14

WHAT SHOULD BE AND WHAT SHOULD NOT BE? DEVELOPING A POS TAGSET FOR KASHMIRI

Aadil Amin Kak

Nazima Mehdi

Aadil Lawaye

INTRODUCTION

Part of speech tagging has been studied extensively in the past two decades. The fundamental problems in PoS tagging task stem from the fact that a word can take different lexical categories depending on its context. The tagger has to resolve this ambiguity and determine the best sequence for a sentence. PoS tagging also known as morphosyntactic categorization or syntactic word class tagging (Halteren 1999) is the process of assigning a part of speech or other lexical class marker to each word in a corpus. Tags are also applied to punctuation markers; thus tagging for natural language is the same process as tokenization for computer languages, although tags for natural languages are much more ambiguous.

BACKGROUND STUDY

Given the prominence of the USA both in linguistics and in computing technologies, the earliest work on tagsets in the 1960s and early 1970s occurred in the US and focused on English. The most important tagsets of this earliest period are those of Klein and Simmons (1963) and Greene and Rubin (1971). Over the course of time, sequence of tagsets for English have been devised such as the Penn tagset and CLAWS tagset including the series $C_{1,}C_{2,}C_{5,}C_{7.}$ The publication of EAGLES recommendations for morphosyntactic annotation of corpora (Leech and Wilson, 1996) was an earliest attempt to develop a common tagset guidelines for several

European languages. The objective of EAGLES guidelines was to standardize the tagsets used in different languages to achieve cross- linguistic compatibility, reusability and interchangeability. For Indian languages several tagsets have been developed also like the one developed under ILMT guidelines, which is designed for specific languages in a flat structure capturing only coarse-level categories. Another tagset which is designed for Indian languages is that of IL- POSTS hierarchical framework. IL-POSTS is a framework for ILs that allows language specific tagsets to be derived from it. An important consideration for its hierarchical structure and decomposable tags is that it should allow users to specify the morphosyntactic information applicable at the desired granularity according to the specific language and task. IL-POSTS framework is laid out in a hierarchy of three levels:

1. Categories

2. Types

3. Attributes

POS tagging is typically achieved by rule-based systems, probabilistic data-driven systems, neural network systems or hybrid systems. For languages like English or French, hybrid taggers have been able to achieve success percentages above 98 %.(Schulze *et al*, 1994).

PART OF SPEECH IN KASHMIRI

Syntactically, Kashmiri is an interesting language showing both verb medial and verb final characteristics. Kashmiri also shows strong V_2 features like Germanic and Icelandic. Words in Kashmiri are either monomorphic or polymorphic.

Three morphological classes have been established in Kashmiri on the basis of morphological and syntactic criteria viz:

a. NOUNS, PRONOUNS AND ADJECTIVES

Examples

NOUNS

کِتاب (book) **کمی□ز** ,(shirt), **کُرسی** (chair)

PRONOUNS

یُہ(He), سو (She) بہ (I), تِم(They)

ADJECTIVES

خراب(bad), جان(Good)

b. VERBS

کھیٚون(to eat), شونٛگُن(to sleep), پَرُن(to read) ,پَکُن (to walk)

c. PARTICLES

Adverbs, Post-positions, conjunctions, Interjections, Emphatic-particles fall under this category.

ADVERBS

وارِ وارِ(slowly), اصٕل پأٹھۍ (in a good manner/ nicely), ہُتیٚتھ(there)

پَتھ کُن(towards back),بوٚن کُن (towards down)

POST POSITIONS

منٛز (inside) اَندر(inside) تَل (under) پیٹھ (on)

CONJUNCTIONS

تہٕ(and) مَگر (but) یا (if) زَنتہٕ(as if)

INTERJECTIONS

آہا(oh) واہ (oh) آلے (oh) واہ واہ (vow)

The above given word class (open and closed both) make the overall part of speech for Kashmiri. However, it’s important to mention over here that there are certain other word classes which fall into the part speech for Kashmiri like equatives and vocatives for example:

EQUATIVES

کھوتہٕ(than)

نِشہٕ(in comparison of)

ہِش(like)

VOCATIVES

ہتسا ہے(hey hello)

ہے بایہٕ(hey brother)

ہیا(hey)

Thus keeping all these word classes into consideration a tentative tagset for Kashmiri is proposed. The tagset proposed here is based on ILMT guidelines keeping in view the requirement for standards and the need for interoperability.

ILMT GUIDELINES

ILMT is a project in which a number of institutes have come together to form a consortium and work towards developing MT systems for various Indian language pairs. The guidelines provided by ILMT are designed in such a way so that they can be easily used for any Indian language. The tagset provided by them is based on three main criteria viz;

1. FINENESS VS COARSENESS IN LINGUISTIC ANALYSIS

It was decided to come up with a set of tags which avoids 'finer' distinctions. The motivation behind this is to have less number of tags since less number of tags lead to efficient machine learning. Further, accuracy of manual tagging is higher when the number of tags is less. The analysis should not be so fine as to hamper machine learning and also should not be so coarse as to miss out important information. It is also felt that fine distinctions are not relevant for many of the applications (like sentence level parsing, dependency marking, etc.) for which the tagger may be used in future.

2. SYNTACTIC FUNCTION VS LEXICAL CATEGORY

In AnnCorra, the syntactic function of a word is not considered for POS 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. In short, it was decided that syntactic and semantic/pragmatic functions were not to be the basis of deciding a POS tag.

3. NEW TAGS VS TAGS FROM A STANDARD TAGGER

The Penn tags have been used as a benchmark for ILMT guidelines. Since the Penn tagset is an established tagset for English, ILMT have used the same tags as the Penn tags for common lexical types. However, new tags have been introduced wherever Penn tags have been found inadequate for Indian language descriptions.

The overall number of tags present in the ILMT tagset is 21. The list of which is given below:

POS TAG SET FOR INDIAN LANGUAGES

S .No.	Category	Tag name	Example
1.1	Noun	NN	
1.2	NLoc	NST	
2.	Proper Noun	NNP	
3.1	Pronoun	PRP	
3.2	Demonstrative	DEM	
4	Verb-finite	VM	
5	Verb Aux	VAUX	

6	Adjective	JJ	
7	Adverb	RB	*Only manner adverb
8	Post position	PSP	
9	Particles	RP	bhI, to, hI, jI, hA.N, na,
10	Conjuncts	CC	bole (Bangla)
11	Question Words	WQ	
12.1	Quantifiers	QF	bahut, tho.DA, kam (Hindi)
12.2	Cardinal	QC	
12.3	Ordinal	QO	
12.4	Classifier	CL	
13	Intensifier	INTF	
14	Interjection	INJ	
15	Negation	NEG	
16	Quotative	UT	
17	Sym	SYM	ani (Telugu), endru (Tamil), bole/mAne (Bangla), mhaNaje (Marathi), mAne (Hindi)
18	Compounds	*C	
19	Reduplicative	RDP	
20	Echo	ECH	
21	Unknown	UNK	

Table 1. IIIT-H Tagset

USING ILMT GUIDELINES FOR KASHMIRI

ILMT guidelines can be used for Kashmiri but there are some idiosyncratic features of the language which need to be included in the tagset for Kashmiri. The idiosyncratic features include:

1. APPEARANCE OF CASES WITH NOUNS AND VERBS

For Example

Nouns

میز	(table)	vs	میزٕس	(dat table)
اِنسان	(man)	vs	اِنسانٕس	(dat man)
کاغز	(paper)	vs	کاغزٕس	(dat paper)
کُلۍ	(trees)	vs	کُلین	(dat trees)
شُرۍ	(children)	vs	شُریو	(erg children)

Example

Verbs

یُن	(to come)	vs	ینُک	(of/ coming)
بناوُن	(to make)	vs	بناوُنک	(of/ making)
روزُن	(to stay)	vs	روزنٕس	(of/staying)
کرُن	(to do)	vs	کرنٕس	(of/doing)

Such verbs (infinitives) change their meaning when they take post position and cases (dative, genetive) with them. In such situations they mostly behave as gerunds and in some cases they retain their category that is verb.

2. TAGS FOR ADVERB OF PLACE AND ADVERB OF TIME

Only manner adverbs are tagged as adverbs not time and place in ILMT guidelines. Time and place adverbs are tagged as PRP(pronouns). But such adverbs don't behave as pronouns at all. So they need to be tagged as adverbs in the proposed tagset.

For example

ADVERBS OF TIME

از (Today) پگاہ (Tomorrow) وونۍ (now) سُبحن (in the morning)

ADVERBS OF PLACE

یێتہِ (here) تتہِ (there) نیٚبرِ (outside) ہیٚرِ (upstairs)

WORD BOUNDARIES (TOKEN DIVISION)

Mostly in Kashmiri a single morphological word is divided into two tokens i.e. a single word is separated by white space and in tagging every orthographic space is considered as a word break even if it occurs within a lexical word.

Example:

مَدِ مُقابلہِ (in opposition)

سَدا بہار (evergreen)

دُپۍ زِ(should say)

مُقہٕ باز (boxer)

بُتھۍ لاگے (face)

This again is a problem and for such words two separate tags are needed in order to tag such words i.e. for such words first part will be tagged according to the category which it possesses and second part of the same word will be tagged as *“POW”* (part of the word).

4. SOME ADDITIONAL TAGS NEEDED FOR A WORD CLASS LIKE VOCATIVES

Example:

ہَتسا ہے (hey hello)

ہے (hey)

ہَیا (hey)

PROPOSED TAGSET FOR KASHMIRI

S.No	Category	Tag name	Example
1	Noun	NN	
2	NLoc	NST	
2.a	Noun with cases	N case/??	
3	Proper Noun	NNP	
4	Pronoun	PRP	
5	Demonstrative	DEM	
6	Verb-finite	VM	
6.a	Verb with cases	V case/N case??	
7	Verb Aux	VAUX	
8	Adjective	JJ	
9	Adverb	RB	

10	Post position	PSP	
11	Particles	RP	
12	Conjuncts	CC	
13	Question Words	WQ	
14	Quantifiers	QF	
15	Cardinal	QC	
16	Ordinal	QO	
17	Intensifier	INTF	
18	Interjection	INJ	
19	Negation	NEG	
20	Sym	SYM	
21	Compounds	XC	
22	Reduplicative	RDP	
23	Echo	ECH	
24	Unknown	UNK	
25	Vocative	Voc??	
26	Equatives	Eqt??	
26	Part of word	POW??	

Table 2. Proposed Tagset for Kashmiri

CONCLUSION

The proposed tagset here is the first experience in developing an annotation scheme for Kashmiri. It is important to mention over here that, while a tagset is the most prerequisite resource for tagging, there are other significant resources which must be in place prior to the development of an automated tagger.

REFERENCES

Bailey, T.G. 1956. *Teach Yourself Urdu*. New York: David McKay Company.

Baker, J.P, A. Hardie, A.McEnery, and B.D.Jayaram. 2003. *Corpus Data for South Asian language processing*. Paper presented at the Corpus Linguistics 2003 Conference. Lancaster.

Barz, RK. 1977. *An Introduction to Hindi and Urdu*. Canberra: Australian National University Press.

Bhatia, T.K. and A. Koul. 2000. *Colloquial Urdu*. London: Routledge.

Brill, E and M. Pop. 1999. "Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging". S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D.Yarowsky. (eds.). *Natural Language Processing Using Very Large Corpora*. Dordrecht: Kluwer Academic Publishers.

Butt, M. 1995. *The Structure of Complex Predicates in Urdu*. Stanford, California: CSLI Publications.

Chae, Y-S and K-S. Choi. 2000. "Introduction of KIBS (Korean Information Base System) Project". M. Gavrilidou, et al. (eds.). *Second International Conference on Language Resources and Evaluation: Proceedings*. Vol. 3: 1731-17353. Athens: European Language Resources Association.

Kellogg, S.H. 1875. *A Grammar of the Hindí Language, in Which are Treated the High Hindí, Braj, and the Eastern Hindí of the Rámáyan of Tulsí Dás*. (Reprinted 1965.) London: Routledge and Kegan Paul.

Khoja, S, R. Garside and G. Knowles. 2001. *A Tagset for the Morphosyntactic Tagging of Arabic*. Paper presented at the Corpus Linguistics 2001 conference, Lancaster.

Leech, G. 1997. “Grammatical Tagging”. R.Garside, G.Leech and A. McEnery. (eds.). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Masica, C.P. 1991. *The Indo-Aryan languages*. Cambridge: Cambridge University Press.

Merialdo, B. 1994. “Tagging English Text with a Probabilistic Model”. *Computational Linguistics*. 20 (2): 155-171.

Schulze B.M. *et al*. 1994. “Comparative State-of-the-art Survey and Assessment of

General Interest Tools”. *Technical Report D1B – I, DECIDE Project*. Stuttgart: Institute for Natural Language Processing.