Intrinsic evaluation of language models for code-switching

Sik Feng Cheong
NUS High School of
Mathematics and Science
20 Clementi Avenue 1
Singapore 129957
h1710019@nushigh.edu.sg

Hai Leong Chieu
DSO National Laboratories
12 Science Park Drive
Singapore 118225
chaileon@dso.org.sg

Jing Lim
DSO National Laboratories
12 Science Park Drive
Singapore 118225
ljing2@dso.org.sg

Abstract

Language models used in speech recognition are often either evaluated intrinsically using perplexity on test data, or extrinsically with an automatic speech recognition (ASR) system. The former evaluation does not always correlate well with ASR performance, while the latter could be specific to particular ASR systems. Recent work proposed to evaluate language models by using them to classify ground truth sentences among alternative phonetically similar sentences generated by a fine state transducer. Underlying such an evaluation is the assumption that the generated sentences are linguistically incorrect. In this paper, we first put this assumption into question, and observe that alternatively generated sentences could often be linguistically correct when they differ from the ground truth by only one edit. Secondly, we showed that by using multi-lingual BERT, we can achieve better performance than previous work on two code-switching data sets. Our implementation is publicly available on Github.[1]

1 Introduction

Code-switching (CS) is the phenomenon where a multilingual speaker alternates between the use of multiple languages in the same utterance (Yow et al., 2018). This practice is common among bilingual speakers, e.g., Chinese (Lyu et al., 2010), Spanish or Hindi-speakers (Khanuja et al., 2020b) who are bilingual in English. To build automatic speech recognition (ASR) systems for CS utterances, a good language model is important, especially when CS training data is often scarce.

Language models (LMs) are an important part of text generation systems in applications such as machine translation and ASR systems. Intrinsic evaluations of LMs are often performed using perplexity, a measurement of how well a LM predicts a test sample. However, previous work has noted that perplexity does not correlate well with ASR performance (e.g. Chen et al., 1998; Gonen and Goldberg, 2019). Extrinsic evaluations of LMs with ASR performance, on the other hand, might be over-fitting to the ASR used in the evaluation. To address this problem, Gonen and Goldberg (2019) proposed a new evaluation metric based on a classification task. First, given a ground truth sentence, they used finite state transducers (FST) to generate a set of phonetically similar sentences. They evaluate LMs based on the task of identifying the ground truth sentence from this set of sentences. LMs can thus be evaluated based on the accuracy of this task, independent of any specific ASR. For this evaluation task, Gonen and Goldberg (2019) proposed a discriminative model that outperformed generative baselines on a English-Spanish (EN-ES) CS test set. However, such an evaluation criteria assumes that all generated sentences are linguistically incorrect. In this paper, we make two contributions

- 还有 (there is) buffalo wings
- high yo buffalo wings
- 还有 (there is) 把 (hold) 发 (hair) 另 (other)
- 海域 (sea area) 巴 (bar) follow wings
- high 由 (by) 把 (hold) follow wings

Figure 1: Examples of alternatives sentences generated to be phonetically similar to the first sentence. Translations provided in brackets.

1. We found that generated sentences with only one modification from the ground truth could be linguistically correct, violating the assumption that such sentences are negative examples. This poses a problem both during training and evaluation. To alleviate this problem, we propose to generate sentences with at least two edits from the ground truth, and

[1] https://github.com/sikfeng/language-modelling-for-code-switching

2. We fine-tuned multi-lingual BERT (mBERT) to improve their published results on their data set. We also experimented with a second English and Chinese (EN-ZH) CS data set.

2 Related Work

The phenomenon of CS is well studied in linguistics (e.g. Poplack, 2000). Earlier work often focus on EN-ES code-switching, studying minorities who claim Spanish as their mother tongue in the English-speaking America. Today, there is also much interest in the Indian languages code-switched with English, due to the large number of Indians using social media (e.g. Khanuja et al., 2020a; Gupta et al., 2021). Recently, researchers in Singapore have collected a CS data set of EN and ZH (Lyu et al., 2010), reflecting the EN-ZH bilingual culture in Singapore and Malaysia.

Research on CS often focuses on downstream NLP tasks: Khanuja et al. (2020b) built a data set called GlueCos to evaluate NLP performance on CS data; Winata et al. (2021) examined the effectiveness of multi-lingual embedding such as mBERT on downstream tasks such as part-of-speech tagging. The first EN-ZH LM was designed by Vu et al. (2012) using statistical machine translation techniques. With deep learning, LMs were also trained directly on CS data (Kumar et al., 2020). Luo et al. (2018) trained an end-to-end EN-ZH CS speech recognition models, using a connectionist temporal classification and attention based model. Qin et al. (2020) used data-augmentation with CS data as an approach to facilitate cross-lingual transfer learning. Winata et al. (2019) used a sequence-to-sequence model to generate CS data for data augmentation. Gonen and Goldberg (2019) pre-trained a BiLSTM (Bidirectional Long Short Term Memory) on monolingual data before fine-tuning on CS data. In this paper, we propose to replace this approach by fine-tuning mBERT (Devlin et al., 2019). We show that our model outperformed BiLSTM (Gonen and Goldberg, 2019) on both EN-ES and EN-ZH data.

Gonen and Goldberg (2019) generated alternative sentences with similar phonemes, and assumed that these generated sentences are linguistically incorrect, using them as negative examples in training and evaluation. However, Dautriche et al. (2017) have found that word form similarity increases with semantic similarity across a number of languages. Hence, making a single phonetically similar edit on a ground truth sentence might result in another correct sentence. We propose to constraint generated sentences to have at least 2 differences from the ground truth to reduce this possibility.

- 她/他/它去 airport (she/he/it goes to the airport)
- no i remember seeing the same/sign
- so we brought her some/son ribs so
- i was talking to your mom/ma'am/man

Figure 2: Examples of sentences with alternatives of a single word separated with /. These generated replacements could be correct without more context. The translation for the first CS sentence is provided in brackets.

3 Generating evaluation data sets

We experimented on two data sets: the EN-ES data used in (Gonen and Goldberg, 2019), created from the Bangor Miami Corpus, and the EN-ZH SEAME corpus (Lyu et al., 2010), collected from residents of Singapore and Malaysia who are bilingual in EN and ZH.

To generate sets of alternative monolingual and code-switched sentences, we took the same approach as Gonen and Goldberg (2019). First, we convert a sequence of language tagged words (either CS or monolingual) into a sequence of the matching phonemes (using pronunciation dictionaries); then we decode the sequence of phonemes into new sentences, which include words from either language (possibly both); During decoding, we allow minor changes in the sequence of phonemes to facilitate the differences between the languages. These steps can be easily implemented using composition of finite-state transducers (FSTs). We used the FST provided by (Gonen and Goldberg, 2019) for the EN-ES data. To build the FST for EN-ZH, we obtained word probabilities for Chinese words from the Mandarin Chinese News Text Corpus, while we used the English word probabilities provided by Gonen and Goldberg (2019). We used the jieba tokenizer[2] to segment Chinese words, and mapped the Chinese pinyin provided by CC-CEDICT.[3]

We observed that if we substitute only one word, the altered sentence could sometimes still be linguistically correct. For example, in the CS sen-

[2] https://github.com/fxsjy/jieba

[3] https://www.mdbg.net/chinese/dictionary?page=cedict

tence *"她/他/它去 airport (translated: he/she/it goes to the airport)"*, the three alternative words are different pronouns that have identical pronunciation. We give a few other examples in Figure 2. Hence, we generate alternative sentences with at least two differences from the original sentence. This increased the probability that the generated sentences are linguistically incorrect, so that they serve their purpose as negative sentences both for training and for evaluation. For each sentence, we generated an alternative set that contains monolingual (for both languages) and CS sentences. The FST is not always able to produce sufficient alternatives for a given sentence: if the FST generates less than 5 alternatives for any type, we discard that set. We capped the number of alternatives of each type to 10, and hence each set has at most 30 sentences. We did a train-validation-test set split such that both the validation and test set contain 1,000 alternative sets each. For EN-ES, the training set consists of 1,021 CS sets, 7,171 EN sets, 3,489 ES sets, while for EN-ZH, it contains 20,689 CS sets, 6,515 EN, and 7,050 ZH sets.

4 Discriminative Training

Language models are usually trained generatively to generate positive sentences. Gonen and Goldberg (2019) proposed a discriminative BiLSTM model by training the BiLSTM to discriminate between ground truth positive sentences with synthetically generated negative sentences. They used as their base model a 2-layer BiLSTM using DyNet (Neubig et al., 2017). The objective of their discriminative training was to assign a higher score to ground truth sentences compared to synthetically generated ones. We denote s_0 as the original ground truth sentence, and $\text{alt}(s_0)$ to be the set of alternative sentences generated for s_0. Each sentence is scored using $\text{score}(s) = w \cdot \text{BiLSTM}(s)$, where $w \in \mathbb{R}^n$ is a learnable parameter vector in the model, with n the dimension of the embedding.

The loss function for each alternative set is

$$\sum_{s \in \text{alt}(s_0)} \max(0, \text{WER}(s_0, s) + \text{score}(s) - \text{score}(s_0)), \quad (1)$$

where $\text{WER}(s_0, s)$ is the word-error-rate between sentences s_0 and s. In our experiments, we followed Gonen and Goldberg (2019) and set the dimensions in the hidden BiLSTM to 650, in a 2-layer BiLSTM. Since the transformer based BERT has a bigger model size, we also experimented with a 3-layer BiLSTM, which should increase the capacity of the model. However, we found that adding a layer to BiLSTM did not improve its performance on our data sets.

4.1 Discriminative mBERT

As BERT (Devlin et al., 2019) has shown to work well for many linguistic tasks, we propose to replace the BiLSTM with the pre-trained multilingual BERT (mBERT). mBERT was pre-trained on the entire Wikipedia dump for 104 languages, including English and Chinese. The mBERT embedding for the sentence is the hidden state of the [CLS] token, with a dimension of 768.

We implemented our approach using the Huggingface transformers library (Wolf et al., 2020), which provides a collection of pre-trained models. We fine-tuned the pre-trained multi-lingual BERT models provided by this library, with the same loss function as Equation 1. We experimented with different networks on the mBERT embedding, such as using 1, 2 and 3 linear layers, as well as using LayerNorm or ReLU on the layers.

5 Experimental Results

In this section, we describe the empirical comparison between BiLSTM and mBERT on the two data sets, and analyze the results.

We experimented with different versions of BiLSTM on the EN-ZH data. For the monolingual data sets, we experimented with using (1) only the alternative sets generated from SEAME, (2) only the alternative sets from Mandarin Chinese News Text Corpus and OpenSubtitles, and (3) all of the above. We obtained the best results with (3). Using a 3-layer BiLSTM did not improve results, which could be due to insufficient training data.

We compare the different versions of mBERT in Table 1. We observe that the models that use 2 linear layers over the mBERT embedding performed the best among the mBERT models. As the results obtained by the variants of mBERT is close, we compare the results of the 2-layer mBERT model with its LayerNorm and ReLU variants, using the McNemar statistical test (Smith and Ruxton, 2020). We found that the results are not significantly different at the p-value of 0.05.

For the EN-ES data, we follow closely the setup of Gonen and Goldberg (2019). Results are shown in Table 2. We re-run the code released by Gonen

Model	Validation				Test			
	CS	EN	ZH	All	CS	EN	ZH	All
BiLSTM	94.50	90.60	91.70	92.27	94.90	92.40	91.40	92.90
$w \cdot h$	98.30	95.80	94.10	96.07	98.60	95.60	95.30	96.50
+ LayerNorm	98.20	95.90	94.60	96.23	98.00	95.50	94.80	96.10
+ ReLU	98.20	95.70	94.70	96.20	98.10	95.90	96.40	96.80
$w_2(w_1 \cdot h + b)$	97.70	96.30	95.10	96.37	98.40	96.00	96.10	96.83
+ LayerNorm	98.30	96.40	95.10	**96.60**	98.50	96.40	95.40	96.77
+ ReLU	98.10	95.90	95.20	96.40	98.70	97.10	95.30	**97.03**
$w_3 \cdot (w_2 \cdot (w_1 \cdot h + b_1) + b_2)$	98.10	94.60	95.10	95.93	98.50	96.10	96.10	96.90
+ LayerNorm	98.10	95.80	94.50	96.13	98.40	95.50	95.80	96.57
+ ReLU	98.10	95.90	94.80	96.27	98.80	95.30	95.80	96.63

Table 1: Results of mBERT with discriminative finetuning for 5 epochs on the EN-ZH data set.

Model	Validation				Test			
	CS	EN	ES	All	CS	EN	ES	All
BiLSTM	71.20	88.82	83.13	83.00	69.60	87.22	82.10	81.50
$w \cdot h$	84.40	94.81	85.94	90.00	82.80	94.73	84.82	89.20
+ LayerNorm	84.40	95.61	85.50	90.30	79.20	94.93	82.49	89.20
+ ReLU	84.00	95.81	85.94	90.40	84.80	95.54	85.99	90.40
$w_2(w_1 \cdot h + b)$	84.40	96.61	84.71	90.60	83.60	96.35	82.88	90.40
+ LayerNorm	85.60	95.81	84.74	90.50	85.60	96.12	84.44	**90.50**
+ ReLU	83.60	96.41	83.13	89.90	79.60	96.35	84.82	89.20
$w_3 \cdot (w_2 \cdot (w_1 \cdot h + b_1) + b_2)$	85.60	96.01	85.54	**90.80**	83.20	96.35	85.99	90.40
+ LayerNorm	83.60	95.41	83.13	89.40	81.20	96.35	81.32	88.70
+ ReLU	82.80	95.21	84.34	89.40	84.80	95.13	84.05	89.70

Table 2: Results of mBERT with discriminative finetuning for 5 epochs on the EN-ES data set.

Language (model)	Validation set	Test set
CS (mBERT)	97.17	97.53
EN (mBERT)	93.70	92.90
EN (BERT)	**95.70**	**95.90**
ZH (mBERT)	94.00	93.40
ZH (BERT)	**96.30**	**95.50**

Table 3: Results of monolingual BERT (bert-large-uncased-whole-word-masking and hfl/chinese-bert-wwm-ext), and mBERT (bert-base-multilingual-cased) models on instances with 30 alternatives per set.

Rank of average scores	Validation set			Test set		
	CS	EN	ZH	CS	EN	ZH
CS>EN>ZH	174	293	7	165	304	5
CS>ZH>EN	428	15	199	465	27	209
EN>CS>ZH	74	619	4	63	588	6
EN>ZH>CS	12	32	4	12	38	1
ZH>CS>EN	287	18	763	284	20	764
ZH>EN>CS	25	23	23	11	23	15

Table 4: Count of the ranks of scores of alternative sentences using mBERT with $w_2 \cdot (w_1 \cdot h + b)$.

and Goldberg (2019) on our generated data sets where alternative sentences are generated with at least 2 differences from the ground truth. We arrive at the same conclusions as before: mBERT does better than BiLSTM, and the different variants of our mBERT models achieved similar results.

5.1 Discussion

Monolingual BERT vs mBERT: we investigate using monolingual BERT instead of mBERT for monolingual only test sets. Table 3 shows that, while the EN and ZH BERT models outperform mBERT, mBERT does reasonably well on both EN and ZH only data, showing that mBERT is robust even on monolingual data, and could be the preferred model on user-created content which might include any CS data.

Ranking of alternative sentences of different types: From Table 4, we can observe that (i) the scores of sentences of the same language as the ground truth are higher, (ii) when the ground truth is monolingual, the CS sentences are often scored higher than sentences of the other language. This

makes sense as generated CS alternatives can still retain words and grammatical structure from the original sentence, while sentences in the other language would need to have all words generated, and (iii) ZH alternatives are ranked higher than EN alternatives for CS sentences. This could be due to the fact that in the SEAME data, the main language is ZH and the switch to EN was often just for one word or phrase (Lyu et al., 2010).

6 Conclusion

In this work, we proposed to modify the evaluation of language models proposed by Gonen and Goldberg (2019). While the evaluation proposed by Gonen and Goldberg (2019) provided for a simple approach to intrinsic evaluation of language models for ASR, we highlighted the problem of using linguistically correct sentences as negative examples, introducing noise both in the training and in the evaluation process. We proposed a simple solution to this problem by requiring a minimum edit distance between the generated samples and the ground truth. In addition, we showed that mBERT outperforms BiLSTM on two code-switching data sets using this evaluation criteria.

Acknowledgments

The work was conducted while the first author is doing an internship at the DSO National Laboratories.

References

Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. 1998. Evaluation metrics for language models.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T Piantadosi. 2017. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, 41(8):2149–2169.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, page 4171–4186, Minneapolis, Minnesota. 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), Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Language modeling for code-switching: Evaluation, integration of monolingual data, and discriminative training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4166–4176.

Akshat Gupta, Sai Krishna Rallabandi, and Alan Black. 2021. Task-specific pre-training and cross lingual transfer for code-switched data.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020a. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16, Marseille, France. European Language Resources Association.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.

V. Kumar, S. Pasari, V. P. Patil, and S. Seniaray. 2020. Machine learning based language modelling of code switched data. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 552–557.

Ne Luo, Dongwei Jiang, Shuaijiang Zhao, Caixia Gong, Wei Zou, and Xiangang Li. 2018. Towards end-to-end code-switching speech recognition. In *INTERSPEECH*. INTERSPEECH.

Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *INTERSPEECH*. INTERSPEECH.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Shana Poplack. 2000. Sometimes i' ll start a sentence in spanish y termino en español: Toward a typology of code-switching. *The bilingualism reader*, 18(2):221–256.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

Matilda Q. R. Pembury Smith and Graeme D. Ruxton. 2020. Effective use of the mcnemar test. *Behavioral Ecology and Sociobiology*, 74:133.

N. T. Vu, D. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E. Chng, T. Schultz, and H. Li. 2012. A first speech recognition system for mandarin-english code-switch conversational speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4889–4892.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *arXiv preprint arXiv:2103.13309*.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 38–45, Online. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics.

W. Quin Yow, Jessica S. H. Tan, and Suzanne Flynn. 2018. Code-switching as a marker of linguistic competence in bilingual children. *Bilingualism: Language and Cognition*, 21:1075–1090.