LEARNING CROSS-LINGUAL WORD EMBEDDINGS WITH UNIVERSAL CONCEPTS

Pezhman Sheinidashtegol and Aibek Musaev[1]

[1]Department of Computer Science, University of Alabama, Tuscaloosa, Alabama

ABSTRACT

Recent advances in generating monolingual word embeddings based on word co-occurrence for universal languages inspired new efforts to extend the model to support diversified languages. State-of-the-art methods for learning cross-lingual word embeddings rely on the alignment of monolingual word embedding spaces. Our goal is to implement a word co-occurrence across languages with the universal concepts' method. Such concepts are notions that are fundamental to humankind and are thus persistent across languages, e.g., a man or woman, war or peace, etc. Given bilingual lexicons, we built universal concepts as undirected graphs of connected nodes and then replaced the words belonging to the same graph with a unique graph ID. This intuitive design makes use of universal concepts in monolingual corpora which will help generate meaningful word embeddings across languages via the word co-occurrence concept. Standardized benchmarks demonstrate how this underutilized approach competes SOTA on bilingual word sematic similarity and word similarity relatedness tasks.

KEYWORDS

Word Embedding Model, NLP, Word Embedding Evaluation Tasks, Universal Concepts, & Bilingual and Cross-lingual Word Embeddings.

1. INTRODUCTION

Previous works have proposed to learn cross-lingual word embeddings by aligning monolingual word embedding spaces. These approaches generally fall into two categories: supervised methods based on various transformations to align word embeddings on distinct languages and unsupervised methods without the use of any parallel corpora. The first method using parallel data supervision was proposed by [15]. A linear mapping from a source to a target embedding space was learned and is shown to be effective for the translation of words between English and Spanish. Another method proposed to normalize the word vectors and constrain the linear mapping to be orthogonal [18]. Recent works by [6] introduced an unsupervised method to align monolingual word embedding spaces using cross-domain similarity local scaling (CSLS). The work by [12] leverages CSLS to improve the quality of bilingual word vector alignment.

In this paper, we propose an alternative approach to the described works that learns cross-lingual word embeddings by injecting the seed universal concepts to the input monolingual texts. Such concepts are notions that are fundamental to humankind and are thus persistent across languages, e.g., the concept of a man or woman, war or peace, good or bad, etc. All words across languages belonging to the same universal concept are replaced with a common unique identifier, such as UUID [13]. The intuition is that the use of universal concepts in monolingual corpora from different languages will help learn syntactic and semantic relationships for words across languages via concept co-occurrence.

Note that this proposed approach results in a unified word embedding space for all languages. This is in contrast with the existing works that rely on aligning monolingual word embeddings on distinct languages. Hence, in this approach, there is no limit on the number of supported languages. Each new language shares the universal concepts with other languages by adding translation word pairs with any of the supported languages. We developed our approach independently of Ammar et al., 2016 [1]. Two important distinctions are: a) use of BFS for generation of universal concepts, and b) our evaluation is based on standard MUSE datasets. These contributions are warranted.

In the proceeding, we introduce our approach and evaluate it using the bilingual word sematic similarity and word similarity relatedness tasks. Afterwards, we obtain robust bilingual embeddings for two language pairs, namely En-Es (English to Spanish) and En-De (English to German). Then, we generate a unified cross-lingual word embedding for all three languages and show that it does not degrade the quality of the generated word vectors. The code for our approach and the cross-lingual word vectors are available at https://github. com/aibek76/universal_concepts.

The rest of this paper is structured as follows. Section 2 briefly describes related publications. Section 3 explains in detail the proposed methods. Section 4 describes the methods for evaluating our generated word embeddings, and our findings. Finally, Section 5 presents our conclusions.

2. RELATED WORK

2.1. Word Representations

There are a number of effective existing approaches to creating monolingual word vectors. Two related models, continuous bag-of-words (CBOW) and skip-gram [14] make up *word2vec*. Both models make use of a context window surrounding the target word to adjust the weights of the projection layer of the network. In the case of CBOW, the context words are used to predict the current word, while skip-gram uses the current word to predict the other words in the context window.

An extension of the skip-gram model under the name *fastText* was proposed by [5]. This model makes use of sub-word information by means of *n*-grams. Vectors are learned for the various *n*-grams in the training text, and word vectors are composed by summing these *n*-gram vectors. The aim of using this approach is to create vector representations which more effectively encode important morphological information, such as word affixes.

GloVe, a portmanteau of Global Vectors, makes use of both the local context of a given word and global word co-occurrence counts [16]. Rather than solely considering the probability that an individual word appears in the corpus at all, the model calculates probabilities that certain words co-occur with the given target word. Words that are highly related to the target word will have a high co-occurrence probability, while words that are unrelated will have a low co-occurrence prob- ability. The authors argue that by considering more global occurrence statistics, their model more effectively captures semantic information than models whose primary focus is the immediate context window.

2.2. Learning Multilingual Word Embeddings Via Alignment

In addition to the available approaches for creating monolingual word vectors, a number of various methods exist for taking sets of monolingual word vectors and aligning their vector spaces to create a new vector space that is multilingual. Methods proposed by [15], [12], and [2]

apply different functions to align the vector spaces. For each of these methods, it is required that there is a small dictionary of bilingual data in order to begin the alignment. Another variation proposed by [17] also uses a bilingual dictionary but additionally presents the possibility of a pseudo-dictionary, which makes the assumption that all strings that are identically spelled in the target and source languages hold similar meanings and aligns the vector spaces using this pseudo-dictionary with considerable success.

A few methods have been proposed that lack the requirement for the grounding bilingual dictionary and are therefore unsupervised. [8] suggests creating a two-agent communication game wherein one agent only understands language A, while the other agent only understands language B. A translation model translates sentences and sends them to each agent which confirms whether the sentences they received make sense in their monitored language. This model may be extended to many languages, so long as a closed loop is created. Another option is using iterative matching as is done in [10,19].

2.3. Core Vocabulary

A concept that is key to supporting the validity of our approach is core vocabulary. The various studies that focus on core vocabulary highlight the fact that a surprisingly small number of words make up a very significant fraction of actual words used in speech and other forms of communication. [4] studied the words used by preschoolers across 3,000 recorded speech samples and found that the top 250 most frequently used words accounted for 85% of the full sample, and furthermore found that the top 25 words accounted for 45% of the sample. A similar study was performed in Australian adults in the workplace and it found that 347 words accounted for 75% of the total conversational sample [3]. The data from these studies validates the usefulness of having a relatively small vocabulary for effective communication within a language. In our proposed method, we use 5k words, which far exceeds the numbers provided in these studies and therefore provides a robust, useful vocabulary for our multilingual vectors.

The word representation models described above are designed to create sets of monolingual vectors. Though there are numerous methods available that attempt to align the distinct vector spaces of sets of monolingual vectors created by these possible models, our method aims to eliminate the need for aligning separate vector spaces. Our method may be implemented with any of the above word representation models to create a unified, multilingual vector space without the need for mapping monolingual vector spaces onto one another. Lastly, this unified vector space is composed of a set of words that are highly useful in communication, as only a very small fraction of words make up a significant percentage of actual words used in communication.

3. PROPOSED APPROACH

3.1. Overview

Here is an overview of the proposed approach. It consists of three steps as follows:

1. Pre-processing step: prepare monolingual corpora
   (a) Build universal concepts using bilingual lexicons
   (b) Generate UUID for each universal concept
   (c) Replace words from universal concepts with corresponding UUIDs
2. Processing step: compute word embeddings using modified corpora
   (a) Compute word embeddings based on word co-occurrence
3. Post-processing step: retrieve vectors for words in universal concepts

(a) Assign vectors computed for UUIDs to all words in corresponding universal concepts

During the pre-processing step, monolingual corpora are modified to encode the seed universal concepts. Then word embeddings based on word co-occurrence are computed using an existing implementation, such as *word2vec* or *fastText*. Finally, during the pre-processing step, the vectors computed for universal concepts are assigned to each word belonging to those concepts.

1.(a) Build universal concepts using bilingual lexicons. We propose to use bilingual lexicons as the seed universal concepts as follows. Each bilingual lexicon contains translations as word pairs between words in the origin language and their translations in the target language. We treat each word as a node and each word pair as an undirected edge or link between nodes. Then we build universal concepts as undirected graphs of connected nodes using breadth first search (BFS). The resulting graphs represent concepts that have the same meaning across languages.

Here is the pseudocode for building universal concepts using Python's syntax:

```
def BuildUniversalConcepts(words, wordPairs):
    nodes = words
    links = wordPairs
    universalConcepts = []
    while nodes is not empty:
        randomNode = PickRandomNode(nodes)
        universalConcept = BFS(
            randomNode,
            links,
            depthLimit = 4
        )
        universalConcepts.append(
            universalConcept
        )
        for node in universalConcept:
            nodes.pop(node)
    return universalConcepts
```

1.(b) Generate UUID for each universal concept. For each universal concept represented as a graph, we generate a unique ID, such as UUID, and associate words in that graph with a corresponding UUID.

1.(c) Replace words from universal concepts with corresponding UUIDs. Given the downloaded monolingual corpora of Wikipedia, in English and Spanish, we combine them into a single corpus. Then we replace the mentions of each word from universal concepts with a corresponding UUID. This ensures that words belonging to the same concept are treated in the same manner across languages.

2.(a) Compute word embeddings based on word co-occurrence. In this step a combined corpus with encoded universal concepts is used as input for computing word embeddings based on word co- occurrence. Note, that any implementation of the word co-occurrence approach, such as *word2vec* or *fastText*, can be applied.

3.(a) Retrieve vectors for words in universal concepts. The vectors computed for UUIDs are assigned to each word in corresponding universal concepts. The resulting word embeddings contain vectors for the words in all languages.

3.2. Illustration of the approach

Here is an example of the input texts in three languages. Note, that we use abbreviations instead of the actual UUID values for clarity:

- English: *brown fox jumps over the lazy dog*
- German: *brauner fuchs springt u¨ber den faulen hund*
- Spanish: *zorro marro´n salta sobre el perro perezoso*

Most of these words are part of 35k words from the full set of the MUSE dataset [6], which was used to generate the universal concepts. The only word that is missing from the training set is *lazy*. The pre-processing step 1 based on this training set will generate the following modified texts:

- English: uuid1 uuid2 uuid3 uuid4 uuid5 *lazy* uuid6
- German: uuid1 uuid2 uuid3 uuid4 uuid5 *faulen* uuid6
- Spanish: uuid2 uuid1 uuid3 uuid4 uuid5 uuid6 *perezoso*

The modified texts are used as input to the processing step 2 that will learn the word embeddings for all tokens including universal concepts and the remaining words. The post-processing step 3 will assign the vectors computed for universal concepts to each word belonging to those concepts, e.g., the same vector will be assigned to the words *fox*, *fuchs*, and *zorro* as they share the same UUID.

Table 1: Comparison of bilingual word semantic similarity performance between bilingual universal concepts word embeddings with window of 10.

Method	En-Es	Es-En	En-De	De-En
RSCLS([11])	51.54	46.77	54.86	11.15
BUCWE	40.95	**53.68**	52.53	**33.53**

Table 2: Performance on word similarities for German and Spanish. Comparing the original mono-lingual embeddings of their mapping to English.

Dataset	Monolingual	BUCWE
De-GUR350	0.65	0.67
Es-WS353	0.54	0.60

4. EXPERIMENTS

4.1. Overview

We tested our model using bilingual word semantic similarity and word semantic relatedness tasks. Both training and evaluation sets are from the MUSE benchmark [6]. To start, we used a pair of bilingual training sets (En-Es and Es-En, De-En and En-De), such that each set had 35k origin words, to build the bilingual universal concepts (BUCWE). Then, we evaluated the training sets performance on a bilingual word semantic similarity task and a word sematic relatedness task. Finally, we combined two sets of bilingual dictionaries with the 35k origin words to build a unified multilingual universal concepts model supporting all three languages (MUCWE) and evaluated it with word similarity task.

4.2. Bilingual Word Semantic Similarity Task

In this task, we used the MUSE test set to evaluate our bilingual universal concepts word embeddings (BUCWE) and bilingual word embeddings provided by the MUSE project. First, we computed a list of the ten most similar words for each word in the origin language. Those words are achieved using semantic word similarity which is defined by the cosine similarity between word vectors. If the translation of the word was in the list, it passed the test, otherwise it failed. Finally, the number of word-pairs that passed the test divided by number of all tested pairs creates a score for this task. This is a deviation of the word translation task used by Mikolov et al. [14]. As is shown in Table 1, our purposed model consistently performs better on XX-En tests. In contrast, RSCLS has had better results on En-XX tests.

4.3. Word Semantic Relatedness Task

In this experiment, we measure if the semantics of word vectors are maintained within languages. We used GUR350 dataset for German and WS353 for Spanish. Each row of these datasets starts with a pair of words and a human judgment relatedness score. For testing our word vector, first, we calculate the cosine similarity between the given word pair vectors. Next, we compute the Spearman correlation coefficients between our computed cosine similarity and the human judgment scores. A greater coefficient word vector means a higher quality word vector. Table 2 shows that our approach does not have a negative impact on the quality of the generated word vectors.

4.4. Multilingual Universal Concepts Word Embeddings (MUCWE)

Next, we ran the word sematic relatedness task for MUCWE to determine the impact of adding more languages to the universal concepts model based on word similarity within each language. Table 3 shows that MUCWE, consisting of three languages, can represent the relatedness of the words within each language as well as the word2vec monolingual word vectors.

Table 3: Comparing the performance of word similarities between multilingual universal concepts using 35k words for English, Spanish and German (MUCWE) and monolingual word2vec word vectors.

Dataset	Es-WS353	De-GUR350
Es-monolingual	0.54	-
De-monolingual	-	0.65
MUCWE	0.54	0.63

5. CONCLUSIONS

This paper introduces implementing learning of cross-lingual word embedding based on the universal concepts approach. Our results show that a cross-lingual word embedding using the seeds of universal concepts has a competitive advantage on bilingual word semantic similarity and word semantic relatedness tasks. We also show that the proposed approach does not degrade the quality of the generated word vectors. Finally, our method features extensibility which is lacking in most of the state-of-the-art methods.

Our future work includes the support of additional languages, both Romance and Asian, in a single model. We also plan to perform a study of the core set of universal concepts across languages through statistical analysis. We are also developing new intuitive intrinsic evaluation task for word embeddings.

ACKNOWLEDGEMENTS

I would like to thank the Computer Science Department at The University of Alabama, because through their funding and support, the faculty and staff had made it possible for me to work on this very interesting subject.

REFERENCES

[1] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. Massively multilingual word embeddings. *CoRR*, abs/1602.01925, 2016.

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas, November 2016. Association for Computational Linguistics.

[3] Susan Balandin and Teresa Iacono. Crews, wusses, and whoppas: core and fringe vocabular- ies of australian meal-break conversations in the workplace. *Augmentative and Alternative Communication*, 15(2):95–109, 1999.

[4] David R. Beukelman, Rebecca S. Jones, and Mary Rowan. Frequency of word usage by nondisabled peers in integrated preschool classrooms. *AAC: Augmentative and Alternative Communication*, 5(4):243–248, 1 1989.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vec- tors with subword information. *CoRR*, abs/1607.04606, 2016.

[6] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herve´ Je´gou. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017.

[7] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. *CoRR*, abs/1805.11222, 2018.

[8] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc., 2016.

[9] Yedid Hoshen and Lior Wolf. An iterative closest point method for unsupervised word trans- lation. *CoRR*, abs/1801.06126, 2018.

[10] Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478. Association for Computational Linguistics, 2018.

[11] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, and Edouard Grave. Improving supervised bilingual mapping of word embeddings. *CoRR*, abs/1804.07745, 2018.

[12] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2979–2984, 2018.

[13] Paul J. Leach, Michael Mealling, and Rich Salz. A universally unique identifier (UUID) URN namespace. *RFC*, 4122:1–32,2005.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[15] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.

[16] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[17] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Of- fline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859, 2017.

[18] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1006–1011, 2015.

[19] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970. Association for Computational Linguistics, 2017.

AUTHORS

Pezhman Sheinidashtegol received his B.S in Computer Engineering from IAU, in 2010 and his M.S. degree in Computer Science from Western Kentucky University, in 2016. He is currently pursuing his Ph.D. in the department of Computer Science at The University of Alabama. His research interests include deep learning, text mining, machine learning, social media, data privacy, and cloud/network security.

Dr. Musaev is an Assistant Professor of Computer Science at the University of Alabama. He received his Ph.D. (2016), M.S. (2000) and B.S. (1999) degrees in Computer Science from Georgia Tech. Before doctoral studies, he founded and managed a software company Akforta that provided enterprise management solutions. He is a recipient of the Scholarship of the President of the Kyrgyz Republic from 1997 to 1999. His primary research interests are in applied data mining of big data, including social media for disaster management and video traffic data for transportation.