Resilient Word-Embedding Alignment for BERT with Character-Based Layer in Noisy Environments

Korn Sooksatra , Alejandro Rodriguez Perez , and Pablo Rivas , *Senior, IEEE*
School of Engineering and Computer Science
Department of Computer Science, Baylor University
Email: {Korn_Sooksatra1, Alejandro_Rodriguez4, Pablo_Rivas}@Baylor.edu

***Abstract*—In the rapidly evolving domain of computational linguistics over the past decade, researchers have been dedicated to uncovering effective mechanisms for translating words into computer systems. A notable breakthrough was the introduction of embedding vectors, assigning unique vectors to words, offering a promising representation avenue. Notably, ELMo introduced bidirectional LSTM, revolutionizing word embeddings by capturing contextual information. The Transformer architecture further transformed the landscape, replacing traditional recurrent neural networks and influencing influential models like BERT and GPT. While BERT excelled across tasks, its word-based tokenizer struggled in noisy environments, leading to numerous unknown tokens. This paper pioneers a novel character-based approach, aligning with BERT's word embeddings to enhance performance and reduce retraining time. Additionally, we introduce adversarial training and vision transformer embeddings to fortify the embedding layer against adversarial alterations, contributing innovative methods to thrive in noisy environments. These approaches can provide more safety and resilience for BERT which is a foundational model. Extensive experiments showcase the efficacy of our approach, making substantial contributions to character-based modifications and fortifications against adversarial alterations.**

***Index Terms*—BERT, Transformer, Natural Language Processing**

I. Introduction

In the dynamic and evolving landscape of computational linguistics, researchers have embarked on a relentless quest to unravel effective mechanisms for translating words into computer systems. A pivotal advancement emerged with the introduction of embedding vectors, assigning a unique vector to each word, thereby presenting a promising avenue for representation; e.g., Skip-Gram [1], Continuous Bag of Words (CBOW) [2], and GloVe [3]. Recognizing the pivotal role of context in shaping word embeddings, Peter et al. introduced the groundbreaking concept of Embedding from Language Models (ELMo), employing bidirectional LSTM to enable each word to encapsulate information about other words within the input context [4].

A transformative juncture materialized in 2017 with Vaswani et al.'s proposal of the self-attention mechanism and the Transformer architecture [5], diverting attention away from conventional recurrent neural networks like LSTM. Embraced by influential models such as ALBERT [6], BERT [7], BART [8], GPT [9], LLaMA 2 [10], and RoBERTa [11], the Transformer architecture revolutionized word embeddings by considering all words in the input, irrespective of their positional distances.

While BERT emerged as a dominant force across various downstream tasks, its tokenizer, reliant on words and subwords, grappled with challenges in noisy environments, resulting in the emergence of numerous unknown tokens. This paper pioneers a novel approach aimed at enhancing BERT's word embedding layer, steering towards a character-based model and building upon the foundations laid in [12]. Drawing inspiration from the success of CharacterBert [13], which leverages convolutional neural networks, our approach involves aligning the output of the new embedding layer with BERT's word embeddings, yielding significant reductions in retraining time while preserving performance.

In acknowledgment of potential adversarial tactics involving character alterations to visually similar ones, we introduce two additional approaches to fortify the embedding layer proposed in [12]. The first approach involves leveraging adversarial training [14], wherein the layer is trained with noisy words. The second approach explores the integration of embeddings derived from state-of-the-art computer vision models, such as vision transformers [15]. In essence, our contributions extend beyond the initial character-based modification, encompassing two innovative methods that fortify the embedding layer against adversarial character alterations. Extensive experiments accompany these contributions, showcasing the efficacy of our approach in navigating and thriving within noisy environments.

The rest of the paper is organized as follows: Section II discusses existing research related to our domain; Section III introduces our alignment methodology in detail; our experiments and results are outlined and discussed in Section IV; In Section V, we discuss the limitations of our model; Finally, broader impact and conclusions are drawn in Sections VI and VII, respectively.

II. Related Works

Researchers have explored various character-centric transformer strategies for enhancing model performance in noisy environments. Al *et al.* [16] introduced a transformer architecture with causal attention, showcasing its effectiveness in language modeling without sequential processing, potentially replacing recurrent neural networks [17]. Gupta *et al.* [18] investigated character-based Transformer models in Neural Machine Translation, highlighting their robustness and superior performance in noisy conditions. Additionally, Banar *et al.* [19] introduced CharTransformer, achieving translation accuracy

comparable to the Transformer with a 34% speed increase. Boukkouri *et al.* [13] and Xue *et al.* [20] revamped BERT and T5's embedding layers [21] for character-based and byte-based architectures, respectively, while Clark *et al.* [22] proposed CANINE, a comprehensive solution for handling UTF-8 characters and reducing computational complexity.

In a unique departure from existing approaches, our work pioneers the exclusive modification of the word embedding layer into a character-based structure for a pre-trained model. This distinctive aspect involves the alignment of only that layer with the word embedding matrix, without necessitating comprehensive retraining of the entire model. Furthermore, we extend our contributions by proposing different alignment strategies explicitly designed to bolster robustness in noisy environments, further distinguishing our work in this domain.

III. Vocabulary Embedding Alignment

BERT, with its vocabulary structured around words and subwords, encounters challenges in noisy environments, leading to the presence of numerous unknown tokens that adversely affect its performance. In response to this, an effective strategy involves the creation of a dedicated embedding layer for vocabularies, offering the flexibility to use either character- or byte-based inputs as visualized in Fig 1. This modification ensures that the model navigates noisy data without encountering unknown tokens, preserving its contextual understanding.

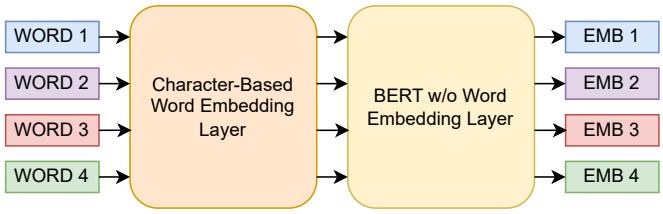


Fig. 1: BERT with character-based word embedding layer. Note that the positional and token-type encodings are omitted here.

While substituting the standard embedding layer with a character-based alternative proves beneficial, the extensive training time required for the entire model hinders efficiency for specific tasks. To address this, our approach involves aligning the output of the character-based embedding layer with that of the original embedding layer. This strategic alignment not only mitigates the time-consuming nature of training the entire model but also allows for a streamlined adaptation to noisy data, optimizing the model's performance in challenging environments.

A. Character-Based Alignment

The innovative approach presented in this study, as introduced by Alejandro *et al.* [12], draws inspiration from the techniques employed in CharacterBert [13]. In this methodology, the authors leverage the same embedding layer structure utilized in CharacterBert, comprising a character embedding layer, a convolutional neural network (CNN), and a highway network. Notably, this character-based embedding layer incorporates specialized vectors tailored for distinct characters, including `[PAD]`, `[CLS]`, `[SEP]`, `[MASK]`, and `[UNK]`. The tokenization process involves breaking down a word into its constituent characters, which are then processed through the character embedding layer. This layer, functioning as a conventional embedding layer, generates embedding vectors corresponding to each character. Subsequently, these vectors undergo processing through the character CNN, which is composed of both a convolutional neural network and a highway network. Ultimately, the output yields the word embedding. For a visual representation of the flow from a word to its embedding vector using this character-based embedding layer, refer to Fig. 2. This intricate yet effective architecture showcases the thoughtful integration of character-based embeddings, laying the groundwork for enhanced linguistic understanding.

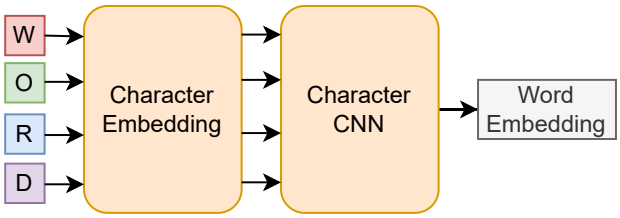


Fig. 2: Character-based word embedding layer's architecture with character-based alignment.

B. Character-Based Alignment with Additional Training

The character-based embedding layer stands out as a potent tool for enabling BERT to exhibit robust generalization, particularly in the presence of noisy data where the issue of unknown tokens is effectively mitigated. Building upon this foundation, our approach draws inspiration from the principles of adversarial training [14], [23] to further enhance the resilience of the embedding layer. The extension of training involves the deliberate introduction of noise into the vocabulary, with a focus on two distinct types: random noise and similar noise. Table I visualizes these types of noises which will be explained in the following.

1) Random Noise: Under the paradigm of random noise, a strategic substitution of some characters within each vocabulary occurs, with these characters being randomly selected during every epoch. While this technique introduces an element of unpredictability, caution is exercised to strike a balance, as an excessive replacement of characters might lead the embedding layer to forget the nuances of the original vocabulary.

2) Similar Noise: Certain domains, such as those associated with human trafficking-related posts, witness adversaries manipulating text by substituting English letters with visually similar Greek or unconventional characters. In response, relying solely on character-based alignment proves insufficient. Here, we introduce the concept of similar noise, where characters within each vocabulary undergo replacement with others that visually resemble them. To identify visually similar characters, we employ a unique approach: converting characters to images

TABLE I: Example of random and similar noises injected into a sample of the WikiText dataset [24]. Noticeably, random characters are selected for the random noise and visually similar characters are used for the similar noise. Note that the noises are in red.

Original text: The game began development in 2010 , carrying over a large portion of the work done on Valkyria Chronicles II . While it retained the standard features of the series , it also underwent multiple adjustments , such as making the game more forgiving for series newcomers . Character designer Raita Honjou and composer Hitoshi Sakimoto both returned from previous entries , along with Valkyria Chronicles II director Takeshi Ozawa . A large team of writers handled the script . The game 's opening theme was sung by May 'n .
Random noise: The game began develMpment in 2010 , carrying over a la3ge portion of the work dene on Valkyria Chronicles II . While it retained the Ytandard features of the series , it also underwent multiple adjustments , such as m+king the rame mory forgiving for series newcomers . Character αesigner Raita Honjou and composer Hitoshi Sakimoto both returned from previous entries , along with Valkyria Chroni8les II director Uakeshi Ozawa . A large team of qriters handled the script . The game 's opening theme was sung by May 'n .
Similar noise: The game began developMent in 2010 , carrying over a large portion of the work done on Valkyria Chronicles II . While it retained the Standard features of the series , it also underwent multiple adjustments , such as making the game more forgiving for series newcomers . Character designer Raita Honjou and composer Hitoshi Sakimoto both returned from previous entries , along with Valkyria Chronicles II director Takeshi Ozawa . A large team of writers handled the script . The game 's opening theme was sung by May 'n .

and leveraging a state-of-the-art computer vision model to generate embeddings for these characters. A subsequent mapper is created to establish a link between a character and its corresponding embedding vector. Through this mapping, we can identify neighbors of a character by evaluating the cosine similarity between their respective vectors, providing a nuanced solution to handle visually similar noise in challenging text environments.

C. Character's Embedding-Based Alignment

This innovative approach serves as an alternative solution specifically tailored to address the challenges posed by visually similar noise in text, as highlighted earlier. Our objective is to ensure that the embedding vectors of two visually similar characters exhibit a high degree of similarity. To achieve this, we employ the character-embedding mapper, a technique previously utilized in our previous approach. However, in this instance, we streamline the process by incorporating a dense layer, strategically employed to reduce the dimensionality of the embedding vectors.

The subsequent steps in this approach closely mirror those of the other methodologies post the character embedding layer. By introducing this refinement, we anticipate that two characters sharing visual similarities will yield corresponding embedding vectors that closely align. The workflow of this enhanced embedding layer is elucidated in Fig. 3, providing a visual representation of the intricate process. Through this nuanced modification, our approach demonstrates a commitment to tackling the intricacies of visually-similar noise, ensuring a more robust and reliable representation of characters in the embedding space.

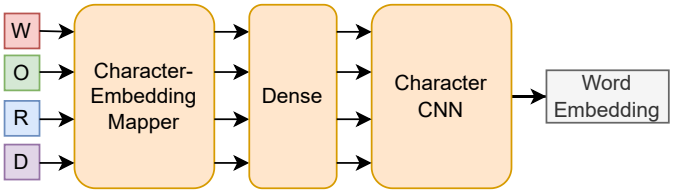


Fig. 3: Character-based word embedding layer's architecture with character's embedding-based alignment.

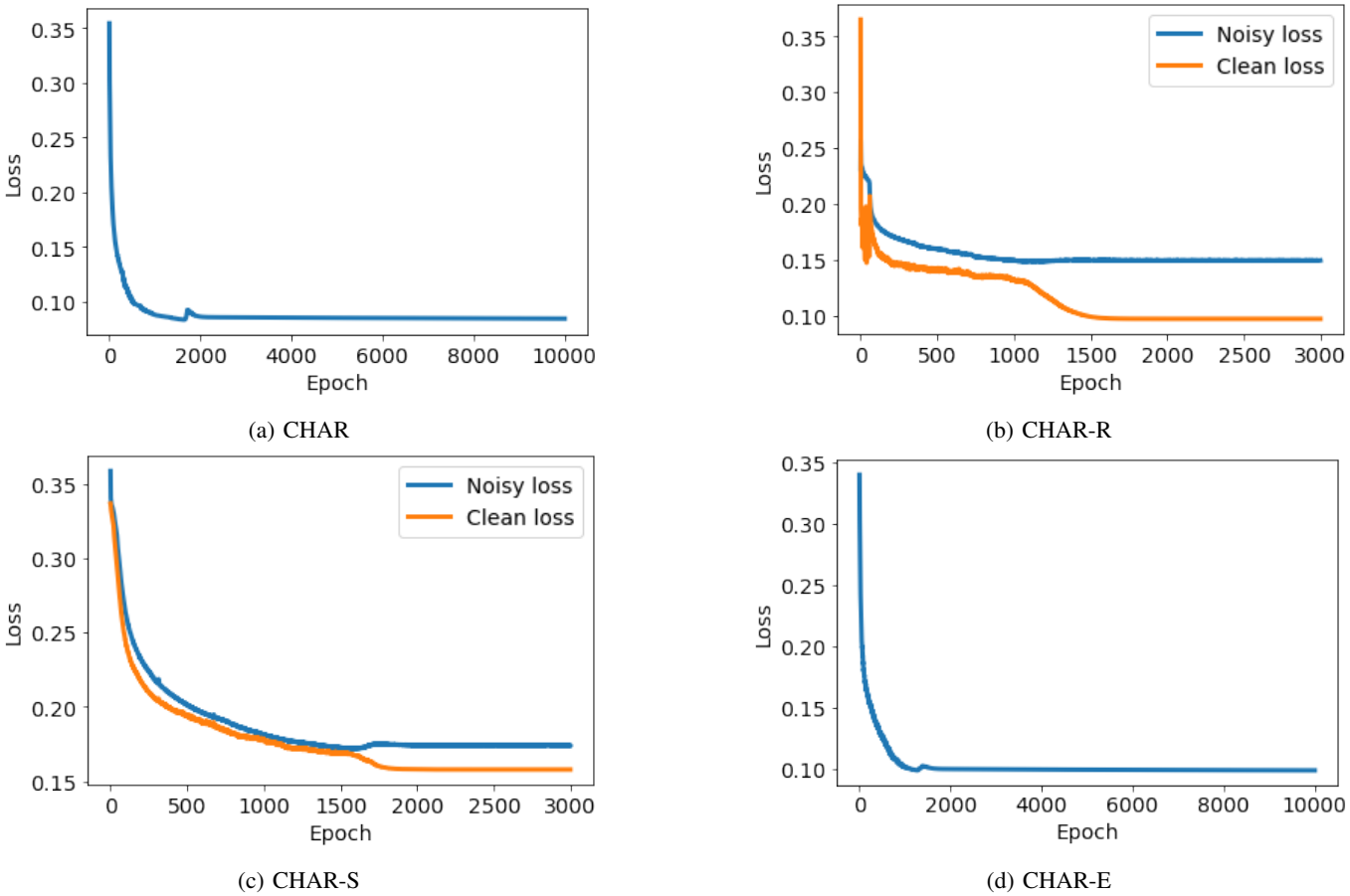
IV. Experiments and Results

We conducted several experiments to demonstrate that our approach is successful; the details of how alignment is characterized, the tokenizer used, the experimental setting, and the dataset used are discussed next, followed by the analysis of the results.

A. Approaches and Alignment Details

We aim to discern the effectiveness of different approaches across varied situations, and as such, we conduct an evaluation involving four distinct approaches outlined in the preceding section. In addition, we include a baseline represented by vanilla BERT. The alignment details for each approach are elucidated as follows:

- **BERT baseline (BERT):** This approach serves as the baseline, employing vanilla BERT without any alignment adjustments.
- **Character-based alignment (CHAR):** In this approach, we align the embedding layer with BERT's word embeddings over 10^4 epochs. The alignment process employs the additive Euclidean-cosine error function, as defined in [12]:
$$|x_1 - x_2| + (1 - \cos(x_1, x_2)),$$
where x_1 and x_2 denote two vectors, and $\cos(\cdot, \cdot)$ represents a cosine similarity function.
- **Character-based alignment with random-noise addition training (CHAR-R):** This approach involves retraining the layer from the character-based alignment by substituting one character in each vocabulary every epoch.



(a) CHAR

(b) CHAR-R

(c) CHAR-S

(d) CHAR-E

Fig. 4: Losses over epochs during the training for each alignment.

The retraining process spans 3000 epochs, utilizing the additive Euclidean-cosine error function.

- **Character-based alignment with similar-noise addition training (CHAR-S):** The layer from the character-based alignment undergoes retraining with the addition of noise. Specifically, one character in each vocabulary is replaced every epoch with a visually similar character randomly selected from five neighbors. The retraining duration spans 3000 epochs, employing the additive Euclidean-cosine error function.
- **Character's embedding-based alignment (CHAR-E):** In this approach, the model is trained using character embeddings obtained from the pretrained Vision Transformer [15], resulting in embeddings with a dimensionality of 768.

This comprehensive evaluation encompasses various alignment strategies, each tailored to address specific considerations, thereby providing a nuanced understanding of their performance across different scenarios.

Fig. 4 provides a comprehensive overview of the training process, capturing the loss for each epoch. Notably, all training instances, including CHAR-R and CHAR-S, showcase convergence at specific loss values. To delve deeper into the training dynamics, we meticulously monitor both noisy and clean vocabulary for CHAR-R and CHAR-S. Remarkably, the losses for these additional training scenarios also exhibit convergence, further affirming the effectiveness of the training methodologies employed. This convergence phenomenon serves as a strong indicator of the thorough and successful training of all approaches under consideration, emphasizing the robustness and reliability of the employed training mechanisms.

B. Tokenizers

The BERT baseline employs its native BERT tokenizer. In contrast, the other approaches deviate from this method, as it becomes impractical to utilize the same tokenizer due to the persistence of unknown tokens. In light of this, we opt for a customized tokenizer that employs white spaces to delineate tokens for the alternative approaches. This adaptation is crucial to ensure effective tokenization and avoid the issue of unknown tokens, providing a tailored solution that aligns with the specific requirements of each approach. The utilization of a customized tokenizer underscores the need for a nuanced and approach-specific preprocessing step, acknowledging the intricacies associated with different alignment strategies.

C. Evironments

In our pursuit of a comprehensive formal evaluation, we construct three distinct environments, each designed to scrutinize

the approaches across varying scenarios:

- **Clean environment (CLEAN):** This environment is simply the original text.
- **Random-noisy environment (RANDOM):** Introducing an element of randomness, this environment involves the substitution of some characters within the text with random noises. The randomness injected serves to emulate the unpredictable nature of noise encountered in real-world scenarios.
- **Similar-noisy environment (SIMILAR):** This environment simulates the presence of visually similar noise, wherein some characters in the text undergo substitution with counterparts possessing visual similarities. This mirrors scenarios where adversaries manipulate text by replacing characters with visually akin alternatives, a common challenge in various real-world applications.

Through the meticulous design of these environments, we aim to subject the evaluated approaches to diverse and representative scenarios, enabling a robust assessment of their efficacy across varying degrees of noise and environmental complexities. This multifaceted evaluation framework enhances our ability to discern the approaches' adaptability and performance under realistic conditions, providing valuable insights for practical deployment.

D. Datasets

In our study, we leverage two diverse datasets, each offering unique insights into natural language processing and classification tasks. The choice of datasets aims to encompass a broad spectrum of challenges and applications within the realm of machine learning and text analysis.

- **WikiText dataset [24]:** serves as a comprehensive corpus, drawing from a vast collection of Wikipedia articles. It spans a wide range of topics, providing a rich and diverse set of textual data for language modeling and understanding. The dataset is often employed in tasks such as language modeling, text generation, and other natural language processing endeavors. This dataset is organized into multiple versions, with varying sizes to accommodate different research requirements. We use the wiki-text-2-raw-v1 version and the train split that includes 36700 samples.
- **AG News Dataset [25]:** In contrast to the broad scope of WikiText, the AG News dataset is tailored for a specific task – news categorization. This dataset focuses on classifying news articles into predefined categories, making it well-suited for text classification and sentiment analysis applications. This dataset comprises news articles collected from the AG's corpus of news articles, encompassing various topics such as world, sports, business, and science. With labeled categories assigned to each article, this dataset facilitates supervised learning tasks, allowing models to learn and generalize patterns in news text across different domains. We use the test split that includes 7600 samples.
- **Yelp Dataset [26]:** This dataset consists of reviews sourced from the Yelp platform, including text data along with associated labels or ratings. Widely used for sentiment analysis, text classification, and related NLP tasks, it offers a diverse range of user-generated content and associated ratings, providing a valuable resource for model training and evaluation. For our experiment, we utilize the test set comprising 50,000 samples.
- **Human Trafficking (HT) Dataset [27]–[29]:** This dataset serves as a specialized resource tailored specifically for the classification of human trafficking activities. It comprises a collection of text-based advertisements sourced exclusively from SkipTheGames, a platform notorious for hosting illicit content related to human exploitation. By providing access to such granular and context-rich data, the dataset enables researchers, law enforcement agencies, and policymakers to develop and refine classification models and analytical tools aimed at identifying, understanding, and ultimately combating instances of human trafficking.

The combination of WikiText's linguistic diversity, AG News' categorical focus, Yelp's user-generated content and HT's noisy text offers a well-rounded foundation for investigating the capabilities of models across various text analysis tasks.

E. Results

To assess the effectiveness of the proposed approaches, we employ the next sentence prediction (NSP) task, a suitable evaluation metric. This choice is motivated by the convenience of utilizing the BERT baseline for alignment directly, eliminating the need to train the entire model on any downstream task. The NSP task allows for a straightforward comparison between the BERT baseline and the alternative approaches without the need for extensive model retraining. It is noteworthy that we refrain from utilizing the mask language model (MLM) task for evaluation. The primary reason behind this decision lies in the disparity of labels between the BERT baseline and the other approaches, rendering the MLM task incomparable across different alignment strategies. This deliberate selection of the NSP task ensures a consistent and fair evaluation framework while emphasizing the importance of task compatibility in assessing the performance of the proposed approaches.

Table II presents the accuracy results on the Wiki dataset for the NSP task across different environments and embedding layers. The environments include CLEAN, RANDOM (with noise percentages of 5% and 10%), and SIMILAR (with noise percentages of 5% and 10%). The types of embedding layers considered are BERT Baseline, CHAR, CHAR-R, CHAR-S, and CHAR-E. The reported accuracy values are the averages of three runs, except for the CLEAN environment, which is static. In the CLEAN environment, the BERT Baseline achieves the highest accuracy of 94.20%. As the noise percentage increases in the RANDOM environment, the accuracy generally decreases for all embedding layers, with the BERT Baseline consistently outperforming the others. In the SIMILAR environment, CHAR-E achieves the highest accuracy at 72.82% and 56.76% for 5% and 10% noise percentage, respectively, and CHAR-E

TABLE II: Accuracy on the Wiki, Agnews, Yelp and HT datasets for NSP task for several types of embedding layers under various environments. The results are the average of three runs, except for the CLEAN environment because it is static.

Dataset	Environment	Noise Percentage %	BERT Baseline %	CHAR %	CHAR-R %	CHAR-S %	CHAR-E %
Wiki	CLEAN	0	**94.20**	93.76	87.79	73.85	88.65
	RANDOM	5	**82.34**	71.54	67.25	54.76	70.60
	RANDOM	10	**61.38**	55.26	53.75	50.86	55.55
	SIMILAR	5	69.21	71.94	60.38	56.06	**72.82**
	SIMILAR	10	54.04	55.88	51.35	51.34	**56.76**
Agnews	CLEAN	0	**61.93**	59.24	57.63	53.55	59.81
	RANDOM	5	**57.25**	53.66	52.7	50.99	54.31
	SIMILAR	5	53.88	52.35	51.55	51.24	**54.47**
Yelp	CLEAN	0	**65.11**	58.67	56.47	51.20	60.36
	RANDOM	5	**53.72**	52.26	51.50	50.39	52.69
	SIMILAR	5	**53.13**	52.03	51.29	50.22	**53.13**
HT	CLEAN	0	**55.7**	53.18	52.28	50.97	54.39
	RANDOM	5	**51.72**	51.17	50.76	50.23	51.22
	SIMILAR	5	**51.57**	51.17	50.61	50.55	**51.57**

and CHAR-R outperform other embeddings for a 10% noise percentage. These results highlight the performance variations of different embedding layers under diverse noise conditions and environments.

Furthermore, on the Agnews dataset, we do not use the 10% noise percentage because the accuracy is very low for all the models. Because BERT was not trained with this dataset, all the accuracy was not very high. In the CLEAN environment, the BERT Baseline achieves the highest accuracy at 61.93%. As noise is introduced in the RANDOM environment, the accuracy decreases for all embedding layers, with the BERT Baseline consistently demonstrating the highest accuracy. In the SIMILAR environment, CHAR-E stands out with the highest accuracy of 54.47%. Overall, these results provide insights into the performance variations of different embedding layers across distinct noise conditions and environments in the Agnews dataset.

Additionally, on the Yelp and HT datasets, across different environments such as CLEAN, RANDOM, and SIMILAR, BERT Baseline consistently outperforms other embedding layers, with its highest accuracy in the CLEAN environment. CHAR-E embedding layer shows competitive performance in the SIMILAR environment, matching BERT Baseline accuracy.

It is evident from our observations that CHAR-R and CHAR-S consistently exhibit suboptimal performance across various scenarios, encompassing different environments and datasets. Despite aligning the embedding layers of these models with noisy vocabulary through the substitution of a single character for each vocabulary, these approaches, namely CHAR-R and CHAR-S, fail to deliver satisfactory results, particularly in noisy environments. The specialized training designed for CHAR-R and CHAR-S, which focuses on aligning with noisy vocabulary, does not prove to be as effective as anticipated. This consistent underperformance underscores the challenges associated with aligning embedding layers with noisy vocabulary and suggests the need for further refinement or alternative strategies to enhance the robustness of these models in noisy scenarios.

V. Limiitations

We also trained our embeddings to align with the embeddings of pretrained BERT for Agnews classification. Fig. 5 shows that we trained those embedding layers until they are converging. Table III outlines accuracy results on the Agnews dataset for a classification task across different embedding layers and environments. In the CLEAN environment, BERT Baseline achieves the highest accuracy of 94.49%, outperforming other embedding layers. However, in environments with noise, particularly in RANDOM and SIMILAR scenarios, BERT Baseline maintains its superiority but experiences a decline in accuracy compared to the CLEAN environment. Overall, CHAR-E exhibits the most competitive performance among our approaches across various noisy environments.

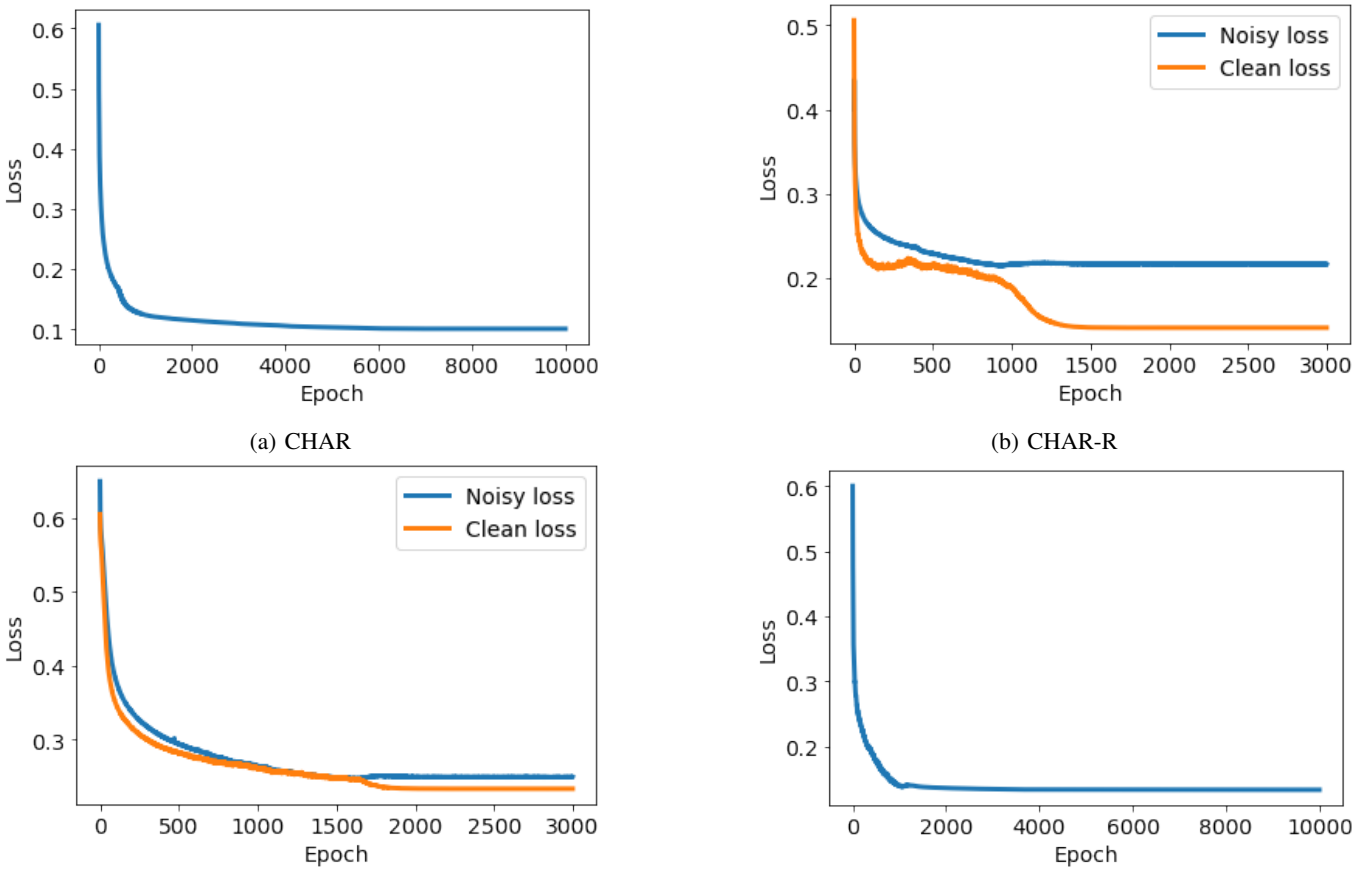
Although our method has demonstrated success with the NSP task, we are still challenged to find a solution to surpass the robustness of the BERT baseline for downstream tasks such as Agnews, where it shows remarkable resilience against noise interference.

VI. Broader Impact

Inspired by CharacterBert's success, we're redefining BERT's word embedding layer with a character-based model. This not only enhances technology but significantly reduces training time, crucial for large-scale models. Our goal is to democratize AI accessibility, fostering a more inclusive landscape, making BERT adaptable to a broader range of applications.

VII. Conclusion

Our study delves into the challenges faced by BERT or other foundational large language models in noisy environments due to unknown tokens and presents innovative strategies to enhance its performance. By introducing a dedicated embedding layer for vocabularies, we enable BERT to effectively navigate noisy data without encountering unknown tokens, thus preserving its contextual understanding. Through various alignment approaches, including character-based alignment and character's



(a) CHAR

(b) CHAR-R

(c) CHAR-S

(d) CHAR-E

Fig. 5: Losses over epochs during the training for each alignment with the Ag News BERT-based classifier's word embeddings

TABLE III: Accuracy on the Agnews dataset for classification task for several types of embedding layers under various environments. The results are the average of three runs, except for the CLEAN environment because it is static.

Environment	Noise Percentage %	BERT Baseline %	CHAR %	CHAR-R %	CHAR-S %	CHAR-E %
CLEAN	0	**94.49**	77.99	84.87	82.06	82.91
RANDOM	5	**92.37**	64.34	79.57	72.12	72.98
RANDOM	10	**87.23**	50.04	70.84	57.84	54.49
SIMILAR	5	**92.69**	61.93	67.36	77.68	73.31
SIMILAR	10	**87.54**	48.19	49.32	70.56	60.35

embedding-based alignment, we strive to optimize BERT's performance across different scenarios.

Our experiments on the WikiText, AG News, and Yelp datasets reveal nuanced insights into the efficacy of these alignment strategies. While the BERT baseline consistently demonstrates strong performance, particularly in clean environments, our proposed approaches exhibit competitive performance, especially in scenarios with noise. Notably, the character's embedding-based alignment approach shows promising results, showcasing resilience against noise interference, particularly in the presence of visually similar noise.

Despite the success observed in the NSP task, our approaches face limitations in surpassing the robustness of the BERT baseline for downstream tasks such as classification, as evidenced by our experiments on the Agnews dataset. Further refinement and exploration of alternative strategies are necessary to address this challenge and enhance the adaptability of our approaches across diverse downstream tasks. Moreover, our approaches should also work on other large language models, and we need to explore them in the future.

Our study sheds light on the importance of addressing the challenges posed by noisy environments in NLP tasks and presents innovative strategies to enhance model robustness. While our approaches show promising results, continued research and development are crucial to further improve their effectiveness and applicability across a wide range of real-world scenarios.

ACKNOWLEDGEMENTS

Part of this work was funded by the National Science Foundation under grants CNS-2210091, CHE-1905043, and CNS-2136961.

REFERENCES

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

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[3] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[4] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," pp. 2227–2237, 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[9] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models, 2023," *URL https://arxiv.org/abs/2307.09288*, 2023.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] A. R. Perez, K. Sooksatra, P. Rivas, E. Q. Caballero, J. S. Turek, G. Bichler, T. Cerny, L. Giddens, and S. Petter, "An empirical analysis towards replacing vocabulary-rigid embeddings by a vocabulary-free mechanism," in *LatinX in AI Workshop at ICML 2023 (Regular Deadline)*, 2023.

[13] H. E. Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, and J. Tsujii, "Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters," *arXiv preprint arXiv:2010.10392*, 2020.

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] R. Al-Rfou, D. Choe, N. Constant, M. Guo, and L. Jones, "Character-level language modeling with deeper self-attention," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3159–3166.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[18] R. Gupta, L. Besacier, M. Dymetman, and M. Gallé, "Character-based nmt with transformer," *arXiv preprint arXiv:1911.04997*, 2019.

[19] N. Banar, W. Daelemans, and M. Kestemont, "Character-level transformer-based neural machine translation," in *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 2020, pp. 149–156.

[20] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, "Byt5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, 2022.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[22] J. H. Clark, D. Garrette, I. Turc, and J. Wieting, "Canine: Pre-training an efficient tokenization-free encoder for language representation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 73–91, 2022.

[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[24] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," 2016.

[25] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *NIPS*, 2015.

[26] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[27] L. Giddens, S. Petter, G. Bichler, P. Rivas, M. H. Fullilove, and T. Cerny, "Navigating an interdisciplinary approach to cybercrime research," *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2023.

[28] A. P. A. Terron, J. Y. Salazar, P. Rivas, and E. Q. C. A. R. Perez, "Task-specific or task-agnostic? a statistical inquiry into bert for human trafficking risk prediction," *LXAI Workshop at NeurIPS*, 2023.

[29] A. R. Perez and P. Rivas, "Combatting human trafficking in the cyberspace: A natural language processing-based methodology to analyze the language in online advertisements," *arXiv preprint arXiv:2311.13118*, 2023.