ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

Rule Based Approach for the Transition of Tagsets to Build the POS Annotated Corpus

Purva S. Dholakia[1], **M. Mohamed Yoonus**[2]

Senior Research Assistant, LDC-IL, Central Institute of Indian Languages, Mysore, India[1]

Senior Lecturer cum JRO, LDC-IL, Central Institute of Indian Languages, Mysore, India[2]

Abstract: Parts-of-Speech (POS) tagging is the one of the most basic as well as challenging task of NLP. Creation of annotated corpus is very indispensable for the technology development for natural languages. In this paper we describe our experience of developing POS annotated Corpus for Guajarati. This paper aims at the comparison of two POS tagsets – Linguistic Data Consortium for Indian Languages (LDC-IL) tagset and the Bureau of Indian Standard (BIS) tagset for Gujarati language and the transition of LDC-IL tagset to BIS tagset. It also focuses on issues which we have come across during POS tagging. Finally, the paper illustrates the results of the rule based tagset transition and its importance in retaining two versions of POS annotated corpus.

Keywords: Corpus, Gujarati, Natural Language Processing, Tagset, POS, Rule-Based Approach, Transition

I. INTRODUCTION

The process of classifying words into their parts-of-speech and labelling them accordingly is known as parts-of-speech tagging, POS tagging, or simply tagging. Parts-of-speech are also known as word class or lexical categories. There are two factors determining the syntactic category of a word (a) the word's lexical probability (word without context) (b) the word's contextual probability. Hence it disambiguates the parts of speech of a word when it occurs in different contexts. For any POS work the tagset of the language has to be developed. It contains the major tags and the morpho-syntactic features called sub tags.

The collection of tags used for a particular task is known as a tagset. There are many different tagsets which are being used for existing corpora; these tagsets vary according to the objectives of specific projects. BIS has come up with a standardized scheme for Indian languages that can be customized according to the characteristics of a language. Therefore, it becomes essential for all ongoing annotation projects to follow these standards. Ignoring the already created resources and annotating new corpora from scratch is not only requires tremendous effort, time and money, but also leads to underutilization of existing resources. Therefore, the need of hour is to utilize the already annotated corpus. Hence, transition from one annotation scheme to another becomes essential.

In this paper, we have tried to bring forth comparative analysis of both tagsets and tagging issues. In some situations, however, we need to first compare and then map the two existing tagsets and use the transition rules to get two kinds of annotated corpus. This paper describes the approach which maps morpho-syntactic tagset (LDC-IL tagset) to a partially layered tagset (BIS Tagset). The transition of tagsets can be through rapidly with the help of computer program written in any programming languages. The transition relies on a manually written set of transition rules, which is automatically transferred from one tagset into another. The rule-based approach is one which uses manually prepared rule list and it assigns the appropriate tags in the given corpora with respect to constraint checking value. The entire of the transition can be accomplished using the tagset transition tool. The tool has been developed in C# using visual studio 2008.

This paper has eight sections. Section 2 gives a brief summary of tagset and its comparison that are available for part of speech. Section 3 describes overview of LDC-IL and BIS POS tagsets. Section 4 gives information about comparison of tagsets. Section 5 addresses POS tagging issues for Gujarati. Section 6 expresses tagset transition rule. Section 7 shows the results of transition. Finally, future directions are briefly considered in the conclusion, section 8.

II. BACKGROUND

Due to the unavailability of large annotated corpus, not much work has been carried out in different Indian languages scenario. The Indian languages are morphologically rich and generating a standard tagset framework for POS tagging is very difficult. Most of the work in POS tagger for Indian Language has been done in Hindi, as described by Shrivastava et al. (2005). Particularly for Gujarati language, the task of POS tagging is carried forward by LDC-IL project, and recently in the project of ILCI (Indian Language Corpora Initiative), the task of POS tagging for Guajarati is in progress.

A POS tagset design should take into consideration all possible morpho-syntactic categories that can occur in a particular language or group of languages (Hardie, 2004). Some effort has been made in the past, including the EAGLES (Expert Advisory Group on Language Engineering Standards) guidelines for morpho-syntactic annotation (Leech and Wilson, 1996) to define guidelines for a common tagset across multiple languages with an aim to capture more detailed morpho-syntactic features of these languages.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

For Indian Languages, several tagsets have been developed. The most prominent among those is that developed under ILMT (Indian Languages Machine Translation) 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 Indian Language Part of Speech Tagsets (henceforth, IL-POST). IL-POST is a hierarchical framework which allows language specific tagset to be derived from it.

III. POS TAGSET: OVERVIEW

The LDC-IL tagset followed for Gujarati POS annotation is based on the ILPOST framework. This framework facilitated language specific customization based on writing conventions, cross linguistic generalizations, reusability across languages as well as application specific customization.

ILPOST-Gujarati is a hierarchical tagset based on the ILPOST framework. The tagset has three layers. The top layer has morphological categories followed by the types of the category in the middle layer. The bottom layer has morpho-syntactic features or attributes of the type of the category. The top layer of the category layer has a fixed set of grammatical classes to which a token to be tagged belongs to. The type layer has a subclasses of the categories based on the form and function of the token.

The attribute layer provides a set of morpho-syntactic features. These attributes are based on the category and its type. Unlike the category and the type layer, attribute layer has multiple morpho-syntactic features. BIS tagset is designed for the standardization in the area of morpho-syntactic annotation for all the Indian Languages. It has category level, sub-type level 1 and sub-type level 2.

IV. TAGSET COMPARISON

Tagset comparison tends to either emphasize the internal quality of a tagset, i.e., whether it can be tagged accurately, or the external quality, i.e., whether it makes important linguistic distinctions (D´ejean, 2000, sec. 2 & 7), and such methods generally either require sophisticated machinery or complete manual evaluation.

LDCIL tagset has 14 main categories while BIS tagset has 11 main categories, out of this 7 categories on top level are similar as Noun , Pronoun, Demonstrative, Adverb, Postposition, Particle, and Residual. In the LDC-IL tagset, under the category of Nominal Modifier, it has Adjective, Quantifier and Intensifier as sub categories while in BIS tagset Adjective and Quantifier are in the separate category and Intensifier is covered under Particle category. Table I shows the Comparison of Gujarati LDC-IL V0.3 and the BIS POS Labels.

TABLE I
Comparison of LDC-IL and the BIS POS Labels

LDC-IL Labels			BIS Labels		
Category	Subcategory	Tag	Category	Subcategory	Tag
Noun		N	**Noun**		N
	Common	NC		Common	NN
	Proper	NP		Proper	NNP
	Verbal	NV		~~Verbal~~	~~NNV~~
	Spatio-temporal	NST		Nloc	NST
Pronoun		P	**Pronoun**		PR
	Pronominal	PPR		Personal	PRP
	Reflexive	PRF		Reflexive	PRF
	Reciprocal	PRC		Reciprocal	PRL
	Relative	PRL		Relative	PRC
	Wh-pronoun	PWH		Wh-word	PRQ
				Indefinite	PRI
Demonstrative		D	**Demonstrative**		DM
	Absolutive	DAB		Deictic	DMD
	Relative Demonstrative	DRL		Relative	DMR
	Wh-demonstrative	DWH		Wh-word	DMQ
				Indefinite	
Verb		V	**Verb**		V
	Main Verb	VM	**Main**		VM
				~~Finite~~	~~VF~~
				~~Non-finite~~	~~VNF~~
				~~Infinitive~~	~~VINF~~
				~~Gerund~~	~~VNG~~
	Auxiliary Verb	VA	**Auxiliary**		VAUX
Nominal Modifier		J	**Adjective**		JJ
	Adjective	JJ			
Adverb		A	**Adverb**		RB
	Manner	AMN			
Postposition		PP	**Postposition**		PSP
	Case	PPC			
	Non-Case	PPNC			
Particle		C	**Conjunction**		CC
				Co-ordinator	CCD
	Co-ordinating	CCD		Subordinator	CCS
Particle		C	**Particle**		RP
	Subordinating	CSB		Default	RPD
	Interjection	CIN		Interjection	INJ
	(Dis)agreement	AGR		Intensifier	INTF
	Emphatic	EMP		Negation	NEG
	Topic	TOP			
	Delimiting	DLIM			
	Honorific	HON			
	Negative	NEG			
	Exclusive	EXCL			
	Terminative	TERM			
Nominal Modifier		J	**Quantifier**		QT
	Quantifier	JQ		General	QTF
				Cardinal	QTC
				Ordinal	QTO
Residual		RD	**Residual**		RD
	Foreign Word	RDF		Foreign word	
	Symbol	RDS		Symbol	
Punctuation		PU		Punctuation	
Unknown		UNK		Unknown	
				Echo words	
Reduplication		RDP			
Additional tags in LDC-IL tagset					
Participle		L			
	Present	LPR			
	Past	LPS			
	Future	LFU			
Numeral		NUM			
	Real	NUMR			
	Serial	NUMS			
	Calendric	NUMC			
	Ordinal	NUMO			

LDC-IL tagset has the Verbal Noun because generally it is derived from verbs and called as gerund. Usually, Verbal Noun, form wise it is a verb but functions as a noun as it is inflected for case, gender, number, and person and also followed by postposition.

But in the BIS tagset, Verbal Noun as a sub category is there only for Dravidian languages, for those languages which have word forms derived from the verb but it is frozen as noun form. For example in Hindi words like 'likhai', 'padhai' are derived from verbs but those have become as frozen form as a noun. We have not found much difficulty in transition of the sub category of noun except verbal nouns.

In LDC-IL tagset, under the category of Pronoun, there is a sub category called Pronominal Pronoun and in BIS tagset we have Personal Pronoun as sub type. Although Pronominal Pronoun has broad connotation while Personal Pronoun reflects specific association.

The rest of subtypes of Pronoun category have been mapped as it as BIS tagset where as we mapped Pronominal Pronoun as Personal Pronoun. In BIS tagset under Pronoun category one more sub type is added that is Indefinite Pronoun.

A demonstrative is that Pronoun that has a deictic function. So, it will always be followed by a Noun, a Pronoun or an Adjective. Under Demonstrative category in LDC-IL tagset, there are sub categories called Absolutive, Relative and Wh-demonstrative. While in BIS tagset we have Deictic, Relative, Wh-demonstrative, and Indefinite Demonstrative. We mapped Absolutive Demonstrative as Deictic Demonstrative. Rest of sub categories have been mapped as it is to BIS tagset.

LDCIL tagset having the two broader subcategories or types of the verb that is verb main and verb auxiliary, within these verbs we are having attributes that covers gender, number, person, tense, aspect, mood, finiteness, honorificity. While in BIS tagset under the Verb category we have customized only subtypes Verb Main and Verb Auxiliary and not attributes.

Not only do the adjectives modify the nouns but Quantifiers and Intensifiers also function as modifiers of Nouns, Adjectives and Verbs too, so that we have mentioned the category types as Adjective, Quantifier and Intensifier under Nominal Modifier. While in BIS tagset Adjective and Quantifier are separate categories and Intensifier has been covered under Particle category.

In LDC-IL tagset, there is a category called Participle which has subtype of present, past and future participle and it has further attributes. Participle category is not there in the BIS tagset.

In the LDC-IL tagset, under the category of Adverb, it has subtype called Adverb of Manner and the Adverb of Time, we are treating it under NST Nouns. In the BIS tagset we follow the same rule we treat Adverb of Manner under the category of Adverb which has no further sub type, and here also Adverb of Time we have covered under NST Nouns.

LDCIL tagset has sub types of case and non-case postpositions in Postposition category because there are some postpositions which change the form as per the gender, number, and case marker. Whereas, in the BIS tagset Postposition does not have any subtypes.

There has been a huge list of particles that had been covered under Particle category as its subtypes in LDC-IL tagset for instance Co-ordinating, Sub-ordinating, Interjection, (Dis) Agreement, Emphatic, Topic, Delimiting, Honorific, Negative, Exclusive, Terminative, Dubitative, Simulative, Inclusive, Comparative, and others. All these above mentioned particles we have found during our tagging so we have categorised it according to above mentioned list in LDCIL tagset. In BIS tagset Default, Interjection, Intensifier and Negation are sub types covered under the Particle category. Rest of the sub types of particle categories from the LDCIL tagset have been mapped as Default Particle in the BIS tagset.

Numeral category is not there in the BIS tagset, there were occurrences of the real numbers in terms of date format, and some time modified form according to its real occurrence based on the context. This tag is used to annotate all those tokens which have a numeric value. That is to say, that number tokens that are not written in words, but rather in numeric values will be annotated under this tag. It has included sub categories called Real [૧, ૨, ૩] Serial [(૧), (૨), (૩)], Calendric [૧૨- ૧૨ - ૨૦૧૧] and Ordinal [બીજો- second, ૪થું- fourth]. At present we are tagging numeral occurrences under Quantifier Cardinal.

Reduplication, Unknown and Punctuations categories we have taken as separate categories as contrasting to BIS tagset where it has been covered under residuals along with Symbol, Foreign words and Echo words as subtypes except for Reduplication. Reduplication category is not there in the BIS tagset. In LDC-IL tagset, Residual category included sub categories called Symbol and Foreign words.

V. TAGGING ISSUES

The tagging issues are as follow: Verbal Noun, Participle and reduplication

A. Verbal Noun: Verbal Nouns are derived from verbs and generally called as gerunds. *–નું(nuM)* suffix is affixed to make Verbal noun but such forms are also infinite verbs. We can distinguish between infinitive form and gerundive form by merely looking at the syntactic context whether it occurs in the verb construction or followed by the postposition. For example, in the first sentence:

મને\PRP ***તરવું\VM*** *છે\VA (manE taravuM chE). Here*

તરવું(taravuM) is Verb Infinitive

Meaning: I want to swim.

Whereas, in the second sentence, the same form functions as a Verbal Noun ***તરવું\NV*** *એ\ PPR સારી\JJ કસરત\NN છે\VA (taravuM E sArI kasarata chE).*

Meaning: Swimming is a good exercise.

As in the BIS tagset for Gujarati there is no category called verb Infinitive, Gerund and Verbal Noun exist. Now

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

it is being tagged as Verb Main only. At present we are tagging all occurrences of Verbal Noun, Gerund, Verb Non-finite, Infinitive, Finite as Main Verb. But the problem arises when this kind of form is inflected for morpho-syntactic features (such as case, gender, number and person). For example the words like *જમવાની? ઉતાવળ/NN (jamavAnI utAvaLa)* earlier in LDC-IL tagset we used to tag word *જમવાની(jamavAnI)* as verbal noun because it is inflected for genitive case and it purely function as a Noun. But in BIS tagset Verbal Noun category is not there, we are tagging it as a Main Verb.

Another example of Verbal Noun: *ખાવા\NV માટેનું/PSP ફળ/NN (khAvA mATEnuM phaLa).* Here, We used to tag the word *ખાવા(khAvA)* as Verbal Noun as it is in oblique form and followed by the postposition *માટેનું/PSP*. But in BIS tagset , we are tagging it as a Main Verb.

B. Participle: For instance the words, *ચઢતી? છોકરી/NN (caDhatI chOkarI)*, meaning (climbing girl), earlier *ચઢતી (caDhatI)* we used to tag is as Participle but now it is being mapped as a main verb as we don’t have category called Participle in BIS tagset. So we are tagging it as a Main Verb but while doing so its adjectival part is not being recognized as here *ચઢતી(caDhatI)* is modifying the noun *છોકરી (chOkarI)* and it can also inflected for gender, number , person and it also can take tense marker.

C. Reduplication: For example in following sentence *હું/PRP ચઢતાં/VM ચઢતાં થાકી/VM ગયો/VAUX (huM caDhatAM caDhatAM thAkI gayO)* , earlier we used to tag the second word *ચઢતાં (caDhatAM)* as reduplication of the verb *ચઢતાં (caDhatAM).* In BIS tagset there is no reduplication category so we are treating the second word *ચઢતાં (caDhatAM)* as a Main Verb only.

D. Complex Predicate: The sentence like *અહીં/NST અનાજ/NN ઉતપન્ન થાય/VM છે/VAUX (ahIM anAja utapanna thAya chE).* It creates confusion what should we tag for the word *ઉતપન્ન (utapanna)* either Adjective or Noun? and one more example *મને/PRP આ/DAB વસ્તુ/NN પ્રાપ્ત/? થઈ/VM (manE A vastu prApta thaI) પ્રાપ્ત (prApta)* is either adjective or noun?. Solution came up as we should tag both words *ઉતપન્ન (utapanna)* and *પ્રાપ્ત (prApta)* as Adjectives.

VI. TAGSET TRANSITION RULES

In the beginning, the user has to make the compatible rules so as to enable it to map the source tagset appropriately to the target tagset. Such transition rules play a vital role in the rule-based approach of algorithmic transition which consists of columns namely, source, target, and attribute level. The source column indicates the source list of LDC-IL tagset by category wise and the target column also indicates the tagset taken from BIS tagset. The final column is a constraint checking value column which contains two groups of values. The first group is known as ‘NIL groups’ and second group is known as ‘non-NIL groups’. In the beginning, the computer programming will check if the value is NIL, and then it will not verify the attribute level of morpho-syntactic feature of source tags and if the value is non-NIL, then it will verify the attribute level. Table II shows the part of rules through which the transition of tagsets have been performed.

TABLE II
Transition Rules

Source	Target	Attribute Level
NC	N_NN	Nil
NP	N_NP	Nil
NST	N_NST	Nil
PPR	PR_PRP	Nil
PRF	PR_PRF	Nil
PRL	PR_PRL	Nil
PRC	PR_PRC	Nil
PWH	PR_PRQ	Nil
DAB	DM_DMD	Nil
DRL	DM_DMR	Nil
DWH	DM_DMQ	Nil
VM	V_VM	Nil
VA	V_VAUX	Nil
JJ	JJ	Nil
JQ	Q_QTF	nmm
JQ	Q_QTC	crd
JQ	Q_QTO	ord
JINT	RP_INTF	Nil
AMN	RB	Nil
CCD	CC_CCD	Nil
CSB	CC_CCS	Nil
CIN	INJ	Nil
CEMP	RP_RPD	Nil
CAGR	RP_RPD	Nil
CDLIM	RP_RPD	Nil
CHON	RP_RPD	Nil
CDED	RP_RPD	Nil
CTOP	RP_RPD	Nil
CHON	RP_RPD	Nil
CEXCL	RP_RPD	Nil
CDUB	RP_RPD	Nil
PPC	PSP	Nil
CEMP	RPD	Nil
UNK	RD_UNK	Nil
PU	RD_PUNC	Nil
NNV	V_VM	Nil
LPS	V_VM	Nil

From the discussion with Guajarati annotator here we have given a rule for verbal noun and participle categories as it should be mapped as a main verb as per the BIS taget.

VII. RESULTS AND DISCUSSION

For this experiment we used LDC-IL Gujarati annotated corpus of size is 26,961. The input of transition system is a LDC-IL tagset corpus. For example

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

Input (e.g.):

એવાં\SIM.neu.pl.dir *પદ્યનો\NC.mas.sg.obl.gen.0.0*

અનુવાદ\NC.mas.sg.dir.0.0.0

કર્યો\VA.mas.sg.3.pst.ipfv.0.fin.0.0.0

છે\VM.mas.sg.3.prs.ipfv.0.fin.0.0.0 .\PU

The output of annotated corpus might be word with BIS label. In some cases, a target result appears with less mixture of LDC-IL tags (source). It is an obvious state that the spelling variation of the input file will affect the result of the output file. For example

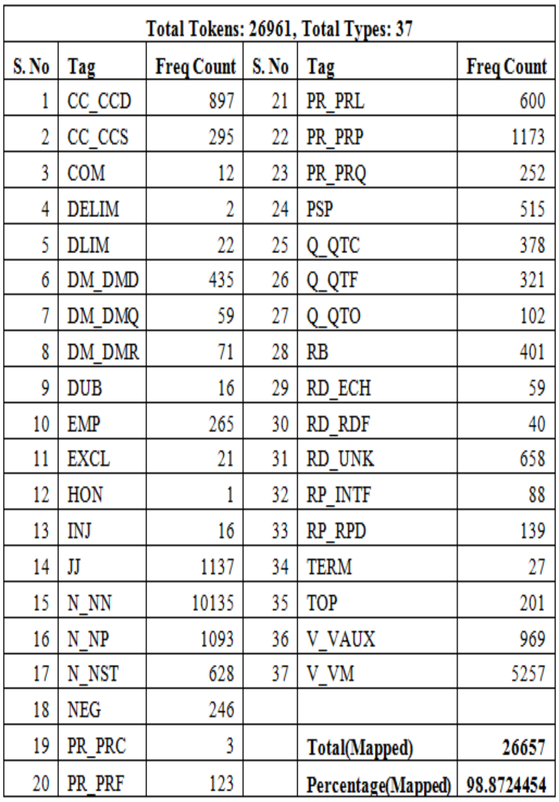
Output (e.g.):

એવાં\SIM.neu.pl.dir *પદ્યનો\N_NN* *અનુવાદ\N_NN*

કર્યો\V_VAUX છે\V_VM .\RD_PUNC

In the above output file, due to the typo error, the *એવાં\SIM.neu.pl.dir* could not be mapped as RP_RPD.

TABLE III
Mapped Tagset Results

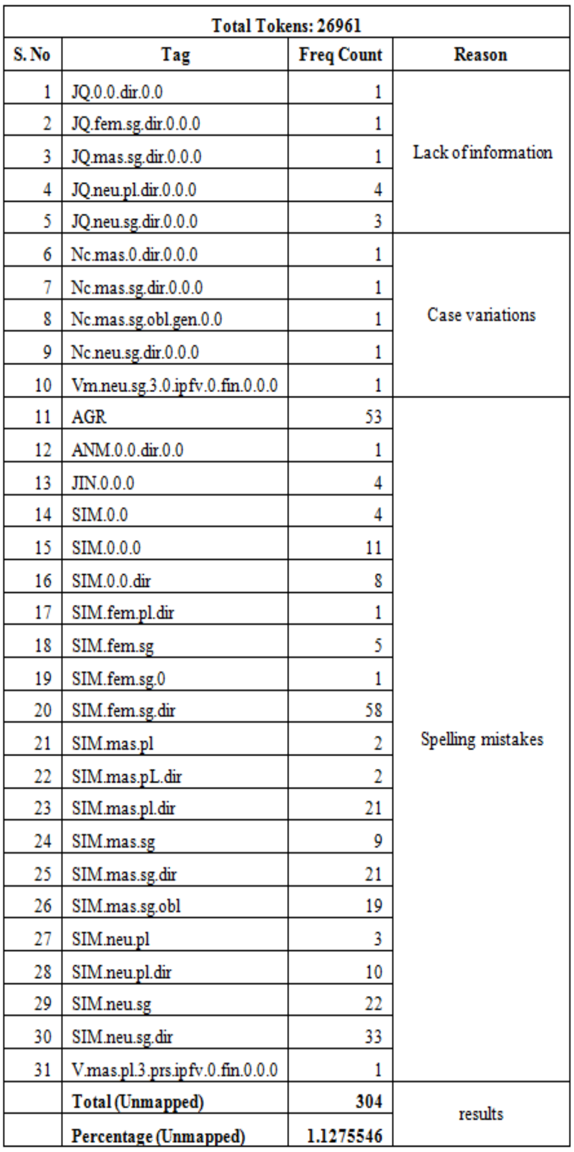


Total Tokens: 26961, Total Types: 37					
S.No	Tag	Freq Count	S.No	Tag	Freq Count
1	CC_CCD	897	21	PR_PRL	600
2	CC_CCS	295	22	PR_PRP	1173
3	COM	12	23	PR_PRQ	252
4	DELIM	2	24	PSP	515
5	DLIM	22	25	Q_QTC	378
6	DM_DMD	435	26	Q_QTF	321
7	DM_DMQ	59	27	Q_QTO	102
8	DM_DMR	71	28	RB	401
9	DUB	16	29	RD_ECH	59
10	EMP	265	30	RD_RDF	40
11	EXCL	21	31	RD_UNK	658
12	HON	1	32	RP_INTF	88
13	INJ	16	33	RP_RPD	139
14	JJ	1137	34	TERM	27
15	N_NN	10135	35	TOP	201
16	N_NP	1093	36	V_VAUX	969
17	N_NST	628	37	V_VM	5257
18	NEG	246			
19	PR_PRC	3		**Total(Mapped)**	**26657**
20	PR_PRF	123		**Percentage(Mapped)**	**98.8724454**

Table III shows that correctly mapped results for Guajarati annotated corpora of size are 26,961 tokens. The 98.87% percentage of tags was correctly mapped from LDC-IL tagset to BIS tagset using rule-based approach.

The table IV shows the unmapped results which are not mapped for the number of reasons including lack of information, spelling mistakes and case-sensitive letters occurred in the annotated corpora. The unmapped result was 1.13% appeared because of above mentioned reasons.

TABLE IV
Unmapped Tagset Results



Total Tokens: 26961			
S.No	Tag	Freq Count	Reason
1	JQ.0.0.dir.0.0	1	Lack of information
2	JQ.fem.sg.dir.0.0.0	1	
3	JQ.mas.sg.dir.0.0.0	1	
4	JQ.neu.pl.dir.0.0.0	4	
5	JQ.neu.sg.dir.0.0.0	3	
6	Nc.mas.0.dir.0.0.0	1	Case variations
7	Nc.mas.sg.dir.0.0.0	1	
8	Nc.mas.sg.obl.gen.0.0	1	
9	Nc.neu.sg.dir.0.0.0	1	
10	Vm.neu.sg.3.0.ipfv.0.fin.0.0.0	1	
11	AGR	53	Spelling mistakes
12	ANM.0.0.dir.0.0	1	
13	JIN.0.0.0	4	
14	SIM.0.0	4	
15	SIM.0.0.0	11	
16	SIM.0.0.dir	8	
17	SIM.fem.pl.dir	1	
18	SIM.fem.sg	5	
19	SIM.fem.sg.0	1	
20	SIM.fem.sg.dir	58	
21	SIM.mas.pl	2	
22	SIM.mas.pL.dir	2	
23	SIM.mas.pl.dir	21	
24	SIM.mas.sg	9	
25	SIM.mas.sg.dir	21	
26	SIM.mas.sg.obl	19	
27	SIM.neu.pl	3	
28	SIM.neu.pl.dir	10	
29	SIM.neu.sg	22	
30	SIM.neu.sg.dir	33	
31	V.mas.pl.3.prs.ipfv.0.fin.0.0.0	1	
	Total (Unmapped)	**304**	results
	Percentage (Unmapped)	**1.1275546**	

The accuracy of transition increases when adding new rules into the existing rules together. For example we found that the categories *AGR*, *JIN* and *SIM* have occurred with spelling mistakes instead of CAGR, JINT and CSIM and the categories *Nc* and *Vm* have occurred with small and capital letters instead of NC and VM in uniform manner. In addition to these, information of non-numeral (nnm), cardinal (crd) and ordinal (ord) was not available in the JQ category. For the solution initially find out issues and then add the corresponding rules to the rule table.

The main categories of Verbal Noun, Participle, Reduplication and the sub categories of Main Verb like Finite verb, Non-finite verb, and Infinitive verb of LDC-IL tags are mapped as Main Verb according to the BIS tagset. The above mentioned LDCIL tagset categories are not in the BIS tagset. Therefore we mapped all those categories into Main Verb of BIS tagset. In addition, the category of Numeral is being mapped as Cardinal under the category of Quantifier.

VIII. CONCLUSION AND FUTURE DIRECTIONS

We have tried to bring forth the comparative analysis of both tagsets (LDC-IL and BIS), for that we also have to keep in mind the design strategy of both tagsets as at what level what particular features are meant to be captured. We have also focused on issues which we have faced while POS tagging as we have worked on both tagsets.

We have developed simple transition approaches for mapping from one tagset to another. In this view, we conclude that rule based approach can be more suitable for deeper layered or hierarchical tagset transition (tagset with attribute level). To retain both versions of POS annotated data, rule based transition approach proves its worth.

Furthermore, the transition system can be applied not only to POS tags, but to other types of tags as well. Quality annotated data is required for the transition system so that it will improve the accuracy of the result.

REFERENCES

[1] M. Shrivastava, N. Agrawal, S. Singh, and P. Bhattacharya. (2005). Harnessing morphological analysis in pos tagging task. In Proceedings of the International Conference on Natural Language Processing (ICON 05), December.
[2] Hardie, A. (2004). The Computational Analysis of Morphosyntactic Categories in Urdu. PhD thesis submitted to Lancaster University.
[3] EAGLES: (Mar, 1996). Recommendations for the Morphosyntactic Annotation of Corpora, EAGLES Document EAG–TCWG–MAC/R.
[4] D´ejean, Herv´e (2000). How to Evaluate and Compare Tagsets? A Proposal. In Proceedings of LREC-00.Athens.
[5] Eric Atwell, John Hughes, Clive Souter:Amalgam: (1994). Automatic Mapping among Lexico-grammatical annotation Models, Internal Paper, CCALAS, Leeds University, Aug.
[6] Reatrice Santorini: (Mar 1991). Part-of-Speech Tagging Guidelines for the Penn Treebank Project. Technical Report. Department Of computer Science and Information Science, University of Pennsylvania.
[7] Sharma, Dipti Misra. (2010). Linguistic Resource Standards. Standards for POS Tag Set for Indian Languages. (Draft) Ms. LRTC, IIIT, Hyderabad.
[8] Sankaran Baskaran et al. (2008). Designing a Common POS-Tagset Framework for Indian Languages, pub, The 6th Workshop on Asian Language Resources.

BIOGRAPHIES

Purva S. Dholakia[1], born in Bhuj- kutch Gujarat is a post graduate student from The Maharaja Sayajirao University, Baroda in English Literature. She is currently working as a Senior Research Assistant in LDC-IL, Central Institute of Indian Languages, Mysore, and has 5 years of research experience in NLP based work with the same institution. She is the author of three papers, has presented several papers in national and international conferences/seminars, and has also delivered lectures in the different parts of the country, on Computational Linguistics and NLP. Corpus Linguistics, Pats-of-Speech tagging, Morphological analyzer & generator, Speech data segmentation and annotation are her specialization areas of work. Apart from academic arena, creative writing, radio announcing (appointed announcer in AIR KUTCH-GUJARAT), and classical dancing (Bharatnatyam) hold as integral interest for her in totality.

M. Mohamed Yoonus[2] was born in Avudaiyar Pattinam Village, Pudukkottai District, Tamilnadu. He has completed M.Sc. degree in Computer Science from the Bharathidasan University, Trichy. He has obtained M.Phil degree in Computer Science from Manonmaniam Sundaranar University, Tirunelveli. He did PG Diploma in NLP from Annamalai University, Chidambaram. He is currently pursuing the Ph.D. degree in Computer Science and Applications at Periyar Maniammai University, Thanjavur. He worked as a Lecturer in Sri Venkatshwara College, Peravurani and Naina Mohamed College, Aranthangi for 5 years. In April 2008, he joined as a Senior Technical Officer with the LDC-IL, Central Institute of Indian Languages, Mysore, wherein he worked as a Lecturer cum Resource Person from July 2010 onwards. Since October 2011, he has been leading his works as a Senior Lecturer cum Junior Research Officer with the same institute. He is the author of 6 publications, more than 10 paper presentations, and more than 15 participations in national and international conferences/seminars. He has delivered many lectures in different parts of the country. He has acted as Resource Person in the "Workshop on Education: Technology, Research and Curriculum" at Mahatma Gandhi Institute (MGI), Mauritius. His research interests include Natural Language Processing (NLP), E-learning, Corpus Based Learning and Computer Applications etc. He has qualified Tamilnadu-State Eligibility Test (SET) for lectureship in November 2012.