Distributed Text Representations Using Transformers for Noisy Written Language

Alejandro Rodriguez Perez, Pablo Rivas Perea and **Gissella Bejarano Nicho**
Baylor University
{alejandro_rodriguez4,pablo_rivas,gissella_bejaranonic}
@baylor.edu

Abstract

This work proposes a methodology to derive latent representations for highly noisy text. Traditionally in Natural Language Processing systems, methods rely on words as the core components of a text. Unlike those, we propose a character-based approach to be robust against our target texts' high syntactical noise. We propose pre-training a Transformer model (BERT) on different, general-purpose language tasks and using the pre-trained model to obtain a representation for an input text. Weights are transferred from one task in the pipeline to the other. Instead of tokenizing the text on a word or sub-word basis, we propose considering the text's characters as tokens. The ultimate goal is that the representations produced prove useful for other downstream tasks on the data, such as criminal activity in marketplace platforms.

1 Introduction

Much work has been devoted to deriving sentence distributed representations (Conneau et al., 2017; Cer et al., 2018; Li et al., 2020). However, all of these approaches suffer from limitations. In the informal contexts we are attempting to deal with, written language tends to be altered to fulfill the writer's need to convey emotions and personality. Also, these contexts are prone to misspelling, and new words and acronyms appear all the time. Usually, non-Latin characters and emojis are used. Word-based natural language processing could hardly cope with these circumstances, as the models are restricted to a finite vocabulary, and handling the variety of scenarios that can appear becomes impractical. These limitations have been highlighted (Tay et al., 2021; Clark et al., 2021; Xue et al., 2021; Sun et al., 2020), for instance, in the context of Language Modeling (Jozefowicz et al., 2016; Kim et al., 2016; Ma et al., 2020; Boukkouri et al., 2020) and Neural Machine Translation (Luong and Manning, 2016), and several alternatives have been proposed. Using models that rely on character information is a core idea behind these proposals.

Another fundamental limitation to obtaining good vector representations for texts is the lack of labeled data. Sentence encoders, for instance, usually rely at least partially on supervised goals to be trained (Cer et al., 2018; Conneau et al., 2017). But generally, the amount of labeled data for the specific purpose is limited and insufficient to optimize such architectures properly. That is why it has become increasingly crucial to pre-train models on general-purpose tasks, either unsupervised or where a lot of labeled data is available (Brown et al., 2020; Devlin et al., 2018), and then fine-tune the weights on a specific downstream task. As an alternative, Zhang et al. use a purely unsupervised method.

The main contribution of this paper is a methodology to derive distributed representations of noisy text using BERT (Devlin et al., 2018). Today, we count with limited (and unlabeled) data. Thus we limit the scope of the current research to propose two unsupervised goals and an artificial supervised task to obtain a pre-trained BERT model. This could help derive vector representation of the input text that will be used in future classification tasks such as criminal activity recognition.

The rest of the document is organized as follows. Section 2 describe the pre-training tasks we propose. Next, Section 3 shows some preliminary results from training this architecture in a reduced scope, and Section 4 describes the path we envision to further developing and testing our proposal. Finally, 5 offers the conclusions of this partial work.

2 Methods

To compensate for the problems that arise when using a word-based approach, we propose to tokenize the text on a character basis. The first advantage is that this results in a much smaller vocabulary.

The second advantage is that taking a character-based input may make the representations much more robust to several phenomena that may appear in everyday contexts (Boukkouri et al., 2020; Ma et al., 2020): intentional syntactical variations, neologisms, use of characters in different alphabets, and spelling errors.

We propose pre-training a BERT model with several unsupervised tasks. The same BERT model is used in each pre-training task, and a head is added after it to perform the corresponding task. The only change to the original architecture is using the positional embeddings defined by (Vaswani et al., 2017) to scale to larger sequences without adding more parameters to the model. The best parameters of BERT during a phase (according to validation) are used as initial parameters for the next phase. After pre-training, the hidden state of the last BERT layer corresponding to the `[CLS]` token of the input text is chosen as its distributed representation. We call it the **text embedding**. Figure 1 depicts the pipeline and models. The three tasks considered are described in more detail below.

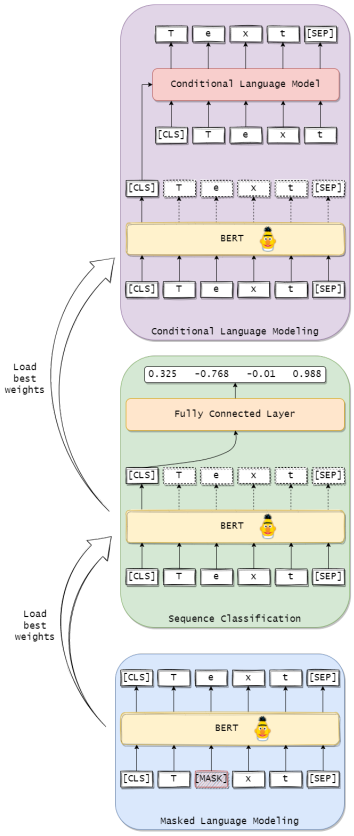


Figure 1: The pre-train pipeline consists of the three steps that can be visualized from bottom up. The weights of the BERT model that achieved better results in the validation set are transferred to the next step.

First, the BERT model is pre-trained in the Masked Language Modeling (MLM) task proposed in (Devlin et al., 2018). Here, some tokens (characters) in the input are masked, and the model is trained to predict the correct character in the masked positions. The same strategy used in Devlin et al. is used to mask tokens.

The second is a Sequence Classification (SC) task that takes the text embedding produced by the BERT model to classify the input text. To solve this task, we added a head consisting of a Fully Connected layer that maps the text embedding to the number of classes. We propose to use the following classes. **CORRECT**: actual text samples from the source data; **ALTERED**: text with some fraction of the characters randomly replaced; **MIXED**: text that has been combined with another piece(s) of text; and **CROPPED** text that has been cropped. The intuition behind the artificial classes is that the model learns to recognize text that has been corrupted. Note how examples of these classes can be constructed without supervision. The **CORRECT** class spans 50% of the training examples, and the other three classes are sampled with equal probability (∼16.7%).

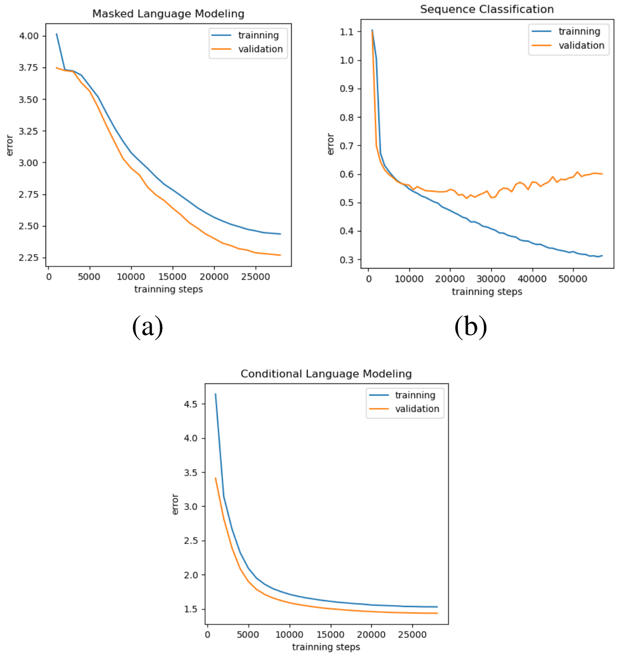
Finally, the model is trained in a Conditional Language Modeling (CLM) task, where the context in which the model conditions the output is the text embedding. Other transformer models have chosen Language Modeling as an unsupervised pre-training task (Radford et al., 2018). For Language Modeling, we use an LSTM-based (Hochreiter and Schmidhuber, 1997) decoder head. To condition the output on the text embedding, each token embedding in the input text is added to the text embedding, and Layer Normalization (Ba et al., 2016) is applied to the result.

3 Experiments

The dataset used was extracted from publicly available Consumer-to-Consumer marketplaces. It contains 2775 different characters. Text samples were cropped and padded to a length of 1024.

We ran an experiment with a small setup to observe the ability of the model to learn the proposed tasks. The size of the BERT model was reduced to 6 hidden layers, 6 attention heads, 500 as intermediate size, and 300 as final embedding size. The Conditional Language Model used 3 stacked LSTMs with hidden size 200 and output size 100 before passing through the final Fully Connected classifier with Softmax that predicts a character in the vocabulary. The three tasks were trained using Categorical Cross-Entropy loss.

Figure 2 shows the error (loss) curves in training and validation sets as a function of the training steps of the pre-train process.



(a)

(b)

(c)

Figure 2: Train and validation error curves as a function of training steps for the three pipelined pre-train tasks. (a) Masked Language Modeling. (b) Sequence Classification. (c) Conditional Language Modeling.

We can observe the training error decrease as more training steps are conducted. In the case of the MLM task, the model seems to tolerate more training. The SC task clearly starts overfitting after a few training steps. Finally, the CLM task training exhibits convergence.

Figure 3 shows a dimensionality reduction plot of the text embeddings for the entire dataset. The method used for dimensionality reduction was UMAP, and the colors represent the classes of the SC task.

Here we observe how the model can distinguish the class that has been added random noise, whereas the other three classes appear interleaved. We consider two possible explanations. First, the three classes **CORRECT**, **CROPPED**, and **MIXED** have many samples cropped [1], and the representations may be reflecting this common pattern. Second, because the changes in these classes affect only the structure of the text, it may be the case that the model is sensitive to the character's distribution but not to the cohesion and completeness of pieces of text.

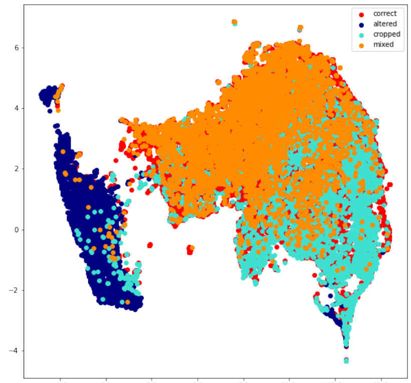


Figure 3: Text embeddings reduced to two dimensions and colored according the classification label in the Sequence Classification task.

4 Future Work

Several paths may be further explored. First, it is paramount to evaluate our pre-trained model on other relevant downstream tasks and compare the performance of standard word or sub-word-based pre-trained language models with ours. Second, it remains for future work to explore alternatives to the models used, like using a text embedding architecture different from BERT, or trying a better Conditional Language Modeling architecture, like adapting the more recent TransformerXL (Dai et al., 2019) to our use case.

5 Conclusions

This work proposed a pipeline to pre-train a BERT-based feature extractor for noisy texts that rely on character representations rather than words or sub-words. Some preliminary results were presented, but there is still plenty to work on. We expect that the resulting representations capture features of the texts that are useful for other downstream tasks in which we will be able to compare our proposal with word and sub-word-based approaches. Mainly, we will be focusing on detecting criminal activity on the Craiglist website based partly on processing the posts' text content.

[1] Either because they are longer than 1024 characters or naturally from the CROPPED class

Acknowledgements

This work was funded in part by the National Science Foundation under grants CNS-2136961 and CNS-2210091.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Junichi Tsujii. 2020. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters. *arXiv preprint arXiv:2010.10392*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2021. Canine: Pre-training an efficient tokenization-free encoder for language representation. *arXiv preprint arXiv:2103.06874*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Minh-Thang Luong and Christopher D Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv preprint arXiv:1604.00788*.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Charbert: Character-aware pre-trained language model. *arXiv preprint arXiv:2011.01513*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training (2018).

Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint arXiv:2105.13626*.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. *arXiv preprint arXiv:2009.12061*.