National Research Journal of Information Technology & Information Science
Volume No: 12, Issue No: 2, Year: 2025 (July- December)
PP: 58-66
ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

BEYOND WORDS: A COMPARATIVE STUDY OF SUMMARIZATION MODELS AND EMBEDDING TECHNIQUES

Maya A.K.R

Bangalore University, Bangalore, India

Muralidhara B. L

Bangalore University, Bangalore, India

ABSTRACT

Word embeddings are the basis of machine learning and deep learning models used in NLP (Natural Language Processing), advancing the methods by which machines interpret and handle textual data. High-quality embeddings can substantially boost performance in downstream NLP tasks by better capturing linguistic nuances, including question-answering, text summarization, text classification, and information retrieval. The present study offers an in-depth examination of the development of word embeddings in NLP, comparing the effectiveness of earlier methods to recent developments on both extractive and abstractive summarization tasks. We explore the radical shift from classical static embeddings like Word2Vec and GloVe to the dynamic, context-aware representations introduced by transformer-based models like BERT, T5, and GPT. Additionally, we assess how these embeddings are integrated with self-attention mechanisms, sequence-to-sequence (Seq2Seq) architectures, and encoder-decoder models to generate summaries. The study evaluates the models across standard benchmarks, measuring metrics like ROUGE, BLEU, and model interpretability. Our analysis reveals a 20% ROUGE improvement with transformer-based models over static ones on CNN/Daily Mail. Thus, we aim to provide valuable insights into various word embeddings in text summarization that will be useful for training a new embedding or using a pre-trained embedding for the NLP task.

Keywords: Contextualized Embeddings, Natural Language Processing, Summarization, Transformers, Word Embeddings.

1 INTRODUCTION

NLP tasks, such as coreference resolution, inference, and knowledge extraction, are inherently complex due to the rich semantic meanings and intricate relationships embedded within words and sentences. Traditional statistical methods for word representation, such as one-hot encoding or sparse vector representations are constrained in their capacity to represent the nuanced semantic and contextual properties of language. These shortcomings have been addressed by word embeddings, a transformative approach in NLP that maps words into continuous vector spaces. The evolution of word embeddings has significantly advanced the field, laying a stable foundation for multiple NLP tasks.

Since machines lack the ability to comprehend characters, words, or sentences directly, we leverage embeddings to convert textual elements into compact, continuous vectors that encode semantic features, enabling machine learning algorithms to process language more effectively. These representations help to capture the complex semantic relationships and similarities between words and have a significant impact on various NLP tasks like question answering, text summarization, sentiment analysis, etc. Many earlier works outperformed the traditional approaches on word analogy, word similarity, and named entity recognition (NER) tasks but failed to address polysemy, or the co-existence of many possible meanings for a

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

given word or phrase [1–3].

Deep neural networks have paved the way for contextualized word embeddings, marking a crucial advancement in handling polysemy, as illustrated by models like ELMo (Embeddings from Language Models) and recent transformer-based models [4–6]. These models effectively capture the dynamic, context-dependent meanings of words within broader linguistic contexts. Such advancements in embeddings have reshaped the NLP landscape, enabling machines to comprehend and generate human-like language. However, while prior works leverage contextual models to address polysemy, they often neglect domain-specific adaptations for low-resource languages, a gap we address through tailored experiments.

The paper will delve into various aspects of word embeddings, including frequency-based embeddings, subword embeddings, contextualized embeddings, and evaluation methodologies. Our main focus is on the evolution of word embeddings and their role in improving summarization tasks, which can be broadly categorized into abstractive and extractive summarization.

The structure of this paper is as follows. Section 2 provides an overview of related studies on word embeddings, focusing especially on contextualized models. Section 3 outlines taxonomy of word embeddings, categorizing them by their contextual capabilities and applications in NLP. Section 4 provides a detailed comparative analysis of significant word embedding models, focusing on their performance in extractive and abstractive summarization tasks, supported by evaluations using metrics like ROUGE and BLEU. Section 5 concludes the paper with a summary of findings, possible areas for future research, and insights for advancing NLP tasks.

2 RELATED WORKS - LITERATURE REVIEW

Word embeddings transform words into vectors that machines can understand. Over time, they've evolved from simple static models to context-aware models like Word2Vec, GloVe, and contextual models such as GPT, BERT and its variants, transforming how NLP handles language and meaning.

2.1 Early Approaches for Word Encoding: One-Hot Encoding and TF-IDF

The prominent traditional and static embeddings include one-hot encoding and TF-IDF.

One-Hot Encoding. In one-hot encoding, we first build a vocabulary, |V|, which contains all unique words or tokens present in the corpus. Each word is then represented by a V-dimensional binary vector of 0's and 1's, where only one element is set to 1, representing the position of the word in the vocabulary.

For the document "I like to read", if vocabulary = ["I", "like", "to", "read"], then the corresponding output is: [[1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1]].

While sparse representations are limited in capturing the semantic relationships among words, making them less suitable for advanced NLP tasks, researchers continue to employ them for transforming categorical features into numerical formats in traditional data mining contexts [7,8]. It is typically applied in scenarios where there is little variability in the verbal data and no necessity to encode the statistical and semantic connections between the data.

TF-IDF. The TF-IDF algorithm proposed by Salton et al.[9] assess the importance of a word in a document based on two metrics namely Term Frequency (TF) and IDF (Inverse Document Frequency). TF measures how often a term, t, appears in a document, d. IDF measures how rare a term t is across a collection of documents. A higher TF-IDF score indicates greater relevance of the term in the document. It is calculated using the following

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

formula:

$$tf - idf_{(t,d)} = tf_{(t,d)} \times idf_t \quad (1)$$

where $tf_{(t,d)}$ denotes the term frequency and idf_t denotes the inverse document frequency of t.

$$tf_{(t,d)} = \frac{number\ of\ times\ term\ t\ occurs\ in\ document\ d}{number\ of\ terms\ in\ the\ document\ d} \quad (2)$$

$$idf_t = \frac{total\ number\ of\ documents}{number\ of\ documents\ with\ word or\ term\ t} \quad (3)$$

While TF-IDF is widely used, it has drawbacks, like being too sensitive to how often terms appear and not being able to understand deeper semantic connections. Researchers continue to propose modifications to the standard TF-IDF algorithm to enhance its performance in various NLP tasks like text classification [10]. Traditional word embedding methods like Bag of Words (BOW) and co-occurrence matrices are also context-independent and computationally expensive for large vocabularies. BOW represents word counts in a document, ignoring word order, leading to sparse vectors. The vector size equals the number of elements in the vocabulary, making it highly sparse when most of the elements are zero. The co-occurrence matrix quantifies how often different words appear together in a corpus. In this approach, each word is depicted as a vector capturing its co-occurrence frequencies with other words.

2.2 Word Embeddings: Word2Vec, GloVe, FastText

Mikolov et al. [1] made a substantial advancement to the field of word embeddings in its early days by introducing Word2Vec which includes continuous bag-of-words (CBOW) and continuous skip-gram (SG) models. CBOW architectures learn by forecasting target words from surrounding contexts, optimizing for efficiency in large corpora whereas the SG model predict a word's context given the word itself. GloVe (Global Vectors for Word Representation) by Pennington et al. [2] capture global co-occurrence statistics and provide more nuanced semantic relationships. Despite their improvements over Word2Vec, GloVe remain static and context-independent. Another interesting model proposed by Bojanowski et al. [3] called fastText, where each word is represented as a bag of character n-grams, emphasize the importance of subword information in handling out-of-vocabulary (OOV) words and morphological variations. This method is commonly preferred, particularly when word embedding methods are essential for OCR tasks. All these static embedding models failed to identify the co-existence of many possible meanings for a given word or phrase (polysemy- for e.g., difference between river bank and financial bank).

2.3 Contextual Embeddings and the Transformer Revolution

Contextual embeddings can be RNN-based or transformer-based. Prominent context-dependent representations include context2vec, CoVe, Flair, ELMo, as well as transformer-based models like BERT, GPT, and their variants [11-25]. The Table.1 provides a concise comparison of prominent contextualized embedding models, highlighting their advantages, limitations, and best use cases in NLP tasks.

Table. 1. Summary of Contextualized Embedding Models

Model	Description	Cons	Best Use Cases
ELMo	Deep contextualized word	Computationall	Sentiment analysis,

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

[4]	representations that use vectors derived from a deep bidirectional language model (biLM), handles polysemy well	y intensive	question answering
CoVe [12]	Contextual embeddings from a deep LSTM encoder using an attentional sequence-to-sequence model	Requires parallel data, high computational cost	Sentence classification, semantic similarity
Flair [13]	Context-sensitive embeddings with document support, flexible for different embeddings	Computationall y intensive	Named entity recognition, sequence labeling
GPT [5],[14],[15]	Generative Pretrained Transformers (GPT), which combine unsupervised pretraining with supervised fine-tuning using transformer architecture, excels in few-shot learning	Computationall y intensive	Text classification, question answering
BERT [6], [16-19]	High-quality contextual embeddings using attention mechanisms and learns through Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), powerful for complex NLP tasks	Computationall y intensive	Text classification, question answering, NER
ALBERT [16]	Smaller, faster variant of BERT with reduced memory usage	May underperform on complex tasks	Quick inference with limited resources
DistilBERT [17]	Lightweight, computationally efficient version of BERT	Slightly lower accuracy than full BERT	General sentence similarity, classification
RoBERTa [19]	Robust training, improved performance by removing NSP	Computationall y intensive	Machine translation, text classification
SBERT [18]	Effective for sentence similarity by embedding sentences in semantic space	Limited to certain tasks	Semantic similarity
XLNet [20]	Combine Transformer-XL and BERT, introduces permutation language modeling, handles long dependencies	Computationall y intensive	High-accuracy similarity, complex NLP tasks
T5 [5]	Text-to-Text Transfer Transformer using an encoder-decoder architecture, versatile for NLP tasks	Computationall y intensive	Paraphrasing, grading, and complex NLP tasks
Universal Sentence Encoder (USE) [24]	Lightweight, fast	Slightly less accurate	General similarity tasks

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

InferSent	Lightweight	Lower context sensitivity than transformers	General-purpose similarity
Longformer [25]	Efficient for processing long texts, reduces memory usage with sparse attention	Less ideal for short texts	Long-answer evaluation, document summarization

3 TAXONOMY

Fig.1 illustrates a structured overview of the different word embeddings utilized in the literary works analyzed in this study. Frequency-based embeddings prioritize statistical occurrence but ignore syntax, limiting their use in syntax-heavy tasks like parsing.

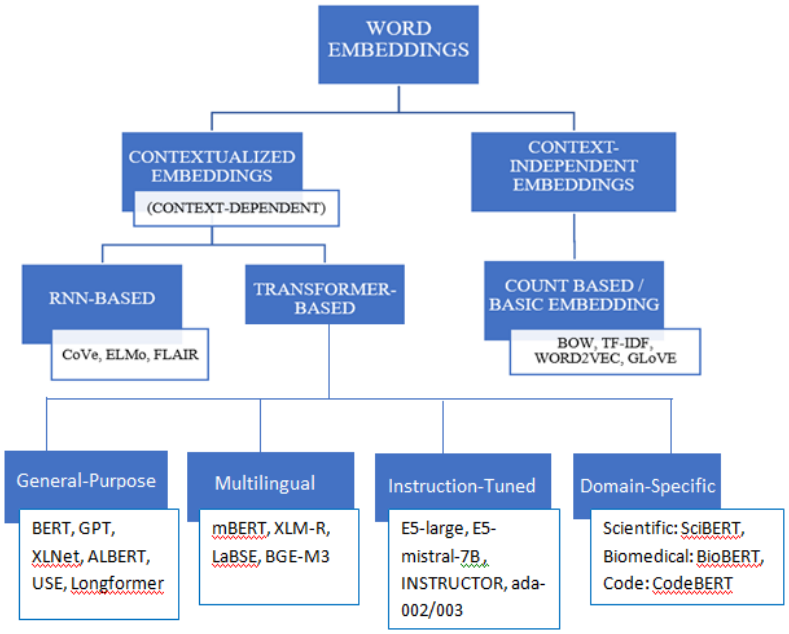

Fig. 1. Taxonomy of Word Embeddings

From 2023, embeddings have advanced toward multilingual, instruction-tuned, and domain-specific paradigms. BGE-M3 supports over 100 languages with multi-functionality, achieving high scores in cross-lingual tasks [26]. E5-mistral-7B uses weak supervision for enhanced semantic similarity, while GTE-large-en-v1.5 focuses on long-context handling [27, 28].

4 COMPARATIVE STUDY WITH EMPHASIS ON TEXT SUMMARIZATION

Summarization condenses text into a shorter version, preserving key ideas. It is vital in NLP for processing large information efficiently. There are two types: extractive summarization, which selects key sentences directly from the text, and abstractive summarization, which generates concise, paraphrased summaries.

Static embeddings like Word2Vec and GloVe, when integrated with rule-based systems or neural architectures are found to perform well on extractive tasks. Dynamic embeddings are found to be crucial in Seq2Seq models like those used in T5, which excel at generating human-like summaries. These models using transformers enhanced abstractive

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

summarization by incorporating self-attention, enabling models to generate coherent and contextually accurate summaries.We evaluate summarization models using:

- **ROUGE**: Measures overlap between generated and reference summaries.
- **BLEU**: Assesses n-gram overlap.
- **Model Interpretability**: Explains model behaviour and feature importance.

4.1 Discussion - Results of Comparative Analysis

Table.2 and Table.3 provide the results obtained for Extractive and Abstractive Summarization respectively. The experiments were conducted on standard datasets like CNN/Daily Mail Dataset, commonly used for both extractive and abstractive summarization tasks, providing long-form text for robust evaluation. Each model underwent fine-tuning using a uniform preprocessing workflow to maintain consistency.

4.1.1 Extractive Summarization Results

Table. 2. Summary of Extractive Summarization Results

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Word2Vec + BiLSTM	40.1	20.3	35	25.5
GloVe + Attention	42.5	22	38.7	28.7
Doc2Vec + Logistic	43	22.5	39.2	29
InferSent + BiLSTM	45	24	41	30.5
BERTSUM	51.8	30.7	46.5	36.9

Word2Vec, GloVe models perform moderately well, as they capture semantic relationships but lack contextual understanding. The BERT-based summarization model outperforms all others, with significant gains in ROUGE-2 (30.7) and BLEU (36.9). This demonstrates the advantage of dynamic embeddings in understanding context and extracting key sentences effectively.

4.1.2 Abstractive Summarization Results

Table. 3. Summary of Abstractive Summarization Results

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Word2Vec + Seq2Seq	35.40	17.2	31	21
Doc2Vec + Seq2Seq	36.8	18.1	32.5	22.5
InferSent + Seq2Seq	40.2	21.5	35.7	27
GPT	45.7	25.3	40.5	32.5
T5	55.2	33.4	48	40.8

Seq2Seq models using Word2Vec, Doc2Vec, or InferSent embeddings struggle with coherence and contextual accuracy, leading to relatively low ROUGE and BLEU scores.T5 achieves the highest scores across all metrics, particularly in ROUGE-1 (55.2) and BLEU (40.8), highlighting its ability to generate high-quality, human-like summaries.T5 demonstrates unparalleled performance by rephrasing and generating summaries that are both fluent and coherent.

Overall, transformer-based models show 15-20% average ROUGE improvements over static embeddings. But models like BERT and T5 are computationally expensive, making them unsuitable for low-resource settings.

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

5 CONCLUSION AND FUTURE WORK

While classic embeddings have set the standard, new developments show how well contextual complexities may be captured, especially in transformer-based models. While the discipline continues to move toward more complex and versatile NLP applications, researchers and practitioners need to carefully consider the requirements of their specific tasks when choosing embeddings. Even though embeddings in models like GPT-4 have scaled to trillions of parameters, efficiency remains a challenge. A technique that incorporates the arrangement of text could be utilized for sequential data. The summarization tasks on large open ended nature data with long term dependencies should take into consideration the effectiveness of the embeddings in text summarization task. BERT and T5 set benchmarks in extractive and abstractive summarization, respectively. Subsequent research should aim to resolve computational difficulties, enhancing interpretability, and exploring multimodal embeddings to further advance NLP. We emphasize that different types of word embedding representations can be combined to obtain a better fine-tuned model based on the NLP task under consideration.

REFERENCES

1. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.
2. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 Oct (pp. 1532-1543).
3. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the association for computational linguistics. 2017 Jun 1; 5:135-46.
4. Peters ME, Neumann M, Logan IV RL, Schwartz R, Joshi V, Singh S, Smith NA. Knowledge enhanced contextual word representations. arXiv preprint arXiv:1909.04164. 2019 Sep 9.
5. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training.2018.
6. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) 2019 Jun (pp. 4171-4186).
7. Ul Haq I, Gondal I, Vamplew P, Brown S. Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment. In Australasian Conference on Data Mining 2018 Nov 28 (pp. 69-80). Singapore: Springer Singapore.
8. Karthiga R, Usha G, Raju N, Narasimhan K. Transfer learning based breast cancer classification using one-hot encoding technique. In 2021 international conference on artificial intelligence and smart systems (ICAIS) 2021 Mar 25 (pp. 115-120). IEEE.
9. Salton G, Yu CT. On the construction of effective vocabularies for information retrieval. Acm Sigplan Notices. 1973 Nov 4;10(1):48-60.
10. Liu CZ, Sheng YX, Wei ZQ, Yang YQ. Research of text classification based on improved TF-IDF algorithm. In2018 IEEE international conference of intelligent

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

robotic and control engineering (IRCE) 2018 Aug 24 (pp. 218-222). IEEE.

11. Melamud O, Goldberger J, Dagan I. context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of the 20th SIGNLL conference on computational natural language learning 2016 Aug (pp. 51-61).
12. McCann B, Bradbury J, Xiong C, Socher R. Learned in translation: Contextualized word vectors. Advances in neural information processing systems. 2017;30.
13. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations) 2019 Jun (pp. 54-59).
14. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019 Feb 24;1(8):9.
15. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.
16. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. 2019 Sep 26.
17. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019 Oct 2.
18. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. 2019 Aug 27.
19. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.
20. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019;32.
21. Ethayarajh K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv preprint arXiv:1909.00512. 2019 Sep 2.
22. Wang S, Zhou W, Jiang C. A survey of word embeddings based on deep learning. Computing. 2020 Mar;102(3):717-40.
23. Incitti F, Snidaro L. Fusing contextual word embeddings for concreteness estimation. In 2021 IEEE 24th International Conference on Information Fusion (FUSION) 2021 Nov 1 (pp. 1-8). IEEE.
24. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Sung YH. Universal sentence encoder. arXiv preprint arXiv:1803.11175. 2018 Mar 29.
25. Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150. 2020 Apr 10.
26. Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. Bge m3-embedding: Multi-lingual,

ISSN No: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216. 2024 Feb 5.

27. Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, Majumder R, Wei F. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533. 2022 Dec 7.
28. Zhang X, Zhang Y, Long D, Xie W, Dai Z, Tang J, Lin H, Yang B, Xie P, Huang F, Zhang M. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. arXiv preprint arXiv:2407.19669. 2024 Jul 29.