of this tag, words erroneously labelled as nn to which it does not really belong. So these tags accurately capture the information present in the words.

6. Conclusion

It is strongly felt that all Indian languages should have the same tag set so that the annotated corpus in corresponding languages may be useful in cross lingual NLP applications, reducing much load on language to language transfer engines. This point can be well explained by taking analogy of existing script representation for Indian Languages. The ISCII and Unicode representations for all Indian languages can be viewed appropriately in the languages we like, just by setting their language code. There is no one-to-one alphabet mapping in the scripts of Indian Languages. For example, the short e,o (ఎ,ఒ)

are present in Telugu, while they are not available in Hindi, Sanskrit etc. Similarly alphabet variations between Telugu and Tamil exist. Even then, all these issues are taken care of, in the process of language to language script conversion. Similarly POS variations across Indian Languages also should be taken care of.

References:

Anne Schiller, Simone Teufel, Christine Thielen. 1995. **Guidelines für das Tagging deutscher Textcorpus mit STTS.** Universitäten Stuttgart und Tübingen.

David Elworthy. 1994. **Automatic error detection in part of speech tagging.** In Proceedings of the International Conference on New Methods in Language Processing, Manchester.

David Elworthy. 1995. **Tagset Design and Inflected Languages**. In Proceedings of the ACL SIGDAT Workshop, Dublin.

Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz. 1993. **Building a Large Annotated Corpus of English: The Penn Treebank.** Computational Linguistics. Volume 19, Number 2, pp. 313--330 (Special Issue on Using Large Corpus).

Greenbaum S. 1992**. The ICE tag set manual.** University College London.

Green B, Rubun G. 1971. **Automated Grammatical Tagging of English**. In Department of Linguistics, Brown University.