IMPROVING HINDI POS TAGGER ACCURACY THROUGH DOMAIN ADAPTATION

Anupama Pandey

The paper presents a comparative evaluation report on multi-domain Hindi taggers. Two taggers are trained in this experiment with the objective of detecting the accuracy rate of the tagger after adapting Cricket domain. The multi-domain tagger, trained as part of ILCI project, includes our major domain (Health, Tourism, Entertainment and Agriculture) presently and adapting Cricket as a new domain was recently proposed in Pandey (2017) which was calculated with a difference of approx. 6% in the tagger accuracy. Statistically, the accuracy of four domain tagger (without Cricket) is 85% and for five domain tagger (with Cricket) is approx. 93% which is 1% lower than the pre-existing Hindi tagger. This paper deals mainly with evaluation of the Hindi tagger (with and without Cricket as one of the domains). Author also attempts at finding the difference in terms of POS tagging issues in the output and the linguistic analysis of the errors found.

Keywords: ILCI Hindi Tagger, Cricket domain, Domain adaptation, POS annotation, domain tagger (DT)

1. Introduction

A part-of-speech tagger is a system which automatically assigns the part-of-speech to contextual information. Potential applications of part-of-speech taggers exist in many areas including Speech Recognition, Speech Synthesis, MT systems, IR and NERs. Words are often ambiguous in their part of speech. For example, English word *store* can either be a noun, a finite verb or an infinitive which can be resolved based on the context of the word in a given sentence. Similar errors were found in Pandey et al. (2017) experiment where author proposed Cricket as a domain for adaptation in a Multi-domain Hindi Tagger. The present work, initially, tests the tagger at two levels- (i) with Cricket as one of the tagger domain and (ii) without Cricket as a tagger domain. Thereafter, it aims at creating a multi-domain Hindi Tagger by justifying the adaptation of Cricket to the corpus and improving on accuracy. It further discusses the error pattern of POS tagger trained for Cricket and the issues overcame under this stage of adaptation.

2. Literature Survey

L. Smith, T. Rindflesch and W.J. Wilbur (2004) develop a MedPost POS tagger for biomedical text with 97% accuracy[1]. This tagger (called MedPost) was developed to meet the need for a high accuracy part-of-speech tagger trained on the MEDLINE corpus. And after that N. Barrett and J. Weber-Jahnke present a POS tagger that is more accurate than two frequently used biomedical POS taggers when trained on a non-biomedical corpus and evaluated on the MedPost corpus. They present a POS tagger for cross-domain tagging called TcT. TcT was more accurate in cross-domain tagging than mxpost. P. Nand, R. Perera and R. Lal (2004) present a HMM POS Tagger for Micro-blogging type texts. They present the results of testing a HMM based POS (Part-Of-Speech) tagging model customized for unstructured texts[2]. They also evaluated the tagger against published CRF based state-of-the-art POS tagging models customized for Tweet messages using three publicly available Tweet corpora. Finally, they did cross-validation tests with both the taggers by training them on one Tweet corpus and testing them on another one. S. Kinoshita, K Bretonel Cohen, P.V. Ogren and L. Hunter (2005) present a BioCreAtIVE Task1A: entity identification with a stochastic tagger. They describe the methods of entity identification (also known as named entity recognition) in the molecular biology domain. The goal of BioCreAtIvE Task1A is to assess the ability of an

[1]https://academic.oup.com/bioinformatics/article/20/14/2320/213968

[2]https://link.springer.com/chapter/10.1007/978-3-319-13560-1_13

automated system to identify mentions of genes in text from biomedical literature.

3. Training the taggers

3.1 Corpus for the experiment

Training corpus**:** The pre-trained Hindi POS tagger includes health, tourism, agriculture and entertainment as four major domains with an approx. 200k tokens. The tagger is then re-trained with approx. 250k tokens in five major domains including Cricket with approx. 50k tokens from each domain. The training data is 80 % of the total corpus and the rest 20 % is the test data. The domain of Cricket comprises miscellaneous data, majorly IPL data.

Test corpus: The taggers have been tested on a set of 21k tokens from Cricket Domain. The same test set is used for both the taggers. The tagger is trained using Support Vector model with no change in the feature selection of the tool for reducing the influence on the result. The test results obtained are discussed in the section below.

4. Test result and evaluation of tagger output

As a result of the experiment, the four domain tagger (4DT[3]), without Cricket data, is calculated to have an accuracy of 85.73% and the accuracy for the five domain tagger (5DT), with Cricket data, is 93.97%. There is a fall noted in the tagger accuracy of the 4DT tagger with noticeable issues when tested for Cricket data. The new accuracy after adaptation (in 5DT) is found 1% lower than the accuracy of pre-existing Hindi tagger (with three domains) which was 94% for random Hindi data as mentioned in Ojha (2015).

4.1 Error report of 4DT

Table 1 shows the tagger generated category wise error report.[4]

Table 1: Accuracy report of 4Domain tagger (4DT)

CC_CCD	509	515	98.8350%
CC_CCS	256	324	79.0123%
DM_DMD	405	486	83.3333%
DM_DMI	63	70	90.0000%
DM_DMQ	1	8	12.5000%
DM_DMR	2	17	11.7647%
JJ	414	614	67.4267%
N_NN	4097	4816	85.0706%
N_NNP	1485	2348	63.2453%
N_NST	391	432	90.5093%
PR_PRC	0	2	0.0000%
PR_PRF	92	92	100.0000%
PR_PRL	137	155	88.3871%
PR_PRP	533	533	100.0000%
PR_PRQ	30	30	100.0000%
PSP	3677	3783	97.1980%
QT_QTC	795	810	98.1481%
QT_QTF	194	259	74.9035%
QT_QTO	96	194	49.4845%
RB	23	47	48.9362%
RD_PUNC	1388	1435	96.7247%
RD_RDF	2	4	50.0000%
RD_SYM	273	285	95.7895%
RP_INTF	55	72	76.3889%
RP_NEG	185	186	99.4624%
RP_RPD	315	316	99.6835%
V_VAUX	1153	1246	92.5361%
V_VM	1624	2143	75.7816%
==========	OVERALL ACCURACY	==========	
HITS	TRIALS	ACCURACY	
18195	21222	85.7365%	

The cases of interrogative demonstrative, relative demonstrative, reciprocal pronoun, ordinal, adverb, and foreign residual are found to have accuracy below 50% in the four domain Hindi tagger. The occurrence of these tags in the test data is also very low (below 200 occurrences). But the errors found with adjectives, common nouns, proper nouns, deictic demonstrative, general quantifier, intensifier and main verbs are the noticeable ones because they have relatively higher occurrences and accuracy between 50-85%. On the other hand, cases of co-ordinator, reflexive pronoun, personal pronoun, interrogative pronoun, cardinal negative and default particle are found to have the highest accuracies (98% and above) by the tagger.

[3] 4DT and 5DT stands for 4 domain tagger and 5 domain tagger, respectively.

[4] The parts of speech categories mentioned in table 1 and 2 are namely: CCD-coordinator, CCS-subordinator, DMD-deictic demonstrative, DMI-indefinite demonstrative, DMQ-interrogative demonstrative, DMR-relative demonstrative, JJ-adjective, NN-common noun, NNP-proper noun, NST-noun locative, PRC-reciprocal pronoun, PRF-reflexive pronoun, PRL-relative pronoun, PRP-personal pronoun, PRQ-interrogative pronoun, PSP-postposition, QTC-cardinal, QTF-general quantifier, QTO-ordinal, RB-adverb, PUNC-punctuation, RDF-foreign residual, SYM-symbol, INTF-intensifier, NEG-negation, RPD-default particle, VAUX-auxiliary verb, VM-main verb.

4.2 Error report of 5DT

The cases of lowest accuracy percentages (below 50%) are found with only interrogative demonstratives, relative demonstratives, and reciprocal pronoun in the 5 domain Hindi tagger which are also noted with lowest occurrences in the test data. Whereas, the cases of adjectives, general quantifier, adverbs and intensifier are having frequency accuracies between 50-85%. On the other hand, the tags with highest accuracy ratio in 5 domain tagger are coordinator, noun locative, reflexive, personal and interrogative pronouns, postpositions, cardinal, punctuations, symbol, negation and default particles with 98% accuracy and above.

Table 2: Accuracy report of 5 Domain Tagger (5DT)

CC_CCD	514	515	99.8058%
CC_CCS	285	324	87.9630%
DM_DMD	433	486	89.0947%
DM_DMI	66	70	94.2857%
DM_DMQ	0	8	0.0000%
DM_DMR	3	17	17.6471%
JJ	519	614	84.5277%
N_NN	4500	4816	93.4385%
N_NNP	2115	2348	90.0767%
N_NST	424	432	98.1481%
PR_PRC	0	2	0.0000%
PR_PRF	92	92	100.0000%
PR_PRL	141	155	90.9677%
PR_PRP	533	533	100.0000%
PR_PRQ	30	30	100.0000%
PSP	3769	3783	99.6299%
QT_QTC	800	810	98.7654%
QT_QTF	196	259	75.6757%
QT_QTO	182	194	93.8144%
RB	36	47	76.5957%
RD_PUNC	1408	1435	98.1185%
RD_RDF	4	4	100.0000%
RD_SYM	281	285	98.5965%
RP_INTF	61	72	84.7222%
RP_NEG	186	186	100.0000%
RP_RPD	315	316	99.6835%
V_VAUX	1195	1246	95.9069%
V_VM	1856	2143	86.6076%
========== OVERALL ACCURACY ==========			
HITS	TRIALS	ACCURACY	
19944	21222	93.9779%	

4.3 Comparative error accuracy report

Based on the output reported above, the graphical representation of the POS tags with lowest accuracy with least occurrences, lowest accuracies with high occurrences and highest accuracies are given in Figure 1, Figure 2, Figure 3 and Figure 4.

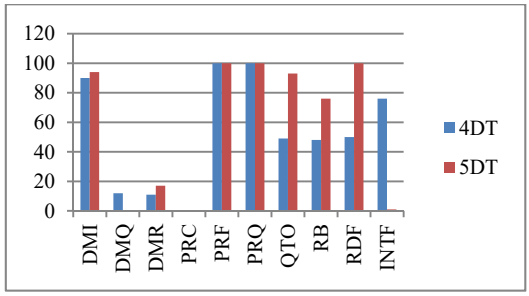

Figure 1: A graph on accuracy with least frequency of occurrence below 200

Among the parts of speech categories with less than 200 occurrences in the test data, the reciprocal pronoun achieved the lowest accuracy while reflexive and interrogative pronoun is found with the highest accuracies reaching 100% for both the taggers. The error rate of ordinals, adverbs and foreign word were found significantly decreasing from 4DT to 5DT whereas intensifiers are found with increasing error rates.

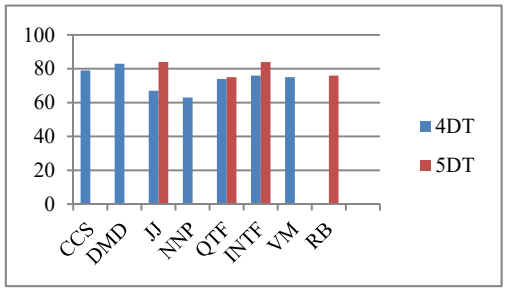

Figure 2: A graph on accuracy between 50-85% with relatively high frequency of occurrence

Figure 3 shows the POS categories achieving accuracies between 50 to 85% in both the tagger. As shown in the figure, adjectives, general quantifiers, intensifier and adverbs fall under this accuracy range in 5DT tagger whereas 4DT includes some more categories like subordinator, demonstrative, proper nouns and main verb whose accuracies have been improved for the 5DT.

Among the tags achieving highest accuracies (85% and above) with high frequency tags, the functional tags are coordinator, postpositions, and punctuations and the content word categories are demonstrative, noun, pronoun, cardinal, and verb

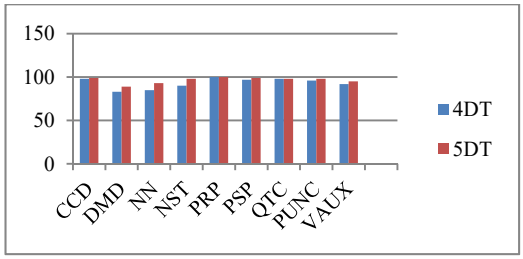

Figure 3: A graph on highest accuracies with higher occurrence.

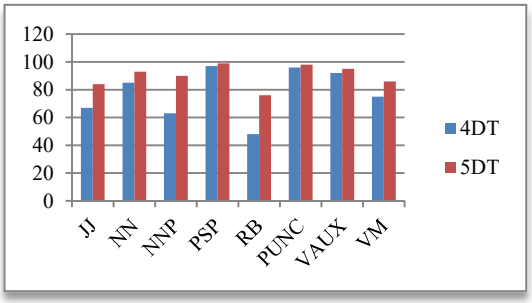

Figure 4: Difference making categories above 500 frequency of occurrence.

As Figure 4 shows, the difference in accuracies of the major POS tags ranges between 2 to 28%. The accuracy difference in adverb is 28%, in adjectives and proper noun it is 27%, in main verb it is 11%, in noun it is 8%, in auxiliary it is 3%, and in postposition and punctuation it is 2%.

5. Comparative analysis of the tagger outputs

5.1 Resolved errors

These are the type of errors mentioned in the Pandey et al. (2017). The similar problem is also reported in the four domain Hindi tagger (except Cricket as a domain) which is found resolved in the five domain tagger (including Cricket as a domain). The outputs of both the taggers are presented below:

Common and Proper nouns**:** Like earlier experiment, the issue of Common versus Proper nouns is still present in the four domain tagger. Names like *India, IPL, Anti-Corruption Unit* etc. are found incorrectly marked by the tagger. Whereas, such errors are barely found in the five domain tagger.

(1) SukravArakokhelAjAnAhai (Itans)
To be played on Friday (English)
SukravAr\N_NNPko\PSP (4DT)
SukravAr\N_NN ko\PSP (5DT)

Although, SukravAr (Friday) is the name of a week day and must be a part of Proper Noun category but the BIS Hindi guideline and the training data counts only complete dates like '2 January' as proper noun. According to the training corpus the 4DT has tagged it incorrectly as common noun which is resolved in the 5DT.

Verb class: The problem of main verbs is found reduced due to the influence of the data from other domain and higher frequency of main verbs in the test data in Hindi tagger. But the problem of auxiliaries in serial verb construction still persists. The problem of main verb is considerably high in the Cricket tagger as compared to the problem of serial verbs. One reason for this might be the lesser frequency of such entries in the training data.

(2) SukravArakokhelAjAnAhai (Itans)
To be played on Friday (English)
khelA\N_NN**jAnA**\V_VAUX **hai**\V_VM(4DT)
khelA\V_VM jAnA\V_VAUX
hai\V_VAUX(5DT)

The serial verb 'to be played' is incorrectly marked as a conjunct (N+V) in the 4DT and the auxiliary *hai* is tagged as the main verb whereas the 5DT has correctly annotated it as a SVC[5].

(3) giraftArakiyegaye the (Itrans)
arrested (English)
giraftAra\N_NN **kiye**\V_VM **gaye**\V_VAUX
the\V_VM (4DT)
giraftAra\N_NN kiye\V_VM gaye\V_VAUX
the\V_VAUX (5DT)

The conjunct verb has also been incorrectly tagged by the 4DT which is not found in the 5DT tagger output.

Noun over adjectives: Words like *dilachaspa* (interesting), *khitAbI* (honoured), *AkhirI* (last) etc. are incorrectly marked as common nouns. This error type is found mainly in the 4DT. Like previous experiment the qualifier *strange and scary* is marked incorrectly in the 4DT.

[5]SVC- Serial Verb Construction

(4) Ajeeba aura bhayAvahasthitI (Itrans)
Strange and scary situation (English)
Ajeeba\N_NN aura\CC_CCD **bhayAvaha**\N_NN sthitI\N_NN (4DT)
Ajeeba\JJ aura\CC_CCD bhayAvaha\JJ sthitI\N_NN (5DT)

Ordinals: The spelled out cardinals are tagged as common noun by the 4DT but not in the 5DT as shown in the example below. Besides this, the ordinals like *donoM* (both) ,*teenoM* (all three) are still found marked as cardinals in the 4DT Hindi tagger which is resolved in the Cricket tagger.

(5) rAjasthAnaroyalsaketInatInakrikeTar(Itrans)
Three cricketers of Rajasthan Royals (English)
rAjasthAna\N_NNProyalsa\N_NNPke\PSP**tIna**\N_NN**tIna**\N_NNkeikeTar\N_NN (4DT)
rAjasthAna\N_NNP royalsa\N_NNP ke\PSP tIna\QT_QTC tIna\QT_QTC keikeTar\N_NN (5DT)

Postpositions: The annotators in the previous experiment came to a consensus of marking cases like ***khilAfa,daurAna***and***alAwA*** as postpositions. While testing, the 4DT is found marking such instances as common noun whereas the 5DT has tagged it correctly. This might be due to the presence of consistent tagging of postpositions in the Cricket corpus.

(6) usakekhilAfakoIkAryawAhInahIMhogI(Itrans)
No action will be taken against him(English)
Usake\PR_PRP **khilAfa\N_NN** (4DT)
Usake|PR_PRPkhilAfa\PSP (5DT)

5.2 Unresolved errors

The errors found in the five domain Hindi tagger (5DT) are not yet resolved and it is under development. Those errors are listed in the present section.

Common and proper noun**:** Some of the proper nouns like names of committee and organisations are found tagged categorically by both the taggers.

(1) aenTIkarapSanayUniT (Itans)
Anti Corruption Unit (English)
aenTI\N_NN **karapSana**\N_NNP **yUniT**\N_NN (4DT)
aenTI\JJ**karapSana**\N_NN**yUniT**\N_NN (5DT)

Anti-Corruption Unit of BCCI is a regulating committee which is incorrectly tagged as common noun phrase by both the taggers.

(2) BevarejaaurasnaekskshetrakIpramukhakampanIpepasikoinDiA (Itrans)
Pepsico India, one of the major companies in the field of snacks and baverages (English)

bevareja\N_NN **aura\CC_CCD** **snaeksa\N_NN**kshetra\N_NN kI\PSP pramukha\JJ kampanI\N_NNP pepasiko\N_NNP inDiA\N_NNP (4DT)

bevareja\N_NNP **aura\CC_CCD** **snaeks\N_NNP**kshetra\N_NN kI\PSP pramukha\JJ kampanI\N_NN pepasiko|N_NNPinDiA|N_NNP (5DT)

The first chunk 'beverage and snacks' is incorrectly marked by 4DT as common noun but the 5DT has correctly tagged it as a proper noun. Whereas, the final chunk of the example 'Pepsico India' is correctly marked by both the taggers.

(3) TI 20 (Itrans)
T 20 (English)
TI\N_NN **20**\N_NN (4DT)
TI\N_NN **20**\QT_QTC (5DT)

T 20 representation in the data stands for the name of a Cricket match like IPL and world cups etc. Sometimes it appears without a space or hyphen in between (like T20 or T-20) otherwise also found written with a space (T_20) as in the example above. The earlier cases without a space are always tagged correctly by the tagger as a Proper Noun but the latter case is mistaken with the tag label category. The 5DT tagger has tagged it category wise but the 4DT tagger has again mistaken a cardinal '20' with a noun tag. Therefore, in order to resolve this kind of error, the data can either be edited in the former format (removing spaces or adding hyphens) which might not be an appropriate solution for other similar unknown occurrences or more similar content can be added to the corpus in order to increase the frequency of such cases.

6. Conclusion

The present paper aimed at evaluating the accuracy rate of the Hindi taggers trained for two types of Corpus- (i) The four domain Hindi tagger excluding Cricket corpus and (ii) five domain

Hindi tagger including Cricket corpus. As a result of the test, the first tagger (4DT) was found exhibiting an accuracy of approx. 85% listing similar nature of errors as previously mentioned in the paper on proposal for adaptation of Cricket to the Hindi tagger. Whereas, the accuracy of second tagger (5DT) is found exhibiting an accuracy of approx. 93%. After adaptation the tagger accuracy is not exact but it is around the initial accuracy (94%) of Hindi tagger trained and tested for multi domain general Hindi Corpus. Towards the end of the paper, a comparative report on the tagger accuracy, as effect of adaptation for Cricket domain, is presented. Noticeable erroneous cases of adjectives, nouns, adverbs and main verbs came into light as major factor affecting the rest 7% accuracy of the 5 domain tagger trained with 250k tokens. The present tagger can be further extended with a larger Cricket Corpus in order to enhance the performance of the tagger.

References

Barrett, N. & Weber-Jahnke J. 2011. A Token Centric Part-of-Speech Tagger for Biomedical Text, *In Proceeding of Conference on Artificial Intelligence in Medicine in Europe*. Retrieved from:https://link.springer.com/chapter/10.1007/978-3-642-22218-4_41 January 4, 2018.

Ekbal A., Bandhopadhyay S. 2009. A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi. Published by CSLI, in Linguistics Issues in Language Technology – LiLT, Volume 2.Retrieved from: https://journals.linguisticsociety.org/elanguage/lilt/article/download/213/213-500-1-PB.pdf December 24, 2017.

Joshi, N., Darbari, H. & Mathur I. 2017. HMM Based Pos Tagger For Hindi, In Proceedings of *3rd International Conference on Computer Science & Information Technology*. Retrieved from: file:///C:/Users/Pc/Downloads/csit3639.pdf December 22, 2018

Kinoshita, S., Cohen, K. B., Ogren, P.V., & Hunter, L. BioCreAtIvE Task1A: entity identification with a stochastic tagger, *Published in Kinoshita et al; licensee BioMed Central Ltd 2005.* Retrieved from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-S1-S4 January 4, 2018.

Miller, J.E., Bloodgood, M., Torri, M., & Shanker V. 2005. Rapid Adaptation of POS Tagging for Domain Specific Uses, *In Proceeding of the HLT-NAACL BoiNLP workshop on Linking Natural Language and Biology, June 2005*. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1411/1411.0007.pdf December 24, 2017.

Miller, J.E., Bloodgood, M., Torri, M., & Shanker V. 2005. Rapid Adaptation of POS Tagging for Domain Specific Uses, In *Proceeding of the HLT-NAACL BoiNLP workshop on Linking Natural Language and Biology, June 2005*. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1411/1411.0007.pdf

Nand, P., Perera, R., & Lal R. 2014. A HMM POS Tagger for Micro-blogging Type Texts, *In Proceeding of Pacific Rim International Conference on Artificial Intelligence*. Retrieved from: https://link.springer.com/chapter/10.1007/978-3-319-13560-1_13 January 4, 2018.

Ojha, A. 2015. Hindi POS Tagger. In *Proceedings of 7th Language & Technology Conference: Human Language Technology as a Challenge for Computer Science and Linguistics, LTC 2015.*

Pandey, Anupama., Singh, Sirshti., Ojha, Atul Kr., &Jha, Girish Nath. 2017. Challenges in Annotation and Domain Adaptation in Hindi POS Tagger: with Reference to Cricket. In proceedings of the International Conference on Applied and Theoretical Computing and Communication Technology (ICATCCT - 2017), pp. 155-159.

Shrivastava, M. &Battacharyya, P. 2008. Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge. In *Proceedings of International Conference on NLP (ICON08)*, Macmillan Press New Delhi. Retrieved from: https://www.cse.iitb.ac.in/~pb/papers/icon08-hindi-pos-tagger.pdf 24 December 2017

Smith, L., Rindflesch ,T., & Wilbur, W.J. 2004. MedPost: a part-of-speech tagger for bioMedical text. available at ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gzRetrieved from: https://www.ncbi.nlm.nih.gov/pubmed/15073016 January 4 2018.

T. Brants. 2000. Tnt – A Statistical Part-of-Speech tagger, In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP*. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.8901&rep=rep1&type=pdf January 4, 2018.

Web Links:

https://www.cse.iitb.ac.in/~pb/papers/icon-08hindi-pos-tagger.pdf

file:///C:/Users/Pc/Downloads/csit3639.pdf