A Study of Efficacy of Cross-lingual Word Embeddings for Indian Languages

Jyotsana Khatri Rudra Murthy Pushpak Bhattacharyya

Indian Institute Of Technology Bombay, Mumbai, India
{jyotsanak, rudra, pb}@cse.iitb.ac.in

Overview

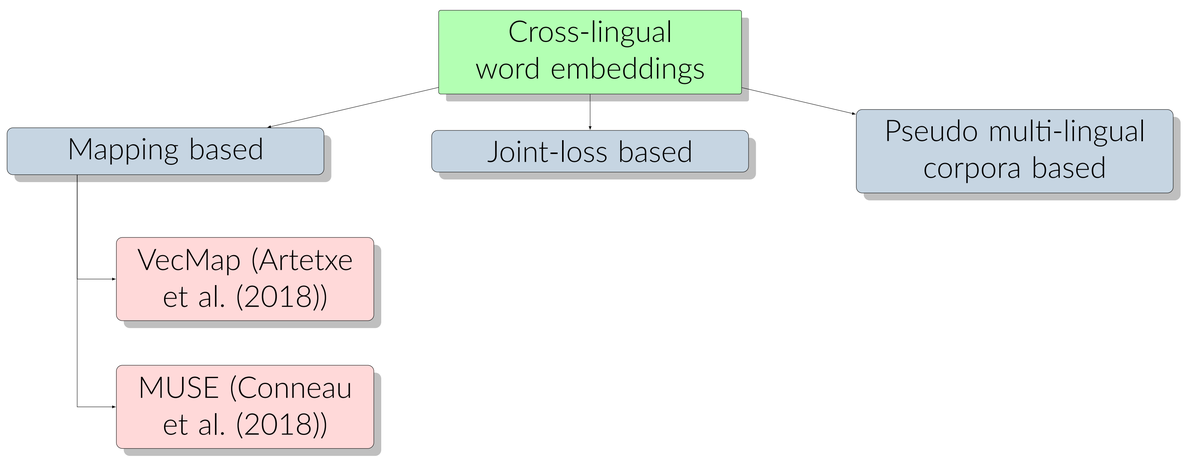
Motivation

- Cross-lingual word embeddings are useful in multi-lingual NLP
- The quality of cross-lingual word embeddings for Indian languages is not explored

Contribution

- Dataset creation for evaluation of cross-lingual word embeddings for 10 Indian language pairs
- Study of mapping based approaches to train cross-lingual word embeddings for Indian languages
- Demonstrate low accuracy of mapping based approaches even for cognate words
- Show performance improvement by training cross-lingual word embeddings using fast-text on combined transliterated corpora of 11 Indian languages

Approaches



VecMap

- Assumes both spaces are isometric
- Learns linear transformations from both spaces to a shared space
- Maps monolingual word embedding spaces of both source and target languages to a shared space using iterative Procrustes refinement with a dictionary initialized in an unsupervised way

MUSE

- Assumes both spaces are isometric
- Learns linear transformation from source to target space
- Maps source monolingual embedding space to target using domain adversarial training and update these using Procrustes refinement

Indic

- **Transliterate** data of all languages in Devnagari Script using IndicNLP (Kunchukuttan et al. (2015)).
- Train model using any algorithm which is used to train monolingual word embeddings on the combined data

Dataset

- We create test dictionaries using **IndoWordnet (Bhattacharyya (2017))**.
- **The words which have at least one common sense are considered as translations.**
- The dataset consists of **test dictionaries** containing words and their multiple translations. We select randomly 10000 words from this large set as test data.

Cognates: Word pairs across languages which have full or partial lexical similarity and have at least one common sense are defined as cognates (true and partial). We select word pairs which have edit distance more than .8 from already selected test data to generate test data of cognate words.

Experimental Details

Language pairs	Hindi-{Assamese, Bengali, Gujarati, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu}
Dataset	Test Dictionaries created using IndoWordNet containing 10000 source words and their translations
Mapping based Approaches	VecMap (Artetxe et al. (2018)) MUSE (Conneau et al. (2018))
Monolingual embeddings	Embeddings provided by Bojanowski et al. (2017) trained using fast-text on wikipedia data, the words in the test dictionary are also included in the monolingual embeddings
Transliteration to Devnagari Script	IndicNLP (Kunchukuttan et al. (2015))
Evaluation	Top-5 accuracy on Bilingual Lexicon Induction
Nearest neighbors	CSLS (Cross-Domain Similarity Local Scaling)

Results

Lang-pair	All words			Cognates (True and Partial)				Word overlap
	VecMap	MUSE	Indic	# cognates	VecMap	MUSE	Indic	
Hi-As	0.02	1.53	**18.92**	946	0.00	3.81	**94.40**	9.46
Hi-Bn	16.83	5.16	**24.32**	1727	32.48	10.89	**90.45**	17.27
Hi-Gu	17.01	6.02	**44.44**	4188	26.81	9.72	**87.11**	41.88
Hi-Ka	14.15	6.18	**16.31**	891	50.17	23.34	**94.05**	8.91
Hi-Ml	0.01	2.67	**7.83**	125	0.00	8.00	**88.80**	1.25
Hi-Mr	19.16	6.88	**29.52**	2325	40.77	14.97	**91.70**	23.25
Hi-Ne	20.67	4.92	**38.02**	3436	39.93	8.93	**91.21**	34.36
Hi-Pa	15.17	7.25	**18.18**	1053	48.24	27.92	**92.12**	10.53
Hi-Ta	**10.02**	4.42	1.58	33	9.09	9.09	**87.88**	0.33
Hi-Te	**8.68**	4.12	7.52	175	25.14	12.0	**96.00**	1.75

Table 1:Top-5 accuracy on Bilingual Lexicon Induction for cross-lingual word embeddings generated using different algorithms for Hindi-X language pairs

Examples

Source Word	True Translation	VecMap		MUSE		Indic	
दीपमाला	दीपमाळ पंचारती दीपमाला	औक्षण वृंदावनी चंद्रज्योती अक्षदा गौराई	होळीत पंचारती महाआरती निरांजने ओवाळी	मगवा अशरा मूठा ठकणे सूचणी	बीईदा तुपग छिणी सरकफास बहकणे	दीपा उझमाला दीपिका दीपेश दीपकळी	दीपशिखा पुष्पमाला दीपिकाला दीपमाला
मिरची	मिरची मिर्ची	मांदेली पिकतो जलेबी सब्जी लिची	मकई सिताफळ दोडका दोडका जामून	गरभरू भुईघर छिणी मंथणे बनजर	तुपगट तिवई बहकणे बीईदा सरकफास	मिरज मिरवीत मिठाची मिरवत मिरवणे	मिरचीची मिर्ची मिरची
पेशकश	प्रस्ताव नजराणा भेट उपहार	मुभाही ग्वाहीही मोफ़त सवलत परवानगी	परवानगी सज्जता घोषणाही ऑफर ऑफर	अनृण अदेय ग्वाहीही देणाऱ्यांची मदतीची	समारंभाची सल्लागाराची देण्यासाठीची सदस्यता घोषणा	पेशंटवर पेशे पेश पेशल पेशंट	पेशू पेशकश
व्याख्याता	लेक्चरर रीडर व्याख्याता	योगशिक्षक अधिव्याख्याता ग्रंथपाल व्याख्याता व्याख्याते	अध्यापक प्राध्यापक	जडाऊ भुईघर तिवई सूचणी भंगट	बहकणे अरस बीईदा गरभरू सरकफास	व्याख्याच व्याख्या व्याख्याने व्याख्यान व्याख्याते	व्याख्यात व्याख्याता

Table 2:Sample outputs generated by Hindi-Marathi cross-lingual word embeddings

These are top nearest neighbors using the cross-lingual word embeddings generated by corresponding algorithms.

Conclusion

- A study of cross-lingual word embeddings for Indian languages.
- Simply training the embeddings using a combined corpus gives good accuracy on bilingual lexicon induction if the word overlap is high.
- Existing mapping based approaches gives low accuracy on bilingual lexicon induction task even for cognate (true and partial) words, if the word overlap is high

Future Work

- Based on insights gained, we would like to propose methods which learns to map both cognate and non-cognate words across languages
- Evaluate cross-lingual word embeddings on downstream NLP tasks

References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789--798.

Bhattacharyya, P. (2017). Indowordnet. In *The WordNet in Indian Languages*, pages 1--18. Springer.

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.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *In Proceedings of ICLR 2018*.

Kunchukuttan, A., Puduppully, R., and Bhattacharyya, P. (2015). Brahmi-net: A transliteration and script conversion system for languages of the indian subcontinent. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 81--85.