Investigations into the Use of Distributed Semantics for Cognate Detection and Phylogenetics

Diptesh Kanojia
(IIT Bombay, India, and Monash University, Australia)

under the guidance of
Prof. Pushpak Bhattacharyya, Prof. Reza Haffari and Prof. Malhar Kulkarni

“Do you speak Indian?”

- India has a total of 22 scheduled languages which primarily belong to the *Indo-Aryan* and the *Dravidian* language families.
- I can speak in Hindi, Punjabi, Marathi, and English, but I can only write in Hindi, and English.
- Similarly, many natives of India can speak and understand multiple Indic languages but can only write a subset of those.

- **I used to wonder why?**

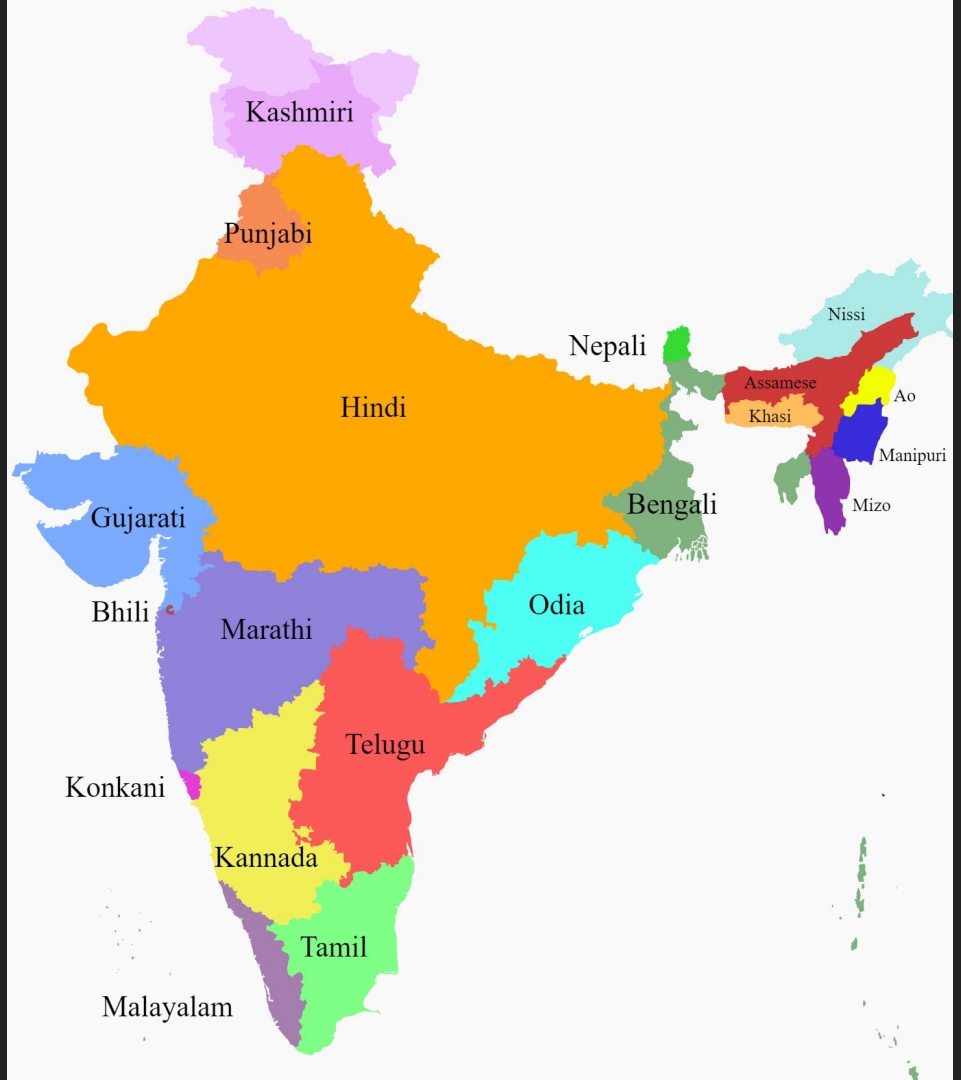


Image Source

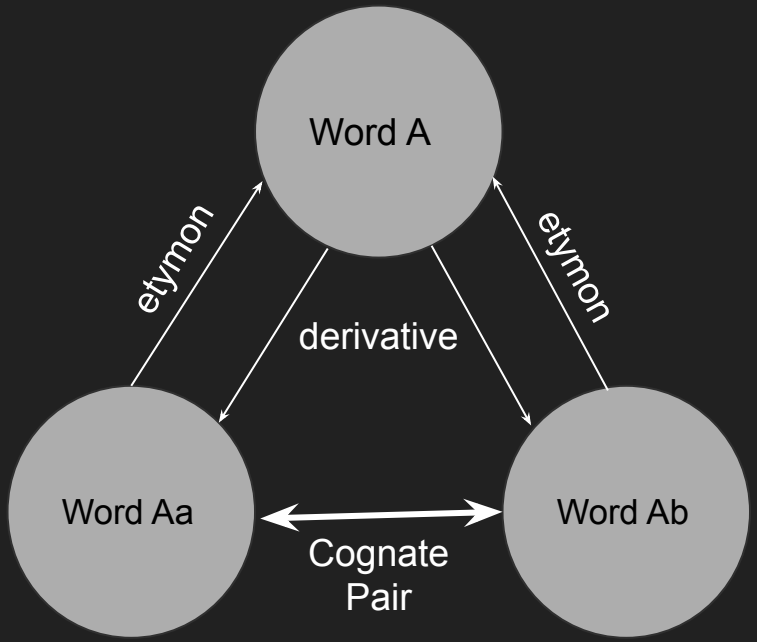
Key Questions

Language relatedness is already exploited in language processing applications such as machine translation. **Do cognate words play a vital role?**

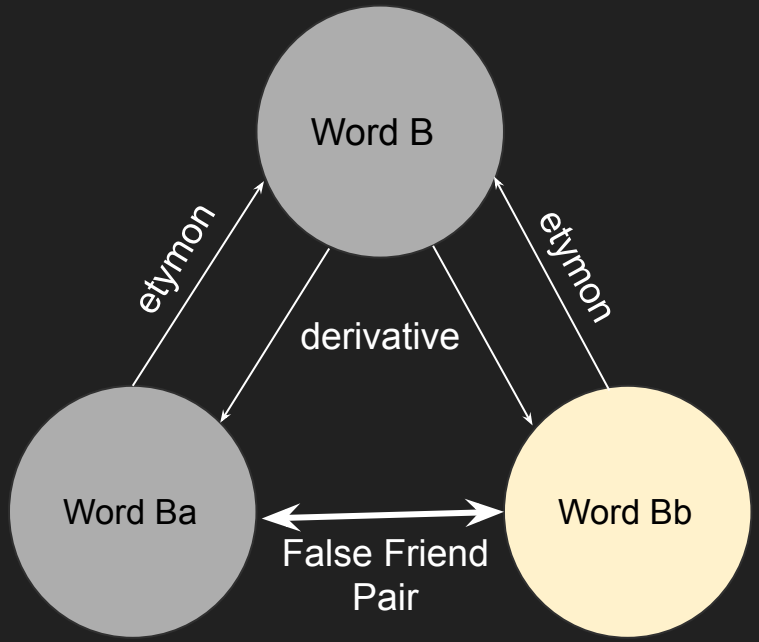
How does one detect these words? What are the challenges?

How are these languages related to each other? Can language relatedness be defined?

See what you (do not) mean!

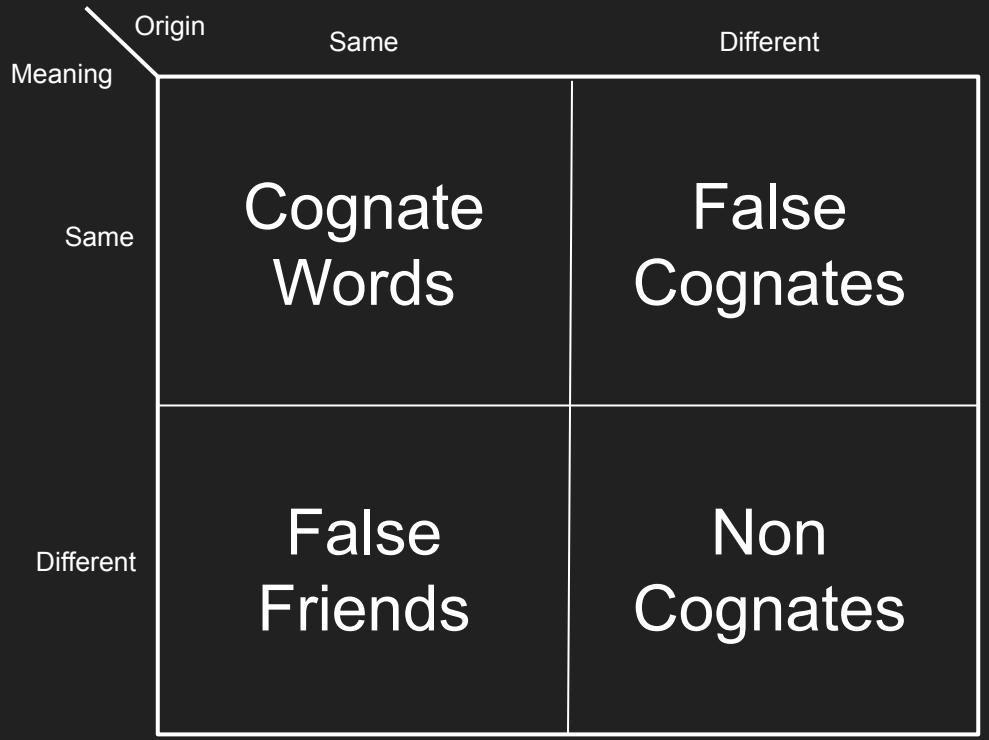


They carry the same meaning

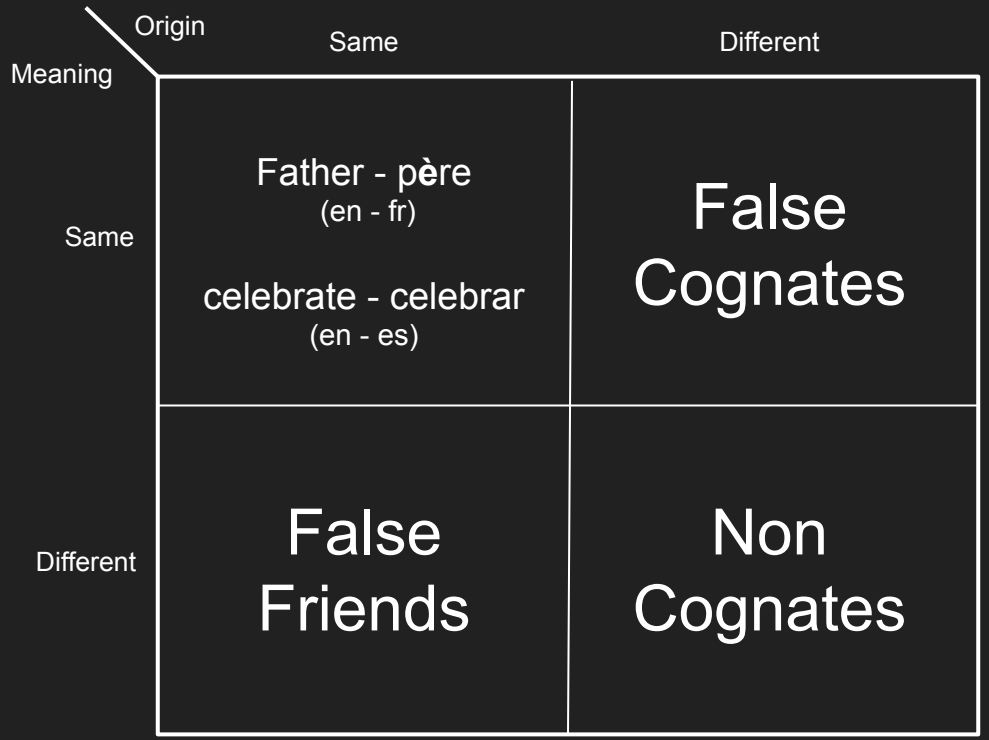


They differ in meaning

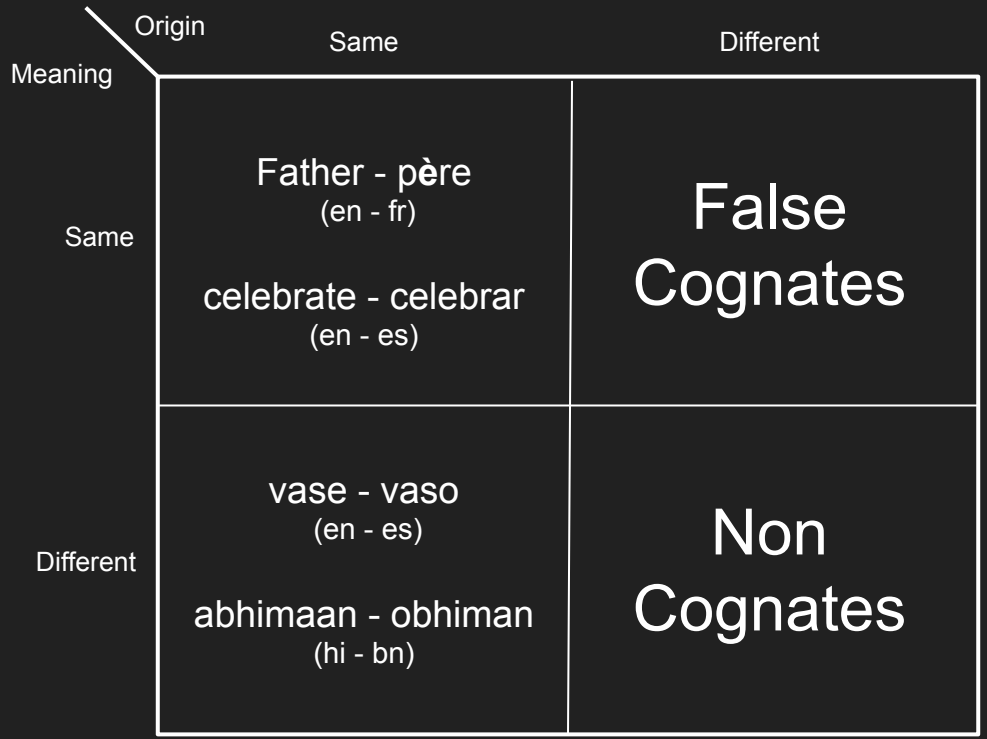
The etymological matrix



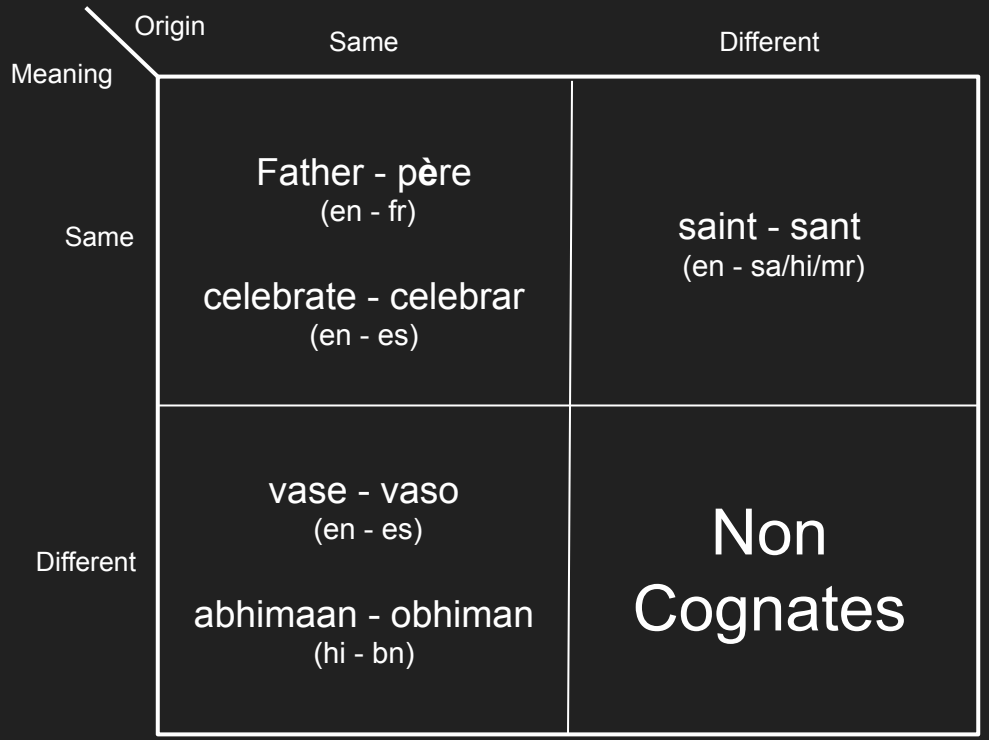
The etymological matrix



The etymological matrix



The etymological matrix



Problem Definition - Cognate Detection

Cognate Detection is defined as the task of identifying whether a given word pair is a cognate pair or not (Mulloni et. al., 2006).

Aim- create a model which learns this identification based on orthographic (spelling) and semantic (meaning) clues, across languages.

Input- two words with their contexts to identify *whether they are cognates or not.*

Problem Definition - False Friends' Detection

The task of False friends' detection is defined as identification of a pair of words which are lexically similar / same, but differ in meaning across languages.

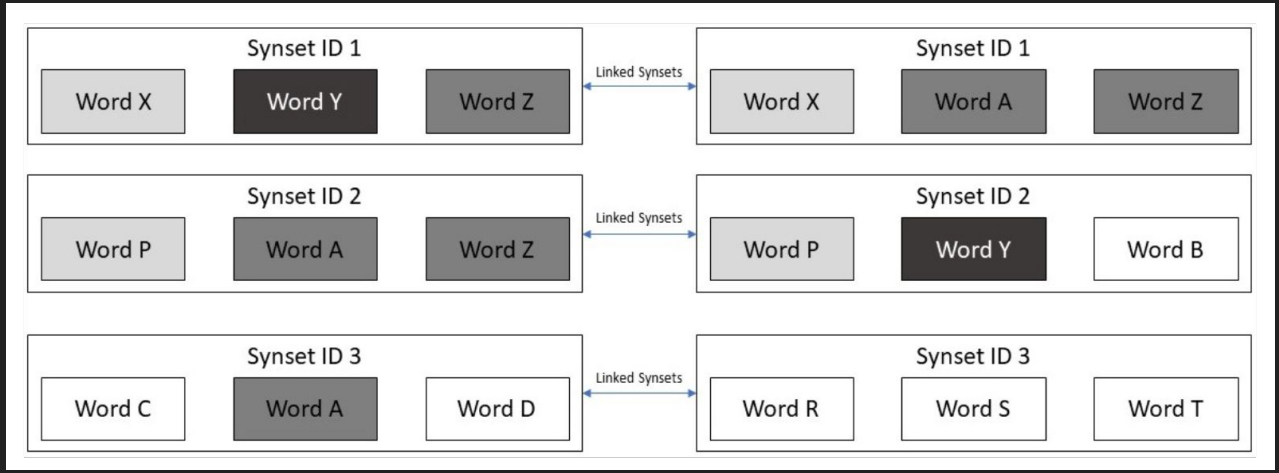
Aim - To create a model given two words and their contextual clues.

Input - two words with their contexts to identify *if these two words are False Friends or not.*

Dataset Construction

IndoWordnet (Bhattacharyya, 2017) is a collection of linked Indian language wordnets.

Our datasets for each study differ in terms of language pairs and hence exact size and number of language pairs are provided when we go into experiment details.



Utilizing Wordnets for Cognate Detection among Indian Languages
(Global Wordnet Conference 2019)

- Dataset used
  - Eleven Indian languages
    - Hindi (Hi), Marathi (Mr), Punjabi (Pa), Sanskrit (Sa), Gujarati (Gu), Bengali (Bn), Malayalam (Ml), Tamil (Ta), Telugu (Te), Nepali (Ne), and Urdu (Ur)
  - Language pairs with Hindi as pivot (Hi-Mr, Hi-Bn, and so on)
  - Data Sources - Indian language corpus (D1) and Wordnets (D2)
- Identify word pairs which belong to the same concept with similar spellings.

Our proposed empirical measure of weighted lexical similarity combines normalized edit distance, q-gram distance and Jaro-winkler similarity.

We use deep neural network based techniques (Feed-forward and Recurrent Neural Network) to create models which learn based such similarities.

Classification Approaches

FFNN - Treat the word as a whole.

Words reside in separate embedding spaces.

The target word passes through the target embedding layer and the output of both embedding lookups is concatenated.

RNN - Treat the word as a sequence of characters.

The embedding spaces contain characters from the source and the target side.

In a similar fashion, the source and target side characters pass through their respective embedding layers and at the end the output is concatenated.

The resulting representations are passed to a fully-connected layer with ReLU activation followed by a softmax layer.

Results

Classifiers trained on the WNData perform better.

RNN outperforms FFNN uniformly and with significant margins.

The highest 5-fold evaluation score achieved was for the classification models on the language pair Hindi-Sanskrit (i.e., 91.66) which are very closely related share a lot of vocabulary.

	FFN		RNN	
	D1	D2	D1	D2
Hi-Mr	69.76	85.76	74.76	89.78
Hi-Bn	65.18	81.04	69.18	86.44
Hi-Pa	73.04	78.50	76.04	83.64
Hi-Gu	61.74	79.16	69.84	89.44
Hi-Sa	61.72	85.87	68.92	91.66
Hi-Ml	56.96	74.77	66.96	79.59
Hi-Ta	55.62	61.70	65.62	68.92
Hi-Te	52.78	65.26	62.78	74.83
Hi-Ne	70.20	83.85	80.20	89.63
Hi-Ur	69.99	73.84	76.99	80.12

Stratified 5-fold validation accuracy over D1 - Corpus Data, and D2 - Wordnet Data

True Cognate Detection using Siamese Deep Neural Networks

(CoDS-COMAD 2020)

Siamese Feed-forward neural network with **monolingual** word-embeddings to detect cognates.

We perform this study with ten Indian languages (nine language pairs) namely, Hindi (Hi), Bengali (Bn), Gujarati (Gu), Marathi (Ma), Punjabi (Pa), Sanskrit (Sa), Malayalam (Ml), Tamil (Ta), Telugu (Te), Nepali (Ne).

Data splits:

- Training - 70%
- Testing - 20%
- Validation - 10%

On a dataset built using linked Indian Wordnets, our approach beats the baseline approach with a significant margin (up to 71%) with the best F-score of 0.85% on the Hindi-Gujarati language pair.

Approach & Results

The intuition for using a siamese feed forward network-based approach is that these networks perform a combined mapping of input vectors into a common target space.

LP	Baseline Approach			Our Approach: Siamese Feed-forward Network (SFN)								
	LSA			MEA (200 dim.)			MEA (300 dim.)			MEA (400 dim.)		
	P	R	F	P	R	F	P	R	F	P	R	F
Hi - Bn	0.39	0.33	0.36	0.80	0.82	0.81	0.81	0.83	**0.82**	0.81	0.80	0.81
Hi - Mr	0.47	0.21	0.29	0.81	0.83	0.82	0.83	0.83	**0.83**	0.82	0.82	0.82
Hi - Gu	0.41	0.16	0.23	0.83	0.84	0.84	0.84	0.86	**0.85**	0.84	0.83	0.84
Hi - Pa	0.29	0.07	0.11	0.78	0.79	0.78	0.82	0.82	**0.82**	0.81	0.80	0.81
Hi - Ml	0.26	0.3	0.28	0.74	0.74	**0.74**	0.73	0.73	0.73	0.73	0.73	0.73
Hi - Te	0.2	0.14	0.16	0.73	0.70	**0.71**	0.70	0.70	0.70	0.70	0.69	0.69
Hi - Ta	0.24	0.17	0.20	0.71	0.71	**0.71**	0.70	0.70	0.70	0.69	0.70	0.70
Hi - Sa	0.41	0.17	0.24	0.82	0.83	0.82	0.81	0.85	**0.83**	0.81	0.81	0.81
Hi - Ne	0.42	0.18	0.25	0.78	0.80	**0.79**	0.78	0.77	0.77	0.78	0.77	0.77

LSA: Lexical Similarity based approach; MEA: Monolingual Embeddings based approach

Challenge Dataset of Cognates and False Friend Pairs from Indian Languages
(LREC 2020)

We describe the creation of three cognacy related datasets for 12 Indian languages.

D1: digitization of a Cognate dictionary and its annotation with linked Wordnet IDs.

Dataset size: 1021 cognate sets with a total of 12252 words. The book consisted of a total of 1556 cognate sets, but during manual validation, 535 were found to be partial cognates and have been ignored from this dataset.

D2: We use linked Indian Wordnets to generate potential cognate lists and create another true cognate dataset with the help of manual annotation.

D3: We create the dataset for false friend pairs by using a similar methodology.

Dataset Annotation

Language Pair	Hi-Bn	Hi-Gu	Hi-Mr	Hi-Pa	Hi-Sa	Hi-Ml	Hi-Ta	Hi-Te	Hi-As	Hi-Kn	Hi-Or
Potential Candidates	50959	81834	47718	25044	33921	18084	5203	16230	14240	12480	54014
Cognates (D2)	15312	17021	15726	14097	21710	9235	3363	936	3478	4103	11894
Percent Agreement	0.9877	0.9849	0.9838	0.9754	0.9617	0.9223	0.9033	0.9553	0.9167	0.9122	0.8833
Cohen's kappa	0.7851	0.7972	0.8628	0.7622	0.7351	0.7046	0.6436	0.7952	0.7591	0.7953	0.8333

Table 1: Number of Potential Cognates, Number of cognates retained on both annotators' agreement [Cognates (D2)], Percent agreement among the annotators and Cohen's kappa score for each language pair in our dataset

Language Pair	Hi-Bn	Hi-Gu	Hi-Mr	Hi-Pa	Hi-Sa	Hi-Ml	Hi-Ta	Hi-Te	Hi-As	Hi-Kn	Hi-Or
Potential Candidates	11128	10378	14430	9062	9285	5192	1018	7149	9374	3384	5011
False Friends (D3)	4380	6204	5826	4489	2193	1076	783	699	3872	926	2602
Percent Agreement	0.8912	0.9122	0.9233	0.9500	0.9018	0.8125	0.9288	0.8492	0.8825	0.9367	0.9133
Cohen's kappa	0.8827	0.8245	0.7815	0.9255	0.9452	0.9064	0.7244	0.8901	0.8432	0.8167	0.9548

Table 2: Number of Potential False Friends, Number of False Friend pairs retained on both annotators' agreement [False Friends (D3)], Percent agreement among the annotators and Cohen's kappa score for each language pair in our dataset

Results of Cognate and False Friends' Detection Tasks

Approaches	Hi-Bn	Hi-As	Hi-Or	Hi-Gu	Hi-Mr	Hi-Pa	Hi-Sa	Hi-Ml	Hi-Ta	Hi-Te	Hi-Kn
Orthographic Similarity	0.36	0.34	0.38	0.25	0.29	0.21	0.24	0.28	0.20	0.16	0.19
Phonetic Similarity	0.42	0.38	0.39	0.29	0.32	0.24	0.25	0.31	0.24	0.22	0.25
Rama et. al. (2016)	0.65	**0.71**	0.61	0.67	**0.72**	0.47	0.53	0.62	**0.53**	**0.65**	0.57
Kanojia et. al. (2019)	**0.68**	**0.71**	**0.62**	**0.75**	**0.72**	**0.73**	**0.72**	**0.66**	**0.53**	0.63	**0.58**

Table 3: Results of the Cognate Detection Task (in terms of F-Scores) for D1+D2. We use the same architecture, features and hyperparameters as discussed in the papers for Rama et. al. (2016) and Kanojia et. al. (2019) and observe that these systems do not perform as well on our dataset, as claimed by the authors.

Language Pairs	Hi-Bn	Hi-As	Hi-Or	Hi-Gu	Hi-Mr	Hi-Pa	Hi-Sa	Hi-Ml	Hi-Ta	Hi-Te	Hi-Kn
Orthographic Similarity	0.36	0.45	0.49	0.51	0.53	0.44	0.52	0.24	0.29	0.30	0.50
Phonetic Similarity	0.60	**0.66**	**0.67**	0.62	0.59	0.69	0.61	0.54	0.48	0.50	0.57
Castro et. al. (2018)	**0.66**	0.64	0.59	**0.65**	**0.69**	**0.73**	**0.72**	**0.65**	**0.52**	**0.69**	**0.64**

Table 4: Results of the False Friends' Detection Task (in terms of F-Scores) for D3. We use the same architecture, features and hyperparameters as discussed in the paper by Castro et. al. (2018) and observe that these systems do not perform as well on our False Friends' dataset.

Injecting Cognates to improve Machine Translation

Language Pair	No. of Cognates	BLEU (baseline)	BLEU (w/ cognates)	Improvement
hi-pa	39458	62.71	62.79	0.08
hi-bn	93395	28.75	30.2	1.45
hi-gu	134919	52.17	52.42	0.25
hi-mr	83783	31.66	32.79	1.13
hi-ta	8615	21.75	21.97	0.22
hi-te	31016	18.62	19.18	0.56
hi-ml	32832	10.4	10.8	0.4

Utilizing Deep Cross-Lingual Word Embeddings to Detect False Friends

(Under Review; ACL 2020)

Cognate identification approaches can confuse a false friend pair to be a cognate if orthographic similarity based techniques are relied upon.

False friends are especially problematic for language learners as learners tend to overgeneralize and assume that they know the meaning of these misleading words.

Hence, we also focus on the task of false friends' detection and propose a novel approach, which can identify false friends' from among possible cognate pairs, by utilizing distributed semantics across languages.

Cross-lingual Vectors and Similarity (CLS)

(Our Approach)

Cross-lingual word embeddings are becoming increasingly important in multilingual NLP.

We obtain vectors for word-pairs and averaged context vectors to create feature sets. We also use angular cosine similarity (Cer et al.,2018) scores for word pairs and their contexts.

For each word pair vector and it's context vectors, we compute the 'word-pair similarity' and 'contextual similarity'.

Our ablation test results show that a combination of orthographic and semantic approach performs the best.

Classification Methodology

We employ both classical machine learning based models and a deep learning based model to detect false friends.

Among the classical machine learning models, we use Support Vector Machines (SVM) and Logistic Regression (LR). We perform a grid-search to find the best hyper-parameter value for C over the range of 0.01 to 1000.

We also deploy a simple Feed Forward Neural Network (FFNN) with one hidden layer. We perform cross-validation with different settings for activation function (tanh, hardtanh, sigmoid and relu) and the hidden layer dimension in the network (30, 50, 100, and 150).

Results

Results of the false friends' detection task, in terms of weighted F-scores for Weighted Lexical Similarity (WLS), Phonetic Vectors and Similarity (PVS), State-of-the-art (Castro et al., 2018), Mono lingual Similarity (MVS) i.e., SoTA w/ FastText,

Cross lingual Similarity (CLS) and a combination of Cross lingual and Weighted Lexical Similarity (CLS+WLS) [Our Approaches]

based features, over all language pairs (LP), and both the datasets (D1 and D2).

		Baseline Approaches				Our Approaches		
	LP	WLS	PVS	SoTA	MVS	CLS w/ SVM	CLS w/ FFNN	CLS+WLS w/ FFNN
D1	Hi-Bn	0.66	0.59	0.58	0.60	0.72	**0.77**	**0.77**
	Hi-Gu	0.60	0.60	0.61	0.63	0.68	**0.89**	**0.89**
	Hi-Ml	0.72	0.65	0.36	0.71	0.77	0.84	**0.87**
	Hi-Mr	0.64	0.64	0.53	0.64	0.72	0.91	**0.92**
	Hi-Ne	0.44	0.57	0.42	0.64	0.65	**0.77**	**0.77**
	Hi-Pa	0.58	0.58	0.37	0.66	0.84	0.84	**0.89**
	Hi-Sa	0.53	0.63	0.21	0.61	0.64	**0.86**	**0.86**
	Hi-Ta	0.71	0.72	0.45	0.69	**0.72**	**0.72**	**0.72**
	Hi-Te	0.56	0.65	0.52	0.56	0.67	0.78	**0.79**
D2	Hi-Bn	0.40	0.38	0.43	0.65	0.70	**0.74**	**0.74**
	Hi-Gu	0.38	0.60	0.56	0.66	0.69	0.70	**0.73**
	Hi-Ml	0.25	0.38	0.33	0.65	0.68	0.68	**0.69**
	Hi-Mr	0.32	0.64	0.58	0.67	**0.71**	**0.71**	0.67
	Hi-Ne	0.14	0.44	0.46	0.55	0.57	**0.66**	0.65
	Hi-Pa	0.17	0.58	0.35	0.58	0.62	**0.65**	**0.65**
	Hi-Sa	0.13	0.53	0.22	0.24	0.31	**0.64**	**0.64**
	Hi-Ta	0.16	0.61	0.39	0.57	0.60	**0.72**	**0.72**
	Hi-Te	0.34	0.56	0.55	0.57	0.64	**0.68**	0.58
	Es-Pt	-	-	0.77	-	0.81	0.84	**0.86**

Problem Definition - Computational Phylogenetics

We define the task of computational phylogenetics as the devising a method which can estimate the relationships between variant of the same text and generate tree. and cluster them such as the variants which are close to each other, are clustered in the same group (**clade**).

We perform the task of phylogenetics by using distance matrix based approaches

- Unweighted Pair Group Method with Arithmetic mean (UPGMA) and,
- Neighbor Joining.

These methods take as input a distance matrix which can be constructed based on the hypothesized distance among the variants.

We propose a novel approach to construct this distance matrix using word embeddings and then use the said matrix to plot the phylogenetic tree, for the variants in question.

Harnessing Deep Cross-lingual Word Embeddings to Infer Accurate Typological Trees
(CoDS-COMAD 2020)

Establishing language relatedness by inferring phylogenetic trees has been a topic of interest in the area of diachronic linguistics.

We hypothesize inter-language distances using our novel approach.

The inter-language distance is computed by:

- Averaging the synset distances among two different wordnets.
- Synset distance is computed by:
    - Averaging the distances among all the word pairs in a parallel synset.

Approaches

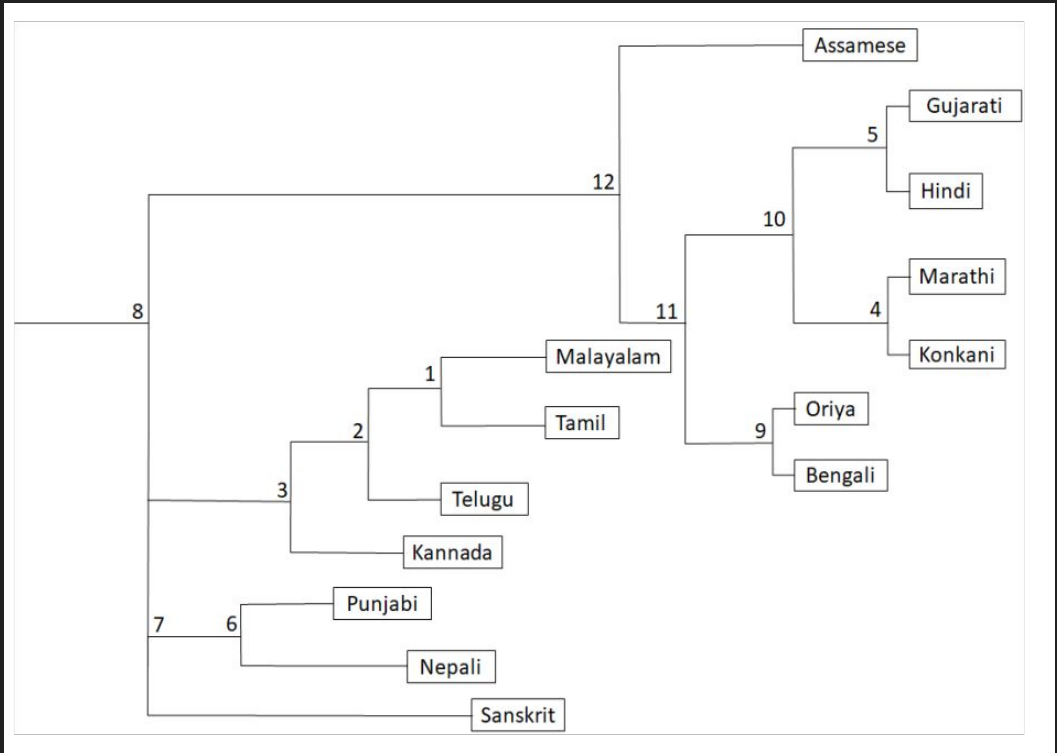
These distances can be computed by both orthographic similarity based approaches, and word embedding based approaches.

We use the orthographic similarity based approach as a baseline.

We also develop a baseline method using lexical similarity-based metrics for comparison and identify that our approach produces better phylogenetic trees which club related languages closer when compared to the baseline approach.

Our novel approach computes the angular cosine distance between all word pairs belonging to the same synset in the common embedding space shared by two languages. Thus, the average over the word-pair distances, and further ‘synset distances’ provides us with a more effective ‘inter-language distance’.

Resultant Tree (Our Approach)



Utilizing Word Embeddings based Features for Phylogenetic Tree Generation of Sanskrit Texts

(ISCLS 2019)

Tracing the root of a text, i.e., the original version of the text, by inferring phylogenetic trees has been a topic of interest in philological studies.

We utilize word embeddings as features to compute the distances among variant manuscripts. We conduct this pilot study on using word embeddings to compute inter-manuscript distances and provide an effective distance matrix to infer phylogenetic trees.

We conduct experiments on the historical Sanskrit text known as Kāśikāvrtti (KV) and infer phylogenetic trees using this approach.

For comparison, we also develop baseline methods using lexical distance-based measures to infer phylogenetic trees for KV.

We show that our methodology produces better trees which club closely related manuscripts together compared to the baseline method.

Dataset Construction

We collect the following data for performing our experiments and tree construction.

KV Dataset

For distance matrix generation, we focus on specific portions of the KV. We collect seventy different versions of the KV on AST 2.2.6. We perform cleaning and manual analysis with the help of philologists. These versions were available in different parts of the country from where we accumulated them in a single repository.

Raw Corpus for obtaining Word embeddings

We obtain raw monolingual Sanskrit corpus from various sources. We download the Sanskrit Wikimedia dump and collate all the articles as a single corpus. We, also, add Glosses and Example sentences from the Sanskrit Wordnet to this corpus. We obtain raw corpus from other sources available online.

We perform cleaning for this corpus by removing any other ASCII characters apart from the Devanagari script. The final cleaned corpus used for creating embeddings contains 5,38,323 lines.

Our Approach for Distance Matrix Construction

We use two approaches for constructing the inter-manuscript distances.

We calculate each inter-variant distance by averaging over ‘Unit Distances’ based on:

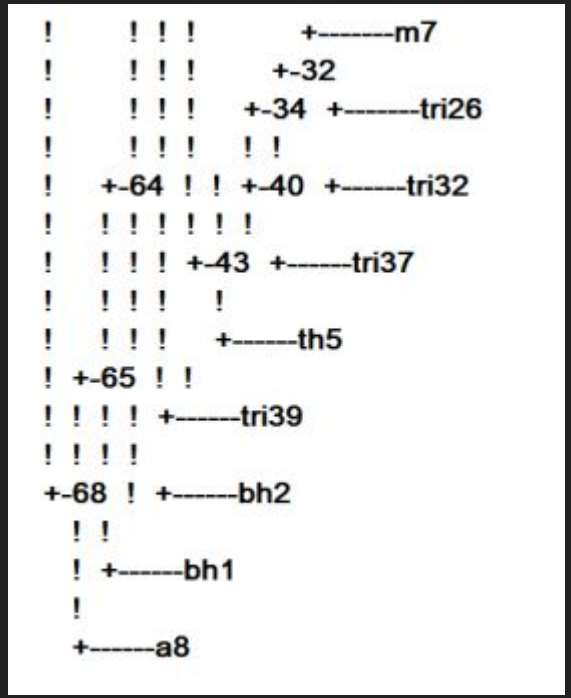
- The baseline approach - utilizes various lexical similarity based measures and later, we also provide weights to them, using empirical approaches, to increase their efficiency.

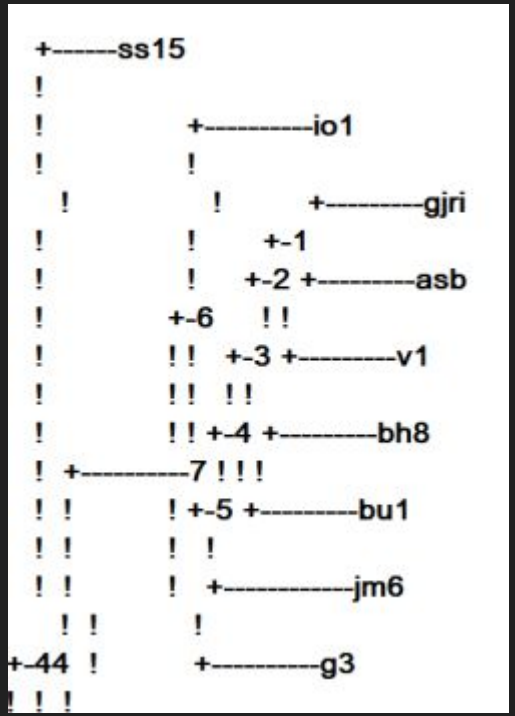
- In our approach, we use word embedding based models and compute distances using vectors obtained from them.

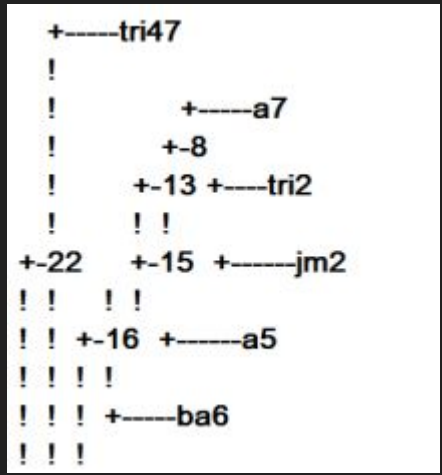
    - Cosine Distance.

    - Angular Cosine Distance (angular cosine distance distinguishes nearly parallel vectors better).

Resultant Clades







Conclusion

We define problem of Cognate Detection, False Friends' Detection, and Computational Phylogenetics.

We use distributional semantics to obtain features and perform the three tasks above using various embedding based methods.

Our novel approaches have shown to perform better than the state-of-the-art with a significant margin for the task of Cognate and False Friends' detection.

We also use word embeddings to compute accurate distance between languages to infer more accurate typological trees.

We apply the same approach to compute distances between variants of a text and generate phylogenetic trees.

Overview

Cognate Detection	False Friends' Detection	Computational Phylogenetics
Utilizing Wordnets for Cognate Detection Among Indian Languages **(GWC 2019)** "Keep Your Dimensions on a Leash" :True Cognate Detection using Siamese Deep Neural Networks **(CoDS-COMAD 2020)** Challenge Datasets of Cognate and False Friend Pairs for Indian Languages **(LREC 2020)**	"I see what you do not mean": Utilizing Deep Cross-Lingual Word Embeddings to Detect False Friends **(ACL 2020; Under Review)**	An Introduction to the Textual History Tool **(ISCLS 2019)** Utilizing Word Embeddings based Features for Phylogenetic Tree Generation of Sanskrit Texts **(ISCLS 2019)** Harnessing Deep Cross-lingual Word Embeddings to Infer Accurate Phylogenetic Trees **(CoDS-COMAD 2020)** Strategies of Effective Digitization of Commentaries and Sub-commentaries: Towards the Construction of Textual History **(SSSU 2020)**

Recommendation Chart of Domains for Cross-Domain Sentiment Analysis: Findings of A 20 Domain Study (LREC 2020)

"A Passage to India": Pre-trained Word Embeddings for Indian Languages (SLTU-CCURL Workshop at LREC 2020)

Future Work

Note to self: “Use the quarantine period to write your thesis. #stayathome”

We shall use cognitive psycholinguistics for the task of cognate detection.

- Using gaze features collected via an eye-tracking machine.

We will inject cognate pairs from our dataset in the Machine translation pipeline and show that they, indeed, help this downstream NLP task. (unpublished)

- Approach 1: Add to parallel corpus
- Approach 2: Use SMT Injection

We will also use cross-lingual word embeddings directly in the translation pipeline which should help the task.

Acknowledgement

Annotators: Dr. Irawati Kulkarni, Dr. Nilesh Joshi and Ms. Lata Popale for digitizing the manuscript data and the cognate dictionary book for us.

Co-authors:

Kevin Patel, Abhijeet Dubey, Aditya Joshi, Akash Sheoran, Sayali Ghodekar, Sai Sravan Munukutla, Yashasvi Mantha, Prof. Eivind Kahrs, Irawati Kulkarni, and Nilesh Joshi.

References

Inkpen, Diana, Oana Frunza, and Grzegorz Kondrak. "Automatic identification of cognates and false friends in French and English." In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, vol. 9, pp. 251-257. 2005.

Mulloni, Andrea, and Viktor Pekar. "Automatic Detection of Orthographics Cues for Cognate Recognition." In *LREC*, pp. 2387-2390. 2006.

Fellbaum, Christiane. "WordNet." In *Theory and applications of ontology: computer applications*, pp. 231-243. Springer, Dordrecht, 2010.

Bhattacharyya, Pushpak. "IndoWordNet." In *In Proc. of LREC-10*. 2010.

Jha, Girish Nath. "The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI)." In *LREC*. 2010.

Diptesh Kanojia

(IIT Bombay, India, and Monash University, Australia)

dipteshkanojia@gmail.com