Does Word2Vec encode human perception of similarity? A study in Bangla

Manjira Sinha
Center for Education Technology
IIT Kharagpur
Kharagpur, India
manjira87@gmail.com

Rakesh Dutta
Dept. of Computer Science and Application
University of North Bengal
Siliguri, India
rakeshhijli@gmail.com

Tirthankar Dasgupta
Innovation Lab
Tata Consultancy Services,
Kolkata, India
iamtirthankar@gmail.com

2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 978-1-7281-5241-7/20/$31.00 ©2020 IEEE 10.1109/ICBSLP47725.2019.201567

***Abstract*—The quest to understand how language and concepts are organized in human mind is a never-ending pursuit undertaken by researchers in computational psycholinguistics; simultaneously, on the other hand, researchers have tried to quantitatively model the semantic space from written corpora and discourses through different computational approaches - while both of these interacts with each other in-terms of understanding human processing through computational linguistics and enhancing NLP methods from the insights, it has seldom been systematically studied if the two corroborates each other. In this paper, we have explored how and if the standard word embedding based semantic representation models represent the human mental lexicon. Towards that, We have conducted a semantic priming experiment to capture the psycholinguistics aspects and compared the results with a distributional word-embedding model: Bangla word2Vec. Analysis of reaction time indicates that corpus-based semantic similarity measures do not reflect the true nature of mental representation and processing of words. To the best of our knowledge this is first of a kind study in any language especially Bangla.**

***Index Terms*—Computational Psycholinguistics, Semantic Priming, Mental Lexicon, Response Time, Degree of Priming, Word2vec model.**

I. Introduction

The representation of words in the human mind is often termed as Mental lexicon (ML) [1]–[3]. Although, words have various degree of association that are governed by different linguistic and cognitive constraints, the precise nature of the interconnection in the mental lexicon is not clear and remains a subject of debate over the past few decades. However, a clear understanding of how words are represented and processed at the mental lexicon will not only help enhance our knowledge of the inherent cognitive processing but may also be used to develop cognitively aware Natural Language Processing(NLP) applications.

A lot of cognitive experiments are being carried out in different languages to study the semantic representation of words in the ML. Typically, semantic priming experiments are used to perform such studies [4], [5]. Priming involves exposure of a stimulus (called prime) that results in quicker recognition of a related stimulus (called the Target). For example, the amount of time required to recognize a word **BRANCH** will be small if it is preceded by a related stimulus like **TREE** as compared to an unrelated stimulus like **HOUSE**. The data collected from such cognitive experiments are then used to develop robust computational models of representation and processing of words in the ML.

Attempts have also been made to provide computational models of representation of semantically similar words in the mental lexicon. Among them, one of the most commonly used is the distributional semantic model of representation. The idea behind distributional semantics is that words with similar meanings are used in similar contexts [6]. Thus, semantic relatedness can be measured by looking at the similarity between word co-occurrence patterns in text corpora. In linguistics, this idea has been a useful line of research for the past couple of decades [7], [8]. Recently, due to the advancements in the area of deep neural networks, a more robust form of distributional semantics models have been proposed in the way of word2vec word embedding. However, most of these works are based on languages like English, French, German, Arabic, and Italian. Very few attempts have been made to perform such studies for Bangla ML. Moreover, an important aspect of a study is to verify the fact of whether the distributional approach of words representation is truly the way human mental lexicon works.

In this research, we have explored how and if the standard word embedding based semantic representation models represent the human mental lexicon. The paper discusses various cases where it has been shown that highly similar words identified by the computational models seldom shows any priming effect by the users. On the other hand, word pairs having low corpus similarity may result in a high degree of priming.

The objectives of the paper are as follows:

- We conduct a semantic priming experiment over a set of 300-word pairs to study the organization and representation of Bangla words in the mental lexicon.
- For each of the 300 Bangla word pairs, we have computed their word embedding using the word2vec

978-1-7281-5241-7/19/$31.00 ©2019 IEEE

technique and computed the semantic distance between each of the word pairs.
- We try to observe whether there exist any correlation between the computed similarity score and the priming effect. Here, the null hypothesis is words having high cosine similarity scores will show a high degree of priming.

II. Semantic Priming Experiment on Bangla Semantically Similar Words

In order to study the effect of priming on semantically related words in Bangla, we have execute the masked priming experiment as discussed in [9]–[11]. In this technique, the prime word (say *chor* (thief)) is placed between a forward pattern mask and the target stimulus (say *polisa* (Police)), which acts as a backward mask. This is illustrated below.

$$prime(1000ms)chora(THIEF) \rightarrow target(2000ms)pulisa(POLICE)$$

Once the target porbe is presented to the participants for a given period of time, they are asked to decide whether the given target word is valid or not. The participants enter their decision by pressing the key 'J' (for valid words) and 'K' (for invalid words) of a standard QWERTY keyboard. The time taken to press any one of the key after the display of the target word (also called the response time (RT)), is recorded by the system timer. We display the same target word once again but with a different visual probe called the CONTROL word. The CONTROL word do not show any semantic, or orthographic similarity with either the prime or the target word. For example, *baYaska (aged)* and *briddha (aged)* is a *prime-target* pair, and the corresponding control-target pair could be *naYana (eye)* and *briddha (aged)*. We use the DMDX software tool[1] to conduct all the experiments conducted in this work.

A. Data and Experiment

We choose 300 word pairs from a Bangla corpus of around 3.2 million unique words [2]. The corpus consist of the literary works of famous Bangla authors; we have also extended the corpus by adding texts from Bangla Wikipedia, News sources, and Blogs. The words are chosen is such a way that they represents a substantial amount of distribution over the entire corpus. This will further be useful to construct the word embedding for computing the semantic similarity.

B. Implicit Perception of Semantic Similarity

1) Methods and Material: There were 300 prime-target pairs classified into two different classes. For each of the targets a *prime* and a *control* word have been chosen. Class-I primes have high degree of relatedness (e.g. **সূর্য**(Surya(sun)) – **অস্ত**(asta(sunset))), where as class-II primes have a low degree of relatedness (**ছাগল**(Chagal(goat)) – **অস্ত**(asta(sunset))).

The controls, do not possess any semantic, orthographic or morphological relationship with the target word. It is important to restrict the subject to make any strategical guess regarding the relation between pairs of words. Thus, the prime-target and the control-target words were mixed with equal number of fillers, which are out of vocabulary words such as non-words.

2) Participants: The experiments were conducted on 100 native Bangla speakers with at least a graduation degree. The age of the subjects varies between 20 to 30 years.

3) Result: Extreme reaction time and incorrect responses of the RT in the lexical decision (about 7.5%) were not included in the latency analysis. We set extreme reaction time for one subject as the median lexical latency of that essence subject. Table I depicts the average reaction time (RT) of some prime-target and control-target word pairs.

Table I
The average reaction time (RT) of the prime-target and control-target word pairs.

Prime Word	*Target Word*	*Control Word*	*Average RT(P-T)*	*Average RT(C-T)*
চোর	ডাকাত	জল	548.62	786.89
লোভ	লোভী	ভয়	636.71	716.80
রোগ	রোগী	বাতাস	700.14	821.81
জল	বায়ু	বই	612.81	726.96
বিচার	বিচারক	জামা	669.99	809.42

Then we calculate the degree of priming (DOP) i.e. the average reaction time (control-target) minus the average reaction time (prime-target) words pairs for each word pair. This is represented as follows:

$$Degreeof Priming(DOP) = Diff(RT(C_T), RT(P_T)) \quad (1)$$

After getting the result of DOP, we again calculate the average of DOP across all users.

III. Distributional Approach towards Measuring Semantic Similarity

Recently, word embedding techniques proposed by Mikolov and et al. (2013a) [12] argued that neural network based word embedding (word2vec) models (see Figure 1) are efficient at creating robust semantic spaces. Typically word embeddings are computed by using two techniques: a) Skip Gram model and b) Continuous Bag of Word model (CBOW). The skip gram model in one hand, considers a central word and tries to predict the context words that best fits the target word. On the other hand, the CBOW model tries to predict a target word given the context words. Consider for example the sentence: **কচ্ছপ একধরনের সরীসৃপ যারা পানি এবং ডাঙা দুই জায়গাতেই বাস করে।**

Let **কচ্ছপ** be the input to the proposed neural network. Our objective here is to predict a target word **সরীসৃপ** using

[1]http://www.u.arizona.edu/ kforster/dmdx/download.htm
[2]obtained from www.snltr.org

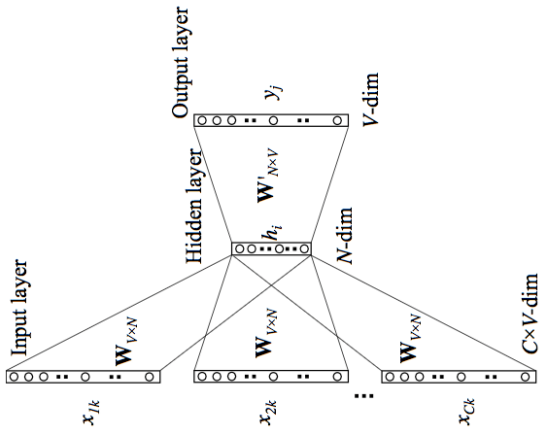

Figure 1. Illustration of the Word2Vec Model

the single context input word **কচ্ছপ**. We use the one hot encoding of the input word and measure the output error compared to one hot encoding of the target word. Thus, the vector representation of the target word is learned by learning to predict the same target word. The basic architecture of the CBOW model is depicted in Figure 1.

We have computed the word2vec embedding of each of the words present in the 300 prime-target, control-target pairs. We use some pre-trained word2vec models i.e model-2 [3] and model-3 [4], and also trained word2vec and generated embedding with different dimension on the corpus discussed in section II-A (model 1). From the different sets of embedding, we have computed the semantic similarity between each of the word pairs using the cosine similarity measure:

$$Cos(w_i, w_j) = \frac{\bar{w_i}.\bar{w_j}}{|w_i||w_j|} \quad (2)$$

Next, we compared the scores with the DOP collected from the priming experiment. Table II depicts the correlation between the different embedding models and the reaction times obtained from the priming experiment.

Table II
Comparison table with different models with respect to the human annotated Data.

Model name	*VOCAB*	*Dimension*	*Co-relation*
Model 1	427261	300	-0.2422
Model 1	427261	400	-0.2104
Model 2	10059	200	-0.0994
Model 3	145350	300	0.0203

From Table II we observed that there is a very week correlation between the Word2Vec models and the Reaction Time data.

[3]https://github.com/Kyubyong/wordvectors
[4]https://fasttext.cc/docs/en/crawl-vectors.html

Table III
Similarity score of Human annotated (DOP) vs Word2vec annotated.

Word1	*Word2*	*Control Word*	*Degree of priming*	*Word2Vec*
চোর	ডাকাত	জল	238.28	0.3236
দেশ	দেশী	জানালা	194.61	0.2549
বিদ্যালয়	ছাএ	পাখি	127.10	0.3771
উদ্দেশ্য	বিধেয়	সাঁতার	309.81	0.2678
ইতিহাস	ভূগোল	বালক	197.67	0.4535
রাম	রাবণ	রাঁধুনি	160.19	0.4060
তাঁতী	কাপড়	রবি	153.19	0.3686
ভোর	আলো	মাঝি	164.45	0.3758
বিদ্যুৎ	গর্জন	দয়া	131.00	0.2584
পিয়ন	চিঠি	চাঁদ	166.10	0.3292
রাত	স্বপ্ন	ডানা	189.92	0.2880
মধুসূদন	রবীন্দ্রনাথ	মাথা	-22.73	0.7273
দ্বীপ	উপদ্বীপ	ফল	-15.11	0.6499
গ্যাস	বাষ্প	বাড়ি	-85.44	0.5529

We observed from Table III two types of anomalies between the similarity score as obtained from word2vec and the priming result (or similarity perceived by human as recorded through psycholinguistics experiments): some word pairs which have high degree of priming but low word2vec similarity score and some word pairs having low degree of priming but high degree of word2vec similarity score. From the types of word pairs it is apparent that some word pairs, which we use together in our daily lives and colloquial use such as, **বিদ্যুৎ** and **গর্জন** . We also tend to foster a strong mental connection among them, that is reflected by the high DOP. On the other hand, as these word pairs are less likely to co-occur in a formal written corpus, they have low cosine similarity scores. Similarly, certain word pairs such as **মধুসূদন** and **রবীন্দ্রনাথ** or **গ্যাস** and **বাষ্প** have high word similarity score as they are likely to co-occur often due to their categorizations (first pair is names of poets and second pair is physics concepts), but not so much in day-to-day usage. Hence, our null hypothesis is proven to be wrong and we can infer that substantial gap still exists between how we *process* words and how computational approaches *think* we do.

IV. Conclusions and Discussion

In this paper we aim to study the representation and processing of semantically similar words in Bangla mental lexicon. Accordingly, we have conducted semantic priming experiment to determine the reaction time of subjects for prime-target and control-target pairs. We further computed the semantic similarity between the same word pairs using the Bangla word2Vec based word embedding model. We have compared the standard word embedding based semantic representation models correctly reflects the organization and processing of the mental lexicon. Analysis of reaction time indicates that corpus based semantic similarity measures does not reflects the true nature of mental representation and processing of words. The paper discusses various cases where it has been shown that highly similar words seldom shows any priming effect by the users

whereas word pairs having low corpus similarity may result in high degree of priming.

It is clear that the existing word embedding models are primarily based on the underline corpus on which they have been trained. Therefore, in order to understand the word representation strategies, ideally the corpus must bear a close resemblance with the human spoken form. However, such an ideal condition must always have to be approximated.

In particular, most of the natural language humans are exposed with belong to the spoken form. In order to use such data, it is required to do manual transcription of the spoken forms into their respective textual representations. This is not only time consuming but requires a huge man-power effort. On the other hand, the typical textual models that presently exists are based on written language that are available in the open web. Although, they are available in plenty, they are often less representative of the actual language input. In a recent work, Brysbaert et al. (2011) [13] showed that word frequency measures based on a corpus of 50 million words from subtitles predicted the lexical decision times of the English Lexicon Project [14] better than the Google frequencies based on a corpus of hundreds of billions words from books. Similar findings were reported for German [15]. In particular, word frequencies derived from non-fiction, academic texts perform worse [16].

Acknowledgment

The graduate students of Vidyasagar University and the twelve standard students of Gopali I.M.high school in India have actively helped us to performed these psycholinguistic experiments. So, we thank to all the participants.

References

[1] J. Grainger, P. Colé, and J. Segui, "Masked morphological priming in visual word recognition," *Journal of memory and language*, vol. 30, no. 3, pp. 370–384, 1991.

[2] E. Drews and P. Zwitserlood, "Morphological and orthographic similarity in visual word recognition." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 5, p. 1098, 1995.

[3] M. Taft, "Morphological decomposition and the reverse base frequency effect," *The Quarterly Journal of Experimental Psychology Section A*, vol. 57, no. 4, pp. 745–765, 2004.

[4] S. Dehaene, L. Naccache, G. Le Clec'H, E. Koechlin, M. Mueller, G. Dehaene-Lambertz, P.-F. van de Moortele, and D. Le Bihan, "Imaging unconscious semantic priming," *Nature*, vol. 395, no. 6702, p. 597, 1998.

[5] J. H. Neely, "Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention." *Journal of experimental psychology: general*, vol. 106, no. 3, p. 226, 1977.

[6] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[7] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behavior research methods, instruments, & computers*, vol. 28, no. 2, pp. 203–208, 1996.

[8] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[9] K. I. Forster and C. Davis, "Repetition priming and frequency attenuation in lexical access." *Journal of experimental psychology: Learning, Memory, and Cognition*, vol. 10, no. 4, p. 680, 1984.

[10] K. Rastle, M. H. Davis, W. D. Marslen-Wilson, and L. K. Tyler, "Morphological and semantic effects in visual word recognition: A time-course study," *Language and cognitive processes*, vol. 15, no. 4-5, pp. 507–537, 2000.

[11] W. D. Marslen-Wilson, M. Bozic, and B. Randall, "Early decomposition in visual word recognition: Dissociating morphology, form, and meaning," *Language and Cognitive Processes*, vol. 23, no. 3, pp. 394–421, 2008.

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

[13] M. Brysbaert, E. Keuleers, and B. New, "Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing," *Frontiers in Psychology*, vol. 2, p. 27, 2011.

[14] D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The english lexicon project," *Behavior research methods*, vol. 39, no. 3, pp. 445–459, 2007.

[15] M. Brysbaert, M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, and A. Böhl, "The word frequency effect," *Experimental psychology*, 2011.

[16] M. Brysbaert, B. New, and E. Keuleers, "Adding part-of-speech information to the subtlex-us word frequencies," *Behavior research methods*, vol. 44, no. 4, pp. 991–997, 2012.