

Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources

Siva Reddy^{1,2}, Serge Sharoff³

¹Lexical Computing Ltd, UK ²University of York, UK

³University of Leeds, UK

Presenter: **Aswarth Abhilash**, IIIT Hyderabad, India

CLIA 2011 @ IJCNLP 2011
Chiang Mai, Thailand
November 13, 2011

Indian Languages

Many Indian languages exhibit similarities in morphology and syntactic behaviour

Indo-Aryan Languages

Hindi, Urdu, Marathi, Nepali, Punjabi, Gujarathi, Rajasthani, Bengali, Oriya, Bihari

Dravidian languages

Telugu, Tamil, Kannada, Malayalam

Some pairs exhibit high similarity

Hindi-Urdu,
Hindi-Marathi,
Tamil-Malayalam,
Telugu-Kannada

...

Kannada and Telugu

Some facts about Kannada and Telugu

- Dravidian family
- Spoken by 35 and 75 million people respectively.
- Telugu was highly influenced by Kannada making them slightly mutually intelligible (Datta, 1998)
- Scripts belong to the same family.
 - Until 13th century both the languages have same script.
- Telugu is relatively resource-richer than Kannada
 - Corpus, Morphological Analyzer, POS Tagger, Dependency Parser

Kannada and Telugu: Similarities at Word level and in Structure

Kannada								
Vivādagāḷa	hinneleyalli	tam'ma	taṇḍada	punarracisalu	aṇṇā	hajāreyavaru	yōjisiddāre	
Telugu								
Vivādāḷa	nēpadhyanlō	tana	jaṭṭunu	punarvyavasthikarin̄cālani	annā	hajārē	yōcin̄cāru	
English Gloss								
Controversies	in view of	his	team	restructuring	Anna	Hajare	planing	

Anna Hazare is planning to restructure his team in view of controversies.

Cross language Tools

Building tools for a Target language using Cross language resources

Kannada Tools from Telugu Resources

Motivation

- Not many resources for Kannada
- Existing resources not as efficient as for other languages
- Telugu relatively resource-richer than Kannada
- Kannada and Telugu are typologically related and exhibit similarities

Our focus is to build POS taggers and Morphological disambiguators/analyzers.

Our Tagset

- Bharati et al. (2006) designed a common POS tagset for all Indian languages
 - e.g. CC, JJ, NN, VM
- We added morphological information to the above tagset
 - e.g. NN.n.f.pl.3.d
 - Main POS Tag, Coarse label, Gender, Number, Person, Case
- Since POS tag contains morphological information, our tagger can also be used as morphological analyzer.

Tagset Statistics

Field	Description	Number of Tags	Tags
	Full Tag	311	NN.n.f.pl.3.d, VM.v.n.sg.3., ...
1	Main POS Tag	25	CC, JJ, NN, VM, ...
2	Coarse POS Category	9	adj, n, num, unk ...
3	Gender	6	any, f, m, n, punc, null
4	Number	4	any, pl, sg, null
5	Person	5	1, 2, 3, any, null
6	Case	3	d, o, null

Table: Fields in each tag and its corresponding statistics. *null* denotes empty value, e.g. in the tag *VM.v.n..3.*, *number* and *case* fields are *null*

HMM Based POS Tagger

$$\operatorname{argmax}_{t_1 \dots t_n} \left[\prod_{i=1}^n P(t_i | t_{i-1}, t_{i-2}) P(w_i | t_i) \right] \quad (1)$$

- $w_1 \dots w_n$ is the word sequence to be tagged
- $t_1 \dots t_n$ denotes the tag sequence
- $P(t_i | t_{i-1}, t_{i-2})$: Tag Transition Probabilities
- $P(w_i | t_i)$ denotes Emission Probabilities

Kannada POS Tagger from Telugu

- Estimate Transition Probabilities and Emission Probabilities from Telugu
- Exploit lexical and syntactic level similarities between Kannada and Telugu

Kannada								
Vivādagaḷa	hinneleyalli	tam'ma	taṇḍada	punarracisalu	aṇṇā	hajāreyavaru	yōjisiddāre	
Telugu								
Vivādāla	nēpadhyanlō	tana	jaṭṭunu	punarvyavasthikarin̄cālani	annā	hajārē	yōcin̄cāru	
English Gloss								
Controversies	in view of	his	team	restructuring	Anna	Hajare	planing	

Anna Hazare is planning to restructure his team in view of controversies.

Steps Involved

- 1 Built large corpora of Kannada and Telugu
 - POS tag the corpus with existing tools
- 2 Determine the transition probabilities of Kannada
 - from Telugu tagged corpus (cross lingual)
 - or from Kannada tagged corpus (mono-lingual)
- 3 Estimate the emission probabilities of Kannada
 - from Telugu tagged corpus
 - or using heuristics combined morphological analyser
 - or from Kannada tagged corpus
- 4 Use the probabilities from the step 2 and 3 to build a POS tagger for Kannada

Step 1: Large Corpus Creation

Corpus Factory (Kilgarriff et al., 2010)

- Build frequency list of Telugu and Kannada from Wikipedia
 - Generate thousands of random tuples from frequency lists
 - Feed them into a search engine and download search hits
 - Clean the pages - remove html markup, language filtering
 - Remove duplicates
-
- Telugu - 4.6 million words corpora [Collected in Dec 2009]
 - Kannada - 16 million words corpora [Collected in June 2011]
 - Differences in sizes due to time difference

Step 1: Large Corpus Creation

- Annotate Telugu Corpus with existing tagger
- We used ILMT consortium tagger built using (Avinesh and Karthik, 2007)
- Tagger is an integration of many tools running in pipeline
 - tokenization, transliteration, morph analyzer, CRF model, transliteration
 - If anything fails in the pipeline, tagger fails
 - Only 70% of the corpus was finally tagged
 - Not scalable, but usable
 - We converted the output to our tagset
- We also tagged Kannada corpus using ILMT Kannada tagger
 - To build mono-lingual taggers and compare performance with cross lingual tagger

Step 2: Transition Probabilities

From Telugu: cross lingual

- Transition Probabilities across typologically related languages are likely to be same (Hana et al., 2004)
- Kannada and Telugu exhibit high similarities in syntactic structure
- Compute transition probabilities from Telugu tagged corpus of Step 1

Kannada						
Vivādagāḷa	hinneleyalli	tam'ma	taṇḍada	punarracisalu	annā	hajāreyavaru
Telugu						
Vivādāla	nēpadhyanlō	tana	jaṭṭunu	punarvyavasthikarin'cālani	annā	hajārē
Tag sequence in Kannada and Telugu is same						
NN.n.n.pl..o	NN.n.n.sg..o	PRP.n.m.sg.3.d	NN.n.n.sg..o	VM.v.any.any.any.	NNP.n.m.sg.3.o	NNP.n.m.sg.3.o

Step 3: Emission Probabilities

From Telugu using Approx. String Matching

- A Telugu-Kannada dictionary can be used but we do not have such dictionary
- Exploit lexical similarities
 - Kannada and Telugu are slightly mutually intelligible (Datta, 1998)
- Edit distance: The minimum number of edits needed to transform one string into the other
- For each word in Kannada, choose the most nearest Telugu word.
- Since Telugu emission probabilities can be known from Telugu tagged corpus, use mappings of Kannada-Telugu words to estimate Kannada emission probabilities

Kannada	Neighbours in Telugu	Result
viBAgavu	(viBAgamu, 0.539) (viBAga, 0.5) (viBAgalanu, 0.467), (viBAgamulu, 0.467)	✓
xAswAnu	('xAswAn', 0.545) ('xAswAru', 0.5) ('rAswAnu', 0.5) ('xAswAdu', 0.5)	✗

Step 3: Emission Probabilities

Telugu tags and Kannada Morphology

- Using Telugu tagged corpus, the mappings of a morphological set to all possible tags are learned
 - Morphological set **n.n.sg..o** is associated with all the tags which satisfy the regular expression ***.n.n.sg..o**
- For every word in Kannada, based on its morphology determined by the morphological analyser, we assign all the tags learned from Telugu.
- Uniform tag distribution is assumed

Pitfall

Explosion of search tags for each word

Step 3: Emission Probabilities

Kannada tags with uniform distribution

- For each word, learn all the possible tag associations from Kannada tagged corpus
- Though we learn from tagged corpus, we do not use frequency information
- We assume uniform distribution of all tags for a word
- Search space is reduced

From Kannada corpus

- Learn emission probabilities directly from the tagged Kannada corpus

Step 4: Tagging Model

- We use TnT (Brants, 2000), an implementation of HMM
- Transition and Emission probabilities from Steps 2 and 3
- Avinesh and Karthik (2007) performance increased when morphological information is used as features for their CRF model.
- Since our tagset includes morphological information, TnT model may benefit from this information.
- TnT Model is known for predicting tags for unseen words
 - a potential morphological analyser for new words

Additional Tools

- For each word form, we learned association between POS Tag, lemma and suffix markers from Kannada annotated corpora [Step 1]
- Our tagger could also be used for lemmatization and suffix prediction

Sample Output

Word	POS Tag	Lemma.Suffix
ಕತೆಯ	NN.n.n.sg..o	ಕತೆ.ಅ
ಪ್ರಕಾರ	NN.n.n.sg..d	ಪ್ರಕಾರ.೦
ಗೆಳೆಯರೊಂದಿಗಿನ	NN.unk....	ಗೆಳೆಯರೊಂದಿಗಿನ.
ಆಟದಲ್ಲಿ	NN.n.n.sg..o	ಆಟ.ಅಲ್ಲಿ
ರಾಜನಾಗಿದ್ದ	VM.unk....	ರಾಜನಾಗಿದ್ದ.
ಚಂದ್ರಗುಪ್ತನು	NNP.unk....	ಚಂದ್ರಗುಪ್ತನು.
ಅಪರಾಧಿಯ	NN.n.m.sg.3.o	ಅಪರಾಧಿ.ಅ
ಪಾತ್ರ	NN.n.n.sg..d	ಪಾತ್ರ.೦
ವಹಿಸಿದ್ದ	VM.v.any.any.any.	ವಹಿಸು.ಇದ್ದ
ಇನ್ನೊಬ್ಬ	QC.unk....	ಇನ್ನೊಬ್ಬ.
ಹುಡುಗನ	NN.n.m.sg.3.o	ಹುಡುಗ.ಅ
ವಿಚಾರಣೆಯನ್ನು	NN.n.n.sg..o	ವಿಚಾರಣೆ.ಅನ್ನು
ಮಾಡಿ	VM.v..pl.2.	ಮಾಡು.೦
ಶಿಕ್ಷೆ	NN.n.n.sg..d	ಶಿಕ್ಷೆ.೦

Evaluation

- Evaluation only for main POS tag.
 - In NN.n.n.sg..o, main POS tag is NN
- Manually annotated Kannada corpora
 - developed by ILMT consortium (licensed)
- The corpus consists of 201,373 words
- No evaluation data for morphological labels

Results

Model	Transition Prob	Emission Prob	Precision	Recall	F-measure
Cross-Language POS Tagger					
1	From Telugu	Approximate string matching	56.88	56.88	56.88
2	From Telugu	Telugu tags and Kannada morphology	28.65	28.65	28.65
3	From Telugu	Kannada tags with uniform distribution	75.10	75.10	75.10
4	From Telugu	Kannada emission probabilities	77.63	77.63	77.63
Mono-Lingual POS Tagger					
5	From the Kannada language	Kannada emission probabilities	77.66	77.66	77.66
6		Avinesh and Karthik (2007)	78.64	61.48	69.01

Table: Evaluation results of various tagging models [only the main Tag]

- Cross language taggers as good as mono-lingual taggers
 - Model 3 and 4 better than existing Kannada Tagger
- Model 3 easy to built since it requires only a Kannada lexicon
- 50% accuracy (Model 1) with almost no resources of Kannada

Summary

- Cross-language resources can be used to build tools for other languages
 - As good as mono-lingual tagger if target lexicon exists
 - at least 50% accuracy if no resources of target language exists
- Promising direction for many resource-poor Indian languages
- POS tagger as a morphological analyser/disambiguator

Tools can be downloaded from <http://sivareddy.in>

Bibliography I

- Avinesh, P. V. S. and Karthik, G. (2007). Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation-Based Learning. In *Proceedings of the IJCAI and the Workshop On Shallow Parsing for South Asian Languages (SPSAL)*, pages 21–24.
- Bharati, A., Sangal, R., Sharma, D. M., and Bai, L. (2006). Anncorra: Annotating corpora guidelines for pos and chunk annotation for indian languages. In *Technical Report (TR-LTRC-31)*, LTRC, IIIT-Hyderabad.
- Brants, T. (2000). Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing, ANLC '00*, pages 224–231, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Datta, A. (1998). *The Encyclopaedia Of Indian Literature*, volume 2.
- Hana, J., Feldman, A., and Brew, C. (2004). A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

Bibliography II

Kilgarriff, A., Reddy, S., Pomikálek, J., and PVS, A. (2010). A corpus factory for many languages. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).