Evaluating Word Embeddings in Natural Language Processing: An Analysis and Synthesis of Current Approaches

José Martín Véliz Zambrano.
Tecnológico de Monterrey, Querétaro, México.

Abstract

Word embeddings have become foundational components in the field of Natural Language Processing (NLP), powering a wide range of applications from sentiment analysis, and machine translation, to question-answering systems. This review aims to elucidate the complex and diversified landscape of word embeddings by examining the consensus on the theoretical underpinnings, the methodological approaches, their implementations, and how they are evaluated. The goal is to provide a synthesised account and an analytical critique of the existing body of literature, identify gaps, and propose avenues for future research.

Introduction

Word embeddings, as Felipe Almeida et al underscored, are distributed representations of words encoded as dense, fixed-length vectors that incorporate semantic and syntactic information. These representations, which are integral to NLP tasks, are primarily generated using either prediction-based or count-based models [1]. The idea stems from the distributional hypothesis that suggests words with similar meanings appear in similar contexts.

Word Embeddings: Theoretical Underpinnings and Methods

The development of word embeddings, the distributed representations of words encoding syntactic and semantic meaning, is anchored by the distributional hypothesis. This hypothesis declares that words with similar meanings are situated in similar contexts within a corpus, therefore forming the foundational premise for constructing meaning through context.

Among the earliest strategies to generate these embeddings are Vector Space Models (VSM) and Latent Semantic Analysis (LSA). VSM represents words as vectors in a high-dimensional space in which the spatial proximity of vectors captures semantic similarity. LSA, leaning towards count-based models, uses Singular Value Decomposition (SVD) on a term-document matrix to decode latent semantic structures.

Neural network-based models, such as Word2Vec, build on these earlier approaches. Word2Vec uses a shallow neural network trained using a continuous bag of words or the skip-gram model. The network is trained to predict the probability of a word given its context (or inversely, the context given the word), which results in a high-dimensional vector space where geometrically proximate words are semantically related.

Global Vectors for Word Representation (GloVe), another prominent neural network model, diverges from Word2Vec by operating on aggregated global word-context co-occurrence statistics from a corpus rather than on local word context.

It employs matrix factorization to the word-context matrix produced from the corpus, capturing the co-occurrence probabilities and thus forming semantically rich word embeddings.

The advances in deep learning then led to the development of contextually aware embedding models like Embeddings from Language Models (ELMo) and the Bidirectional Encoder Representations from Transformers (BERT). ELMo makes use of character-level encoding and bidirectional LSTM to produce context-dependent word embeddings. It presents every single word with a unique embedding that is influenced by the word's contextual use. ELMo word embeddings can take into account polysemy, where a single word possesses multiple meanings depending on how it is used within a sentence.

BERT represents a further evolution of embedding techniques. Like ELMo, it captures context, but it does so in both directions, using transformer architectures. BERT's bidirectional training allows the model to learn the context of a word based on all of its surroundings, making it capable of understanding the full context of a word.

The expansion and refinement of these methods witness varying strengths from capturing simple co-occurrence statistics to exploiting the potential of neural networks and considering the complexities of word usage and context. The evolution of word embedding techniques continues, with growing interest in enhancing dimensionality reduction, optimizing computational efficiency, and refining the ability to capture more nuanced semantic relationships.

Evaluation of Word Embeddings

The effectiveness of word embeddings is gauged through a series of evaluation strategies that are classified broadly into intrinsic and extrinsic methods. Intrinsic methods evaluate word embeddings directly through lexical analogy or word similarity tasks. In contrast, extrinsic evaluations are completed through downstream NLP tasks like named entity recognition, sentiment analysis, and text classification [2].

It was proposed by Schnabel et al a data and model-driven approach to constructing query inventories to evaluate unsupervised word embeddings. The study demonstrated that the standard cosine similarity measure could be swayed by frequency-based effects [3]. It recommended novel evaluation methods that compared embeddings concerning chosen queries to diminish bias and glean more valuable insights.

Future Directions

Amid the flourishing development of word embeddings, challenges persist, most prominently in the field of evaluation. Lack of consensus on preferred evaluation methods, subjectivity in human assessments, deficiency of suitable training data, the absence of persistent correlation between intrinsic and extrinsic methods, and bias in embeddings are only some of the issues that require further investigation [2]. Notably, the performance of different embeddings is inconsistent across different

tasks, emphasizing the need for task-dependent selection[3].

Moreover, the need for word embeddings capable of handling multilingual and multi-sense contexts, bias management, and the creation of language-independent evaluation datasets are areas ripe for exploration. There is also an opportunity to delve deeper into unsupervised word embeddings and their potential for more advanced NLP applications.

Conclusion

Word embeddings are an integral part of today's natural language processing efforts. While the current generation of word embeddings like BERT, GloVe, and Word2Vec have progressed substantially, iterative improvements are needed to manage existing limitations and challenges. This review underscores the need for a vigorous and integrative study of word embedding methods, their evaluation, and practical implementation. Continued research in more optimized and precise embeddings would invariably enhance the performance of NLP tasks across domains and technologies. The direction of future research calls for a multi-perspective and multifaceted approach to the evaluation and development of word embeddings.

6. References

[1]Almeida, F., & Xexeo, G. (2023). Word Embeddings: A Survey. arXiv preprint arXiv:1901.09069.

[2]Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. Institute for System Analysis of Russian Academy of Sciences (ISA RAS).

[3]Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings.

[4]Worth, P. J. (2023). Word Embeddings and Semantic Spaces in Natural Language Processing. International Journal of Intelligence Science.

[5]Turing. (s.f). A Guide on Word Embeddings in NLP. https://www.turing.com/kb/guide-on-word-embeddings-in-nlp