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

Indo-European Lexical Cognacy Database (IELex)

- 163 Languages, 225 Meanings
- 5000 Cognate Sets
- IPA transcription

Parallel Corpora

- Hindi-Marathi (TDIL)
- English-French (Europarl)

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. H. Bradley and G. Kondrak. "**Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists.**"
    - Orthographic word similarity features
    - Language pair similarity features

2. T. Rama. "**Automatic cognate identification with gap-weighted string subsequences.**"
    - String subsequences based features

3. T. Rama. "**Siamese convolutional networks based on phonetic features for cognate identification.**"
    - Representing words as images

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

$$\Phi_1(s_1, s_2) = \{\phi_u(s_1) + \phi_u(s_2); \forall u \text{ present in } s_1 \text{ and } s_2\}$$

Additive Model

$$\Phi_2(s_1, s_2) = \{\phi_u(s_1) + \phi_u(s_2); \forall u \text{ present in } s_1 \text{ or } s_2\}$$

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

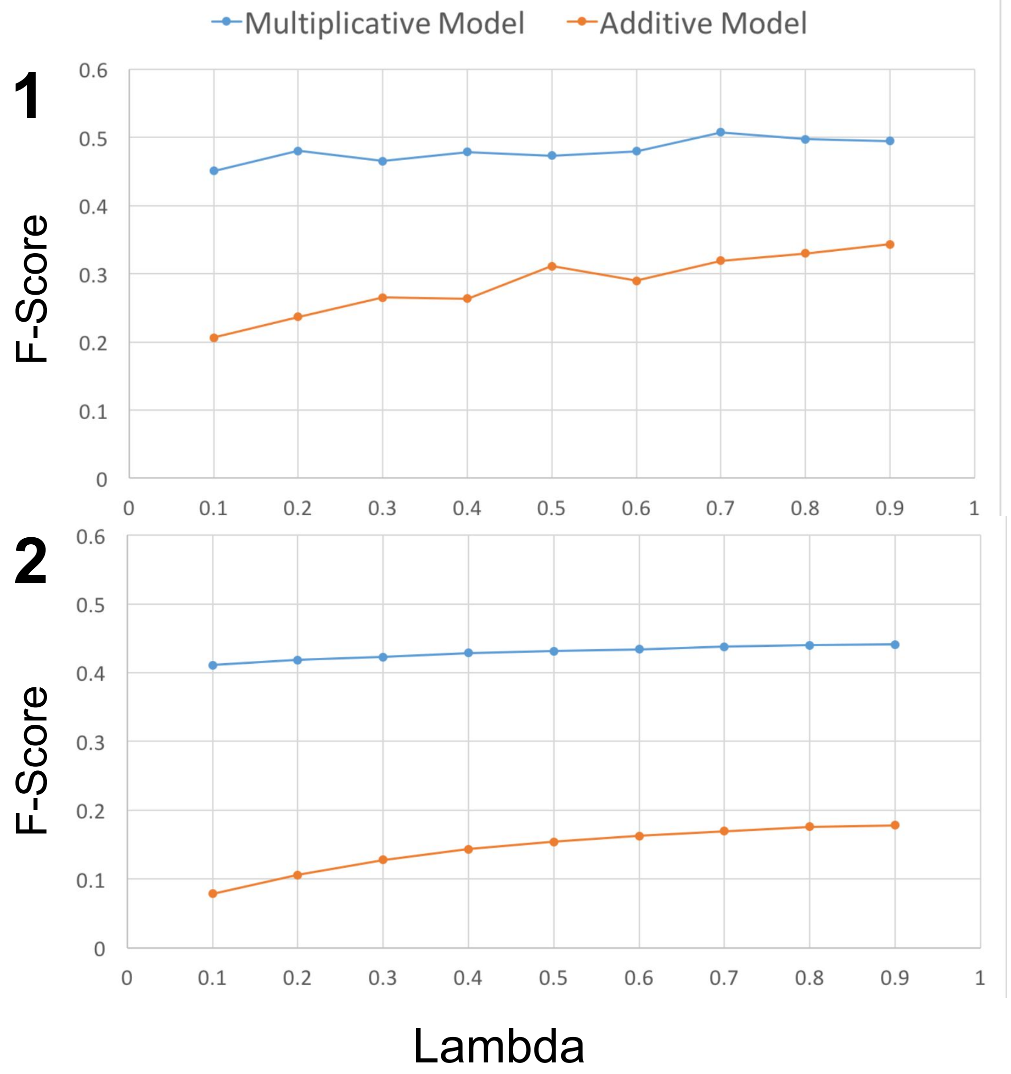
Testing Methods

1. Simple 5-Fold Cross Validation

- Training samples taken from all concepts
- No common words between Training and Testing set

2. Cross-Concept 5-Fold Cross Validation

- Training samples taken from certain concepts and Testing samples from remaining concepts
- Learn general trends of sound change in the languages



Error Analysis

Performance over Broad Categories of Samples

	Testing Data From		
Training Data From	Adjectives	Nouns	Others
Adjectives	0.513	0.330	0.160
Nouns	0.422	0.490	0.208
Others	0.350	0.380	0.360
All	0.522	0.495	0.351

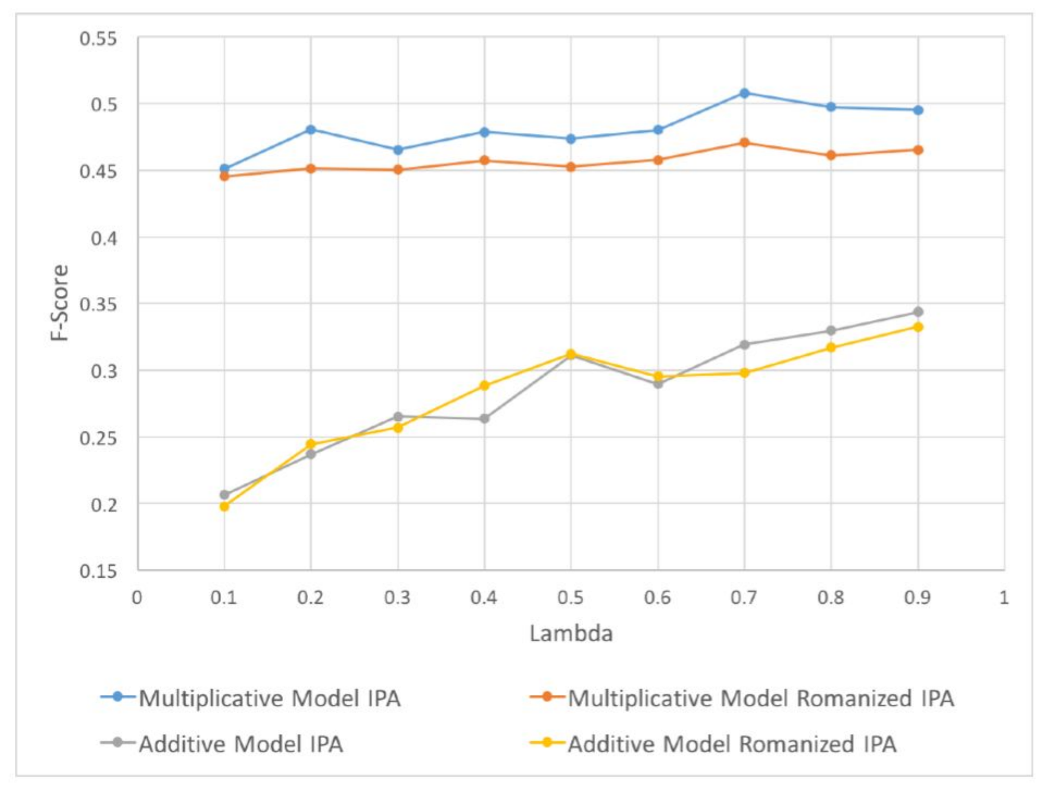
Multiplicative model, $\lambda = 0.7$, $p = 3$

Performance over Individual Concepts

Concept	Precision	Recall	F-Score	Cognate Classes
CHILD	99.98	79.99	0.888	24
TOOTH	99.99	76.92	0.869	5
BLACK	85.70	85.70	0.856	14
LAKE	81.81	89.99	0.856	22
...				
WHEN	99.98	7.59	0.141	8
HOW	79.98	7.69	0.140	8
WHAT	99.95	5.49	0.103	5
IN	59.98	3.99	0.074	12

Multiplicative model, $\lambda = 0.7$, $p = 3$

Role of Transcription In Detecting Sound Change

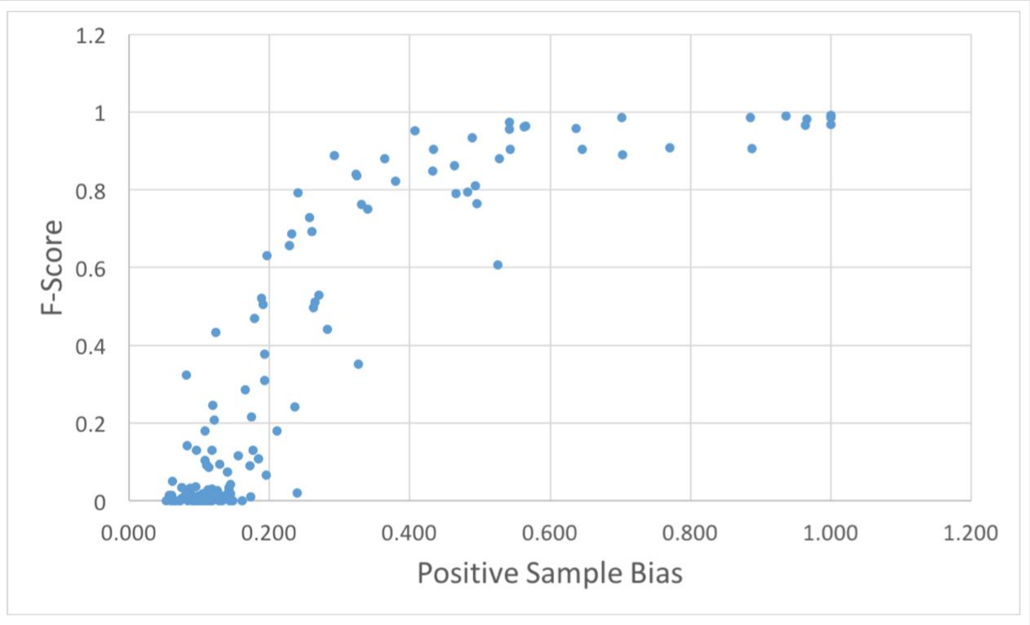


Cognates for the concept 'WHAT'

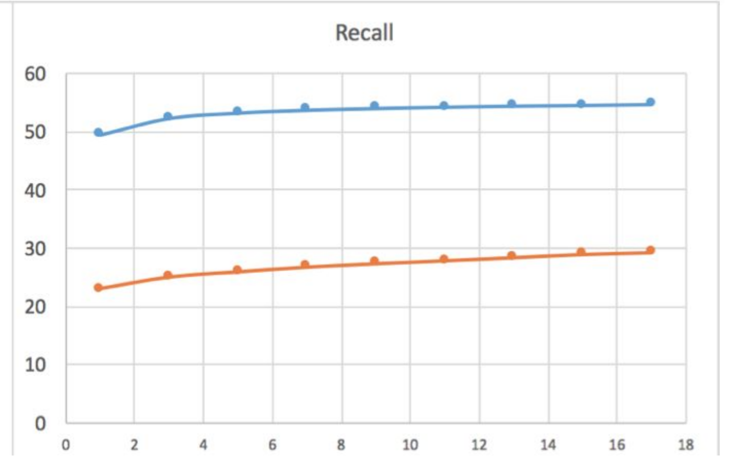
Language	**IPA**	**Romanized IPA**	***CV* String**
Danish	va	HVAD	*CCVC*
English	wɒt	WHAT	*CCVC*
French	kə	QUE	*CVV*
Marathi	kaj	KAY	*CVV*
Slovak	tʃɔ	CO	*CV*
Slovenian	kǎːj	KAJ	*CVC*
Spanish	ke	QUE	*CVV*
Swedish	vɑːd	VAD	*CVC*

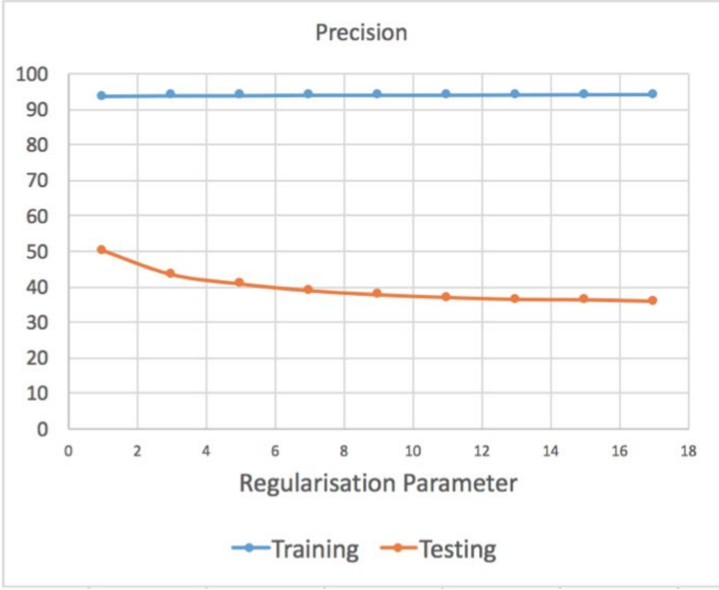
Analysis of Additive Model

- Overfitting on training data
- Large gap between training and testing constant with varying regularisation penalty
- High f-score on samples with high positive sample bias



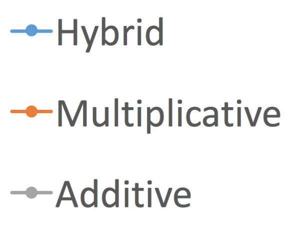
$\lambda = 0.7, p = 3$

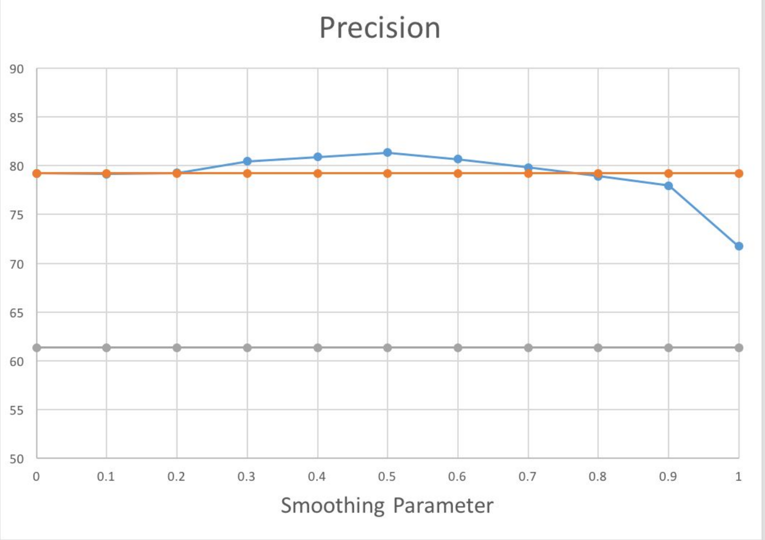


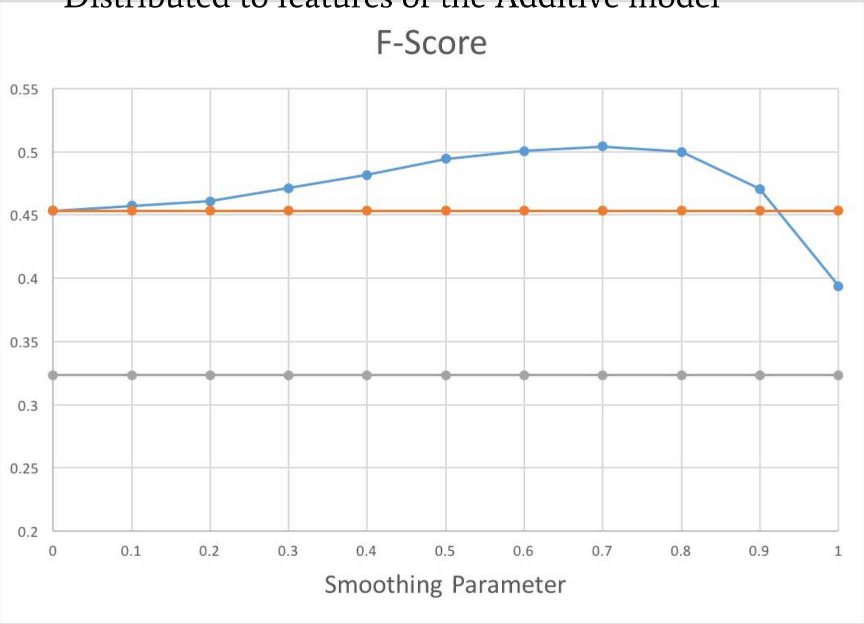


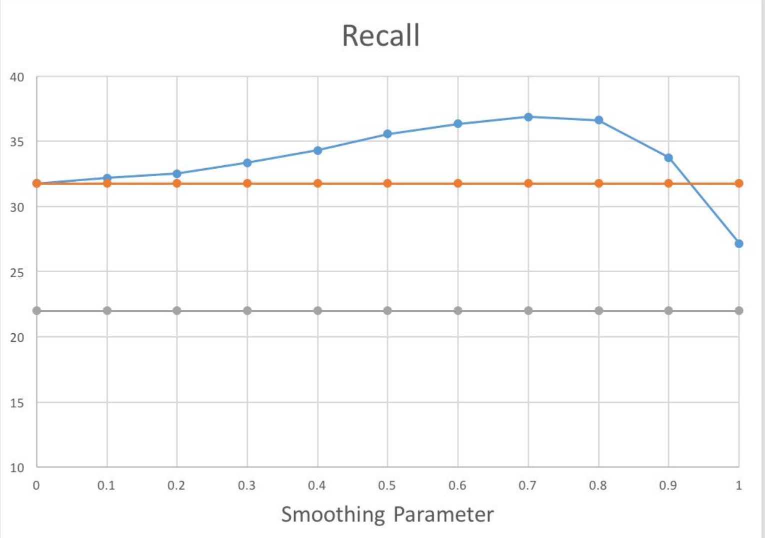
Hybrid Model

- Hybrid model between *Multiplicative* and *Additive* models
- Smoothing used to form hybrid features
  - Weight stolen from positive feats of the Multiplicative model
  - Distributed to features of the Additive model









Romanized IPA Dataset, $\lambda = 0.7$, $p = 3$

Future Work

- Word Embedding based features
  - Introduce semantic features
  - Context information
  - PolyGlot - Distributed word representation for multilingual NLP
    - Word embeddings for 117 Languages and 100K Vocabulary size
    - Trained on processed Wikipedia text dumps
- Character level RNN based model
  - Character encodings to help against transcription problem
  - Attention models
- Apply model to the domain of Hindi-Marathi and Hindi-Punjabi

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. Hauer, Bradley, and Grzegorz Kondrak. "**Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists.**" *IJCNLP*. 2011.

4. 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.

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