“Keep Your Dimensions on a Leash” True Cognate Detection using Siamese Deep Neural Networks

Sravan Munukutla[1], Sayali Ghodekar[1], Diptesh Kanojia[1,2,3], Pushpak Bhattacharyya[1], Malhar Kulkarni[1]

[1]IIT Bombay, [2]IITB-Monash Research Academy, [3]Monash University
{sravanmunukutla, sayalighodekar26, dipteshkanojia}@gmail.com, pb@cse.iitb.ac.in, malhar@hss.iitb.ac.in

Introduction

- Cognates are words from different languages sharing a common etymological origin and having lexical, phonetic and semantic similarities. For example **university - universität** (English-German), **doctor-docteur** (English-French).
- The study of cognates plays a crucial role in applying comparative approaches for historical linguistics, in particular, solving language relatedness and tracking the interaction and evolvement of multiple languages over time.
- The task of detecting cognates using computational methods and algorithms is known as Automatic Cognate Detection. It helps NLP tasks of Machine Translation (Al-Onaizan et al., 1999), Information Retrieval (Meng et al., 2001) and Phylogenetics (Rama et al., 2018).
- Although the Cognate identification for Latin languages is a well established research problem, the study in Indian languages is a relatively new research domain.

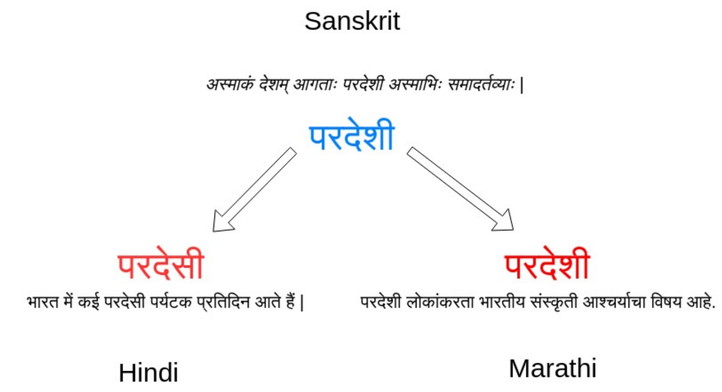

Figure 1: *Hindi Marathi cognate word-pair derived from Sanskrit*

Contributions

- Word embeddings (Mikolov et al., 2013) capture the relevant context of words in a vector space and group the words based on similarities in their context and semantics. Thus we propose a pipeline for a Cognate Identification system using Siamese Feed-Forward networks, utilizing these embeddings.

Background and Related Work

- Existing approaches to Automatic Cognate Detection consider only the phonetic and orthographic information.
- Jäger et al. (2017) use SVM for phonetic alignment and perform cognate detection for various language families.
- Rama (2016) implement a phoneme level Siamese convolutional networks for the task of pair-wise cognate identification. This network learns phoneme level feature representations and language relatedness from raw words for cognate identification.
- Ciobanu and Dinu (2015) perform the task by employing an orthographic alignment inspired by the sequence alignment of computational biology.

Datasets

- **Dataset 1: WordNet Data**
  - To build the candidate true cognate list, we make use of linked IndoWordnets (Bhattacharyya, 2010), which are used to obtain words expressing similar concepts.
  - A pair of words extracted from parallel synset in a linked language pair WordNet, exhibiting high lexical similarity measures can be classified as cognates. Lexical similarity/distance can be expressed by Normalized Edit Distance (Nerbonne and Heeringa, 1997) or Cosine Similarity.
  - We build potential true cognate list for 9 language pairs, namely, Bengali (Bn), Marathi (Mr), Gujarathi (Gu), Punjabi (Pa), Sanskrit (Sa), Malayalam (Ml), Telugu (Te), Tamil (Ta) and Nepali (Ne) with Hindi (Hi) being the source language.
- **Dataset 2: Corpora for Word Embeddings**
  - Word embeddings require a large amount of monolingual corpora for efficient training of a usable model with high accuracy.
  - We obtain monolingual corpora from various sources which ranges 439K lines (Ta) to 48124K lines (Hi).
  - We also analyze the lexical richness of the corpora using metrics like Type Token ratio(TTR) and Moving Average Type-Token Ratio (MTTR) (Covington and McFall, 2010).
  - We use this corpora to create monolingual word embeddings of 200, 300 and 400 dimensions using the fastText (Bojanowski et al., 2017) library.

Siamese Feed Forward Network

- Siamese Neural Networks were introduced to solve the problem of signature verification as an image matching problem (Bromley et al., 1994)
- Siamese nets are two identical networks that accept distinct inputs but are joined in by a function that calculates a distance metric between the outputs of the two nets.
- 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.

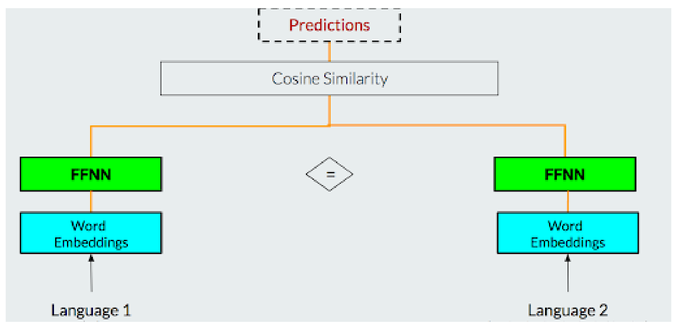

Figure 2: *Siamese Feed Forward Neural Network*

Approach

- **Lexical Similarity Based Approach : Baseline**
  - This approach makes use of a **Weighted Lexical measure (WLS)** of **Normalized Edit Distance (NED)** and **Cosine Similarity(CoS)**. The weighted lexical score is given by,

$$WLS = (NED * 0.75) + (CoS * 0.25)$$

  - We calculate the Weighted lexical score of each of the words and averages the score over length of the gloss. The weighted score is calculated over words(score1) and the gloss(score2) and finally score1 and score2 are averaged to obtain a final WLS score.
- **Siamese Feedforward Approach**
  - For the Siamese Feed Forward Neural Network, we use fastText word embeddings of 200, 300 400 dimensions.
  - We provide word embeddings as an input to the Siamese feed-forward layer and compute the semantic distance using cosine similarity between two comparable output features. This distance determines the class (cognate/non-cognate) of the word pairs.

Results

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.*

- As we can see from Table 1, the monolingual embeddings approach significantly outperforms the lexical similarity(baseline) approach for all 9 language pairs.
- For languages sharing lexical similarity with Hindi eg: Bengali (Bn), Marathi (Mr), Punjabi (Pa), Gujarati (Gu) and Sanskrit (Sa), the F-score increases with and increase in word embedding dimensionality from 200 to 300.
- For Dravidian languages eg: Malayalam (Ml), Tamil (Ta), Telugu (Te), the highest F-score is for 200 dimension embeddings, however it doesn't improve over 300 dimension embeddings. This observation can be accounted due to the fact that Hindi and Dravidian languages have relatively less cognate pairs.
- Thus the 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.

Conclusion and Future Work

- Monolingual word embeddings outperform approaches based on lexical similarity-based metrics.
- 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.

References

Al-Onaizan, Y., Curín, J., Jahr, M., Knight, K., Lafferty, J. D., Melamed, D. M., Och, F. J., Purdy, D. S., Smith, N. E., and Yarowsky, D. (1999). Statistical machine translation: Final report.

Bhattacharyya, P. (2010). Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems 6*, pages 737–744. Morgan-Kaufmann.

Ciobanu, A. M. and Dinu, L. P. (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)*, pages 431–437, Beijing, China. Association for Computational Linguistics.

Covington, M. A. and McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

Jäger, G., List, J.-M., and Sofroniev, P. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multilingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain. Association for Computational Linguistics.

Meng, H. M., Wai-Kit Lo, Berlin Chen, and Tang, K. (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.*, pages 311–314.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nerbonne, J. and Heeringa, W. (1997). Measuring dialect distance phonetically. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.

Rama, T. (2016). Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027, Osaka, Japan. The COLING 2016 Organizing Committee.

Rama, T., List, J.-M., Wahle, J., and Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?

Presented at the ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD) 2020 (7th ACM IKDD CoDS and 25th COMAD), Hyderabad, India, January 5-7, 2020.