Cognate Discovery
For Bootstrapping Lexical Resources

Shantanu Kumar
2013EE10798

Supervisors

Sumeet Agarwal Ashwini Vaidya

Motivation

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

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

- ➔ Cognate identification is essential in historical linguistics.
- ➔ 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].
- ➔ Potentially used to bootstrap lexical resource creation in a low resource language.

*Sourced from [Hauer et al., 2011]

Objective

- ➔ Automatically discovering cognate pairs between closely related South-Asian language pairs like Hindi-Marathi and Hindi-Punjabi

- ➔ Linguistic analysis of cognates. Distinguish between cognates that are semantically similar or dissimilar.

Datasets

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

- 84 Languages
- 200 Meanings

Indo-European Lexical Cognacy Database (IELex)

- 163 Languages
- 225 Meanings
- 5000 Cognate Sets

Parallel Corpora

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

Previous Work

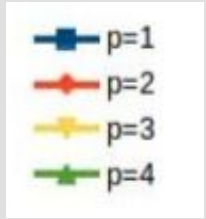
1. H. Bradley and G. Kondrak. "**Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists.**"
    - Orthographic word similarity features
    - Language pair similarity features
    - Clustering into cognate groups

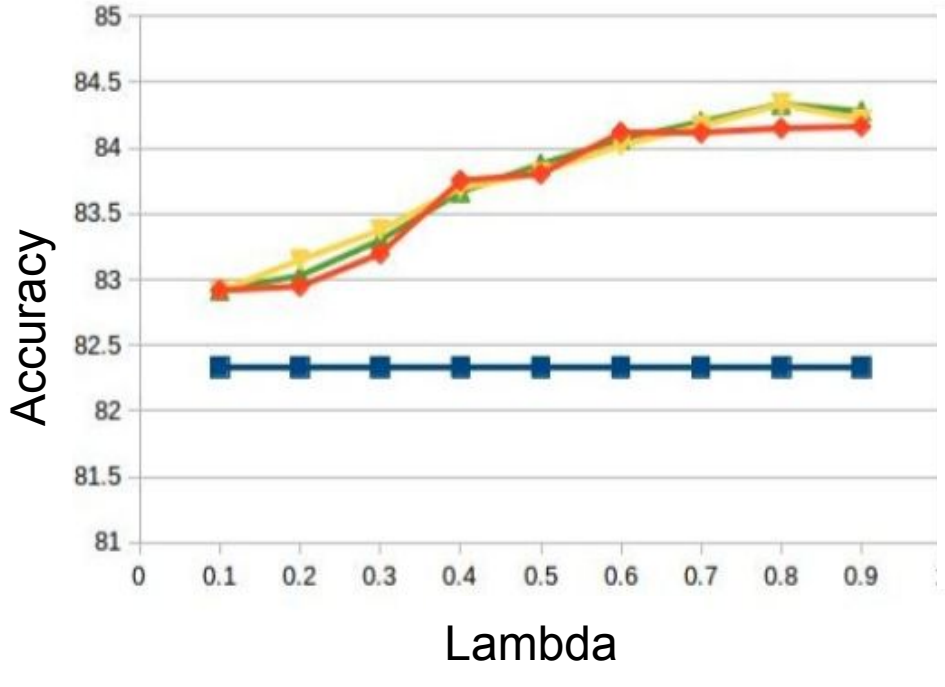
2. T. Rama. "**Automatic cognate identification with gap-weighted string subsequences.**"
    - String subsequences based features

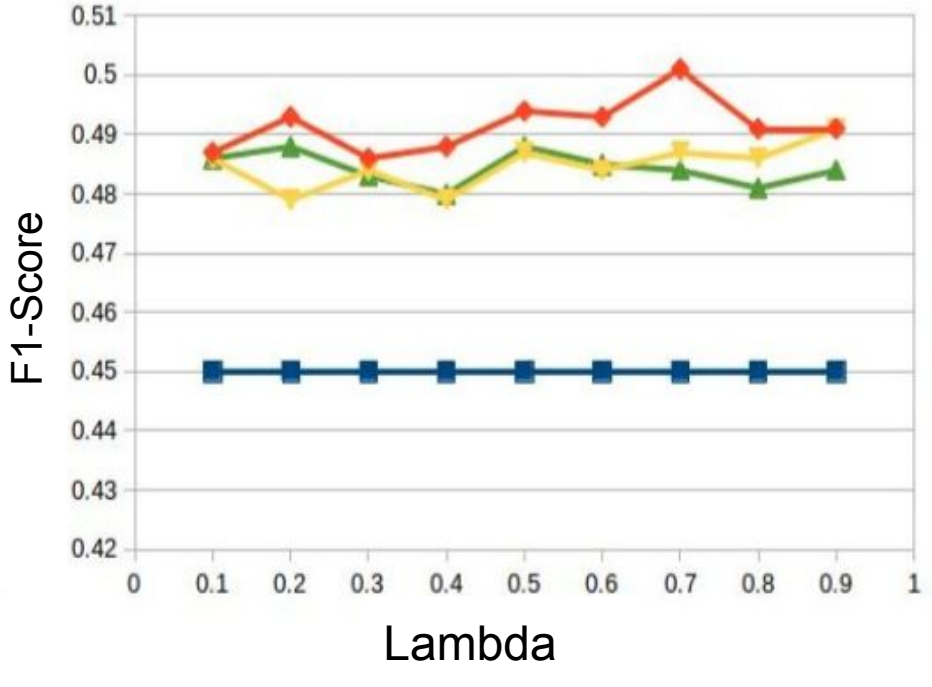
Results (As Stated in Paper)

Max Accuracy : 84.4%

Max F-Score : 0.50





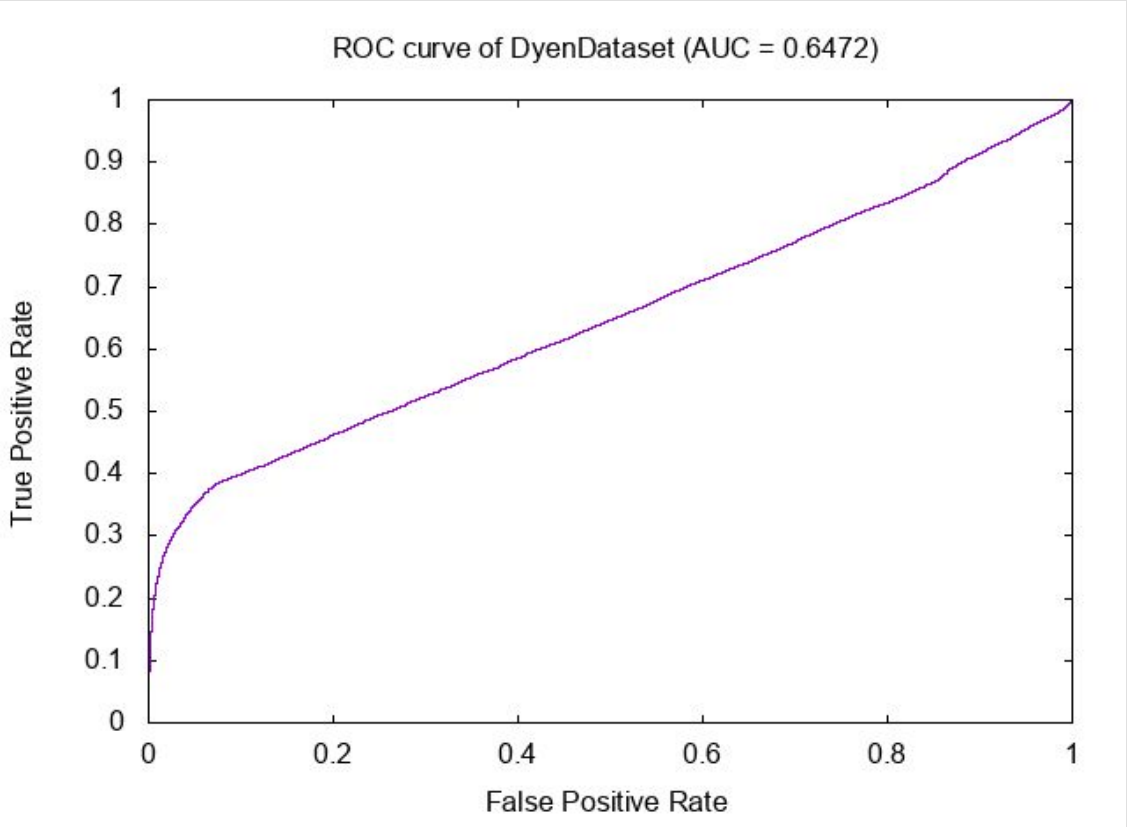


Images sourced from paper (T. Rama, 2015)

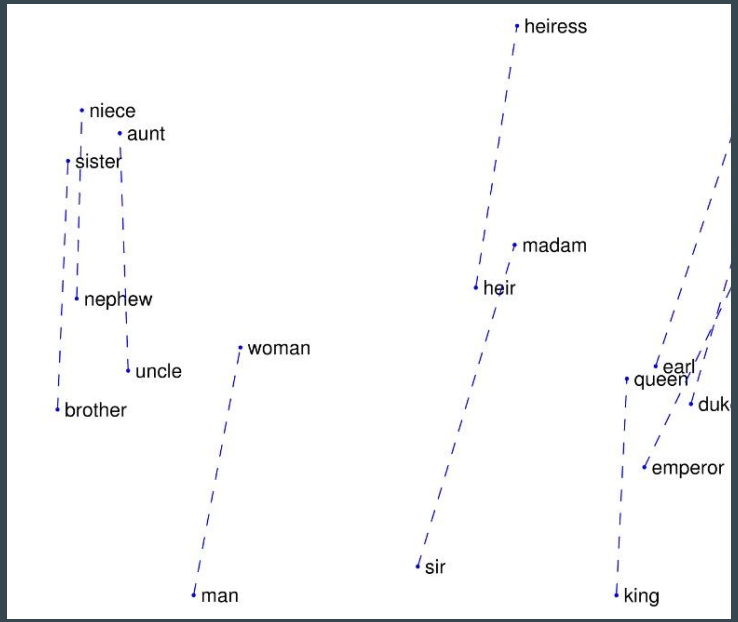
Results (Replicated Model)

(10 Fold Cross Validation)

Precision	:	64.0%
Recall	:	38.7%
F-Score	:	0.48



Initial Solution Approach



- ➔ Implement Baseline model
- ➔ Word Embedding based features
  - Introduce semantic features
  - Context information
  - Increase dimensionality
- ➔ PolyGlot : Distributed word representation for multilingual NLP
  - Word embeddings for 117 Languages and 100K Vocabulary size
  - Trained on processed Wikipedia text dumps
- ➔ Apply model to the domain of Hindi-Marathi and Hindi-Punjabi

Challenges

- Non-uniformity of data transcription format

ANIMAL - Hindi	
Romanized Phonetic Alphabet	JANVER
International Phonetic Alphabet	dʒanʊər
Devanagari	जानवर

- Evaluation and Ground truth
  - Sample the set of findings for manual evaluation
  - Explore/search options for automatic evaluation

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. Navlea, Mirabela, and Amalia Todirascu. "**Using Cognates in a French-Romanian Lexical Alignment System: A Comparative Study.**" *RANLP*. 2011.

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

5. 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.