"Keep Your Dimensions on a Leash" : True Cognate Detection using Siamese Deep Neural Networks

Diptesh Kanojia[†], Sravan Munukutla, Sayali Ghodekar, Pushpak Bhattacharyya, Malhar Kulkarni[*]
Center for Indian Language Technology, Indian Institute of Technology, Bombay

ABSTRACT

Automatic Cognate Detection helps NLP tasks of Machine Translation, Information Retrieval, and Phylogenetics. Cognate words are defined as word pairs across languages which exhibit partial or full lexical similarity and mean the same (*e.g.*, **hund-hound** in German-English). In this paper, we use a Siamese Feed-forward neural network with word-embeddings to detect such word pairs. Our experiments with various embedding dimensions show larger embedding dimensions can only be used for large corpora sizes for this task. 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.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Natural language processing*; • **Information systems** → Information retrieval.

KEYWORDS

Cognate Detection, Indian Languages, Word Embeddings

1 INTRODUCTION

Automatic Cognate Detection helps the task of Machine Translation [1], Information Retrieval [7] and Computational Phylogenetics [9]. Existing approaches to Automatic Cognate Detection consider only the phonetic [6, 8] and orthographic information [4, 5] neglecting the semantic information. In this paper, we propose the use of word embeddings for detecting cognates. Further, we describe the use to linked Wordnets as a dataset for building candidate cognate lists.

We build our word lists using the linked IndoWordnets [2] for ten Indian langauges namely Hindi (Hi), Bengali (Bn), Gujarati (Gu), Marathi (Ma), Punjabi (Pa), Sanskrit (Sa), Malayalam (Ml), Tamil (Ta), Telugu (Te), Nepali (Ne). We compare words among parallel synsets and store the words which exhibit partial or full lexical similarity. Our word pair list sizes range from 656 (Hi-Ta) to 9472 (Hi-Gu). We obtain monolingual corpora from various sources which ranges ~439K lines (Ta) to ~48124K lines (Hi).

2 APPROACHES

In all our approaches, we report the results from performing 5-fold cross-validation on WNData. We consider 70% of the data for training, 20% for testing, and the remaining 10% as validation split. The results are calculated over the test split. In the baseline lexical similarity based approach (LSA), we use a weighted lexical similarity to find out the lexical distance between the *context* of both words (score1) and their respective contexts (*i.e.*, bag of words based score2). From each set of bag of words, we compute similarity scores for *every word from the source side* with *every word on the target side* and average them. **The intuition for harnessing a siamese feed forward** network-based approach is that these networks perform a combined mapping of input vectors into a common target space. These networks find a function such that a simple distance in the target space approximates the "semantic" distance or distance in the meaning, from the input space. In the input layer, we provide the embeddings of a word pair. In the output layer, we use cosine similarity and a sigmoid function to predict the class of the word-pair. The network utilizes cross-entropy loss as its loss function. *An important contribution of our work is that we perform this classification based on various embedding dimensions.* We build embedding models using the sub-word enriched fastText [3] approach. We show reproducible results[1] of our approach in Table 1.

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

Table 1: Results in terms of Precision (P), Recall (R) and F-Score (F) for LSA vs. SFN for various dimension sizes.

3 CONCLUSION AND FUTURE WORK

In this paper, we successfully utilize monolingual word embeddings and outperform approaches based on lexical similarity-based metrics. We experiment with various embedding dimensions and show that larger embedding dimensions can be used only when a large corpus size is available to help reduce the ambiguity among the distributional similarity based sense clusters. We establish a use case for the utilization of word embeddings for the detection of cognates among Indian languages. In future, we would like to utilize cross-lingual word embeddings to project the distribution of senses into a common space to perform the task of cognate detection.

[*]The first three authors have equal contributions. [†]diptesh@cse.iitb.ac.in

[1] http://www.cfilt.iitb.ac.in/cognateSiamese

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371207

REFERENCES

[1] Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A Smith, and David Yarowsky. 1999. Statistical machine translation. In *Final Report, JHU Summer Workshop*, Vol. 30.

[2] Pushpak Bhattacharyya. 2017. IndoWordNet. In *The WordNet in Indian Languages*. Springer, 1–18.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[4] Alina Maria Ciobanu and Liviu P Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 99–105.

[5] Alina Maria Ciobanu and Liviu P Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 431–437.

[6] Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1. 1205–1216.

[7] Helen M Meng, Wai-Kit Lo, Berlin Chen, and Karen Tang. 2001. Generating phonetic cognates to handle named entities in English-Chinese cross-language spoken document retrieval. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.* IEEE, 311–314.

[8] Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 1018–1027.

[9] Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *arXiv preprint arXiv:1804.05416* (2018).