A Simple Yet Effective Hybrid Pre-trained Language Model for Unsupervised Sentence Acceptability Prediction

Yang Zhao and **Issei Yoshida**
IBM Research - Tokyo
19-21 Nihonbashi Hakozaki-cho, Chuo City, Tokyo, 103-8510 Japan
yangzhao@ibm.com, issei@jp.ibm.com

Abstract

Sentence acceptability judgment assesses to what degree a sentence is acceptable to native speakers of the language. Most unsupervised prediction approaches rely on a language model to obtain the likelihood of a sentence that reflects acceptability. However, two problems exist: first, low-frequency words would have a significant negative impact on the sentence likelihood derived from the language model; second, when it comes to multiple domains, the language model needs to be trained on domain-specific text for domain adaptation. To address both problems, we propose a simple method that substitutes Part-of-Speech (POS) tags for low-frequency words in sentences used for continual training of masked language models. Experimental results show that our word-tag-hybrid BERT model brings improvement on both a sentence acceptability benchmark and a cross-domain sentence acceptability evaluation corpus. Furthermore, our annotated cross-domain sentence acceptability evaluation corpus would benefit future research.

1 Introduction

Sentence acceptability judgment aims to assess to what degree a sentence is acceptable to native speakers of the English Language. An effective sentence acceptability scorer is beneficial for many applications, such as ranking outputs from a dialogue system to pick the most fluent and natural response, or being used as an English fluency checker to help identify grammar issues.

Previous unsupervised works mainly exploit either ngram-based or neural-based language model's Negative Cross Entropy (NCE) (Kann et al., 2018) and its variants such as Syntactic Log-Odds Ratio (SLOR) (Pauls and Klein, 2012; Lau et al., 2017) to obtain the sentence acceptability score. However, two problems exist when employing a language model to estimate sentence acceptability: 1) First, low-frequency words greatly impact a sentence probability (or perplexity) from a language model. Although subword tokenizers attempt to alleviate this problem by splitting rare or unknown words into subwords, some subwords are still infrequent in their original context, leading to a considerable increase in sentence-level perplexity. 2) Second, cross-domain adaptation inefficiency. Many terminologies in specific domains affect sentence acceptability prediction and it is often a common practice to select in-domain text to do continual pretraining of the language model, which is time-consuming and inefficient.

To address the aforementioned two problems, we present a simple frequency-based method (Section 2) to substitute low-frequency words with the English-specific Part-Of-Speech (POS) tag, XPOS, in sentences that are used for continual pretraining of the BERT model. Notably, we are interested in the following research questions, RQ1: *how much percentage of low-frequency words should be substituted to obtain the best performance on sentence acceptability judgment task?* RQ2: *Can we train one model tackling cross-domain sentence acceptability tasks to avoid pretraining for each domain?* The experimental results demonstrate that the word-tag-hybrid BERT improves the correlation with human rating on the English sentence acceptability benchmark. To establish sentence acceptability evaluation in cross-domains and to overcome the lack of evaluation corpus, we annotated 3,000 pairs of acceptable and unacceptable sentences for financial, law, and biomedical domains. The proposed hybrid BERT outperforms the baselines upon the cross-domain sentence acceptability benchmark.

The contributions of this work are as follows: (1) We investigate a word-tag-hybrid training schema for a masked language model with a adjustable substitution rate. The experimental results validate the effectiveness of the proposed method on sentence acceptability evaluation benchmarks; (2) we

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 194–201
November 20–23, 2022.

annotated 3,000 pairs[1] of acceptable and unacceptable sentences in the financial, law, and biomedical domains.

2 Methodology

We herein describe how to construct a training data set for our word-tag-hybrid BERT model. Our strategy is to replace low-frequency words in a sentence of the corpus with more abstract, broader tags to mitigate the issue of low frequency. We give a detail of each step of construction and assume that the corpus C is a (large) set of sentences that is available for masked language model training.

Step 1 is to build a set of low-frequency words V_{Low}. To identify which words should be included in V_{Low}, we use the whole of Wikipedia entries (say W) for the target language. We apply a standard NLP pipeline to split each entry into sentences and tokenize each sentence to get a list of words of the sentence. Let V be the set of all distinct words in W. Then, inspired by the idea of "frequency binning" in Mikolov et al. (2011), we sort all the obtained words in descending order according to their occurrence frequencies in W, and assign an index for each word from 1 to $|V|$, as shown in Figure 1. The sum of all words' frequencies is

$$F = \sum_{i=1}^{|V|} f_i, \tag{1}$$

where f_i is the frequency of i-th word in V, so $f_1 \geq f_2 \geq \cdots$. Then, we determine the "boundary" word with index m with respect to the substitution rate α (a fixed threshold between 0 and 1) so that the following inequalities hold.

$$\frac{\sum_{i=1}^{m} f_i}{F} < 1 - \alpha < \frac{\sum_{i=1}^{m+1} f_i}{F} \tag{2}$$

We select all the words whose index are greater than m to build up V_{Low}.

Step 2 is to create the training data from C. For each sentence s in C, apply the same NLP pipeline in Step 1 to s to obtain a sequence of words $w_1, \cdots, w_n$ and their corresponding POS[2] tags $p_1, \cdots, p_n$, where n is the number of words in s and p_i is the POS tag of w_i. Then, we replace w_i with p_i in s if $w_i \in V_{Low}$ to yield a new sentence s'. As shown in Figure 2, when α increases, more words are replaced with their POS tags. We will use the set of all s' to do continual pre-training of masked language models.

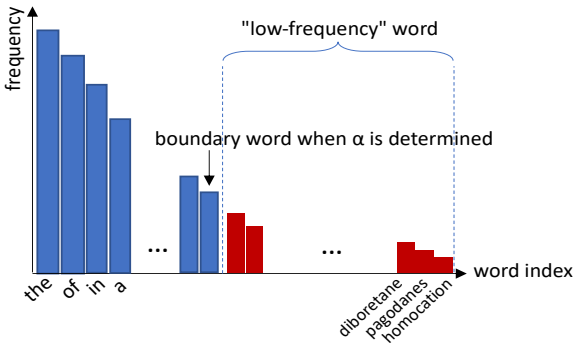

Figure 1: Words in descend order according to their frequencies.

Step 3 is to add all POS tags into the vocabulary of the masked language model to ensure that these POS tags will not be split by the subword tokenizer during masked language model training.

3 Experiment

3.1 Training Details

We employ bert-base-cased[3] as the base model for continual pre-training. As for the corpus C, we use the WikiText-103 Benchmark dataset[4], which is widely used in language model training. After preprocessing the raw data, we yielded 3.6 million sentences in the training set and 7.7k sentences in the validation set. The validation set is used to early stop the training. The continual pre-training uses 8 V100 GPUs. The α varies from 0.00, 0.05, 0.10, 0.15, 0.20, 0.40, 0.60, 0.80, and 1.00. We apply Stanza[5] NLP pipeline to tokenize all the sentences in the training and validation set of WikiText-103 and obtain the XPOS tag for each word and use 43 XPOS tags[6] to substitute words.

3.2 Evaluation Benchmark

We use a sentence acceptability benchmark containing 2,918 pairs of sentences and human acceptability ratings in Toutanova et al. (2016). The average rating range goes from the worst (1.0 points) to the best (3.0 points). Given a sentence, the BERT

[1] https://github.com/codenlp22/data

[2] We use XPOS, a set of language-specific part-of-speech tags, in our experiment

[3] https://huggingface.co/bert-base-cased

[4] https://blog.salesforceairesearch.com/the-wikitext-long-term-dependency-language-modeling-dataset/

[5] https://stanfordnlp.github.io/stanza/

[6] Please refer to Appendix A for the full list of XPOS.

α	sentence with word substitution
0.00	The liver is an organ only found in vertebrates which detoxifies various metabolites , synthesizes proteins and produces biochemicals necessary for digestion and growth .
0.10	The liver is an organ only found in NNS which VBZ various NNS , VBZ proteins and produces NNS necessary for NN and growth .
0.40	The NN is an NN only found in NNS which VBZ various NNS , VBZ NNS and VBZ NNS JJ for NN and NN .
1.00	DT NN VBZ DT NN RB VBN IN NNS WDT VBZ JJ NNS , VBZ NNS CC VBZ NNS JJ IN NN CC NN .

Figure 2: Sample sentences with different substitution rate α. Words are replaced with XPOS tag in red color.

model outputs a Negative Cross Entropy (NCE), i.e., the log probability normalized by sentence length. Following Kann et al. (2018), the Pearson correlation was calculated as the evaluation metric.

Domain-Specific Sentence Acceptability Corpus To overcome the lack of sentence acceptability benchmark in a specific domain where there is a significant amount of low-frequency words and terminologies, we collect 3,000 sentences from specific domains. They are respectively financial domain[7], law[8] domain, and biomedical[9] domain and there are 1,000 sentences in each domain. We take each sentence as an acceptable sentence and corrupt the acceptable sentence to construct an unacceptable sentence by using three operations as follows respective:

1. Delete: removing the ROOT node word in the dependency tree of the acceptable sentence to make an unacceptable sentence.

2. Shuffle: swapping the order of a randomly selected bigram in the acceptable sentence to make an unacceptable sentence, as Févry and Phang (2018) did.

3. Insert: randomly sampling one additional word from our constructed dataset, and then randomly insert the newly sampled word into the acceptable sentence to make an unacceptable sentence, similar to what Févry and Phang (2018) did.

We assume that each operation will make the sentence ill-formed and unnatural, which leads to three evaluation sub-datasets: (i) Deletion Dataset with 1k instances (ii) Shuffle Dataset with 200 instances (iii) Insert Dataset with 200 instances. Note that (ii) and (iii) come from the same 1k source sentence in each domain as (i) did. We only annotated a small portion of (i) to investigate other sentence corruption operations due to the annotation capacity. Then, we asked two human annotators to manually check whether the corrupted sentence does have syntactic and semantic violations by following the annotation criteria[10], similar to the one in the previous work (Warstadt et al., 2019). As a result, annotators removed a small number of invalid unacceptable sentences. Table 1 shows the statistics of the annotated data.

Delete	Financial	Law	Biomedical
# of sentences	1k	1k	1k
ave. of tokens	23.0	21.7	19.2
Shuffle	Financial	Law	Biomedical
# of sentences	200	200	200
ave. of tokens	22.5	21.9	18.1
Insert	Financial	Law	Biomedical
# of sentences	200	200	200
ave. of tokens	22.3	21.5	18.6

Table 1: Statistics of annotated corpora in financial, law, and biomedical domain.

Accuracy is used in domain-specific sentence acceptability judgment: let $\text{PPL}_{LM}(X)$ be the sentence-level perplexity of a masked language model where X is an input sentence. For a pair of acceptable sentence X_{acc} and unacceptable sentence X_{unacc}, if $\text{PPL}_{LM}(\text{X}_{acc}) < \text{PPL}_{LM}(\text{X}_{unacc})$, then the prediction is correct; otherwise, it is incorrect.

4 Result and Analysis

Table 2 shows the Pearson correlation result when training and testing the hybrid BERT model with different α. Our observations are as follows:

1. When α is set to 0.00, the BERT model is training on sentences of WikiText-103, a subset of Wikipedia article used originally for training vanilla BERT (Devlin et al., 2018). The correlation result of hybrid BERT (#3)

[7]Company's financial news
[8]Law case text from U.S. supreme court.
[9]Articles from American National Institutes of Health.
[10]Refer to appendix B for our annotation instruction.

	Pearson
#1 WP-NCE (Kann et al., 2018)	0.413
#2 Word-SLOR (Kann et al., 2018)	0.454
#3 WP-NCE hybrid w/ α = 0.00	0.442
#4 WP-NCE hybrid w/ α = 0.05	0.452
#5 WP-NCE hybrid w/ α = 0.10	**0.503**†
#6 WP-NCE hybrid w/ α = 0.15	0.468
#7 WP-NCE hybrid w/ α = 0.20	0.460
#8 WP-NCE hybrid w/ α = 0.40	0.434
#9 WP-NCE hybrid w/ α = 0.60	0.459
#10 WP-NCE hybrid w/ α = 0.80	0.434
#11 WP-NCE hybrid w/ α = 1.00	0.393

Table 2: Pearson correlation result between masked LM outputs and human ratings. WP refers to the word piece obtained by subword tokenizer; WP-NCE refers to word piece-based NCE. Best results in bold. † significantly better than #1 and #2 with $p < 0.01$, one tailed, (Diedenhofen and Musch, 2015).

improves compared with vanilla BERT (#1) but is lower than the previous best result (#2). We herein do not experiment with the SLOR because SLOR is a post-processing method of language model output while our focus is on language model output itself.

2. When α is set to 1.00, the BERT model is essentially continually training on POS tag sequences. We observed the lowest correlation performance (#11), which is because that if all words are substituted with their corresponding XPOS tags, the vocabulary size will dramatically reduce from 30k to 43, lossing rich linguistic information of words, and tag itself is too coarse-grained for sentence acceptability prediction.

3. The hybrid BERT with α equal to 0.10 (#5) correlates with human rating the best. The correlation performance drops as α increases from 0.1 to greater values, indicating that hybrid BERT with α equal to 0.10 achieves the best trade-off between words and POS tags.

To further investigate how the word-tag-hybrid BERT performs on multiple domains, we apply the word-tag-hybrid BERT with α equal to 0.1 to pairs of acceptable and unacceptable sentences in financial, law, and biomedical domains. Note that there is no training data and only three evaluation datasets. Table 3 shows the accuracy result. We observed the followings:

1. Compared to the vanilla BERT model (&3), hybrid BERT with α equal to 0.1 (&4) obtained accuracy improvements across domains, validating the effectiveness of integrating XPOS substitution in training.

2. To investigate whether the word substitution is effective or word substitution with XPOS is effective, we replace all 10% (α=0.1) low-frequency words with a special token, [UNK], in the evaluation data for each domain. (&4) v.s. (&2) as well as (&3) v.s. (&1) show that XPOS substitution is better than [UNK] substitution probably because XPOS contains richer linguistic information that is of help to sentence acceptability prediction.

3. Surprisingly, for BERT with [UNK] (&1) and hybrid BERT with [UNK] (&2), the latter shows significantly better accuracy results across domains, implying that the word-tag-hybrid training is beneficial to [UNK] substitution even there is no POS tag in testing data.

4. With respect to the shuffle operation (Table 4) and insert operation (Table 5), the overall performance of hybrid BERT is better than or comparable to that of BERT, suggesting that there is still an advantage of replacing the low-frequency words with XPOS for other type of unacceptable sentences (i.e., insert-based and shuffle-based sentences).

Due to the space limitation, we refer readers to Appendix C - case study - for an intuitive illustration of how word-tag-hybrid BERT alleviates the low-frequency effect on perplexity.

5 Related Work

There are two research lines. Ek et al. (2019) view sentence acceptability prediction as a supervised learning problem where they extracted many features such as POS tags and semantic tags to improve the LSTM prediction performance. On the other hand, (Lau et al., 2015, 2017; Kann et al., 2018) model sentence acceptability prediction as an unsupervised problem similar to ours where their focus is to transform the language model output into other variants such as SLOR. In contrast, we aim to investigate trade-offs between word and XPOS to improve language model outputs such as perplexity.

Delete (ROOT)	Financial domain	Law domain	Biomedical domain
&1 BERT+[UNK] (α=0.10)	77.6	72.8	75.2
&2 hybrid BERT+[UNK] (α=0.10)	80.5	77.1	84.6
&3 BERT (Devlin et al., 2018)	86.8	86.1	88.1
&4 hybrid BERT (α=0.10)	**88.6**	**89.5**	**93.8**

Table 3: Accuracy on Deletion dataset of sentence acceptability judgment task in financial domain, law domain, and biomedical domain. Best results are in bold.

Shuffle (bigram)	Financial domain	Law domain	Biomedical domain
#1 BERT (Devlin et al., 2018)	90.5	92.5	93.5
#2 hybrid BERT (α=0.10)	90.5	**93.0**	**95.0**

Table 4: Accuracy on Shuffle dataset of sentence acceptability judgment task in financial domain, law domain, and biomedical domain. Best results are in bold.

Insert	Financial domain	Law domain	Biomedical domain
$1 BERT (Devlin et al., 2018)	82.5	**88.0**	88.5
$2 hybrid BERT ($\alpha$=0.10)	**83.0**	87.5	**89.5**

Table 5: Accuracy on Insert dataset of sentence acceptability judgment task in financial domain, law domain, and biomedical domain. Best results are in bold.

6 Conclusion

In this work, we investigate leveraging XPOS to substitute low-frequency words in the training data of pre-trained masked language model and found model with 10% word substitution rate achieved the better correlation and accuracy on the sentence acceptability evaluation corpora. In the future, we plan to expand our method to other languages in sentence acceptability prediction task.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. We also thank our colleagues, Hiroshi Kanayama from IBM Research Watson NLP and Akihiro Nakayama from IBM Watson development for their helpful discussions.

References

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*.

Birk Diedenhofen and Jochen Musch. 2015. cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, 10(4):e0121945.

Adam Ek, Jean-Philippe Bernardy, and Shalom Lappin. 2019. Language modeling with syntactic and semantic representation for sentence acceptability predictions. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 76–85.

Thibault Févry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2011. Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5528–5531. IEEE.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 959–968.

Kristina Toutanova, Chris Brockett, Ke M Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *EMNLP*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

A Full XPOS List

We use 43 XPOS tags (ID: 1 – 43) out of 49 XPOS tags by excluding 6 punctuation XPOS tags (ID: 44 – 49).

ID	XPOS
1	NNPS
2	NN
3	RBR
4	NNP
5	NFP
6	EX
7	IN
8	SYM
9	FW
10	WDT
11	VBP
12	UH
13	RBS
14	LS
15	JJR
16	GW
17	PRP
18	-LRB-
19	PRP$
20	PDT
21	RB
22	VBN
23	RP
24	ADD
25	WRB
26	AFX
27	VB
28	-RRB-
29	JJS
30	NNS
31	WP
32	CC
33	VBD
34	TO
35	POS
36	VBG
37	WP$
38	CD
39	VBZ
40	JJ
41	HYPH
42	MD
43	DT
44	”
45	$
46	,
47	.
48	:
49	“

Table 6: XPOS list from Stanza POS tagger.

B Unacceptable Sentence Annotation Instruction

We made three modifications (i.e., delete, shuffle, and insert) to generate an unacceptable sentence. Then, we asked two human annotators to examine whether the generated unacceptable sentence has semantic and syntactic violations. The purpose is to remove sentences that are still acceptable after three modifications. Here is the instruction:

Please read the following sentences and judge whether each sentence is acceptable to you by using two criteria:

1. is this sentence syntactically correct?

2. is this sentence semantically correct?

If either of them is false, assign label 0 to the sentence; otherwise, assign label 1 to the sentence.

After the annotation, we select the sentences both annotators assign label 0 as unacceptable sentences.

C Case Study

Figure 3 shows one example sentence in the financial domain. The acceptable sentence is *The most significant challengers in the market are Logset and Sampo - Rosenlew*, while the unacceptable sentence is *The most significant challengers in the market are and Sampo - Rosenlew*. The unacceptable sentence is ungrammatical due to the lack of a root word, *Logset*. An ideal language model should be able to assign **lower** perplexity (PPL) to the acceptable sentence and **higher** PPL to the unacceptable sentence. Herein we experiment with two models, the vanilla BERT model and hybrid BERT model with substitution rate α equal to 0.10.

The number below each token is the log probability (NCE). The lower the value is, the less probable this token should appear in the context. Our observation is that (1) The vanilla BERT assigns lower PPL to the unacceptable sentence but higher PPL to the acceptable sentence because there exist several low-frequency words such as *Logset*, *Sampo*, and *Rosenlew*. These words lead PPL to increase rapidly. (2) On the other hand, as for the hybrid BERT model, the low-frequency words have been replaced with *NNP*, a proper noun tag, which alleviates the low-frequency effect on PPL.

Sentence	The	most	significant	challengers	in	the	market	are	Logset	and	Sampo	-	Rosenlew					
(a) BERT_acceptable																		
PPL (141.2)	-0.12	-0.19	-3.01	-8.62	-0.22	-0.25	-0.49	-7.90	-0.42	-4.36	-8.40	-8.55	-0.02	-5.29	-6.12	-2.45	-10.82	-21.85
(b) BERT_unacceptable																		
PPL (119.6)	The -0.12	most -0.20	significant -3.20	challenger -7.55	##s -0.12	in -0.28	the -0.54	market -6.65	are -2.06	and -6.05	Sam -5.35	##po -6.02	- -3.04	Rosen -10.45	##lew -20.12			
(c) hybrid_acceptable																		
PPL (27.7)	DT -15.68	most -2.16	significant -4.03	NNS -2.05	in -0.44	the -0.50	market -8.74	are -0.38	NNP -0.37	and -1.44	NNP -0.58	- -4.50	NNP -2.33					
(d) hybrid_unacceptable																		
PPL (64.3)	DT -14.62	most -1.74	significant -4.35	NNS -2.14	in -0.48	the -0.65	market -7.85	are -1.95	and -6.93	NNP -2.69	- -4.48	NNP -2.09						

Figure 3: Case study in financial domain.