The Illiterati

Part-of-speech tagging for Magahi and Bhojpuri without even knowing the alphabet

Thomas Proisl, Peter Uhrig, Philipp Heinrich, Andreas Blombach, Sefora Mammarella, Natalie Dykes, Besim Kabashi

Chair of Computational Corpus Linguistics
Friedrich-Alexander-Universität Erlangen-Nürnberg

September 11, 2019

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

NLP Solutions for Under Resourced Languages (NSURL 2019)

Tasks 9 and 10: Part-of-speech tagging for Magahi and Bhojpuri

	Magahi	Bhojpuri
tag set	18	33
training	61.435	94.692
test	8.205	10.582

Our strategies

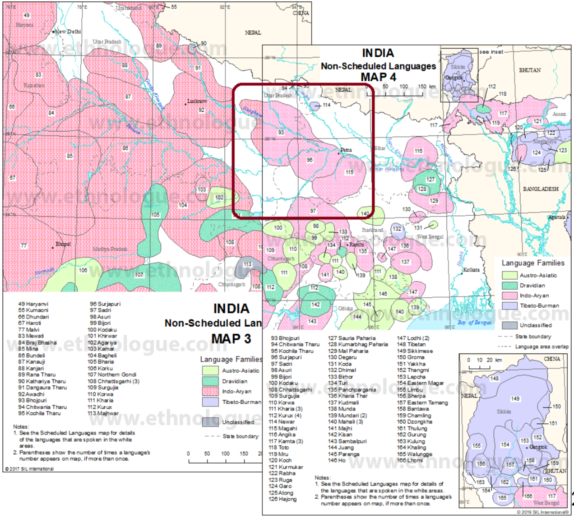
- use customisable off-the-shelf taggers
- include additional resources (transfer learning)
  - Brown clusters
  - word embeddings
  - tagged corpora of related languages

Language Map: Bhojpuri in India

 [1]

[1] source:https://commons.wikimedia.org/wiki/File:Bhojpuri_Speaking_Region_in_India.png

Language Map: Magahi and Bhojpuri



[2]

---

[2] sources: https://www.ethnologue.com/map/IN_03, https://www.ethnologue.com/map/IN_04

About Magahi and Bhojpuri

- principal languages of the Bihari group (West-Eastern Indo-Aryan language)
  - Magahi
  - Bhojpuri
  - Maithili
- Bhojpuri: approx. 51 million native speakers (2011 census), spoken in Western Bihar, Eastern Uttar Pradesh, and in Southwest Nepal
- Magahi: approx. 13 million native speakers (21 million if Khortha, a prominent dialect, is also included), mainly spoken in Southern Bihar

Interesting features w.r.t. POS tagging

- SOV order
- rich verb morphology
- extensive use of postpositions
- Magahi: unusual agreement system (verb has to agree with subject *and* object)

Three off-the-shelf POS Taggers

- SoMeWeTa[3] (Proisl, 2018)
  - based on averaged structured perceptron
  - supports domain adaptation and external resources
- BiLSTM+CRF sequence tagger (Guillaume Genthial, Riedl and Padó (2018)[4])
  - based on character and word embeddings
  - supports transfer learning
- The Stanford Tagger[5] (Toutanova et al., 2003)
  - based on a maximum entropy cyclic dependency network
  - hyperparameter tuning

[3] https://github.com/tsproisl/SoMeWeTa
[4] https://github.com/riedlma/sequence_tagging
[5] https://nlp.stanford.edu/software/tagger.html

Additional Resources

freely available resources used in addition to training data

- Hindi UD treebank (HDTB; ca. 352,000 tokens)[6]
- Two Magahi corpora[7]
  - POS-tagged Magahi corpus (KMI-Mag; ca. 46,000 tokens)
  - Corpus of untagged Magahi texts (ca. 2.8 million tokens)
- Plain text extracted from Wikimedia dumps[8]
  - Hindi (ca. 34.7 million tokens)
  - Bihari (ca. 700,000 tokens)
- Brown clusters (Brown et al., 1992) computed from Wikimedia dumps and untagged Magahi corpus
- Pre-trained fastText embeddings for Hindi and Bihari[9]

[6] https://github.com/UniversalDependencies/UD_Hindi-HDTB
[7] https://github.com/kmi-linguistics/magahi
[8] https://dumps.wikimedia.org
[9] https://fasttext.cc/docs/en/crawl-vectors.html

Experiments Using SoMeWeTa

Focus on Brown clusters and transfer learning

Cross-validation results on training data (selection)

model	accuracy
Bhojpuri (no additional resources)	91.62 (±0.97)
Bhojpuri (hi)	91.79 (±1.00)
Bhojpuri (hi+mag)	91.99 (±0.83)
Bhojpuri (hi+bh+mag)	92.04 (±0.80)
KMI-Mag → Bhojpuri (hi+bh+mag)	*92.06 (±0.94)*
Magahi (no additional resources)	88.92 (±1.24)
Magahi (mag)	89.12 (±1.23)
Magahi (hi+mag)	*89.32 (±1.15)*
Magahi (hi+bh+mag)	89.15 (±1.17)
KMI-Mag+Bhojpuri → Magahi (hi+mag)	89.30 (±1.14)

Brown clusters beneficial, additional transfer learning not

Experiments Using SoMeWeTa

Focus on Brown clusters and transfer learning

Cross-validation results on training data (selection)

model	accuracy
Bhojpuri (no additional resources)	91.62 ($\pm$0.97)
Bhojpuri (hi)	91.79 ($\pm$1.00)
Bhojpuri (hi+mag)	91.99 ($\pm$0.83)
Bhojpuri (hi+bh+mag)	92.04 ($\pm$0.80)
KMI-Mag $\rightarrow$ Bhojpuri (hi+bh+mag)	*92.06 ($\pm$0.94)*
Magahi (no additional resources)	88.92 ($\pm$1.24)
Magahi (mag)	89.12 ($\pm$1.23)
Magahi (hi+mag)	*89.32 ($\pm$1.15)*
Magahi (hi+bh+mag)	89.15 ($\pm$1.17)
KMI-Mag+Bhojpuri $\rightarrow$ Magahi (hi+mag)	89.30 ($\pm$1.14)

Brown clusters beneficial, additional transfer learning not

Experiments Using the BiLSTM-CRF Tagger

Focus on embeddings and transfer learning

Cross-validation results on training data

model	accuracy
Magahi (Hindi embeddings)	88.97 ($\pm$1.14)
Magahi (Bihari embeddings)	89.09 ($\pm$1.00)
HDTB $\rightarrow$ Magahi (Hindi embeddings)	89.85 ($\pm$0.99)
KMI-Mag $\rightarrow$ Magahi (Hindi embeddings)	*90.70 ($\pm$0.92)*
Bhojpuri (Hindi embeddings)	90.78 ($\pm$0.55)
Bhojpuri (Bihari embeddings)	90.80 ($\pm$0.57)
KMI-Mag $\rightarrow$ Bhojpuri (Hindi embeddings)	*91.23 ($\pm$0.68)*

Using Hindi embeddings and pretraining on KMI-Mag works best for both languages!

Experiments Using the BiLSTM-CRF Tagger

Focus on embeddings and transfer learning

Cross-validation results on training data

model	accuracy
Magahi (Hindi embeddings)	88.97 ($\pm$1.14)
Magahi (Bihari embeddings)	89.09 ($\pm$1.00)
HDTB $\rightarrow$ Magahi (Hindi embeddings)	89.85 ($\pm$0.99)
KMI-Mag $\rightarrow$ Magahi (Hindi embeddings)	*90.70 ($\pm$0.92)*
Bhojpuri (Hindi embeddings)	90.78 ($\pm$0.55)
Bhojpuri (Bihari embeddings)	90.80 ($\pm$0.57)
KMI-Mag $\rightarrow$ Bhojpuri (Hindi embeddings)	*91.23 ($\pm$0.68)*

Using Hindi embeddings and pretraining on KMI-Mag works best for both languages!

Experiments Using the Stanford Tagger: Magahi

parameter	default	value/range
closedClassTags	(none)	ADP AUX CCONJ DET NUM PART PRON SCONJ PUNCT
arch/macro	generic	generic, left3word, bidirectional5words
arch/further unknown-words option	(none)	naacl2003unknowns
arch/unicode shapes for rare words	(none)	(-2,2), (-1,1), (0), (none)
iterations	100	100
learnClosedClassTags	false	false
curWordMinFeatureThresh	2	1..4
minFeatureThresh	5	1..5
rareWordMinFeatureThresh	10	1..10
rareWordThresh	5	1..8
veryCommonWordThresh	250	100, 150, 200, 250

- 76,800 hyperparameter combinations
- 2 runs per parameter combination (first and last 20% as test data)

153,600 training cycles (FAU HPC)

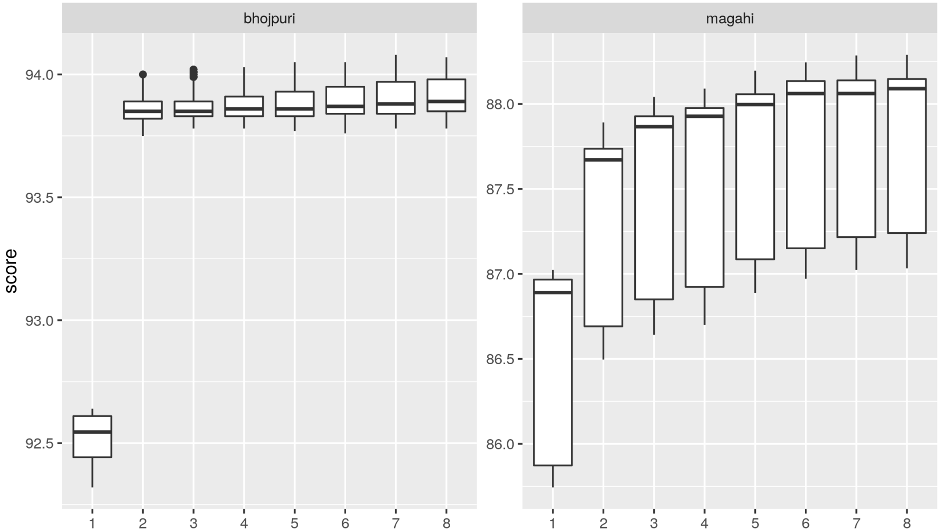
Experiments Using the Stanford Tagger: Bhojpuri

parameter	default	value/range
closedClassTags	(none)	(none)
arch/macro	generic	generic, left3word, bidirectional5words
arch/further unknown-words option	(none)	naacl2003unknowns
arch/unicode shapes for rare words	(none)	(-2,2), (-1,1), (0), (none)
iterations	100	100
learnClosedClassTags	false	true
closedClassTagThreshold	40	40
curWordMinFeatureThresh	2	1..4
minFeatureThresh	5	1..5
rareWordMinFeatureThresh	10	1..10
rareWordThresh	5	1..8
veryCommonWordThresh	250	100, 150, 200, 250

- 76,800 parameter combinations
- full 10-fold crossvalidation

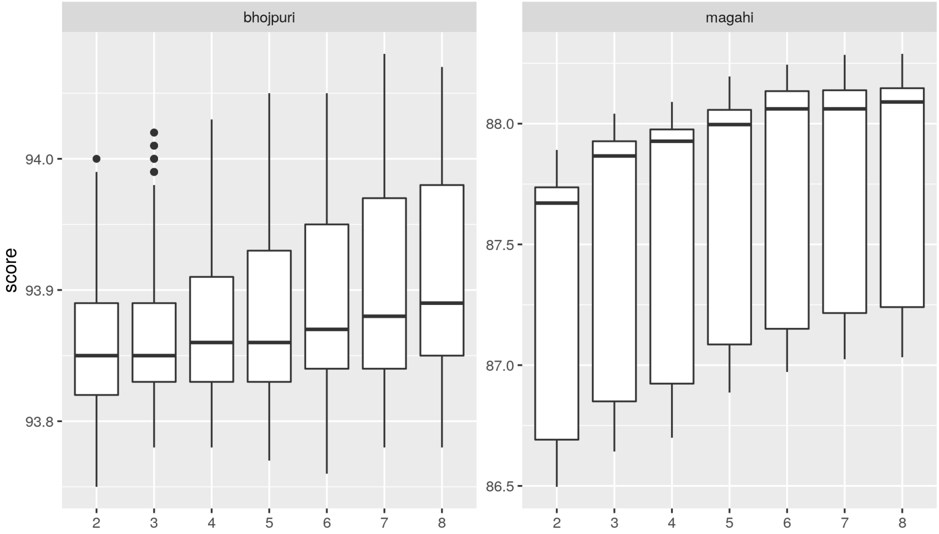
768,000 training cycles (FAU HPC)

Parameter Analysis: rareWordThresh



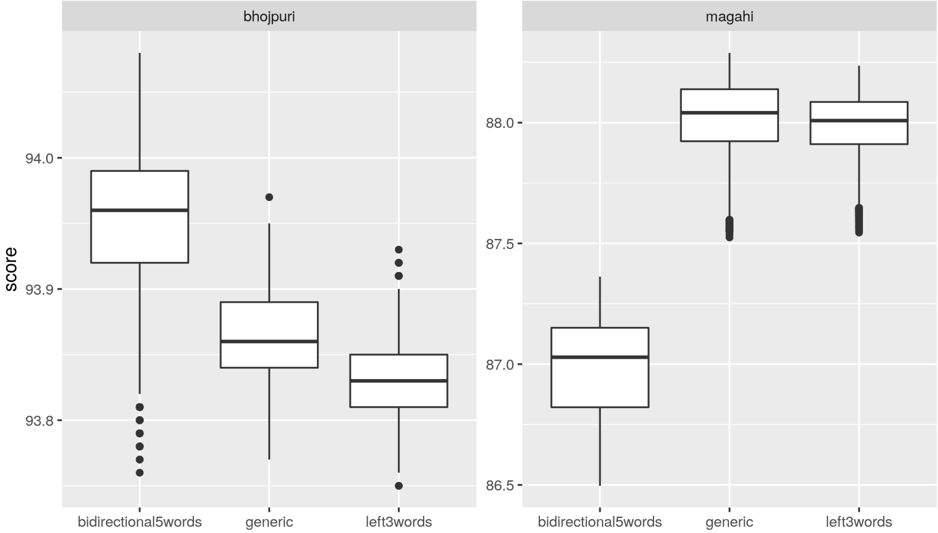
Parameter Analysis: rareWordThresh

exluding rareWordThresh=1



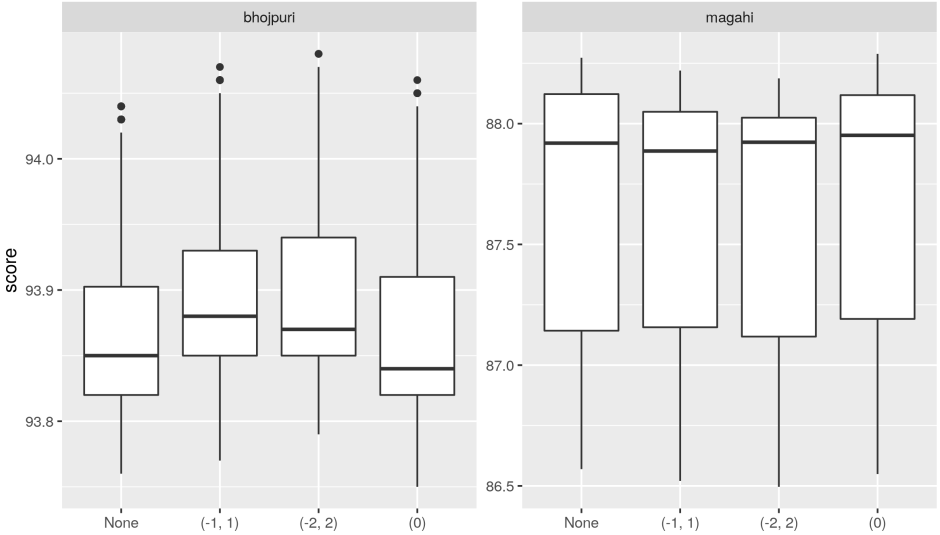
Parameter Analysis: macro

exluding rareWordThresh=1



Parameter Analysis: unicodeshape

exluding rareWordThresh=1



Parameter Analysis: Summary

- performance decreases abruptly when *rareWordThresh* is set to 1 (at least hapax legomena should be treated as rare words)
- performance was insensitive to variation in *veryCommonWordThresh* (this option was ignored by the system)
- macro has most influence
  - Bhojpuri: `bidirectional5words`
  - Magahi: `generic` and `left3words` training data annotation?
- *rareWordThresh* explains most of the remaining variation

Official Results

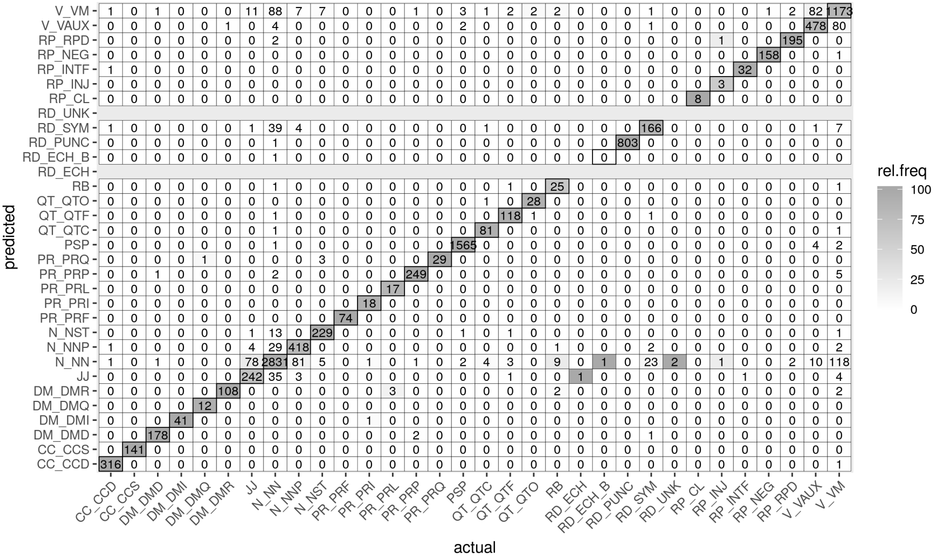
	submission	F_1	acc.
1	**Stanford**	95	94.78
1	NITK-NLP_SUB1	95	
2	**SoMeWeTa**	93	92.76
3	**BiLSTM-CRF**	92	92.01
4	NITK-NLP_SUB2	89	

Table: Results for Bhojpuri

	submission	F_1	acc.
1	NITK-NLP_SUB2	79	
2	**BiLSTM-CRF**	77	78.86
2	**SoMeWeTa**	77	78.68
3	**Stanford**	74	76.57
4	NITK-NLP_SUB1	73	

Table: Results for Magahi

Bhojpuri: Confusion Matrix for BiLSTM

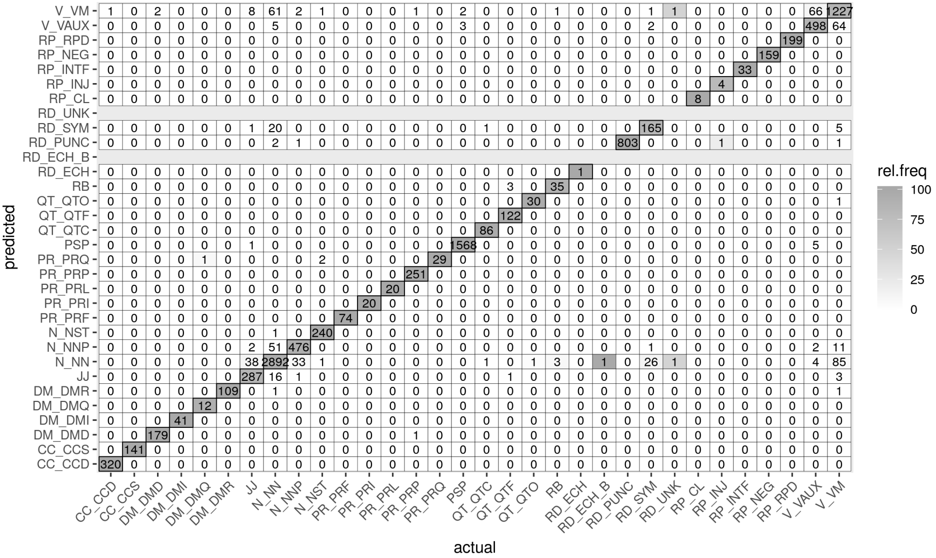


Bhojpuri: Confusion Matrix for SoMeWeTa

predicted \ actual	CC_CCD	CC_CCS	DM_DMD	DM_DMI	DM_DMQ	DM_DMR	JJ	N_NN	N_NNP	N_NST	PR_PRF	PR_PRI	PR_PRL	PR_PRP	PR_PRQ	PSP	QT_QTC	QT_QTF	QT_QTO	RB	RD_ECH	RD_ECH_B	RD_PUNC	RD_SYM	RD_UNK	RP_CL	RP_INJ	RP_INTF	RP_NEG	RP_RPD	V_VAUX	V_VM
V_VM	1	0	1	0	0	0	7	86	1	4	0	1	0	3	0	2	0	2	0	0	0	0	0	2	1	0	0	0	0	4	80	1179
V_VAUX	0	0	0	0	0	1	0	6	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	481	66
RP_RPD	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	195	0	0
RP_NEG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	159	0	0	1
RP_INTF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	1
RP_INJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
RP_CL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0
RD_UNK																																
RD_SYM	0	0	0	0	0	0	1	29	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	158	0	0	0	0	0	0	0	5
RD_PUNC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	803	2	0	0	0	0	0	0	1	0
RD_ECH_B																																
RD_ECH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
RB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	30	0	0	0	0	0	0	0	0	0	0	0	0
QT_QTO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	1
QT_QTF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	117	1	0	0	0	0	0	0	0	0	0	0	0	0	0
QT_QTC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
PSP	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1567	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3
PR_PRQ	0	0	0	0	0	0	0	0	0	2	0	1	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PR_PRP	0	0	1	0	0	0	0	1	0	0	0	0	0	249	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
PR_PRL	0	0	0	0	0	0	0	1	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PR_PRI	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PR_PRF	0	0	0	0	0	0	0	0	1	0	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
N_NST	0	0	0	0	1	0	1	5	0	232	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
N_NNP	0	0	0	0	0	0	3	43	449	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	8
N_NN	1	0	1	0	0	0	74	2851	57	5	0	3	0	0	0	0	3	3	1	6	0	1	0	32	1	0	1	0	0	0	6	115
JJ	0	0	0	0	0	0	250	24	4	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	10
DM_DMR	0	0	0	0	0	108	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	3
DM_DMQ	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DM_DMI	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DM_DMD	0	0	178	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC_CCS	0	141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC_CCD	319	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

rel.freq: 100, 75, 50, 25, 0

Bhojpuri: Confusion Matrix for Stanford Tagger



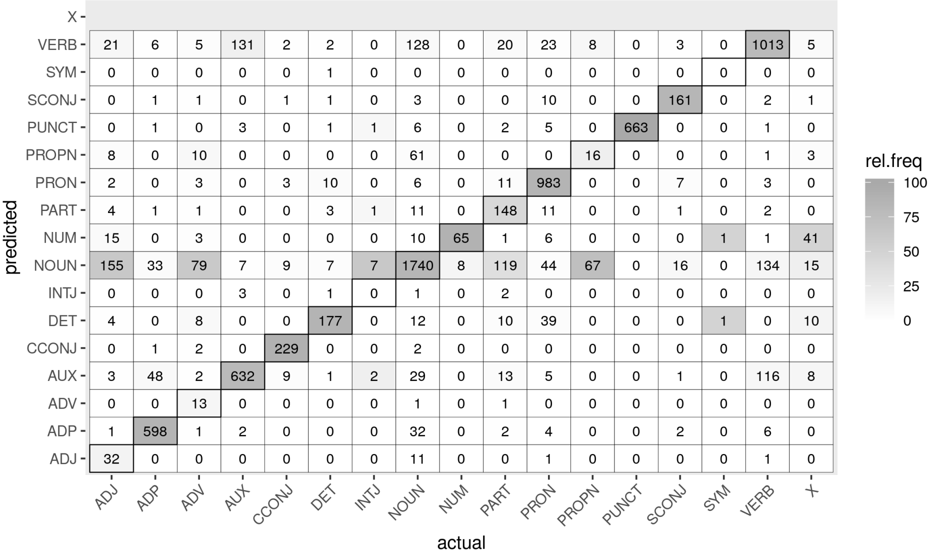
predicted \ actual	CC_CCD	CC_CCS	DM_DMD	DM_DMI	DM_DMQ	DM_DMR	JJ	N_NN	N_NNP	N_NST	PR_PRF	PR_PRI	PR_PRL	PR_PRP	PR_PRQ	PSP	QT_QTC	QT_QTF	QT_QTO	RB	RD_ECH	RD_ECH_B	RD_PUNC	RD_SYM	RD_UNK	RP_CL	RP_INJ	RP_INTF	RP_NEG	RP_RPD	V_VAUX	V_VM
V_VM	1	0	2	0	0	0	8	61	2	1	0	0	0	1	0	2	0	0	0	1	0	0	0	1	1	0	0	0	0	0	66	1227
V_VAUX	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	2	0	0	0	0	0	0	498	64
RP_RPD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	199	0	0
RP_NEG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	159	0	0	0
RP_INTF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0
RP_INJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
RP_CL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0
RD_UNK																																
RD_SYM	0	0	0	0	0	0	1	20	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	165	0	0	0	0	0	0	0	5
RD_PUNC	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	803	0	0	0	1	0	0	0	0	1
RD_ECH_B																																
RD_ECH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
RB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	35	0	0	0	0	0	0	0	0	0	0	0	0
QT_QTO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	1
QT_QTF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	122	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QT_QTC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PSP	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1568	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
PR_PRQ	0	0	0	0	1	0	0	0	0	2	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PR_PRP	0	0	0	0	0	0	0	0	0	0	0	0	0	251	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PR_PRL	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PR_PRI	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PR_PRF	0	0	0	0	0	0	0	0	0	0	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N_NST	0	0	0	0	0	0	0	1	0	240	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N_NNP	0	0	0	0	0	0	2	51	476	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	11
N_NN	0	0	0	0	0	0	38	2892	33	1	0	0	0	0	0	0	1	0	1	3	0	1	0	26	1	0	0	0	0	0	4	85
JJ	0	0	0	0	0	0	287	16	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3
DM_DMR	0	0	0	0	0	109	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
DM_DMQ	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DM_DMI	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DM_DMD	0	0	179	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC_CCS	0	141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC_CCD	320	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Magahi: Confusion Matrix for Stanford Tagger

predicted \ actual	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
X																	
VERB	11	6	4	246	4	2	2	87	1	23	7	2	0	1	0	978	5
SYM																	
SCONJ	0	1	1	0	0	0	0	0	0	0	3	0	0	158	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	663	0	0	0	0
PROPN	0	0	0	0	0	0	0	6	0	0	0	3	0	0	0	0	6
PRON	1	0	1	0	3	6	0	1	0	5	994	0	0	16	0	1	0
PART	3	0	0	1	0	2	0	1	0	110	5	0	0	0	0	1	0
NUM	7	0	1	0	0	0	0	1	57	0	3	0	0	0	1	0	40
NOUN	188	46	100	216	5	12	3	1929	15	175	84	86	0	14	0	235	22
INTJ	0	0	0	0	0	0	6	1	0	0	0	0	0	0	0	0	0
DET	10	2	12	0	0	182	0	9	0	10	31	0	0	0	1	1	10
CCONJ	0	0	1	0	240	0	0	0	0	0	0	0	0	0	0	0	0
AUX	1	15	0	313	1	0	0	4	0	5	3	0	0	2	0	61	0
ADV	0	0	7	0	0	0	0	1	0	0	1	0	0	0	0	2	0
ADP	1	619	1	2	0	0	0	12	0	1	0	0	0	0	0	1	0
ADJ	23	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

rel.freq: 100, 75, 50, 25, 0

Magahi: Confusion Matrix for BiLSTM



predicted \ actual	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
X																	
VERB	21	6	5	131	2	2	0	128	0	20	23	8	0	3	0	1013	5
SYM	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
SCONJ	0	1	1	0	1	1	0	3	0	0	10	0	0	161	0	2	1
PUNCT	0	1	0	3	0	1	1	6	0	2	5	0	663	0	0	1	0
PROPN	8	0	10	0	0	0	0	61	0	0	0	16	0	0	0	1	3
PRON	2	0	3	0	3	10	0	6	0	11	983	0	0	7	0	3	0
PART	4	1	1	0	0	3	1	11	0	148	11	0	0	1	0	2	0
NUM	15	0	3	0	0	0	0	10	65	1	6	0	0	0	1	1	41
NOUN	155	33	79	7	9	7	7	1740	8	119	44	67	0	16	0	134	15
INTJ	0	0	0	3	0	1	0	1	0	2	0	0	0	0	0	0	0
DET	4	0	8	0	0	177	0	12	0	10	39	0	0	0	1	0	10
CCONJ	0	1	2	0	229	0	0	2	0	0	0	0	0	0	0	0	0
AUX	3	48	2	632	9	1	2	29	0	13	5	0	0	1	0	116	8
ADV	0	0	13	0	0	0	0	1	0	1	0	0	0	0	0	0	0
ADP	1	598	1	2	0	0	0	32	0	2	4	0	0	2	0	6	0
ADJ	32	0	0	0	0	0	0	11	0	0	1	0	0	0	0	1	0

Magahi: Confusion Matrix for SoMeWeTa

predicted \ actual	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
X	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
VERB	6	4	5	89	5	1	2	153	0	99	7	21	0	2	0	950	5
SYM																	
SCONJ	0	1	2	1	2	1	0	5	0	1	6	0	0	159	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	663	0	0	0	0
PROPN	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
PRON	2	0	3	3	5	8	0	6	0	10	1010	0	0	10	0	5	0
PART	9	0	1	1	0	2	0	5	0	107	6	0	0	0	0	2	0
NUM	22	0	0	1	0	0	0	10	69	1	5	2	0	1	1	2	41
NOUN	158	31	79	96	5	7	6	1802	4	79	47	66	0	9	0	175	15
INTJ	0	0	0	0	0	0	3	0	0	0	0	0	0	1	0	0	0
DET	8	1	14	0	0	180	0	15	0	12	40	1	0	0	1	2	10
CCONJ	0	0	2	0	235	1	0	3	0	0	1	1	0	3	0	0	0
AUX	1	13	1	581	1	1	0	14	0	13	4	0	0	2	0	136	12
ADV	0	0	19	0	0	0	0	4	0	1	3	0	0	0	0	0	0
ADP	1	639	1	6	0	0	0	28	0	4	2	0	0	4	0	1	0
ADJ	38	0	1	0	0	3	0	7	0	1	0	0	0	0	0	6	0

rel.freq: 100, 75, 50, 25, 0

Error Analysis: Summary

- Bhojpuri ($F_1 \in [89, 95]$)
  - errors very much what one would expect
  - rare categories show low recall
  - frequent tags (N_NN, V_VM) go-to labels for misclassifications
  - confusion of similar morphosyntactic categories (V_VM $\leftrightarrow$ V_AUX, N_NN $\leftrightarrow$ N_NNP)
- Magahi ($F_1 \in [73, 79]$)
  - major problems:
    - recall for PROPN, X
    - ADJ (15.5%), ADV (14.8%), PART (32.5%)
  - go-to labels NOUN and VERB
  - obvious confusion: VERB $\leftrightarrow$ AUX
  - different tag distributions in test and training data
    - 602 ADJ and 16777 NOUN in the training set
    - 245 ADJ and 2053 NOUN in the test set

Conclusion

- results for Bhojpuri very satisfying
  - close to 95% accuracy on a tagset (33 tags, 100,000 tokens training data)
  - a bit of a downer: mindless parameter-tuning yields best results
  - differences in system performance probably not significant
- results for Magahi very disappointing
  - problems with a-priori tag distribution
  - here: use of additional resources outperforms mere parameter-tuning

Thanks for listening.

Questions?

References

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

Thomas Proisl. SoMeWeTa: A part-of-speech tagger for German social media and web texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki, 2018. European Language Resources Association.

Martin Riedl and Sebastian Padó. A named entity recognition shootout for German. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Volume 2: Short Papers*, pages 120–125, Melbourne, 2018.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252–259, Edmonton, 2003.