Capturing Semantic Similarities from Bangla and English Word Embeddings Trained on Monolingual Corpora

Samya Sunibir Das
Student, North South University
samya.das@northsouth.edu

Dr. Mohammad Ashrafuzzaman Khan
Assistant Professor, North South University
mohammad.khan02@northsouth.edu

Abstract

In this paper, we explore the concept of capturing semantic similarities from word vectors obtained from monolingual corpora using a three-step method. Instead of training Bangla and English word embeddings in a single vector space in a Cross-lingual manner, we separately train Bangla word embeddings using the largest Bangla newspaper dataset and utilize the pre-trained Word2vec model trained on Google News Corpus for English word embeddings. Our three-step approach demonstrated we could extract the deeper semantic relationship between word pairs by comparing them directly and exploring their neighbouring words.

1 Introduction

Despite three decades of research, and Bangla being one of the most spoken languages globally, it is considered a low-resource language in the NLP (Alam et al., 2021) domain. However, ongoing works are being done to improve Bangla NLP by leveraging large size of corpus. The researchers that have developed BanglaBERT, (Bhattacharjee et al., 2022) managed to crawl and use 27GB of Bangla text data for their work. Such volume of Bangla text data is scarce to find and conduct research on. In this paper, we used the largest available Bangla newspaper dataset (Biswas, 2022), partially to train Bangla word embeddings.

Word embeddings, which are real-valued vectors for words, can capture their semantic and syntactic properties. Word embeddings can be learned from large-scale monolingual corpora using various algorithms, such as skip-gram or continuous bag-of-words (CBOW).

A vast amount of research work has been conducted thus far utilizing multilingual and cross-lingual vector spaces using different methodologies.(Ammar et al., 2016) However, exploring the deeper semantic relationships between languages where the word embeddings are trained independently on monolingual vector spaces is still a research area where much work is left to be desired.

Cross-lingual word embeddings are effective for low-resource language to high-resource language machine translation tasks because they can transfer knowledge from high-resource to low-resource languages. This is done by learning a common representation for words in different languages so that words with similar meanings are close to each other in the vector space; this is considered applicable for tasks such as machine translation and cross-lingual information retrieval. (Adams et al., 2017; Berlot and Kaplan, 2020)

Goal of our Paper

Building on the concept of Cross-Lingual Word Embeddings of being able to transfer learning across languages by capturing semantic similarities(Espinosa-Anke et al., 2021), in our paper, but instead of training both Bangla and English word embeddings in a single vector space, we train the Bangla portion separately from the sourced dataset to build the Word2Vec model and use the model from (Mikolov et al., 2013a) that was trained on the Google News corpus. We show despite them being independently trained on monolingual corpora, the word embeddings demonstrate deeper and meaningful semantic relationships among them, which are evaluated by our three-step methods.

2 Related Work

Previous works conducted with Bangla embeddings contributed to the understanding of Bangla embeddings, providing insights into their performance and intrinsic evaluation techniques performance of different word embedding models. One study (Sultana Ritu et al., 2018) evaluated the models in terms of accuracy and efficiency using a shared Bangla dataset, and highlighted the advan-

tages of dynamic word clustering models over N-gram models in terms of processing time and memory efficiency. Another work (Sadman et al., 2019) focused on intrinsic evaluation of Bangla word embeddings created with CBOW and Skipgram models. These evaluation tasks included: analogy prediction, semantic relatedness, synonym and antonym detection, and concept categorization.

One research conducted by (Pandit et al., 2019) explored semantic similarity between words in Bangla by applying path-based and Word2Vec methods. They achieved good results and cosine similarity scores between words by incorporating cross-lingual resources from English.

There was another work conducted by (Periñán-Pascual, 2021)proposed a model to estimate the strength of associative words that can be semantically related or not. The model combined corpus- and network-based word embeddings and used the weighted average of the cosine-similarity coefficients to measure word associations.

3 Data

3.1 Bangla

The dataset used in this study for handling the Bangla word embeddings is sourced from the largest available Bangla Newspaper Dataset - (Biswas, 2022). It contains 2.2M news articles from three of the most reliable/popular newspapers that are available online. The size of the entire dataset is 22.81GB. We were able to utilize roughly 12GB of the entire data and had utilized 291446363 (291 Million) words.

3.2 English

On the other hand, the English word vectors were obtained from the word2vec model that was trained on the Google News Corpus (Mikolov et al., 2013a), where 100 billion words of 300 dimensions were utilized, with 3 million vectors. [1]

4 Methodology

4.1 Data Preprocessing

In the data preprocessing phase, the Bangla text underwent several steps to prepare it for training. Stopwords, punctuation marks, non-Bangla words, and special characters were removed to eliminate noise. The dataset was structured in JSON format with the genre types of news, headlines, body text, tags, and additional information. All the texts were combined to utilize for training. Finally, the text was tokenized.

4.2 Building the Word2vec model

After pre-processing, the task was to build a Word2vec model. From (Mikolov et al., 2013b), Skip-gram works well with small amount of training data and is found to represent rare words well. On the other hand, CBOW is faster and has better representations for more frequent words. Initially, we went with the CBOW architecture, as it is set by default on gensim and as we are using a newspaper dataset where many words would be of high frequency and be repeating. [2] The word dimensions were kept to 300, the same as the Word2vec model that was trained on the Google News dataset.

4.3 Approaches Taken

In our complete analyses, we chose to sample the word embeddings with 505 Bangla words. We adopted three different strategies to conduct our analyses to capture semantic similarities by calculating the cosine similarity [3] scores between the word embeddings.

Method 1: We took a Bangla word, took its direct English translation, then compared the cosine similarity score of the word embeddings between them. In this same method, our approach was to study the semantic similarities between the Bangla word and the corresponding English words synonyms and antonyms as well. For analyzing the synonyms, we took the nearest neighbouring word (most similar words) [4] of the corresponding English word from the googlenews Word2vec model. Our target was to find positive scores of cosine similarity between a Bangla word and its English translation, positive scores between Bangla word and the nearest neighbouring English word, and negative scores between Bangla word and the antonym of the English word.

Method 2: In this method, we took a Bangla word, similarly as before, took its direct English translation, then compared the cosine similarity scores between those, then found out the first 10 nearest

[1] Gensim Data

[2] CBOW Gensim

[3] Cosine Similarity Wiki.

[4] Word2vec most similar words

neighbouring words of the English word, and compared those with the Bangla word. Our target was to find a positive match between the Bangla word and the corresponding English translation and the first 10 neighbouring words. This approach was taken because we recognized that many words can have multiple meanings, and we aimed to identify the exact word that closely resembles our target word.

Method 3: In our final method, we first extracted the first 2 neighbouring Bangla words from our model for Bangla word embeddings. Then, we compared the similarity scores for each of those 2 neighbouring Bangla words with the first 10 neighbouring English words.

5 Results and Analysis

Method 1: After following this method, we found 288 True Positives (Positive score between Bangla - English Translations and Nearest Neighbouring words both), 217 False Negatives (Negative score between Bangla - English Translations and Nearest Neighbouring words both), 238 True Negatives (Negative score between Bangla - English Antonyms), 267 False Positives (Positive score between Bangla - English Antonyms). Overall, it yields a 52% accuracy based on this approach.

Method 2: After conducting this strategy, we found 479 words out of the 505 sampled Bangla words do, in fact, have a positive matching score either between the direct translated English word or any of its 10 neighbouring words. Altogether, we had conducted 5050 experiments for all the words. Bangla words that do not have any positive scores between their English counterpart or its 10 neighbours include: খোঁজা,সাহায্য,দুঃখ,ঘৃণা,লুকানো,স্নেহ,অবিবাহিত,উপকার, তিক্ততা,হত্যা,নেয়া.

. Table 2 exhibits that taking the Bangla words that did fail to exhibit any meaningful semantic relation in Table 1 and the first method, the same words' English counterparts' neighbours can exhibit a semantic relationship with the target Bangla word.

Method 3: Based on this method, we observed that 497 of the 505 original Bangla words have at least one positive match when comparing their neighbours with the neighbours of the corresponding English words. The Bangla words that did not have any positive scores were: দুঃখ, গলানো, প্রতিশ্রুতি, কেন্দ্রীয়, পক্ষ, ছয়, সংস্থা, রক্ষা. Overall, 10100 experiments were conducted and we observed 5247 positive scores and 4853 negative scores.. In the previous method, we observed that we achieved the highest positive results when matching Bangla words with their English translations and their respective neighbours. In this approach, we noticed that the maximum positive results were obtained when comparing neighbours of the Bangla words with the neighbours of the English words. This observation suggests that these two models have significant semantic similarity.

Edge Cases: In exploring the word embeddings, we found out expected positive scores implying semantic similarity to be appearing as negative. Several factors could contribute to this observation; firstly, the nuances in translation - direct translation between languages may not fully encapsulate the original words' subtle nuances, idiomatic expressions, and cultural connotations. Even closely related translations might not align perfectly in meaning, leading to divergent vector representations and resulting in negative cosine similarity scores. We observed many noisy data while generating neighbours from the googlenews Word2Vec model, such as underscored words, misspelled English words or even website links. Another reason could be inconsistency in data pre-processing, we have observed while testing the word frequencies, most of the words and phrases were scaled to about 3 million in the googlenews Word2Vec model, such as the word 'time' appears 2994536 times in their model. While the word 'সময়' appears in our model 35284 times.

6 Conclusion Future Work

In conclusion, this paper demonstrated the effectiveness of separately training Bangla and English word embeddings using Word2vec models; while the embeddings are trained on monolingual corpora, they exhibited meaningful semantic relationships, as evidenced by exploring the scores of the words as well as their neighbours. As this work is still in progress, we aim to analyze using other different news data like CNN news corpus as well in the future for better data and we aim to improve our methodologies to find even more efficient approaches. We also aim to use other multi-lingual or cross-lingual approaches to better enrich our research in the future.

Bangla Word	English	Score	Nearest Neighbour	Score	Antonym	Score
ঝঞ্ঝা	tempest	0.04818	firestorm	0.019006	breeze	-0.05042
পানি	water	0.00867	sewage	0.02129	solid	-0.00593
সময়	time	0.05348	day	0.06087	untimely	0.01169
সুযোগ	opportunity	0.05409	chance	-0.00039	difficulty	-0.03196
সক্ষম	able	-0.07791	trying	-0.02526	unable	-0.07486
বারবার	repeatedly	-0.01175	frequently	0.01268	once	-0.05207
গলানো	melt	-0.00620	melting	0.01809	harden	-0.01945
দুঃখ	sadness	-0.06699	sorrow	-0.02437	happiness	0.05859
বসা	sit	-0.06789	sitting	-0.06622	stand	-0.01909
মুক্তি	freedom	-0.02311	freedoms	-0.02171	captivity	0.03259

Table 1: The table for the first method, comparing the similarity scores between the target Bangla word and it's English translation, the target Bangla word and it's English translation's nearest neighbour, and the target Bangla word and it's English translation's antonym.

Bangla Word	Best Match with English Neighbour	Score
সক্ষম	can	0.02198
বারবার	frequently	0.01268
গলানো	melting	0.01809
দুঃখ	sorrow	-0.02437
বসা	sit	0.01630
মুক্তি	democracy	0.02859

Table 2: The table for the second method, best match English neighbour is shown from the 10 that were compared with the target Bangla word. Words shown that were fail cases in Table 1, but we extracted deeper semantic relation by exploring the neighbours.

Bangla Word	Bangla Neighbour	English Word	English Neighbour	Score
ঝঞ্ঝা	ঝড়ঝঞ্ঝা	tempest	firestorm	0.07236
সময়	সময়ে	time	periods	0.08760
পানি	পানিও	water	groundwater	0.03213
দুঃখ	দুঃখও	sadness	grief	-0.0986
মুক্তি	মুক্তিও	freedom	liberties	0.01102
বসা	বসেছিলেন	sit	sitting	0.01500

Table 3: Table for the third method, exploring deeper to find semantic relations between Bangla neighbours with their corresponding English words' neighbours.

Limitations

Our methodology includes rigorous pseudo-brute forcing, so this is not exactly efficient, we may need to look for even better or more effective methods. We also faced memory-related issues while training our Bangla data.

Acknowledgements

We would like to thank the dataset creator for enriching Bangla NLP domain by making such large corpus of Bangla news data public and to conduct research on. We would also like to thank the exciting and ever-growing Bangla NLP community for creating an exciting environment to conduct research on Bangla language.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain. Association for Computational Linguistics.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings.

Marco Berlot and Evan Kaplan. 2020. Machine translation with cross-lingual word embeddings.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla.

Enam Biswas. 2022. Bangla newspaper dataset - ebd.

Luis Espinosa-Anke, Geraint Palmer, Padraig Corcoran, Maxim Filimonov, Irena Spasić, and Dawn Knight. 2021. English–welsh cross-lingual embeddings. *Applied Sciences*, 11(14).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality.

Rajat Pandit, Saptarshi Sengupta, Sudip Naskar, Niladri Dash, and Mohini Sardar. 2019. Improving semantic similarity with cross-lingual resources: A study in bangla—a low resourced language. *Informatics*, 6:19.

Carlos Periñán-Pascual. 2021. Measuring associational thinking through word embeddings. *Artificial Intelligence Review*, 55(3):2065–2102.

Nafiz Sadman, Akib Sadmanee, Md. Iftekhar Tanveer, Md. Ashraful Amin, and Amin Ahsan Ali. 2019. Intrinsic evaluation of bangla word embeddings. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5.

Zakia Sultana Ritu, Nafisa Nowshin, Md Mahadi Hasan Nahid, and Sabir Ismail. 2018. Performance analysis of different word embedding models on bangla language. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5.