Cognate Discovery
For Bootstrapping Lexical Resources

Shantanu Kumar
(2013EE10798)

Supervisors

Prof. Sumeet Agarwal
Dr. Ashwini Vaidya

Motivation

Cognates : Cross-language words which originate from a common ancestral language.

Night (English)	Nacht (German) *
Father (English)	Pater (Greek)
Star (English)	Tara (Hindi)

- ➔ Essential for historical linguists.
- ➔ Successfully applied to NLP tasks like as Sentence Alignment [Simard et al., 1993][Navlea et al., 2011] and Statistical Machine Translation [Kondrak et al., 2003].
- ➔ Assist in lexical resource creation.

*Sourced from [Hauer et al., 2011]

Datasets

1. Indo-European Dataset (IELex)
   - 52 Languages, 208 Meanings
   - Romanized and IPA transcription
2. Austronesian Dataset
   - 100 Languages, 210 Meanings
   - ASJP transcription
3. Mayan Dataset
   - 30 Languages, 100 Meanings
   - ASJP transcription

- Parallel Sentence-aligned Corpora
  - Hindi-Marathi (TDIL)

Part of Wordlist Used

Concepts

Languages

	ALL	BIG	ANIMAL
English	All	Big	Animal
French	Tut	Grand	Animal
Marathi	Serve	Motha	Jenaver
Hindi	Seb	Bara	Janver

[From Dataset by Dyen et al.]

Previous Work

- SURFACE SIMILARITY
  - B. Hauer and G. Kondrak. "**Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists.**" - IJCNLP 2011
    - Orthographic word similarity features
  - T. Rama. "**Automatic cognate identification with gap-weighted string subsequences.**" - HLT-NAACL 2015
    - Gap-weighted common string subsequences
- DEEP REPRESENTATION
  - T. Rama. "**Siamese convolutional networks based on phonetic features for cognate identification.**" - COLING 2016
    - Representing words as 2D matrices

Common Subsequence Model

$$\phi_u(s) = \Sigma_{\forall I, s[I]=u} \lambda^{l(I)}$$

$$l(I) = i_{|u|} - i_1 + 1$$

$$\Phi(s) = \{\phi_u(s); \forall u \in \cup_{n=1}^{p} \Sigma^n\}$$

Multiplicative Model

$$\Phi_{Mul}(s_1, s_2) = \{\phi_u(s_1)+\phi_u(s_2); \forall u \text{ present in } s_1 \text{ and } s_2\}$$

Additive Model

$$\Phi_{Add}(s_1, s_2) = \{\phi_u(s_1) + \phi_u(s_2); \forall u \text{ present in } s_1 \text{ or } s_2\}$$

Hybrid Model

$$\Phi_{Avg}(s_1, s_2) = (1 - \alpha) \cdot \Phi_{Mul}(s_1, s_2) + \alpha \cdot \Phi_{Add}(s_1, s_2)$$

Subsequence Vector Example

PATER

'ae':	0.14925373,
'ar':	0.10447761,
'at':	0.21321962,
'er':	0.21321962,
'pa':	0.21321962,
'pe':	0.10447761,
'pr':	0.07313433,
'pt':	0.14925373,
'te':	0.21321962,
'tr':	0.14925373

FATHER

'ae':	0.07763300,
'ah':	0.11090429,
'ar':	0.05434310,
'at':	0.15843470,
'er':	0.15843470,
'fa':	0.15843470,
'fe':	0.05434310,
'fh':	0.07763300,
'fr':	0.03804017,
'ft':	0.11090429,
'he':	0.15843470,
'hr':	0.11090429,
'te':	0.11090429,
'th':	0.15843470,
'tr':	0.07763300

Siamese ConvNet Model

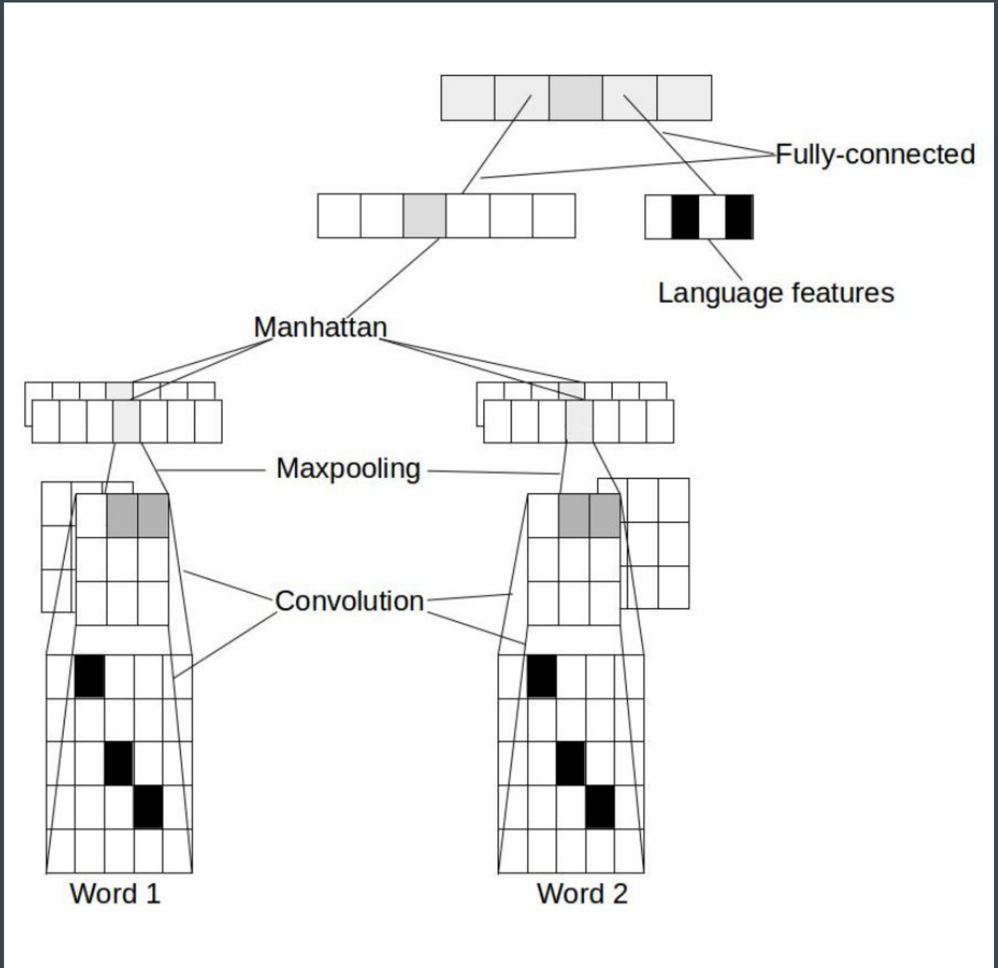
- Inspired from networks used for detecting similarity in images

- Manually defined character embeddings based on phonological properties

- No hand engineered features on word level

Drawbacks

- Character vectors defined by phonetic classes, convolution does not seem intuitive
- Both words encoded independently



*Sourced from [T.Rama, 2016]

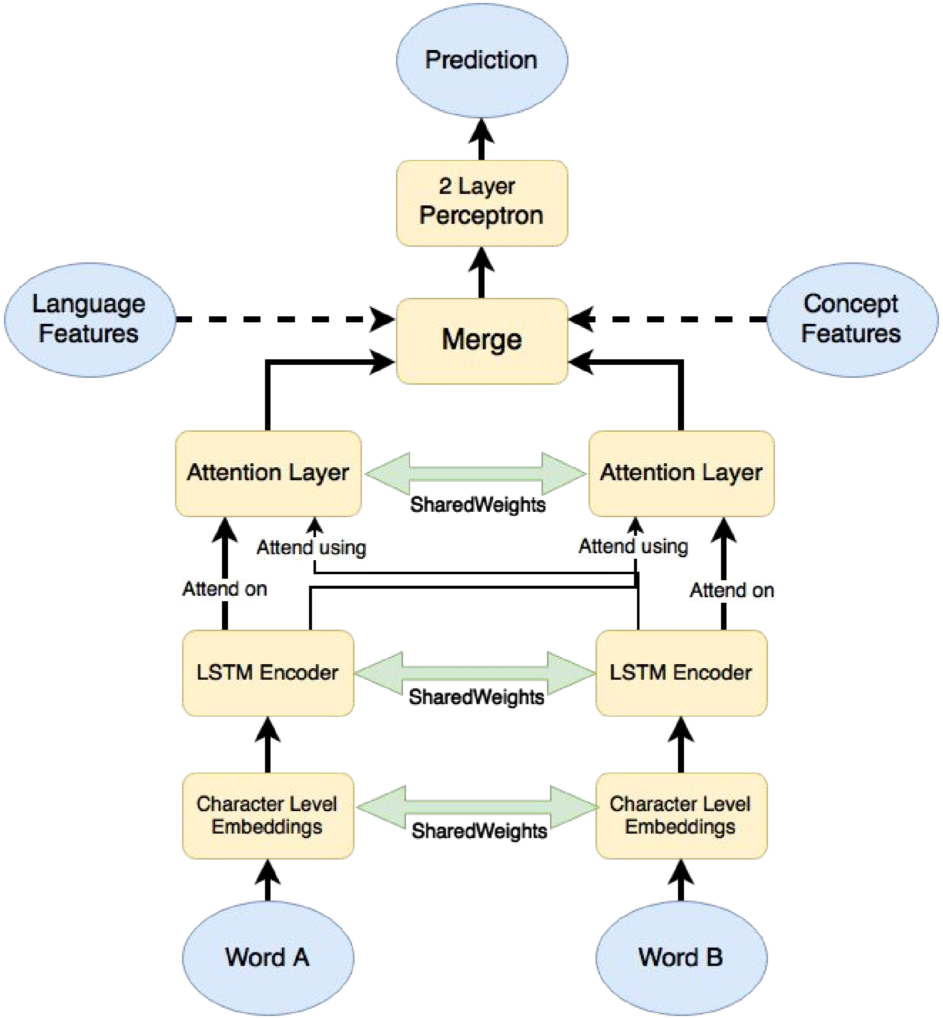
Character Embeddings Used

Features	p	b	f	v	m	8	4	t	d	s	z	c	n	S	Z	C	j	T	5	k	g	x	N	q	G	X	7	h	l	L	w	y	r	!	V
Voiced	0	1	0	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1
Labial	1	1	1	1	1	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	0	0	0	0
Dental	0	0	1	1	0	1	1	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
Alveolar	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Palatal/Post-alveolar	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Velar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
Uvular	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
Glottal	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	1	0	0	0	0	0	0	0
Stop	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0
Fricative	1	1	1	1	0	1	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0
Affricate	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nasal	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Click	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	0	0	1	0
Approximant	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	0	1	1	0	0	0
Lateral	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	1	0	0	0	0	0
Rhotic	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	0	1	0	0

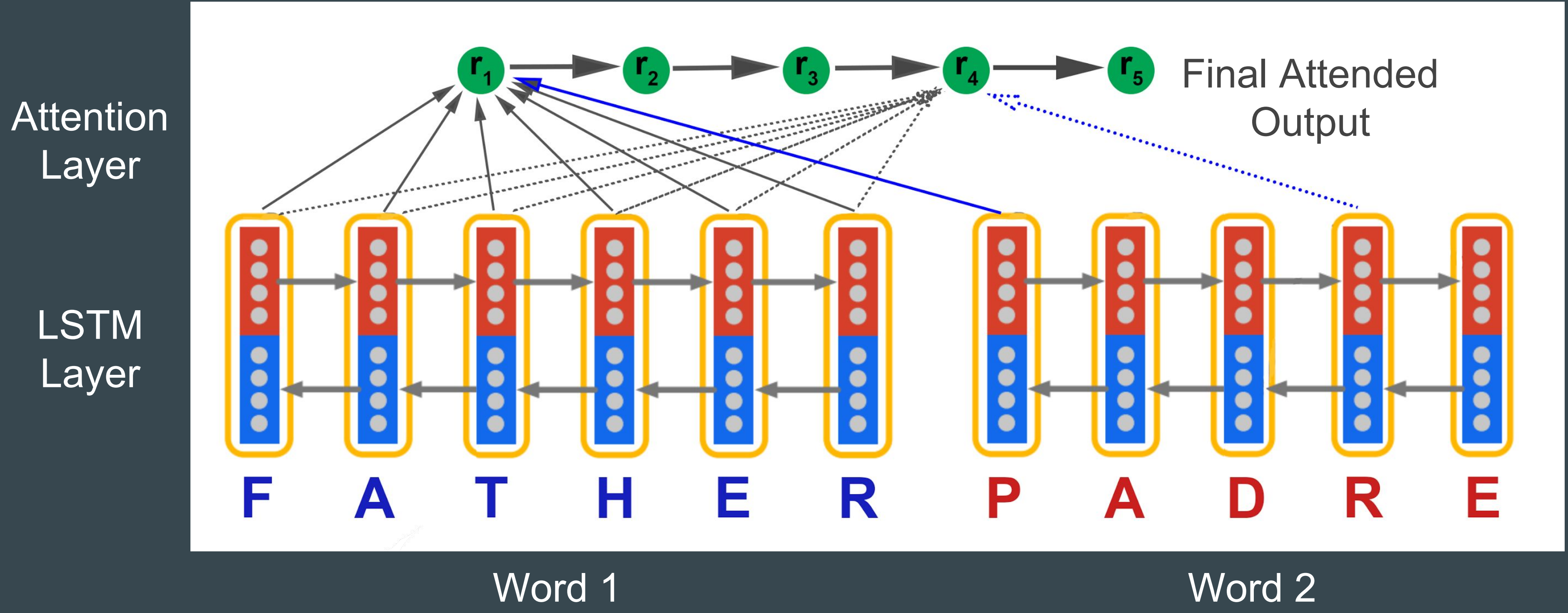
*Sourced from [T.Rama, 2016]

Co-Attention LSTM Model

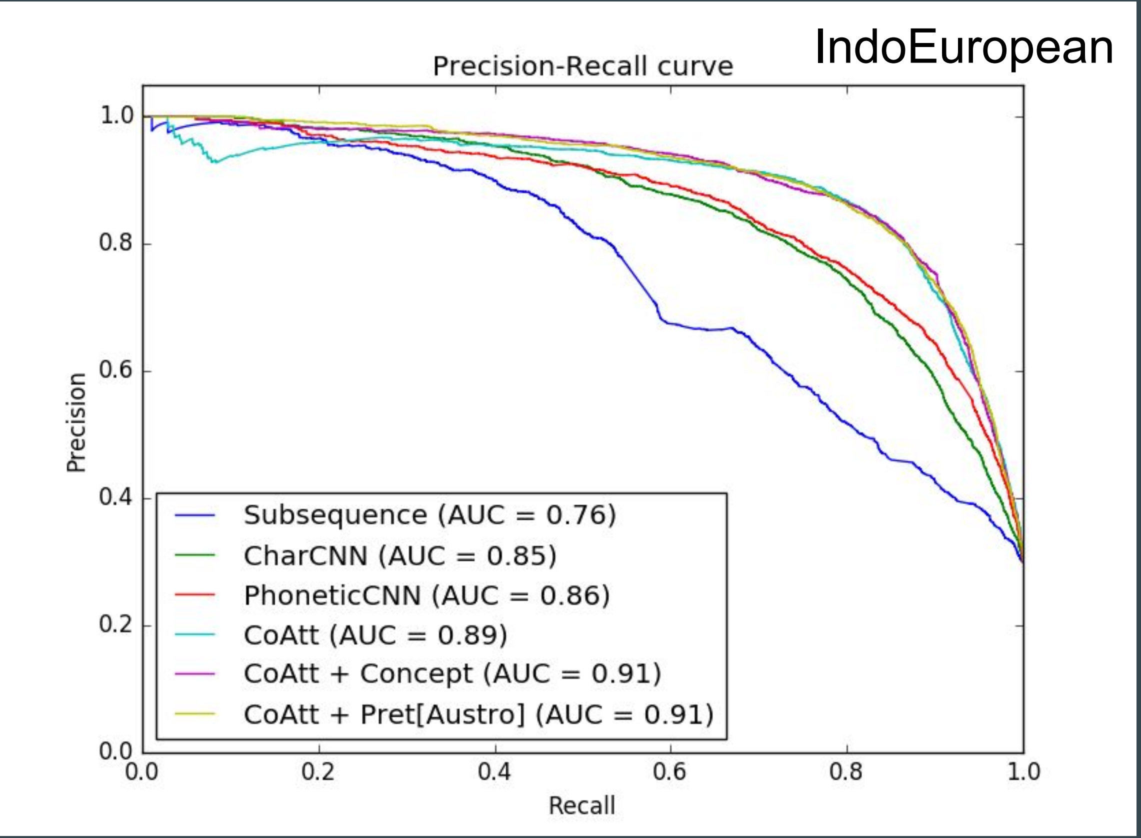
- Symmetric model with co-attention
- Shared LSTM encoder to encode each word
- Character-level encodings learnt
- Additional *Concept Features* added using Glove embeddings of the concept
- Network pre-trained across different language families



Recurrent Attention Layer (T. Rocktäschel et al. 2016)

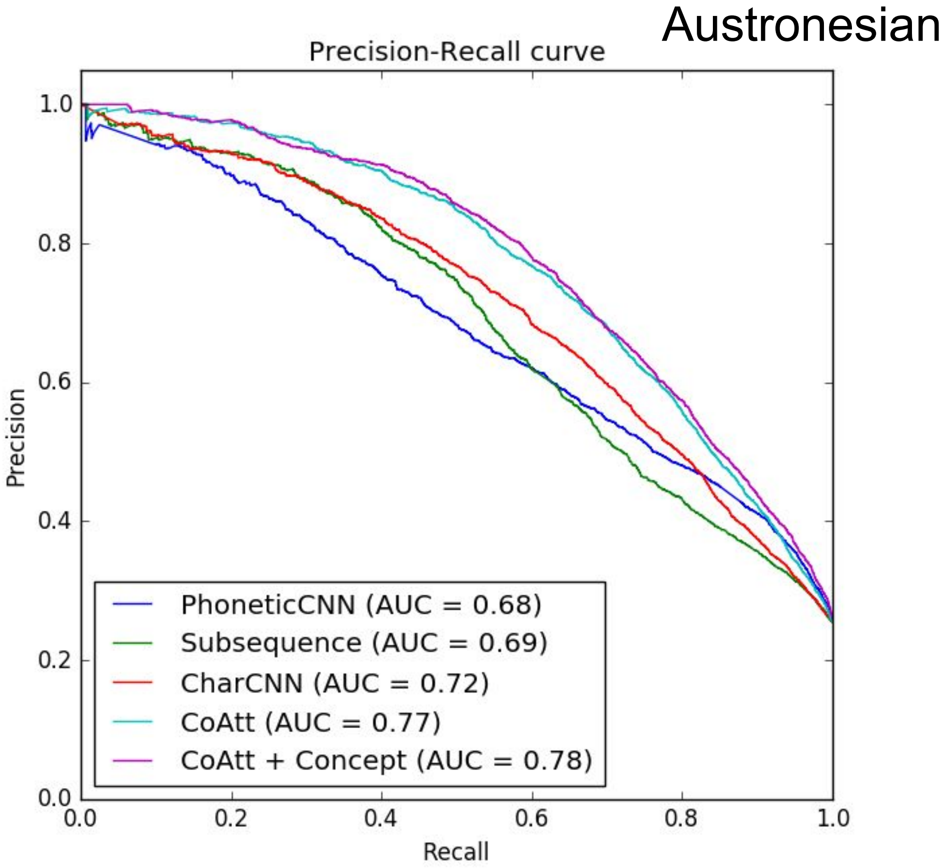


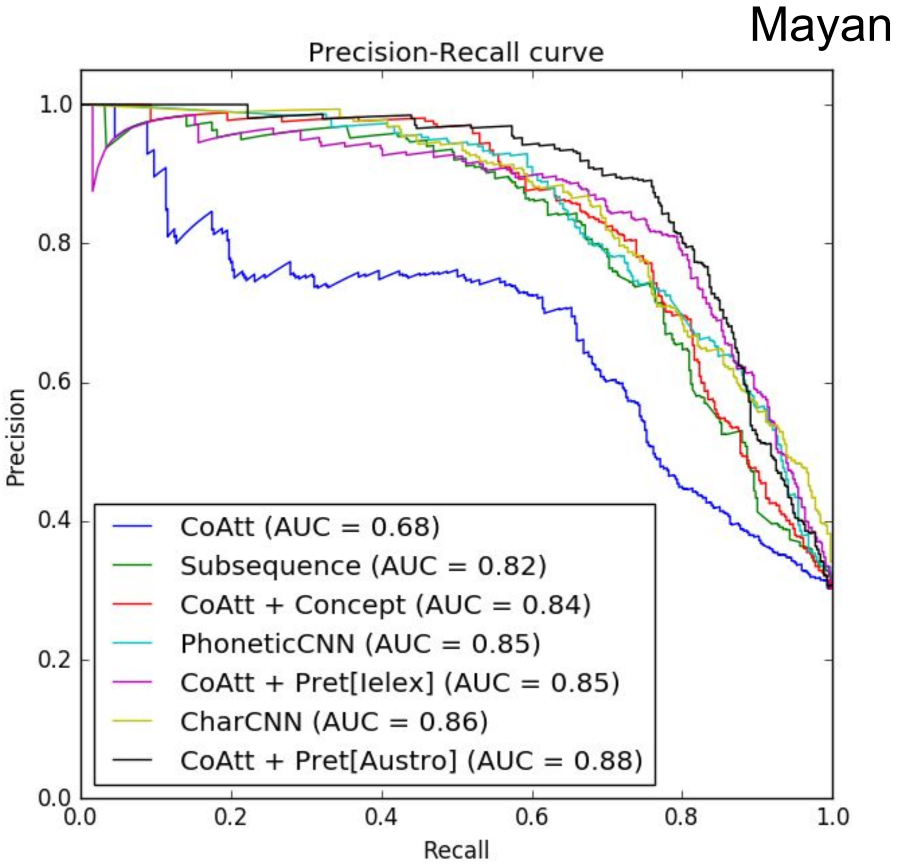
Results - Cross Language Evaluation



- LSTM model gives significantly improved results than ConvNet
- Adding Concept Features Improves the AUC curve
- Pretraining the model on Austronesian dataset before training on Indo-European also improves the performance

Results - Cross Language Evaluation





Results - Cross Language Evaluation

	Indo-European		Austronesian		Mayan	
	Total	Positive	Total	Positive	Total	Positive
Training Samples	218,429	56,678	333,626	96,356	25,473	9,614
Testing Samples	9,894	2,188	20,799	5,296	1,458	441

Table 5.1: Data size for Cross Language Evaluation

Model	**Indo-European**		**Austronesian**		**Mayan**	
	F-Score	*AUC*	*F-Score*	*AUC*	*F-Score*	*AUC*
Gap-weighted Subsequence	59.0	75.5	58.8	68.9	71.8	81.8
PhoneticCNN	73.7	86.1	54.6	68.0	72.8	85.0
PhoneticCNN + Language Features	62.2	85.4	46.8	67.0	66.4	84.0
CharacterCNN	75.3	85.3	62.2	71.6	75.9	85.7
CharacterCNN + Language Features	70.7	82.6	61.4	70.1	61.1	82.2
CoAtt	83.8	89.2	69.0	77.5	67.1	67.7
CoAtt + Concept Features	**83.5**	90.5	**68.9**	**77.9**	76.2	84.2
CoAtt + Pre-training (Austro)	83.2	**90.6**	-	-	**80.4**	**88.3**
CoAtt + Pre-training (IELex)	-	-	-	-	79.6	85.2

Table 5.2: Cross Language Evaluation Results

Concept Wise Analysis - From Indo-European Dataset

Concept	# Cognate Classes	CoAtt	CoAtt + Concept	CoAtt + Pret(Austro)	Phonetic CNN	Char CNN	Sub Sequence
WHO	2	**0.914**	0.900	0.872	0.429	0.646	0.044
WHAT	2	0.955	0.905	**0.966**	0.414	0.686	0.196
HOW	3	**0.902**	0.875	0.800	0.269	0.615	0.163
WHERE	4	**0.857**	0.780	0.820	0.435	0.531	0.054
THERE	8	0.900	0.842	0.837	0.778	**0.976**	0.174
EAT	10	0.686	0.722	0.800	0.778	**0.872**	0.429
IN	14	0.353	**0.565**	0.383	0.300	0.353	0.000
AT	18	0.250	0.143	**0.500**	0.421	0.258	0.000
IF	20	0.316	0.316	**0.462**	0.333	0.200	0.000
BECAUSE	37	**0.286**	0.000	0.000	0.000	0.000	0.000

Transcription Tests

- ➔ Indo-European dataset is available in IPA and ASJP formats
- ➔ ASJP is a coarser transcription with a smaller character vocabulary
- ➔ IPA Model predicts false negatives
  - Very fine transcription, correspondence not learnt
- ➔ Adding *Concept Features* corrected the mistake
  - Different threshold for different concepts

ASJP		IPA	
Word 1	**Word 2**	**Word 1**	**Word 2**
swim	sinda	swɪm	ˈsɪnd̥a
swim	zwem3n	swɪm	zwɛmən
swim	svim3n	swɪm	ʃvɪmən
swem3	sinda	ˈsʊø̀m:ə	ˈsɪnd̥a
swem3	zwem3n	ˈsʊø̀m:ə	zwɛmən
swem3	svim3n	ˈsʊø̀m:ə	ʃvɪmən
sinda	zwem3n	ˈsɪnd̥a	zwɛmən
sinda	svim3n	ˈsɪnd̥a	ʃvɪmən
zwem3n	sima	zwɛmən	ˈsim:a
sima	svim3n	ˈsim:a	ʃvɪmən

Cognates pairs from the concept **SWIM** for various languages

Model	**Indo-European (ASJP)**		**Indo-European (IPA)**	
	F-Score	*AUC*	*F-Score*	*AUC*
CoAtt	83.8	89.2	82.2	89.1
CoAtt + Concept Features	83.5	90.5	82.1	90.7

Cross Concept Evaluation

- Cross concept seems to be a tougher task
- Words from different concepts have different sequence structures
- Different concepts also have varied degrees of cognacy due to different frequency of use

Model	Indo-European		Austronesian		Mayan	
	F-Score	*AUC*	*F-Score*	*AUC*	*F-Score*	*AUC*
Gap-weighted Subsequence	51.6	62.0	53.1	64.5	61.0	75.4
PhoneticCNN + Language Features	**66.4**	**73.2**	57.8	66.6	**80.6**	88.1
CharacterCNN + Language Features	63.5	70.5	**60.9**	**70.2**	79.6	**89.1**
CoAtt	64.8	69.8	57.1	61.0	70.5	74.8
CoAtt + Language Features	65.6	70.8	57.3	62.0	69.6	71.9
CoAtt+ Concept Features	64.1	70.6	58.0	63.1	71.9	78.6
CoAtt + Pre-training (Austro)	65.8	71.0	-	-	71.1	78.4
CoAtt + Pre-training (IELex)	-	-	-	-	71.2	79.0

Table 5.5: Cross Concept Evaluation Results

Hindi-Marathi Tests

HINDI	MARATHI	SCORE
चोटयिाँ	समोर	0.0086563
रूम	खोलीतील	0.0086572
छंतोली	पालखीवर	0.0086578
यात्रा	लोक	0.008658
मुमकनि	लोकसंख्या	0.008658
कुदरत	नसिर्ग	0.008662
मलि	क्षेत्रास	0.0086626
गहराइयों	तलाव	0.0086626
नगिम	महामंडळाच्या	0.0086626
नाम	पॅलेस	0.0086633
इमारतें	नवाबांनी	0.0086637
प्रकृति	नसिर्ग	0.0086646
कलि	मशदि	0.0086649
मूर्ति	रुपे	0.0086649
तड़के	२७ला	0.008665

Noun Pairs Negatives

HINDI	MARATHI	SCORE
फूलों	फुलांनी	0.985515
फूलों	फुलांचे	0.985508
फूलों	फुलांच्या	0.985507
फ्लाइंग	फ्लाईंग	0.985507
लाइन	लाइनची	0.985505
कस्बा	कसबा	0.985502
बाइक	बाईक	0.985439
फूलों	फ़ुलांसाठी	0.985436
शांति	``शांती	0.985435
शांति	शांती	0.985435
भागों	भागांवर	0.985435
बसें	बसेंस	0.985429
रोजाना	रोज	0.985425
रेखाओं	रेषांनी	0.985424
पर्वत	पर्वत	0.985225

Noun Pairs Positives

HINDI	MARATHI	SCORE
बाँध	बांधू	0.985483
बनवा	बनवून	0.985482
बाँध	बांधून	0.985473
बँधे	बांधून	0.98547
बनाने	बांधण्यास	0.985444
बना	बनवा	0.985438
बनाना	बनवणे	0.985435
भर	भरून	0.985432
बना	बनवत	0.985429
स्थिति	साकार	0.985426
बनने	बनण्यास	0.985423
बनाने	बनवू	0.985421
बना	बनवतो	0.985417
भिगोकर	भिजवून	0.985415
बनवाने	बनवण्याच्या	0.985414

Verb Pairs Positives

Conclusion & Future Direction

- Surface similarity measure fail infront of Deep structure representation in capturing phonological evolution to predict cognates
  - LSTM and CNN models capture complementary information. Hybrid model to exploit best of both models.

- Additional *Concept Features* and *Cross-Family Pretraining* helps to improve performance, especially on the smaller Mayan dataset.

- Recurrent model lags behind the CNN model for Mayan dataset on Cross-Concept evaluation
  - Analysis of advantages of the CNN model in the task

- Applied the model to domain of Hindi-Marathi word pairs
  - Sample and evaluate the performance
  - Apply to other language pairs

References

1. Simard, Michel, George F. Foster, and Pierre Isabelle. "**Using cognates to align sentences in bilingual corpora.**" *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2.* IBM Press, 1993.

2. Kondrak, Grzegorz, Daniel Marcu, and Kevin Knight. "**Cognates can improve statistical translation models.**" *Proceedings of the 2003 Conference of NAACL on Human Language Technology* 2003.

3. Rama, Taraka. "**Automatic cognate identification with gap-weighted string subsequences.**" *Proceedings of the 2015 Conference of NAACL on Human Language Technologies* 2015.

4. Rama, Taraka. "**Siamese convolutional networks based on phonetic features for cognate identification.**" *COLING* 2016.

5. Rocktäschel, Grefenstette, Hermann, Kočiský and Blunsom. "**Reasoning about Entailment with Neural Attention**" *International Conference on Learning Representations (ICLR)* 2016.