Comparison and Mapping of Two Tagsets for Gujarati Language

Purva S. Dholakia
Senior Research Assistant
purvadholakia@gmail.com

Mohamed Yoonus. M
Senior Lecturer/JRO
yoonussoft @gmail.com

Linguistic Data Consortium for Indian Languages (LDC-IL)
Central Institute of Indian Languages – Mysore

Content

- Introduction
- POS Tagset: Overview
- Tagset Comparison
- Tagging Issues
- Mapping
- Results and Discussion
- Conclusion

Introduction

•The process of classifying words into their parts-of-speech and labeling them accordingly is known as parts-of-speech tagging, POS tagging, or simply tagging.

a) word's lexical probability
b) the word's contextual probability

Continued...

- The collection of tags used for a particular task is known as a tagset.

- Tag sets vary according to the objectives of specific projects.

- In some situations, however, we need to first compare and then map the two existing tag sets and use the mapping to get two kinds of tagged corpus.

Objectives

•This paper aims at the comparison of two POS tag sets - LDC-IL Tag set and the BIS Tag set for Gujarati language.

•Discusses the tagging issues for Gujarati Language

•Describes mapping approach which maps morpho-syntactic tagset(LDC-IL tagset) to a partially layered tagset(BIS Tagset).

POS Tagset: Overview

POS Tagset: Overview

- LDC-IL tagset is a hierarchical tagset based on the ILPOST framework.
- The tagset has three layers.
  - Top layer - morphological categories
  - Middle layer - types of the category
  - Bottom layer - morpho-syntactic features or attributes of the type of the category.

- BIS tag set (Bureau of Indian Standard) 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.

Tagsets Comparison

Tagsets Comparison

•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 we had 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.

Tagset Comparison

BIS Labels			**LDC-IL Labels**		
Category	*Sub-category*	*Tag*	*Category*	*Subcategory*	*Tag*
NOUN		**N**	**NOUN**		**N**
	Common	NN		Common	NC
	Proper	NNP		Proper	NP
	~~Verbal~~	~~NNV~~		Verbal	NV
	Nloc	NST		Spatio-temporal	NST
PRONOUN		**PR**	**PRONOUN**		**P**
	Personal	PRP		Pronominal	PPR
	Reflexive	PRF		Reflexive	PRF
	Reciprocal	PRL		Reciprocal	PRC
	Relative	PRC		Relative	PRL
	Wh-word	PRQ		Wh-pronoun	PWH
	Indefinite	PRI			
DEMONSTRATIVE		**DM**	**DEMONSTRATIVE**		**D**
	Deictic	DMD		Absolutive	DAB
	Relative	DMR		Relative Demonstrative	DRL
	Wh-word	DMQ		Wh-demonstrative	DWH
	Indefinite				

VERB		**V**	**VERB**		**V**
MAIN		VM		Main Verb	VM
	~~Finite~~	~~VF~~			
	~~Non-finite~~	~~VNF~~			
	~~Infinitive~~	~~VINF~~			
	~~Gerund~~	~~VNG~~			
AUXILIARY		VAUX		Auxiliary Verb	VA
ADJECTIVE		**JJ**	**NOMINAL MODIFIER**		**J**
				Adjective	JJ
ADVERB		**RB**	**ADVERB**		**A**
				Manner	AMN
POSTPOSITION		**PSP**	**POSTPOSITION**		**PP**
				Case	PPC
				Non-Case	PPNC
CONJUNCTION		**CC**	**PARTICLE**		**C**
	Co-ordinator	CCD			
	Subordinator	CCS		Co-ordinating	CCD

PARTICLE		**RP**	**PARTICLE**		**C**
	Default	RPD		Subordinating	CSB
	Interjection	INJ		Interjection	CIN
	Intensifier	INTF		(Dis)agreement	AGR
	Negation	NEG		Emphatic	EMP
				Topic	TOP
				Delimiting	DLIM
				Honorific	HON
				Negative	NEG
				Exclusive	EXCL
				Terminative	TERM
QUANTIFIER		**QT**	**NOMINAL MODIFIER**		**J**
	General	QTF		Quantifier	JQ
	Cardinal	QTC			
	Ordinal	QTO			
RESIDUAL		**RD**	**RESIDUAL**		**RD**
	Foreign word			Foreign Word	RDF
	Symbol			Symbol	RDS
	Punctuation		**PUNCTUATION**		**PU**
	Unknown		**UNKNOWN**		**UNK**
	Echo words				
			REDUPLICATION		**RDP**

EXTRA TAGS IN LDC-IL TAGSET			
	PARTICIPLE		**L**
		Present	LPR
		Past	LPS
		Future	LFU
	NUMERAL		**NUM**
		Real	NUMR
		Serial	NUMS
		Calendric	NUMC
		Ordinal	NUMO

- Occurrences of Numeral tags(LDC-IL Tagset)
- Real - [૧, ૨, ૩]
- Serial [(૧), (૨), (૩)],
- Calendric [૧૨- ૧૨ -૨૦૧૧] and
- Ordinal [બીજો- second, ૪થું- fourth].

Tagging issues

Tagging Issues

- *Verbal Nouns:* Verbal Nouns are derived from verbs and generally called as gerunds. In Gujarati –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 post-position.

- મને\PRP **તરવું\VM** છે\VA (manE taravuM chE). Here તરવું(taravuM) is verb infinitive

Meaning: I want to swim.

તરવું\NV એ\ PPR સારી\JJ કસરત\NN છે\VA (taravuM E sArI kasarata chE). And here તરવું(taravuM) is verbal noun.

Meaning: Swimming is a good exercise.

Tagging Issues

- ***Inflected for case:***

- ex – જમવાની /? ઉતાવળ /NN (jamvAnI utAvaLa)

- ex- જમવાનામાં/? મીઠું/NN નાખજો/VM (jamvAnAmA mIthUn nAkhajO) meaning - add salt in the food.

- ***Followed by post position :***

- ખાવા /? માટેનું / PSP ફળ/NN (khAvA mATEnuM phaLa) meaning - Fruit for eating.

Tagging Issues

- *Participle:* Ex. 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 it's modifying element 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.

Tagging Issues

- *Reduplication:* for example in following the sentence હું/PRP ચઢતાં/VM ચઢતાં થાકી/VM ગયો/VAUX (huM caDhatAM caDhatAM thAkI gayO) earlier we used to tag the second word ચઢતાં as reduplication of the verb ચઢતાં.

- There is no reduplication category in BIS tagset so we are treating the second word ચઢતાં as a main verb.

- 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?

Mapping

Mapping

•User ⟶ the compatible rules ⟶ Mapping

•The mapping rule plays vital role in constraint-based approach of mapping algorithm which consists of columns namely source, target and attribute level.

•The source column indicates the source list of LDC-IL tagset

•The target column indicates the tagset list of BIS tagset.

• The final column is a constraint checking value column which contains two groups of values.

Continued…

- The first group is known as ‘NIL groups’ and second group is known as ‘non NIL groups’.

- Computer program will check if the value is NIL 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.

Constraint-Based Approach: Rule format

Source List	Target List	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
VM	V_VM	Nil
VA	V_VAUX	Nil
JJ	JJ	Nil
JQ	Q_QTF	nnm
JQ	Q_QTC	crd
JQ	Q_QTO	ord

Results and Discussion

- For this experiment we used LDC-IL Gujarati annotated corpus
- Corpus size is 26961.
- Correctly mapped 98.87% percentage
- Unmapped 1.13% percentage
- **Reasons:**
  - lack of information
  - spelling mistakes
  - case-sensitive letters

Mapped Tagset Results

Total Tokens: 26961, Total Types: 37					
SNo	**Tag**	**Freq Count**	**SNo**	**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		**Average(Mapped)**	**98.8724454**

Unmapped Tagset Results

Total Tokens:		**26961**	
SNo	Tag	Freq Count	
1	JQ.0.0.dir.0.0	1	Lack of informati on
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
	Average (Unmapped)	**1.127**	

Results and Discussion

•The accuracy of mapping 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.

Results and Discussion

•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 verb main of BIS tagset.

•In addition, the category of numeral is being mapped as cardinal under the category of quantifier.

Conclusion

•We have tried to bring forth the comparative analysis of both tagsets (LDC-IL and BIS).

•Design Strategy

• We have also focused on issues which we have faced while pos tagging as we have worked on both tagsets.

•We have developed simple mapping approach for mapping from one tagset to another.

Continued..

- Constraint based approach is more suitable for deeper layered or hierarchical tagset mapping.

- Furthermore, Apart from POS level , the mapping system can be applied to other levels .

- Quality annotated data is required for the mapping system so that it will improve the accuracy of the result.

References

- Babarczy, Anna, John Carroll, and Geoffrey Sampson: (2006). *Definitional, personal, and mechanical constraints on part of speech annotation performance*. Journal of Natural Language Engineering 12, 77–90.
- 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.
- Eric Atwell, John Hughes, Clive Souter:Amalgam: (1994). *Automatic Mapping among Lexico-grammatical annotation Moels,* Internal Paper, CCALAS, Leeds University, Aug.
- Sankaran Baskaran et al. (2008). *Designing a Common POS-Tagset Framework for Indian Languages*, pub, The 6th Workshop on Asian Language Resources
- EAGLES: (Version of Mar, 1996). *Recommendations for the Morphosyntactic Annotation of Corpora*, EAGLES Document EAG–TCWG–MAC/R.

Thank You

Purva S. Dholakia

&

Mohamed Yoonus. M

Linguistic Data Consortium for Indian Languages (LDC-IL)

Central Institute of Indian Languages – Mysore