IITB-Monash Research Academy

Indian Institute of Technology Bombay | MONASH University

An Indian-Australian research partnership

Harnessing *Cross-lingual Features* to **Improve Cognate Detection** for Low-resource Languages

Diptesh Kanojia, Raj Dabre, Shubham Dewangan, Pushpak Bhattacharyya, Gholamreza Haffari, & Malhar Kulkarni

Questions?

diptesh@cse.iitb.ac.in
dipteshkanojia@gmail.com

Video presentation for the 28th International Conference on Computational Linguistics (COLING 2020) to be held from December 8th to December 11th, 2020.

Automatic Detection of Cognates

- Cognates: Words in different languages with common roots
  - Liberté - Liberty (Fr-En), Night - Nuit (En-Fr), जीवन (jeevana) - জীবন (Jībana) [meaning life], *etc.*
  - The notions of Orthographic Similarity, Phonetic Similarity, and Semantic Similarity.
  - Help NLP tasks- Machine Translation (Kondrak et. al., 2005, 2003), Cross-lingual Information Retrieval (Makin et. al., 2008; Meng et. al., 2001), Cross-lingual Question Answering, *etc.*
- Classification or Clustering based approaches for cognate detection
  - We use the binary classification-based approach.
  - Features obtained from orthographic similarity, phonetic vectors, cross-lingual embedding models.
- For low-resource Indian languages
  - Same language family for most of them (also same linguistic area).
    - The 'Sanskrit Connection'!
  - Focus on resource constrained NLP tasks.
    - Pre-trained models on monolingual corpora to the rescue.

Previous Work

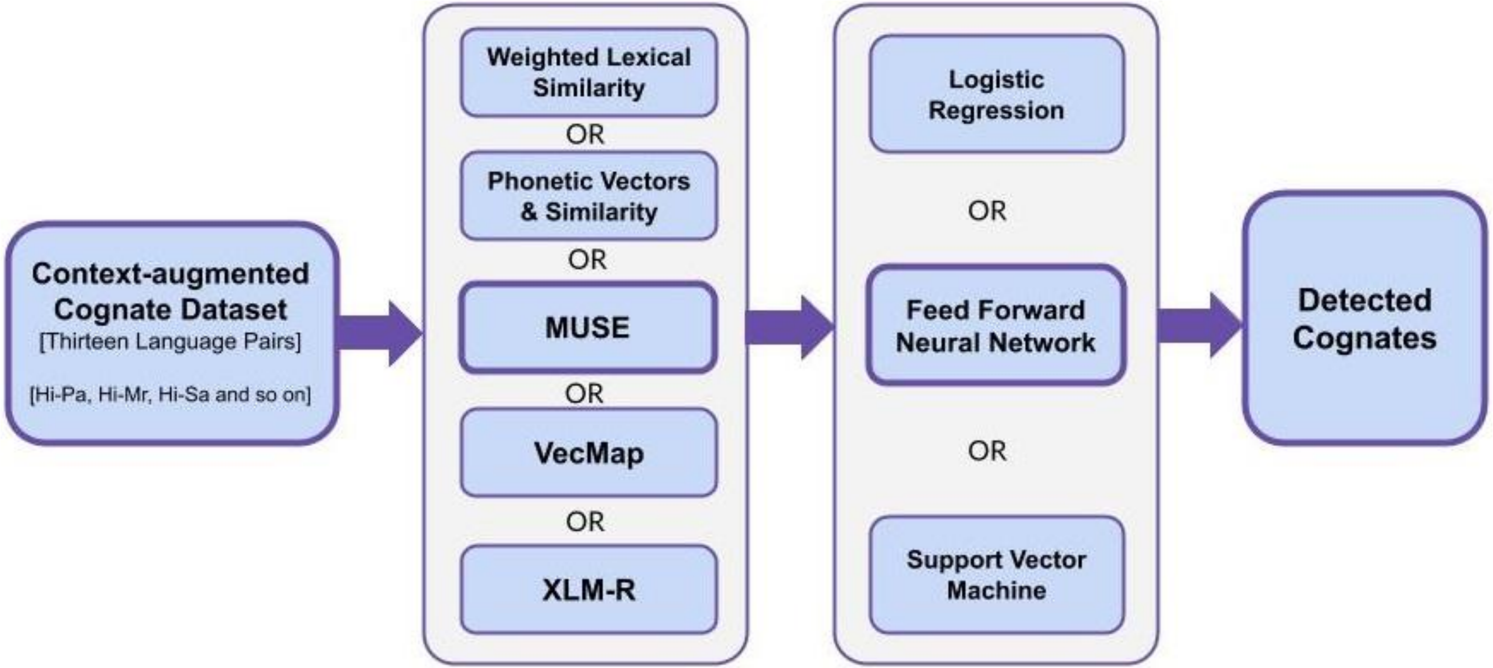
- Computation of a similarity score between potential candidate pairs.
- Orthographic similarity (Jager et al., 2017; Melamed, 1999; Mulloni and Pekar, 2006).
- Phonetic similarity (Rama, 2016; List, 2012; Kondrak, 2000).
- Distance measure with the scores learned from an existing parallel set (Mann and Yarowsky, 2001; Tiedemann, 1999).
- Rama (2016) employ a Siamese convolutional neural network.
  - Phonetic features jointly with language relatedness for cognate identification.
- Jager et al. (2017) use SVM for phonetic alignment and perform cognate detection for various language families.

Key Question & Contributions

"Can semantic information be leveraged from Cross-lingual models to improve cognate detection amongst low-resource languages?"

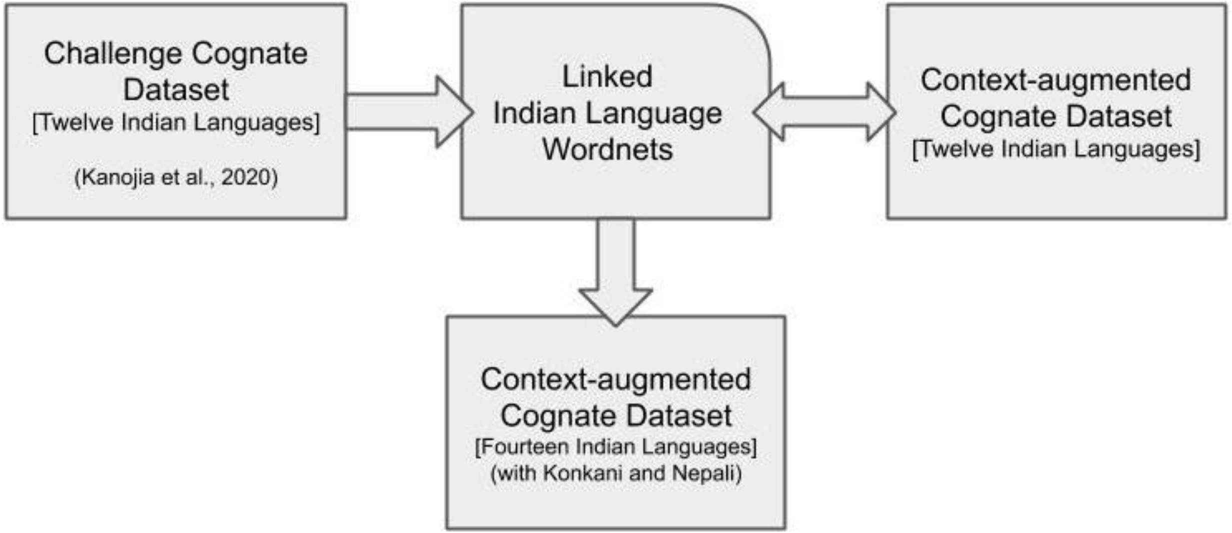
- Utilizing cross-lingual features for the automatic cognate detection task.
- Improvements shown using the cross-lingual features for all the language pairs.
- Improvements shown over baseline Neural Machine Translation (NMT-BPE) system by induction of detected cognates.

Our Idea: Cross-lingual Features For Cognate Detection



Dataset and Pre-processing

- Challenge Cognate Dataset by Kanojia et. al., 2020.
    - We add two new languages, Konkani and Nepali to this dataset.
- Indian languages are written in various scripts.
    - Preprocessing step: Unicode-offset based Transliteration



Results

LP	Baseline Approaches									Cross-lingual Embeddings based Approaches											
	WLS w/ FFNN			PVS w/ Siamese CNN (Rama, 2016)			WLS w/ RNN (Kanojia et al., 2019)			XLM-R w/ FFNN			MUSE w/ FFNN			VecMap w/ FFNN			MUSE + WLS w/ FFNN		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Hi-Bn	0.51	0.28	0.36	0.68	0.62	0.65	0.67	0.69	0.68	0.81	0.76	**0.78**	0.77	0.75	0.76	0.72	0.74	0.73	0.80	0.75	0.77
Hi-As	0.48	0.26	0.34	0.72	0.71	0.71	0.72	0.70	0.71	0.70	0.72	0.71	0.80	0.75	**0.77**	0.74	0.73	0.73	0.84	0.75	**0.79**
Hi-Or	0.51	0.30	0.38	0.65	0.58	0.61	0.66	0.58	0.62	0.65	0.61	0.63	0.72	0.68	**0.70**	0.67	0.70	0.68	0.81	0.69	**0.75**
Hi-Gu	0.43	0.16	0.23	0.70	0.65	0.67	0.81	0.71	0.76	0.80	0.73	0.76	0.80	0.84	**0.82**	0.77	0.74	0.75	0.83	0.85	**0.84**
Hi-Ne	0.50	0.16	0.24	0.72	0.84	0.78	0.78	0.73	0.75	0.75	0.75	0.75	0.86	0.83	**0.84**	0.78	0.73	0.75	0.86	0.83	**0.84**
Hi-Mr	0.51	0.20	0.29	0.70	0.68	0.69	0.74	0.70	0.72	0.76	0.71	**0.73**	0.70	0.73	0.71	0.71	0.71	0.71	0.72	0.73	0.72

- Use of **cross-lingual features improves task performance**,
  - Contextual embeddings (XLM-R) are not always the best except for two language pairs (Hi - Bn and Hi-Mr).
- A **combination of MUSE + WLS features** outperforms all other feature combinations.
- Best F-scores obtained by Hi-Gu and Hi-Ne language (very linguistically close wit high cognate sharing)
- *Kindly refer to paper for scores for all languages and detailed analyses*

Improving Downstream Task (Neural MT)

- Seven language pairs (Hi-Pa, Hi-Bn, Hi-Gu, Hi-Mr, Hi-Ta, Hi-Te, & Hi-Ml)
- 50k parallel sentence from the ILCI parallel corpus.
  - 46277 Sentences for *Training*, 2000 Sentences for *Test*, & 500 Sentences for *Development*.
  - Injected detected cognate pairs into corpus as training sentence (word) pairs.
- RNN-NMT model with sub-words (Bahdanau et al., 2014 + Sennrich et al., 2015)
  - 2,500 BPE Merge Operations (optimal for low-resource set; empirically determined)
  - Hidden size of the model was 500 units
  - SGD optimizer to train for 150,000 steps of 1024 sentence pair batches (8000 warm-up steps)

Approaches / LP	Hi-Pa	Hi-Bn	Hi-Gu	Hi-Mr	Hi-Ta	Hi-Te	Hi-Ml
NMT-BPE Baseline	62.79	28.75	52.17	31.66	13.78	19.18	10.4
Cognate-aware NMT-BPE	**65.55**	**29.43**	**52.39**	**32.41**	**13.85**	**19.58**	**11.18**

Discussion

- Consistent improvements over the strongest baseline (Kanojia et. al., 2019b)
    - 9% points (highest being 18% points for the Hi-Ta language pair)
- Improvements observed in the MT systems
    - 2.76 BLEU points for the Hi-Pa language pair (with 15001 cognate pairs)
    - With the lowest number of cognate pairs*, i.e.,* 930, an improvement of 0.4 BLEU score is observed.
    - Maximum number of cognates induced for Hi-Mr language pair (15834), but only slight improvement observed*, i.e.,* 0.75 BLEU points.
    - Probable reason: Better sub-word segmentation as BPE segmentation is cross-lingually consistent
- Examples of detected cognate pairs (undetected via previous approaches)
    - धकेलना - ધકેલવું (dhakelna-dhakelavun) (Hi-Gu) [both meaning “to push”]
    - जब्त - ਕੁਰਕੀ (jabta-kurki) (Hi-Pa) [both meaning “seizure”]
    - कटुक - കടുപ്പ് (katuk-kaduppa) (Hi-Ml) [both meaning “bitter”]

Conclusions & Future Work

- Harnessed cross-lingual embeddings to improve cognate detection- thirteen Indian language pairs.
- Used a linked knowledge graph to augment a publicly released cognate dataset.
- Significant improvements in cognate detection quality (up to 18%).
- Cognate-aware NMT-BPE results also show a consistent improvement in translation quality.
- Future work
    - Further investigation to improve the performance of contextual embeddings for this task.
    - Adding more sources for potential cognates and improving the challenge dataset.
    - Experiments within Indo-European language family to seek improvements.

Thank you! :)

Kindly reach out to us if you have queries!

diptesh@cse.iitb.ac.in

References

- Grzegorz Kondrak. 2005. Cognates and word alignment in bitexts. Proceedings of the tenth machine translation summit (mt summit x), pages 305–312.
- Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers, pages 46–48.
- Ranbeer Makin, Nikita Pandey, Prasad Pingali, and Vasudeva Varma. 2008. Experiments in cross-lingual IR among Indian languages. Advances in Multilingual and Multimodal Information Retrieval. Springer Berlin/Heidelberg.
- 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., pages 311–314. IEEE.
- Gerhard Jager, 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, volume 1, pages 1205–1216.
- Girish Nath Jha. 2010. The TDIL program and the Indian langauge corpora intitiative (ILCI). In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Language Resources Association (ELRA).
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. ICLR 2015.
- I Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. Computational Linguistics, 25(1):107–130.
- Andrea Mulloni and Viktor Pekar. 2006. Automatic detection of orthographics cues for cognate recognition. In LREC, pages 2387–2390.
- Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1018–1027.
- Johann-Mattis List. 2012. Lexstat: Automatic detection of cognates in multilingual wordlists. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 117–125. Association for Computational Linguistics.
- Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pages 288–295. Association for Computational Linguistics.

References (contd.)

- Gideon S Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, pages 1–8. Association for Computational Linguistics.
- Jorg Tiedemann. 1999. Automatic construction of weighted string similarity measures. In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.
- Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2020. Challenge dataset of cognates and false friend pairs from indian languages. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 3096–3102.
- Diptesh Kanojia, Kevin Patel, Pushpak Bhattacharyya, Malhar Kulkarni, and Gholemreza Haffari. 2019b. Utilizing wordnets for cognate detection among indian languages. In Global Wordnet Conference (2019).
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.
- Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. ¨ Neural Comput., 9(8):1735–1780, November.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.