# 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

## Indo-European Dataset (Dyen et al., 1992)

- 84 Languages, 200 Meanings
- Romanized transcription (34 characters)

## Indo-European Lexical Cognacy Database (IELex)

- 52 Languages, 207 Meanings
- IPA transcription (536 characters)

## Parallel Corpora

- Hindi-Marathi (TDIL)

### Part of Wordlist Used

**Concepts**

**Languages**

| | ALL | AND | ANIMAL |
|---|---|---|---|
| English | All | And | Animal |
| French | Tut | Et | Animal |
| Marathi | Serve | Ani | Jenaver |
| Hindi | Sara | Or | Janver |

[From Dataset by Dyen et al.]

# Previous Work

1. B. Hauer and G. Kondrak. "**Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists.**" - IJCNLP 2011
    - Orthographic word similarity features

2. T. Rama. "**Automatic cognate identification with gap-weighted string subsequences.**" - HLT-NAACL 2015
    - Gap-weighted common string subsequences

3. 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 |

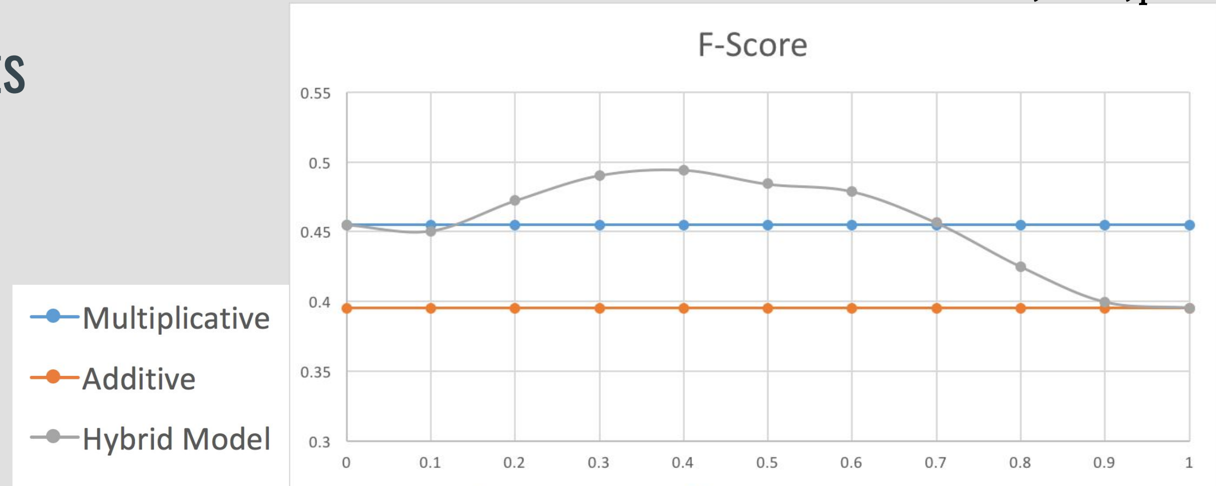
**FATHER**

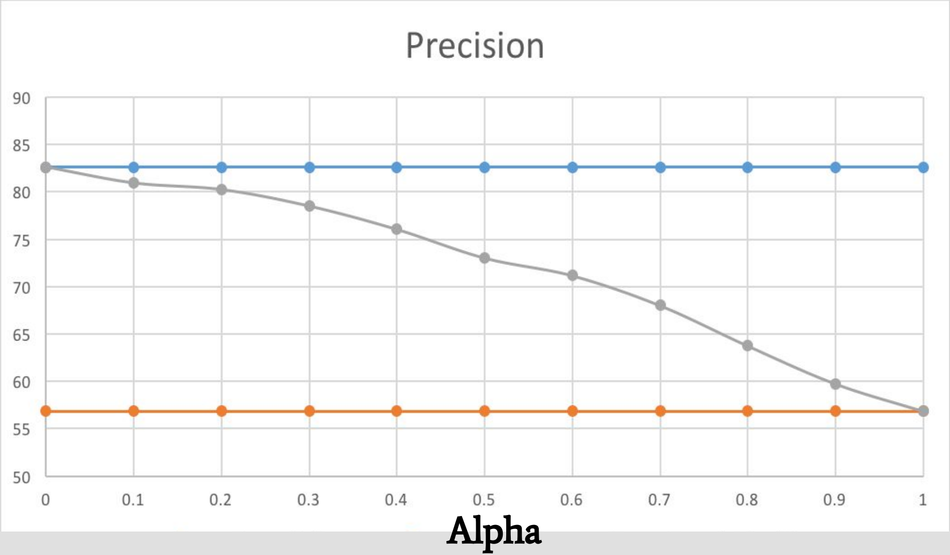
| | |
|---|---|
| 'ae': | 0.07763300, |
| 'ah': | 0.11090429, |
| 'ar': | 0.05434310, |
| 'at': | 0.15843470, |
| 'er': | 0.15843470, |
| 'fa': | 0.15843470, |
| 'fe': | 0.05434310, |
| 'fh': | 0.07763300, |
| 'fr': | 0.03804017, |
| 'ft': | 0.11090429, |
| 'he': | 0.15843470, |
| 'hr': | 0.11090429,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

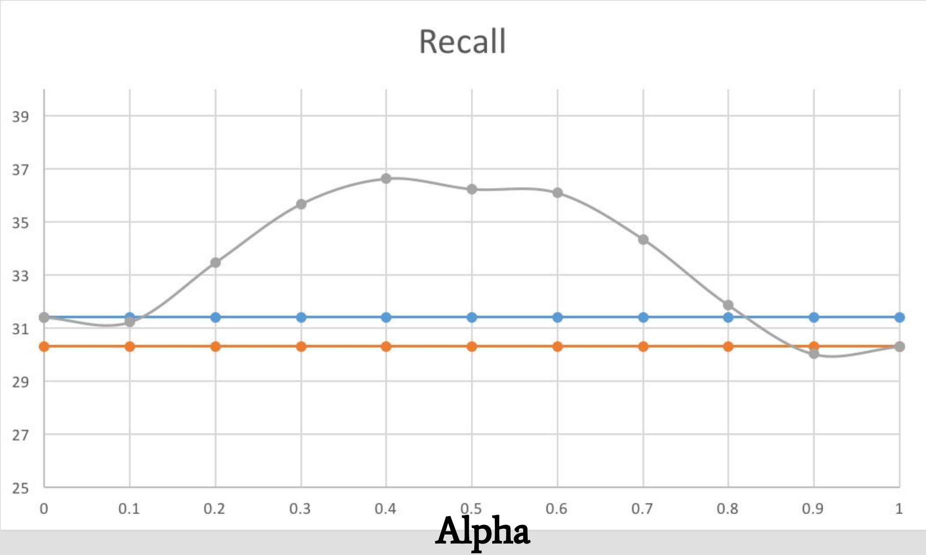
IELex Dataset, $\lambda = 0.7$, $p = 3$

# Subsequence Model Results

- *Hybrid* model between *Multiplicative* and *Additive* models
- Inspired from *Smoothing* of sparse vectors



# Siamese ConvNet Model

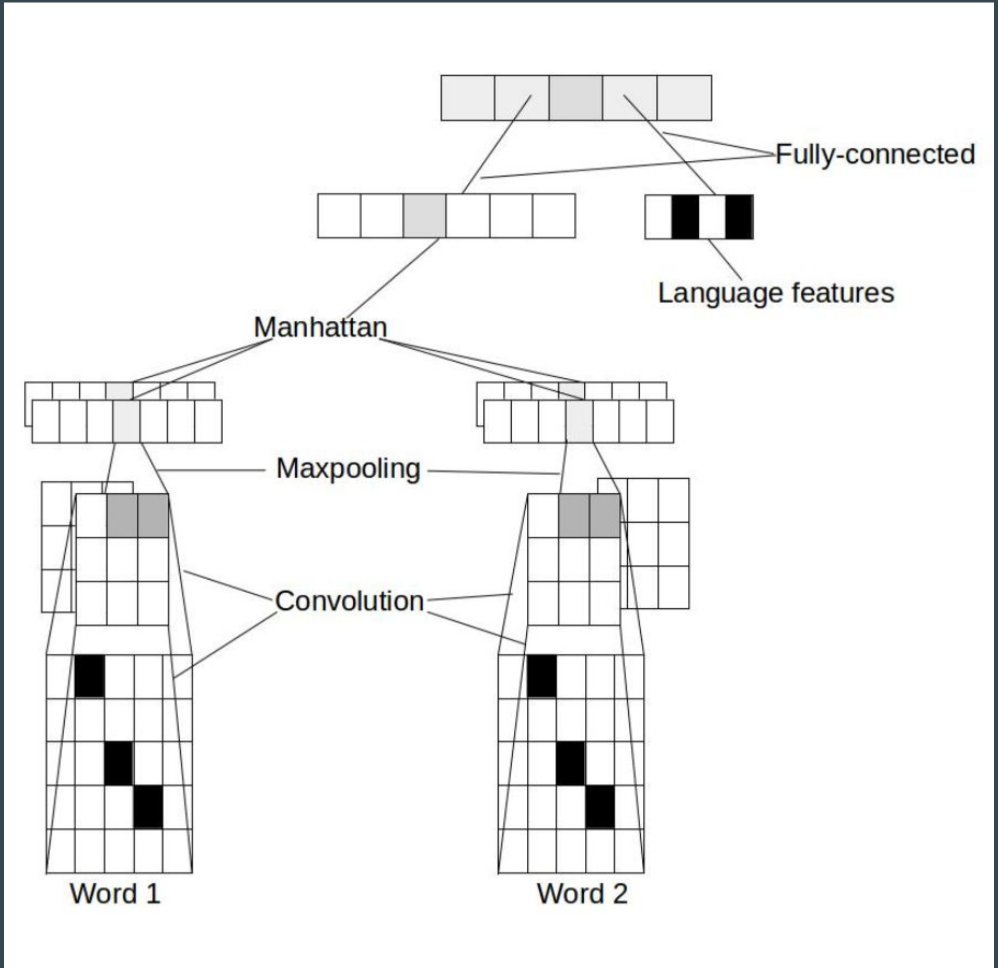
- Inspired from networks used for detecting similarity in images

- Manually defined character embeddings based on phonological properties

- No hand engineered features on word level

## Drawbacks

- Variable length words padded/clipped
- Character vectors defined by phonetic classes, convolution does not seem intuitive




*Sourced from [T.Rama, 2016]

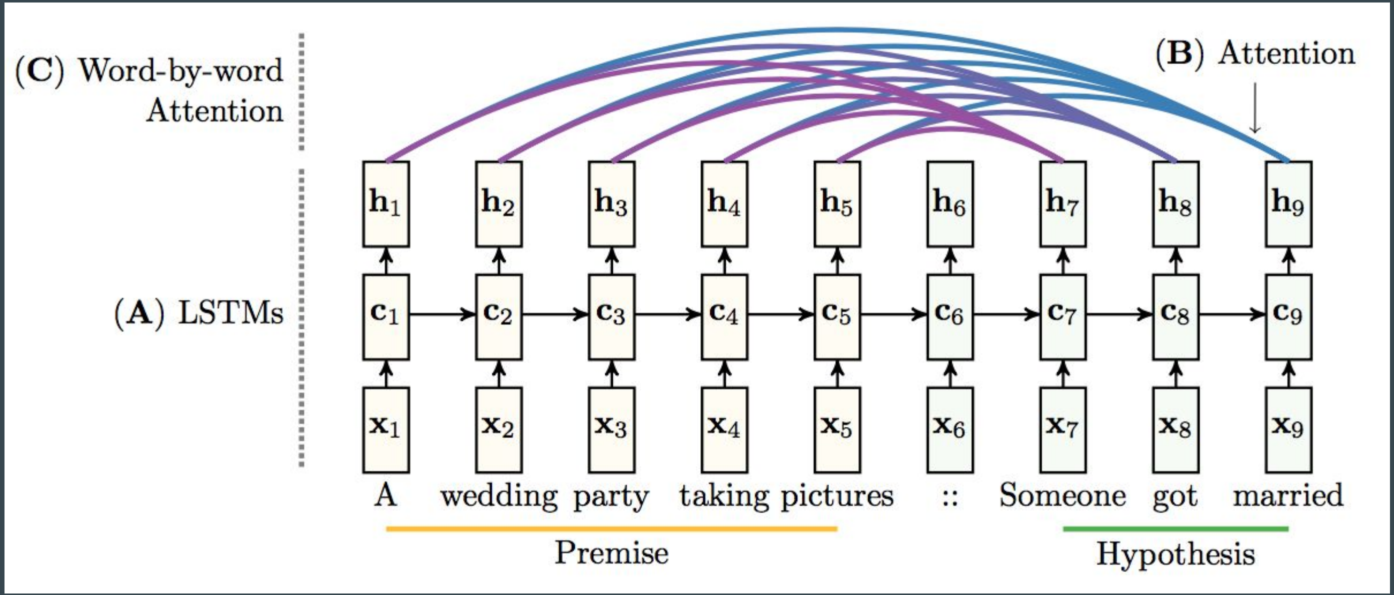
# Character Embeddings Used

| 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]

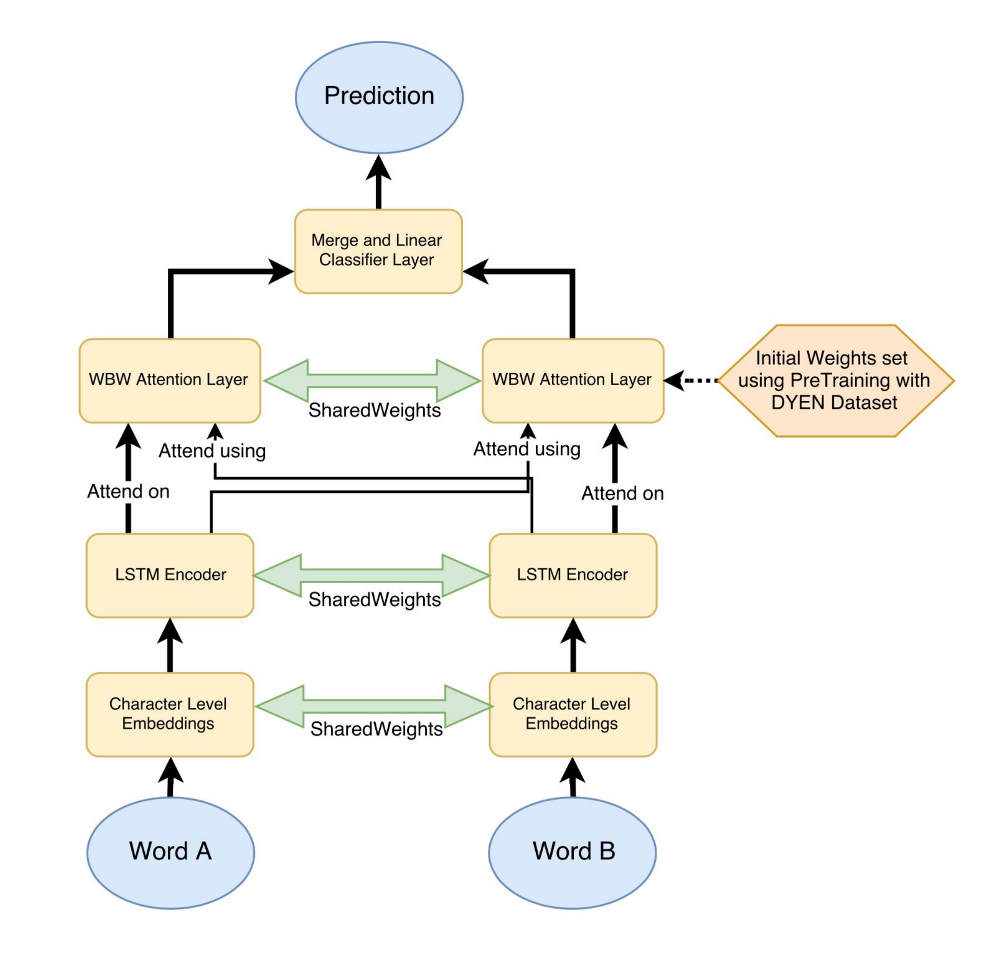
# Recurrent Model with Attention

- Rocktäschel, Tim, et al. "Reasoning about entailment with neural attention." - ICLR, 2016

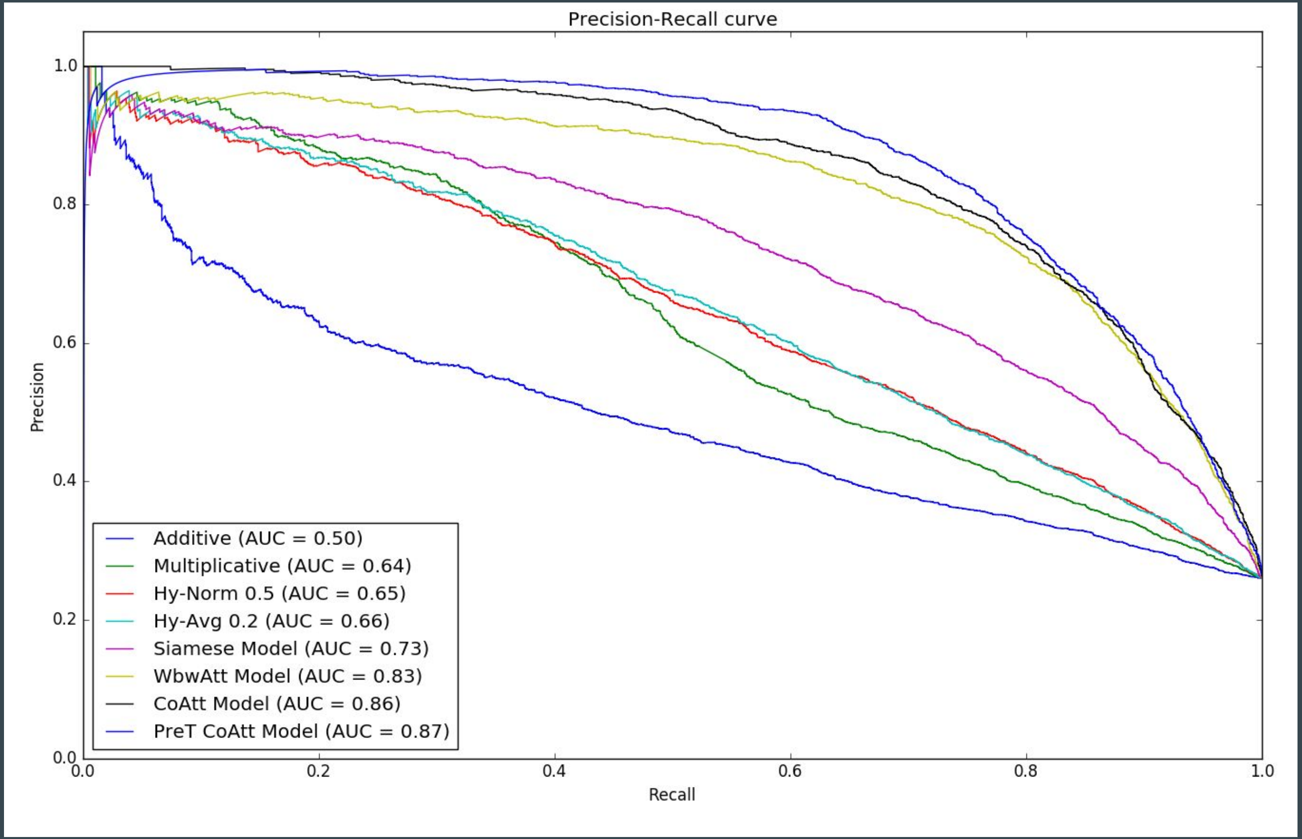


# Co-Attention LSTM Model

- Symmetric model with co-attention
- Shared LSTM encoder to encode each word
- Character-level encodings learnt
- Weights of attention layer pretrained using the romanized IPA dataset to improve performance



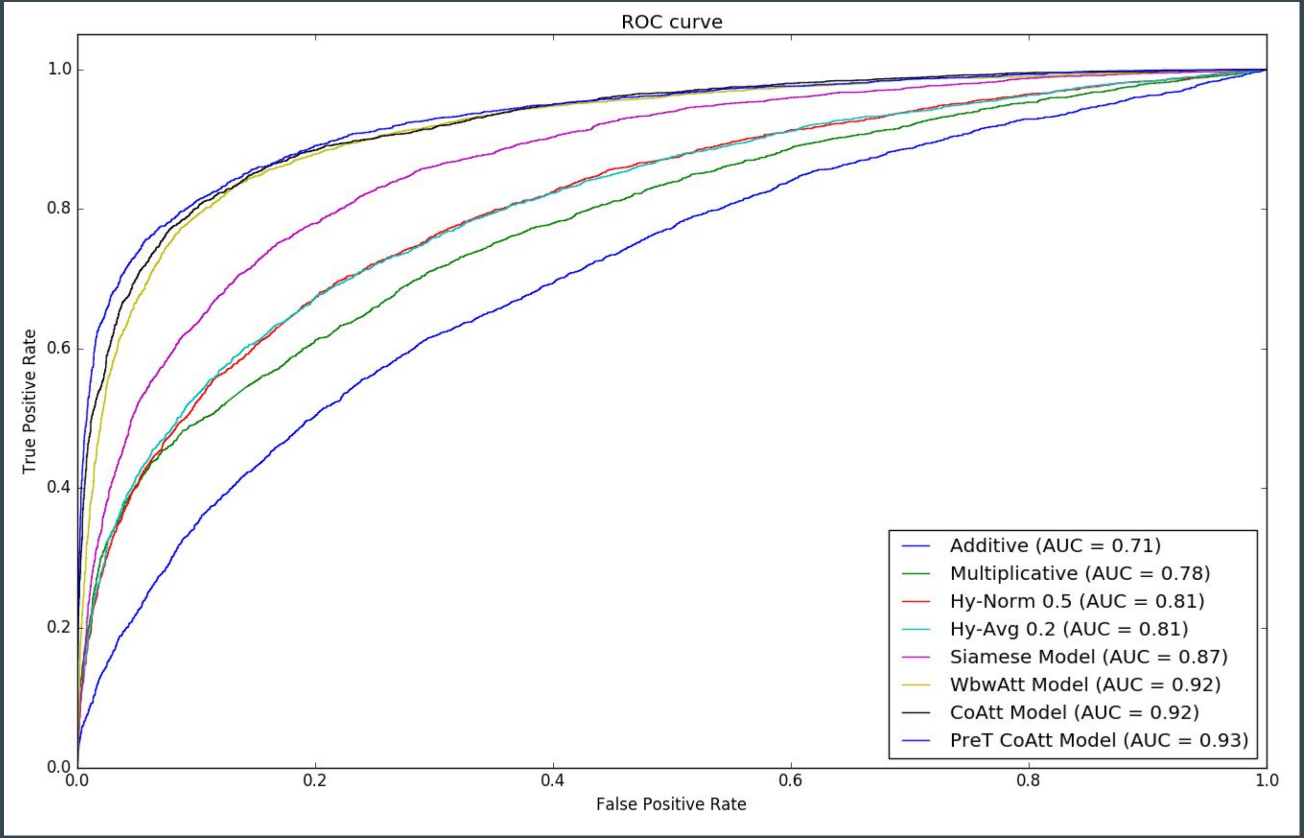
# Results - Cross Language Evaluation




- LSTM model gives significantly improved results than ConvNet

| | |
|---|---|
| tini | trʲi |
| tini | trɛɪ |
| tini | tres |
| tosk | tanː |
| tosk | ˈdant |
| tosk | dans |
| ˈmargir | ˈmnɔɦɔ |
| ˈmargir | mɪ̀kːɛl |
| kena | kome |
| kena | kəise |
| kena | kɛse |

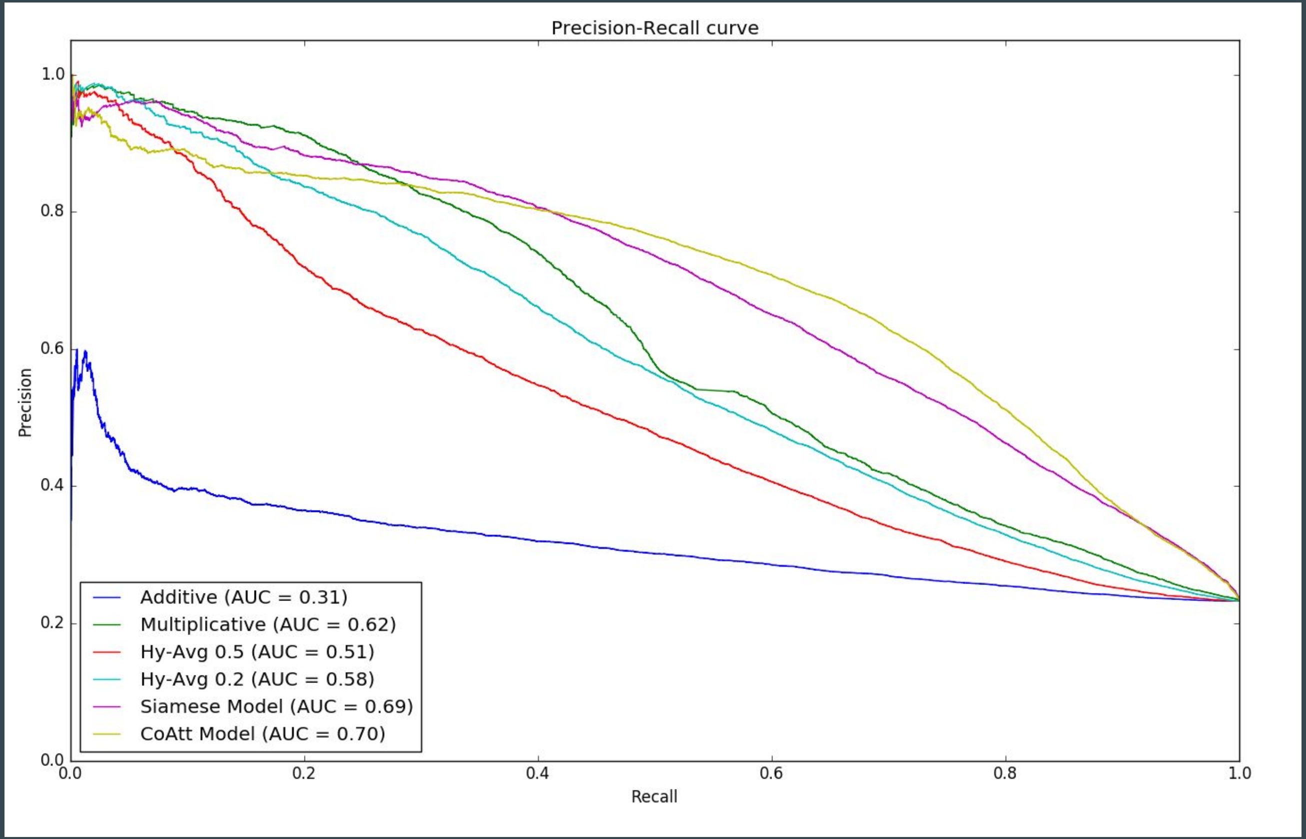
# Results - Cross Language Evaluation




- LSTM model gives significantly improved results than ConvNet

| | |
|---|---|
| **tini** | **trʲi** |
| **tini** | **trɛɪ** |
| **tini** | **tres** |
| **tosk** | **tanː** |
| **tosk** | **ˈdant** |
| **tosk** | **dans** |
| **ˈmargir** | **ˈmnɔɦɔ** |
| **ˈmargir** | **mɪ̀kːɛl** |
| **kena** | **kome** |
| **kena** | **kəise** |
| **kena** | **kɛse** |

# Results - Cross Concept Evaluation



# Future Work

- ➔ Analysis on the performance of Recurrent Model
  - Reasons for improved performance over ConvNet and Subsequence model
  - Performance over different sets of words
- ➔ Analysis of Character Embeddings learnt
  - Do the character level embeddings learnt represent the different phonetic classes
- ➔ Apply model to the domain of Hindi-Marathi
  - Cognate extraction from aligned texts by testing extracted noun-pairs on model
  - Manual evaluation
- ➔ Word-level and Language-level features
  - Semantic features can be introduced using word embeddings
  - Previous works have tried to encode language level features to improve performance

# 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 the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Association for Computational Linguistics, 2003.

3. Rama, Taraka. "**Automatic cognate identification with gap-weighted string subsequences.**" *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015.

4. Rama, Taraka. "**Siamese convolutional networks based on phonetic features for cognate identification.**" *arXiv preprint arXiv:1605.05172*(2016).

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