Building Large Scale POS Annotated Corpus For Urdu & Hindi

Shahid Mushtaq Bhat
Shahid.bhat3@gmail.com
Dr. Richa
ldc-richa/ciil@ciil.stmpy.soft.net

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

OUTLINE

1. Abstract
2. Introduction
3. Tagsets: An Overview
4. Process of POS Tagging
5. Statistics of the Sample
6. POS Annotation Issues
7. Conclusion

ABSTRACT

- In this presentation, we are going to share our experience of developing 69.723k POS annotated corpus of Hindi & 66.488k POS annotated corpus of Urdu, using the BIS Tag-set.
- We didn't annotate corpus directly either using LDCIL manual tool or by using the LDC-IL POS Tagger rather we made a transition from LDCIL Annotation Scheme (based on ILPOST) to the contemporary BIS Scheme (inspired by ILMT).
- This transition resulted in the afore mentioned quantum of annotated corpus as per BIS Standards.

INTRODUCTION

- POS Tagging is the process of labeling words in the running text corpus with their grammatical categories and optionally with the associated grammatical features.
- It is essentially a classification problem, where we have to classify the set of words in a text as per some predetermined scheme.
- For some languages (with split-orthography) it is also a mapping-problem which involves mapping of the arrays of tokens (words, chunks or sentences) on the arrays of tags in proper agreement with the syntactic structure of a language.
- In the entire pipe-line of NLP, it plays a limited role of syntactic category disambiguation.

TAGSETS: AN OVERVIEW

- POS Tag-set is minimal set of categories & sub-categories that can be used to classify all the words of a language with maximum precision.
- The initial efforts in POS Annotation resulted in tag-sets that were simple inventories of tags corresponding to the morpho-syntactic features such as Brown & Upenn (Hardie, 2004).
- It was CLAWS2 tag-set (Sartoni, 1987) which is a landmark in the history of tag-set designing.
- It marked an important change in the structure of tag-sets, from a flat-structure to a hierarchical-structure.
- The term "hierarchical", when used for a tag set, means that the categories in that tag set are structured relative to one another. A hierarchical tag set will contain a small number of categories, each of which contains a number of sub-categories, each of which may contain sub-sub-categories, and so on, in a tree-like structure (Hardie 2003).
- For example:
- LDC-IL Tagsets [Hierarchical but Fine grained]
- BIS Tagsets [Hierarchical but Coarse grained]

So, far various Tag-sets have been developed for ILs.

1. AU-KBC Tamil Tag set (2001)
2. Hardie's Tag-set for Urdu (Hardie (2005).
3. IIIT-Hyderabad Tag-set for Hindi (Bharati et al. 2006)
4. Micro-Soft Research of India (MSRI) IL-POST for Hindi & Bangla (Baskaran et al. 2008)
5. MSRI-JNU Sanskrit Tag-set
6. CSI-HCU for Telugu (Sree R.J et al. 2008)
7. IIT-Kharagpur Tag-set for Bangla
8. Nelrlac Tag-set for Nepali
9. LDCIL Tag-sets for all ILs (2009/2010)
10. BIS Tag-sets for all ILs (210/2011)

Note: Many of the afore mentioned tag-sets were strictly/ loosely following the guidelines of EAGLES (*Expert Advisory Group for Language Engineering Standards)* for morpho-syntactic annotation (Leech & Wilson 1999).

For example:- 1, 2, 4, 5, 9

PROCESS OF POS TAGGING

- The entire process of POS Annotation for the current work was completed in 3-phases
- Manual Annotation using LDCIL Tool & Tag-set (customized).
- Automatic POS Annotation using LDCIL POS Tagger.
- Automatic Mapping from LDCIL to BIS

MANUAL POS ANNOTATION/TAGGING

- Manual annotation cum validation of 50k corpus (in XML format) was done with the help of a stand alone, GUI & easily customizable tool developed using VB.NET.

- Three annotators carried out the annotation work.

- 0.3 version of LDC-IL guideline for Hindi & Urdu were followed.

LDC-IL POS Annotation Tool

CONT…


Annotation Tool Urdu Version 0.2
File Edit Format Options
Linguistic Data Consortium for Indian Languages (LDC-IL)
Common
Proper
Verbal
Spatio-temporal
Noun
Pronoun
Demonstrative
Nominal Modifier
Verb
Adverb
Postposition
Particle
Numeral
Reduplication
Residual
Unknown
Punctuation
Sentence/Paragraph Selection
Sent/Para No.
<<Prev
Next>>
Join Next Sentence
<<Prev UnTag
Next UnTag>>
start
Microsoft PowerPoint ...
Developing a Hierarc...
Urdu Statistics - Micro...
LDCIL Works
untitled - Paint
Annotation Tool Urdu...
EN
11:38 PM

EVALUATION: INTER-ANNOTATOR AGREEMENT

c:\Tagged\Urdu\Mansoor\ TagGSUrdu1.xml(462)	c:\Tagged\Urdu\Shahid\ TagGSUrdu1.xml(451)	c:\Tagged\Urdu\Rushda\ TagGSUrdu1.xml(448)	Category	Attrib.
نصراللہ\NP.mas.0.dir.0	نصراللہ\NP.mas.0.dir.0	نصراللہ\NP.mas.sg.dir	100	66.667
بیگ\NC.mas.0.dir.0	بیگ\NP.0.0.dir.0	بیگ\NP.mas.sg.dir	0	0
ابتدا\NC.0.0.obl.0	ابتدا\NST.obl	ابتدا\NC.fem.sg.obl.0	66.667	0
میں\PP.0.0.loc	میں\PP.0.0.loc	میں\PPC.0.sg.loc	66.667	66.667
واقعی\CIN	واقعی\AMN.obl	نو\JJ.0.sg.dir	0	0
مربٹوں\NC.mas.pl.obl.0	مربٹوں\NC.mas.pl.obl.0	واقعے\AMN.obl	66.667	66.667
کی\PP.fem.sg.gen	کی\PP.fem.sg.gen	مربٹوں\NC.mas.pl.obl.0	66.667	66.667
ملازمت\NC.fem.sg.obl.0	ملازمت\NC.fem.sg.dir.0	کی\PPC.fem.0.gen	66.667	0
میں\PP.0.0.loc	میں\PP.0.0.loc	ملازمت\NC.fem.sg.obl.0	66.667	66.667
تھے\VM.mas.pl.3.pst.0.0.fin.y	تھے\VA.mas.sg.0.pst.0.0.fin.y	میں\PPC.0.0.loc	0	0
PU\۔	PU\۔	تھے\VM.mas.sg.3.pst.0.0.fin.y	66.667	66.667
تاہم\CSB	تاہم\CSB	PU\۔	66.667	100
انہوں\PPR.0.pl.3.obl.0.n.n.0.n	انہوں\PPR.0.sg.3.obl.0.n.n.dst.y	تاہم\CX.y.n	66.667	0
نے\PP.0.0.erg	نے\PP.0.0.erg	انہوں\PPR.mas.sg.3.obl.0.n.y.prx.y	66.667	66.667
بروقت\ALC.dir.n	بروقت\AMN.dir	نے\PPC.0.0.erg	0	0
ننے\JJ.0.pl.dir	ننے\JJ.0.pl.dir	بروقت\NC.fem.sg.dir.0	66.667	66.667
آقا\NC.mas.pl.dir.0	آقا\NC.mas.0.dir.0	ننے\JJ.mas.sg.dir	66.667	0
تلاش\NC.0.0.dir.0	تلاش\VM.0.0.0.0.0.0.nfn.n	آقاتلاش\NC.mas.pl.dir.0	0	0

CONT....

VM.0.0.0.0.0.0.nfn.n\کر	VM.0.0.0.0.0.0.nfn.n\کر	VM.0.0.0.0.0.0.nfn.n\کر	100	100
VA.0.pl.3.0.prf.0.fin.n\لیے	VA.mas.pl.0.0.prf.0.fin.n\لیے	PP.0.0.bnf\لیے	66.667	0
PU\۰	PU\۰	PU\۰	100	100
NC.mas.pl.obl.0\مرہٹوں	NP.mas.pl.obl.0\مرہٹوں	NC.mas.pl.obl.0\مرہٹوں	66.667	100
PP.0.0.gen\کے	PP.0.0.gen\کے	PPC.0.0.gen\کے	66.667	100
NC.0.sg.obl.0\لشکر	NC.mas.sg.obl.0\لشکر	NC.mas.sg.obl.0\لشکر	100	0
PP.0.0.loc\میں	PP.0.0.loc\میں	PPC.0.0.loc\میں	66.667	100
بڑھتے\ VM.mas.pl.3.0.prf.0.fin.n	بڑھتے\ VM.mas.pl.0.0.prf.0.fin.n	بڑھتے\ VM.mas.pl.3.0.ipfv.0.fin.n	100	0
VA.0.pl.0.0.prf.0.fin.n\ہوئے	ہونے\ VA.mas.pl.0.0.prf.0.fin.n	ہونے\ VA.mas.pl.3.0.prf.0.fin.n	100	0
NC.0.pl.obl.0\تنازعوں	NC.mas.pl.obl.0\تنازعوں	NC.mas.pl.obl.0\تنازعوں	100	0
PP.0.0.ins\سے	PP.0.0.ins\سے	PPC.0.0.ins\سے	66.667	100
NC.mas.pl.obl.0\انگریزوں	NC.mas.pl.obl.0\انگریزوں	NC.mas.pl.obl.0\انگریزوں	100	100
PP.0.0.erg\نے	PP.0.0.erg\نے	PPC.0.0.erg\نے	66.667	100
JJ.fem.sg.0\بڑی	JQ.fem.n.obl.nnm\بڑی	JQ.fem.y.obl.nnm\بڑی	0	0
JJ.fem.sg.obl\ہوشیاری	AMN.obl\ہوشیاری	JJ.fem.sg.obl\ہوشیاری	66.667	66.667

CONT….

- 1. In Urdu/Hindi, the postposition को/ko/کو is either ‘dative’ or ‘accusative’ case marker. For example,

raam ko bhuuk lagii ‘Ram is hungry.’ [Dative]

me ne raam ko dekhaa ‘I saw Ram.’ [Accusative]

But, sometimes the use of को/ko/کو is different than the above.

For example,

Raam ko jaana hai ‘Ram has to go’. Here, को/ko/کو provides some kind of modal information.

me itvaar ko jaavongaa ‘I will come on Sunday’. Here, को /ko/کو denotes location in time.

- 2. In Hindi/Urdu, the postposition से/se/سے is either ‘instrumental’ or ‘ablative’ case marker. For example,

me chakuu se seb kaaTtaa hon ‘I cut the apple with the knife.’ [Instrumental]

Yehan se baahir mat jao ‘Do not go outside from here.’ [Ablative]

But, in some cases, से /se/سے denotes the superlative as well as the comparative degree. For example;

Sab se uunchii choTii ‘The highest peak’

maam shyam se behtar hai ‘Ram is a better boy than Shyam.’

VALIDATION-1

- Finally after clearing the inter-annotator disagreement, the validation of this 50k annotated corpus was carried out.
- This was the gold standard POS annotated corpus.
- The gold standard is generally used for training a tagger.
- We couldn't use this data for training as it was fine grained. But experiments were carried out which revealed that machine learning was almost negligible.

Removal Of Features

- Finally to train the tagger properly the features were removed and the 50k fine grained POS annotated data was rendered with only POS Categories & Sub-categories.

For example:

اُٹھایا/**VM.mas.sg.3.0.prf.0.fin.n** “uThayaa”

The label in the above annotated word was trimmed to remove features **[.mas.sg.3.0.prf.0.fin.n]**

The above annotated was rendered as given

اُٹھایا/**VM**

AUTOMATIC POS TAGGING

- The coarse grained data so obtained (by trimming features) was used for training purpose & more 20k corpus was automatically annotated by using the LDCIL POS Tagger.
- Again the 50k + 20k = 70k data was validated.
- The 70k validated data was almost ready for next turn of training & tagging but then we had to follow BIS.

Now the Problem was;

How to convert the LDCIL Tagged data into BIS Tagged data?

Automatic Mapping

- This is really a horrible thing when you have completed annotation & validation of 70k and in between you have to change you annotation scheme.
- To avoid the manual labor, we formulated simple mapping algorithms.

- Accordingly, 70k data tagged as per modified LDCIL tag-set (without features) was converted into the data tagged with BIS standards.

VALIDATION-2

- Since, mapping didn’t solve the problem fully, all the categories of LDCIL Tagset couldn't be mapped on the BIS Tagset.
- There were some big differences so we had to start the next phase of validation in which mainly those elements were corrected which were left by the mapping algorithm.
- Rest we had to validate the entire 70k data once more which we are currently doing.
- This phase of validation is almost over for Hindi but for Urdu it is still going on.

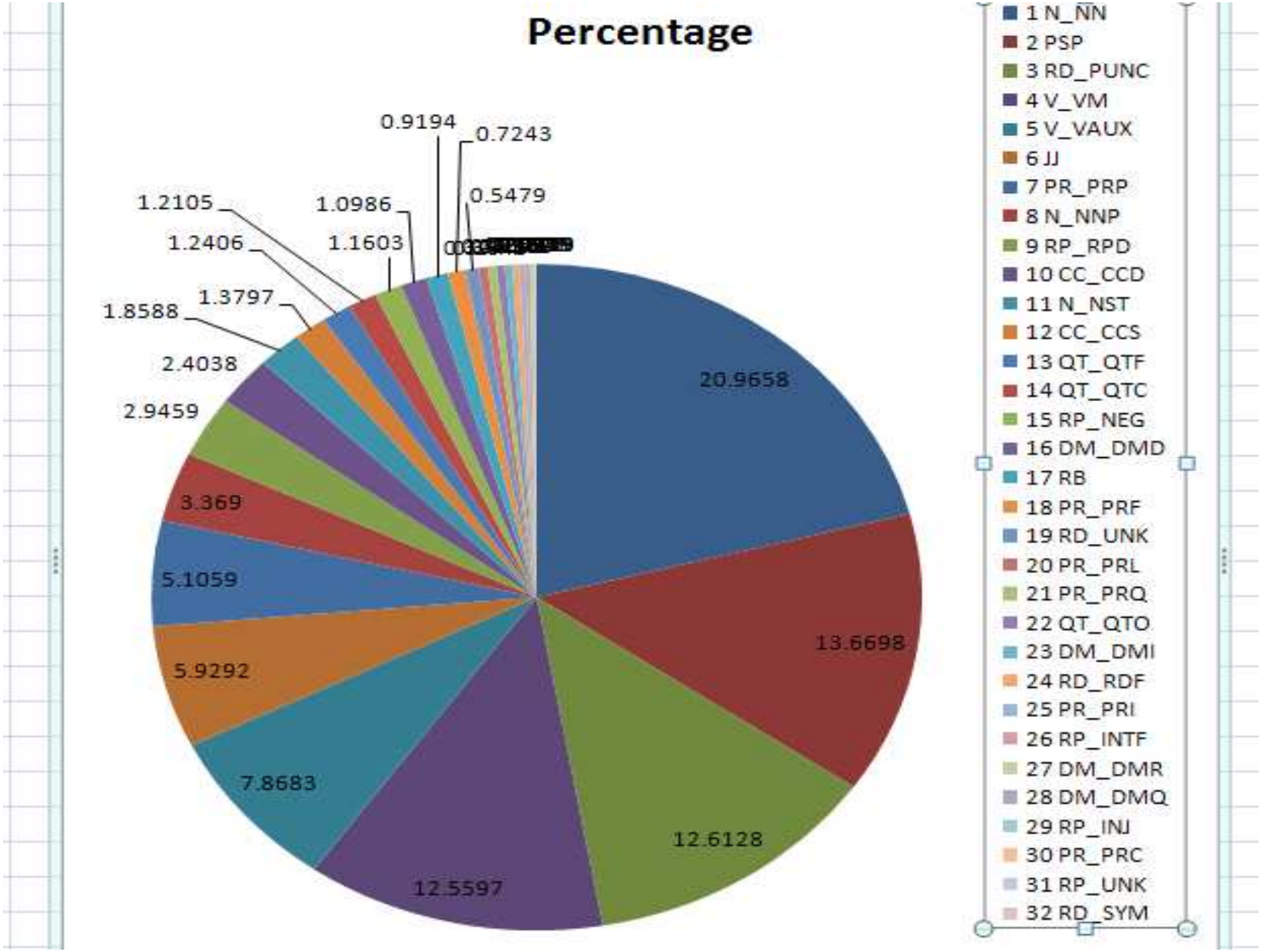
HINDI DATA: SOME STATISTICS

S.No.	Tag	Freq Count	Percentage
01	N_NN	14618	20.9658
02	PSP	9531	13.6698
03	RD_PUNC	8794	12.6128
04	V_VM	8757	12.5597
05	V_VAUX	5486	7.8683
06	JJ	4134	5.9292
07	PR_PRP	3560	5.1059
08	N_NNP	2349	3.3690
09	RP_RPD	2054	2.9459
10	CC_CCD	1676	2.4038
11	N_NST	1296	1.8588
12	CC_CCS	962	1.3797
13	QT_QTF	865	1.2406
14	QT_QTC	844	1.2105
15	RP_NEG	809	1.1603
16	DM_DMD	766	1.0986

H_CONT…….

17	RB	641	0.9194
18	PR_PRF	505	0.7243
19	RD_UNK	382	0.5479
20	PR_PRL	275	0.3944
21	PR_PRQ	254	0.3643
22	QT_QTO	248	0.3557
23	DM_DMI	216	0.3098
24	RD_RDF	207	0.2969
25	PR_PRI	176	0.2524
26	RP_INTF	125	0.1793
27	DM_DMR	105	0.1506
28	DM_DMQ	37	0.0531
29	RP_INJ	25	0.0359
30	PR_PRC	15	0.0215
31	RP_UNK	9	0.0129
32	RD_SYM	2	0.0029
	Total	69723	100.0000

H_CONT…..


Percentage
0.9194
0.7243
0.5479
1.2105
1.0986
1.2406
1.1603
1.3797
1.8588
2.4038
2.9459
3.369
5.1059
5.9292
7.8683
12.5597
12.6128
13.6698
20.9658
1 N_NN
2 PSP
3 RD_PUNC
4 V_VM
5 V_VAUX
6 JJ
7 PR_PRP
8 N_NNP
9 RP_RPD
10 CC_CCD
11 N_NST
12 CC_CCS
13 QT_QTF
14 QT_QTC
15 RP_NEG
16 DM_DMD
17 RB
18 PR_PRF
19 RD_UNK
20 PR_PRL
21 PR_PRQ
22 QT_QTO
23 DM_DMI
24 RD_RDF
25 PR_PRI
26 RP_INTF
27 DM_DMR
28 DM_DMQ
29 RP_INJ
30 PR_PRC
31 RP_UNK
32 RD_SYM

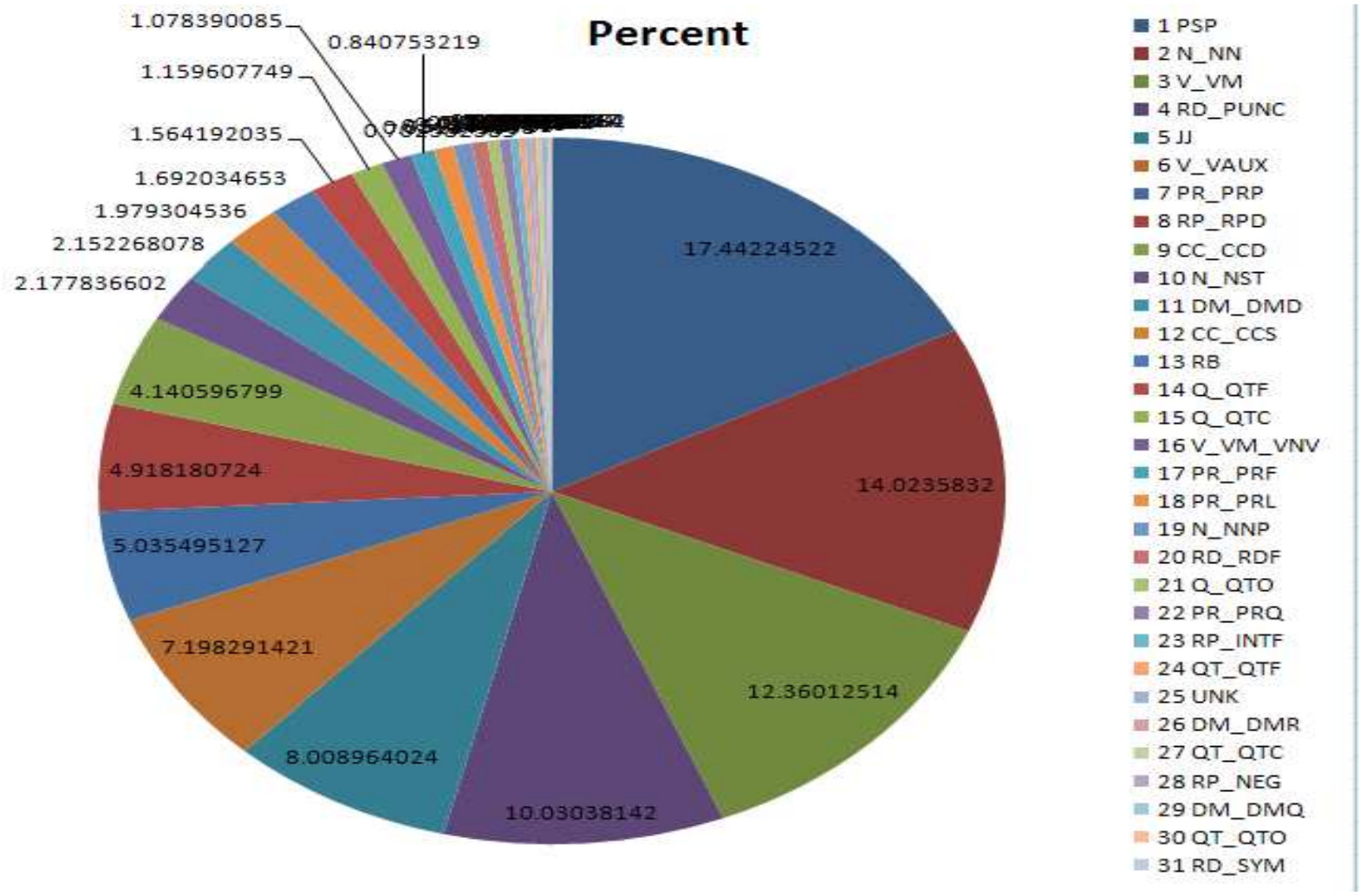
URDU DATA: SOME STATISTICS

S.No	Tag	Frequency	Percentage
01	PSP	11597	17.44225
02	N_NN	9324	14.02358
03	V_VM	8218	12.36013
04	RD_PUNC	6669	10.03038
05	JJ	5325	8.008964
06	V_VAUX	4786	7.198291
07	PR_PRP	3348	5.035495
08	RP_RPD	3270	4.918181
09	CC_CCD	2753	4.140597
10	N_NST	1448	2.177837
11	DM_DMD	1431	2.152268
12	CC_CCS	1316	1.979305
13	RB	1125	1.692035
14	Q_QTF	1040	1.564192
15	Q_QTC	771	1.159608
16	V_VM_VNV	717	1.07839

U_CONT…

17	PR_PRF	559	0.840753
18	PR_PRL	467	0.702382
19	N_NNP	433	0.651245
20	RD_RDF	380	0.571532
21	Q_QTO	278	0.418121
22	PR_PRQ	269	0.404584
23	RP_INTF	191	0.28727
24	QT_QTF	143	0.215076
25	UNK	141	0.212068
26	DM_DMR	125	0.188004
27	QT_QTC	123	0.184996
28	RP_NEG	109	0.163939
29	DM_DMQ	80	0.120322
30	QT_QTO	40	0.060161
31	RD_SYM	12	0.018048
	Total	66488	100

U_CONT…


Percent
1.078390085
0.840753219
1.159607749
1.564192035
1.692034653
1.979304536
2.152268078
2.177836602
4.140596799
4.918180724
5.035495127
7.198291421
8.008964024
10.03038142
12.36012514
14.0235832
17.44224522
1 PSP
2 N_NN
3 V_VM
4 RD_PUNC
5 JJ
6 V_VAUX
7 PR_PRP
8 RP_RPD
9 CC_CCD
10 N_NST
11 DM_DMD
12 CC_CCS
13 RB
14 Q_QTF
15 Q_QTC
16 V_VM_VNV
17 PR_PRF
18 PR_PRL
19 N_NNP
20 RD_RDF
21 Q_QTO
22 PR_PRQ
23 RP_INTF
24 QT_QTF
25 UNK
26 DM_DMR
27 QT_QTC
28 RP_NEG
29 DM_DMQ
30 QT_QTO
31 RD_SYM

POS Annotation Issues

It is usual experience for any annotator to face lots of problems while annotating data.

- Such problems/issues need to be documented properly.
- And discussed properly. It takes time to resolve them.
- In many cases there will be easy solution.
- But in some cases we need to take a decision rather to find a solution because there are fuzzy areas in every natural language where you don't have a categorical answer, true-false logic doesn't work here.

1. Fuzzy Items

- Hindi-Urdu Complex predicates are generally comprised of NN/JJ + Light Verb

For example:

- kush honaa
- khaDa honaa
- baDa karnaa
- haasil karnaa
- praapt karnaa
- pedaa honaa

If we try to find out some feature in the last three elements we won't fine any.

2.Case Syncretism

- When there is mismatch between a form & function of a postposition/case marker. It is called case syncretism.
- In Urdu & Hindi there is syncretism in dative-accusative & instrumental-ablative
- For details refer (slide 13)

3.Zwitter Ion Of Languages [verbal noun]

- It is an analogy taken from chemistry. Ion is a basically a charged particle. It can be positive, negative.
- But Zwitter ion carry both positive & negative charges simultaneously!!! So, it is hard to classify.
- Similarly, verbal noun/gerund carry verbal & nominal features simultaneously.
- But functionally they play the role of arguments in a sentential construction.

"kitaab paDne se imtihaan denaa aasaan hojata hai" [verbal root + case/case marker]

- Are they nouns or verbs?

4.Paradox Of Corpus Annotation

- Does form determines function or function determines form?
- Corpus linguistics is a methodology which tries to capture the functional aspect of corpus rather the formal one.
- But there is no categorical decision on the form-function aspect in any POS schemes for ILs. BIS (inspired by ILMT) is no exception in this case.

For example:

Verbal Nouns/Gerunds play a clear cut nominal function but as per recommendations we have to classify it under verb. The decision is motivated by formal aspect.

Even standards are now laid down for ILs form-function duality is yet to be resolved

5. Complex Items/Expressions

- Complex Postpositions is yet another problematic area in POS Annotation. In Urdu-Hindi such expressions consist of PSP + NST/NN/JJ + (PSP)

For example:

- ke aagE
- ke baare mein
- kii wajeh se
- kii tarah
- ke laayak
- ke zaryE

6.IZAFAT [DARD-E DIL]

- This construction is basically a Persian construction predominantly used in Urdu Corpus.
- It is typically NN + NN, NN + JJ combination.
- The two categories are actually two separate tokens but there is a marker called izafat (e.g.E).
- The izafat marker performs function of “genitive” or simply a linker in Urdu.

Shall we consider it MWE?

7.Split-Orthography (Urdu)

- The term split-orthography is actually used due to the unavailability of any technical term in the existing literature to denote the splitting/joining tendency in the Perso-Arabic script due to which affixes and roots are written separately; even some lexical items are written in two tokens, forming multi-token words.
- The term is, in a way, a new coinage to describe this tokenization problem of Urdu and Kashmiri.
- تہزیبیافتہ تہذیب یافتہ
- حیرتانگیز حیرت انگیز

CONT…..

Suffix

1.	mand	e.g. akl mand
2.	nigAr	e.g. mazmOn nigAr
3.	yAftah	e.g. tAlIm yafta/yaftah
4.	badr	e.g. muluk badr
5.	angaiz	e.g. hairat angaiz
6.	khAnah	e.g. kutub khAnah
7.	nawAz	e.g. hindustAn nawAz
8.	dAr	e.g. kwar dAr
9.	pazIr	e.g. taraqI pazIr
10.	deh	e.g. ArAm deh
11.	nAk	e.g. khof nAk
12.	guzAr	e.g. nimAz guzAr
13.	war	e.g. nAm war
14.	shud	e.g. khatm shudah
15.	kardah	e.g. hasil kardah
16.	bharI	e.g. ras bhari
17.	gAr	e.g. gunah gAr
18.	war	e.g. qasUr wAr
19.	gU	e.g. kush gU
20.	talab	e.g. ArAm talab

Prefix

1)	bA	e.g. bA himmat
2)	bAi	e.g. bAi kasUr
3)	ham	e.g. ham kayAl
4)	naw	e.g. naw umar
5)	gair	e.g. gair zaruri/ mohram/hazir

CONCLUSION

- In this presentation we have summarized our experience of developing 50k LDC-IL Fine Grained, 70k LDCIL Coarse Grained & 70k BIS Coarse grained annotated corpus of Hindi-Urdu.
- Also we shared our experience of transition from ILPOST to BIS which was very tough job for us.
- Finally, we highlighted some issues & paradoxes that we came across during the annotation process.

REFERENCES

- Hardie, A . 2004. The Computational Analysis of Morpho-syntactic Categories in Urdu. PhD thesis submitted to Lancaster University.
- Leech, G and Wilson, A. 1996. Recommendations for the Morpho-syntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R.
- Habash, N. & Owen Rambow. 2005. Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In Proceedings of the Conference of American Association for Computational Linguistics (ACL05).
- Baskaran S. et al. 2007. Framework for a Common. Parts-of-Speech Tagset for Indic Languages. (Draft) http://research.microsoft.com/~baskaran/POSTagset/
- Cloeren, J. 1999. Tagsets. In Syntactic Wordclass Tagging, ed. Hans van Halteren, Dordrecht.: Kluwer Academic.
- Garside, R. 1987 The CLAWS word-tagging system. In The Computational Analysis of English, ed. Garside, Leech and Sampson, London: Longman.
- Leech, G & Wilson, A. 1999. Standards for Tag-sets. In Syntactic Wordclass Tagging, ed. Hans van Halteren, Dordrecht: Kluwer Academic.
- Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS- 90-47, Department of Computer and Information Science, University of Pennsylvania
- IIT-tagset. A Parts-of-Speech tagset for Indian languages. http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf
- (Brill 93) E. Brill. A corpus-based Approach to Language Learning.1993.
- AU-KBC tagset. AU-KBC POS tagset for Tamil. http://nrcfosshelpline.in/smedia/images/downloads/Tamil_Tagset-opensource.odt
- (Hardie 03) A. Hardie. Developing a tagset for automated part-ofspeech tagging in Urdu. Proceedings of the Corpus Linguistics 2003 conference, 16, 2003.
- (Leech & Wilson 99) G. Leech and A. Wilson. Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R, 1999.

THANK YOU

Questions/ Comments