IMPROVING SPOKEN LANGUAGE UNDERSTANDING BY ENHANCING TEXT REPRESENTATION

Thai Binh Nguyen

Interactive Systems Lab, Karlsruhe Institute of Technology, Germany
thai-binh.nguyen@kit.edu

ABSTRACT

Language Model (LM) which is commonly trained on a large corpora has been proven the robustness and effectiveness for tasks of Natural Language Understanding (NLU) in many applications such as virtual assistant or recommendation system. These applications normally receive outputs of automatic speech recognition (ASR) module as spoken form inputs which generally lack both lexical and syntactic information. Pre-trained language models, for example BERT [1] or XLM-RoBERTa [2], which are often pre-trained on written form corpora perform decreased performance on NLU tasks with spoken form inputs. In this paper, we propose a novel model to train a language model namely CapuBERT that is able to deal with spoken form input from ASR module. The experimental results show that the proposed model achieves state-of-the-art results on several NLU tasks included Part-of-speech tagging, Named-entity recognition and Chunking in English, German, and Vietnamese languages.

***Index Terms*—** Spoken Language Understanding, BERT, CapuBERT, ASR

1. INTRODUCTION

Pre-trained LM has demonstrated an effective strategy to learn the contextual representation of language, which is beneficial for downstream NLP tasks. Most of pre-trained LMs developed so far just concentrate on written text that includes the features of punctuations and capital letters. For the ASR's output, which is spoken form and lacks of such features, these pre-trained LMs become ineffective. Hence, it poses challenges when carrying out downstream NLP tasks on such kind of text. The major differences between existing pre-trained LMs can be categorized based on three perspectives: model architectures, training dataset types, and pre-training objectives.

Model architectures: In the previous studies, pre-trained LMs are mostly based on well-known model architectures such as LSTM [3] and Transformer [4] network. BERT [1] and its variants (e.g. ALBERT [5], RoBERTa [6], etc.) use a multi-layer bidirectional Transformer encoder. In this study, we leverage the RoBERTa$_{base}$ architecture.

Training dataset types: Pre-trained LMs usually learn the contextual representation from general large-scale written text corpora without external domain knowledge and modality information. In this study, we assume the mismatch between spoken text (output from ASR) and written text (input for pre-trained LMs) is mostly caused by the lack of capital letters and punctuations (henceforth, capu). We create a spoken corpora by removing all capu information from the written corpus since a large-scale spoken corpus is not available yet.

Pre-training objectives: The pre-training objectives are crucial for learning the contextual representation of language. The *masked language model* (Mask LM) is first adapted to overcome the drawbacks of the standard unidirectional LM. The *next sentence prediction* objective is also proposed to use in BERT [1] to enhance text representation, but RoBERTa [6] shows that it is not necessary. ALBERT [5] model proposed *sentence ordering objective* task to replace the next *sentence prediction objective*. For adapting with spoken text, we propose a novel pre-training objective, named Mask Capu that guides the model to restore capu labels for spoken input.

In summary, our contribution in this paper is twofold: First, we propose a novel CapuBERT language model that is able to jointly learn the contextual representations and external knowledge such as capitalization and punctuation. Second, our proposed CapuBERT model are optimized to obtain state-of-the-art performance on spoken text. In addition, we theoretically and empirically evaluate the effectiveness of CapuBERT by experimenting on three downstream tasks: Part-of-Speech (POS) tagging, Named Entity Recognition (NER), Chunking in three languages: English, German, and Vietnamese. The experimental results consistently show that our proposed model outperforms all baseline models in several experiment setups.

2. RELATED WORK

Capu information turns out to be very important for NLU tasks [7, 8, 9, 10]. Without it, NLU tasks' performance, such as NER, Chunking, and POS tagging, down in a significant way. Many studies attempt to handle the capu restoration problem since it can reduce the mismatch between written text

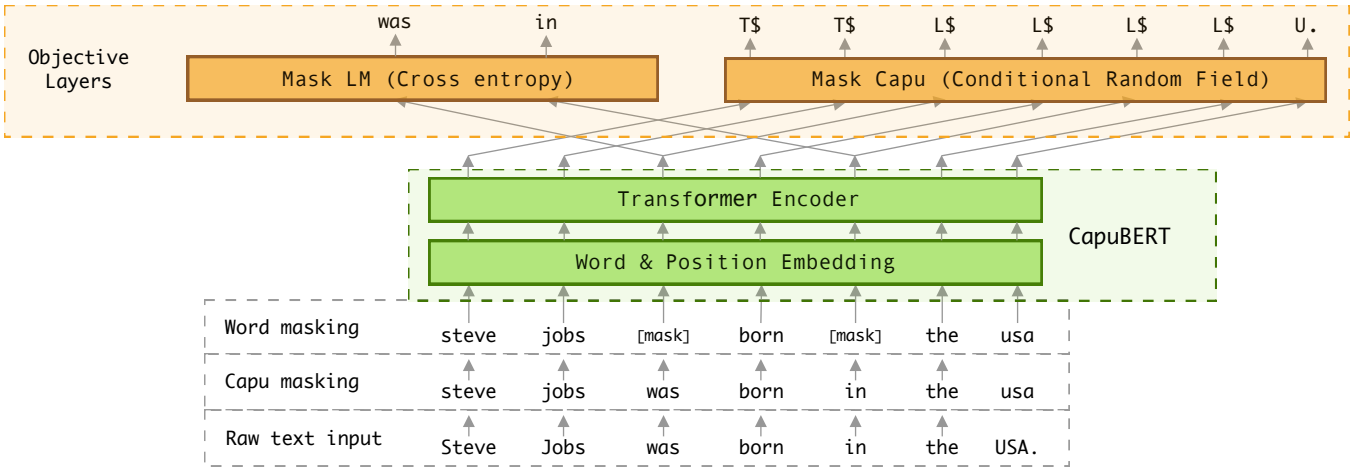

Fig. 1. Overview of the pre-training CapuBERT model. The raw text input '*Steve Jobs was born in the USA.*' firstly goes through the capu masking process to remove capu information: '*steve jobs was born in the usa*' and create capu labels: '*T$ T$ L$ L$ L$ L$ U.*'. Word masking process then randomly masks 15% words: '*steve jobs [mask] born [mask] the usa*'. Model is trained to detect masked words using Mask LM layer and predict capu labels using Mask Capu layer.

and spoken text. Nguyen *et al.* (2019) [11] proposed overlap chunk merging method to recover capu information for infinite length spoken text. Sunkara *et al.* (2020) [12] fine-tuned a pre-trained BERT model in the medical domain data to recover capu. Nguyen *et al.* (2020) [9] constructed a pipeline models to extract named-entity from speech, in which they deal with the mismatch between written text and spoken text by using a capu restoration model. However, these can make the SLU pipeline more complex.

Pre-trained LMs BERT [1] and its variants are well known as a best way to embed the text information in space vector. Many objective functions were proposed to help BERT archive contextual representation. However, current pre-trained LMs cannot deal with the spoken text since they lack of capu information. In this study, we show that by combining the Mask LM objective with a capu prediction task (Mask Capu objective), the pre-trained LMs can embed both contextual representation of language and capu features. By that, we only need a single pre-trained LM to handle spoken text.

3. CAPUBERT LANGUAGE MODEL

In this section, we describe the architecture of CapuBERT model, then discuss in detail the objective functions and optimization setup.

3.1. CapuBERT architecture

We use RoBERTa$_{base}$ model to learn the contextual representations. The outputs then are fed into two different layers to predict masked words and capu labels. For capu labels prediction, since capu labels is related (e.g. following dot usually is a uppercase word), after some experiment we choose Conditional Random Field layer to get the best performance. The architecture of CapuBERT model is illustrated in the Figure 1.

Language	Case types			Punctuations			
	U	L	T	$	.	,	'!, ?'
English	7%	75%	18%	90%	4%	5%	1%
Vietnamese	5%	76%	19%	90%	5%	4%	1%
German	5%	65%	30%	87%	6%	6%	1%

Table 1. Capitalization and punctuation statistics

3.2. Objective function

CapuBERT model is trained by using two unsupervised tasks which are Mask LM and Mask Capu tasks. For each step, the model loss is an equal sum of the loss values of two tasks. The Mask LM is same as in BERT model [1]. For Mask Capu task, we consider three different types of word forms which are *uppercase all letters in a word* (**U**), *uppercase first letter in a word* (**T**), and *lowercase all letters in a word* (**L**). Each punctuation belongs to the word right before it. These punctuations are *period* ('**.**'), *comma* ('**,**'), *exclamation mark* ('**!**'), *question mark* ('**?**') and a *special blank* ('**$**') indicates non-punctuation. In total we have 15 classes of capu labels. The Mask Capu task leverages the Conditional Random Field layer to predict the capu label of each word.

For a sequence $\mathbf{x}$ of T tokens, we first construct a corrupted version $\hat{\mathbf{z}}$ by implementing capu masking and word masking process. Let the masked tokens be $\bar{\mathbf{z}}$ and capu labels

be **c**. The training objective is to reconstruct $\bar{\mathbf{z}}$ and predict **c** from $\hat{\mathbf{z}}$:

$$\max_{\theta} \quad \log p_\theta(\bar{\mathbf{z}}, \mathbf{c}|\hat{\mathbf{z}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(z_t|\hat{\mathbf{z}}) + \log p_\theta(\mathbf{c}|\hat{\mathbf{z}}) \tag{1}$$

where $m_t = 1$ indicates z_t is masked word, otherwise $m_t = 0$.

3.3. Optimization

We rely on the $fairseq$ [13] framework to implement CapuBERT model. We set batch size of 4096 and a peak learning rate of 0.0002. The model is then optimized by using Adam [14] optimization. We train each experimented models approximately in 500k steps by using 8 GPUs V100.

4. EXPERIMENTS

4.1. Spoken data corpora

We employ the English, German, and Vietnamese corpus from the CC100 dataset [15]. Statistically, the original English, German, and Vietnamese corpora contain 55.6B tokens (300GB), 10.3B tokens (66GB), and 24.7B tokens (137GB) respectively. In our work, these corpora are reprocessed by keeping only alphabet character, number, and punctuation. The corpora are then converted to spoken form by removing capu information. We use BPE [16] algorithm to tokenize input sentence to subword units.

Table 1 shows the statistics of case types and punctuation proportion for each corpus. While the distribution of case types is similar in both English and Vietnamese, the proportion of uppercase words in German (35%) is higher than the proportion of English and Vietnamese (25%). Roughly 10% of words are followed by punctuations.

4.2. Experimental setup

4.2.1. Downstream tasks

We experiment CapuBERT model on three downstream tasks: Part-of-speech (PoS) tagging, Named-entity recognition (NER) and Chunking (Chunk). For English and German, we use the CoNLL-2003 dataset [17] while we make use of the VLSP 2016 [18] dataset for Vietnamese. The original version of these datasets is in the written form. Therefore, in order to verify the effectiveness of the proposed model, we transform all input text to spoken form by performing the capu masking process as in the Figure 1.

To concrete the efficacy of the proposed in practice, we extend experiment used the real Speech NER dataset. We choose a rich resource language and a low resource language to build the Speech NER dataset. For English, 9 people read the CoNLL-2003 NER English test set (including 3453 samples with total 47k words) and produce 11 hours of audio. For Vietnamese, we use dataset was proposed in [9] (4272 samples with total 242k words, 26 hours of audio). Processing Speech NER data requires ASR to transcript speech to spoken text. ASR module can introduce errors in text output, which make it different from the reference text (that have ground-truth NER tag). To align them, we use algorithm proposed in [9]: If ASR output is right, the hypothesis entity tag was remained. If the error type is deletions or substitutions, the hypothesis tag of this word will become **O**. Else if error type is insertions, the tag will be removed.

4.2.2. Pre-trained LM baselines

To show the effectiveness of our pre-trained CapuBERT on the spoken text, we set up different evaluation scenarios with other well-known pre-training LMs by using RoBERTa model [6] with $fairseq$ [13] framework. Pre-training LMs are trained on different types of corpus (*cased*, *uncased*, *uncapu*) in three different languages. Specifically, *cased* means the corpus with all capu information; *uncased* means the corpus with lowercase words only, and *uncapu* means the corpus which is in the spoken form. Both RoBERTa$_{uncapu}$ and the proposed CapuBERT are trained on the *uncapu* dataset, while CapuBERT trains jointly with the additional Mask Capu objective. For the *cased* dataset, we utilize pre-trained RoBERTa [6] and PhoBERT [19] models for English and Vietnamese languages respectively, while we train remained models from scratch on *uncased* and *uncapu* datasets.

4.2.3. Downstream task models

Following [1], in all three tasks of POS tagging, NER, and Chunking, we append a linear prediction layer on top of pre-trained LM since these tasks can be handle like a sequence tagging problem. With *cased* model, we perform an additional step is *capu restoration* on the spoken text corpus of downstream tasks. We employ SOTA *capu restoration* model proposed in [9] to recover capu information for the text corpus in the downstream tasks. Table 3 shows performances of these capu restoration model in three languages.

4.3. Experimental results

Table 2 shows evaluation results of different pre-trained LMs on three downstream tasks across three languages. The spoken text makes previous pre-trained LMs challenging to perform on downstream tasks. While RoBERTa$_{cased}$ (row 1) performed worse on all three tasks via all three languages, RoBERTa$_{uncased}$ (row 3) obtained a better results than the RoBERTa$_{cased}$ since it can get familiar with lowercase text. The RoBERTa$_{uncapu}$ (row 4), even trained on the spoken text with its original objective function, had a tiny gap compared

Model	English			German			Vietnamese		
	Chunking	NER	PoS Tagging	Chunking	NER	PoS Tagging	Chunking	NER	PoS Tagging
RoBERTa$_{cased}$	81.19	85.22	91.42	86.43	63.26	92.89	90.79	78.25	89.41
RoBERTa$_{cased}$ + capu restoration	88.28	86.51	92.04	90.20	70.73	93.05	90.79	83.20	89.18
RoBERTa$_{uncased}$	88.20	85.93	91.90	87.12	69.62	93.95	92.58	82.99	90.78
RoBERTa$_{uncapu}$	88.25	85.70	91.80	87.40	69.60	94.00	92.60	83.03	90.50
CapuBERT (our)	**88.94**	**87.22**	**92.97**	**92.01**	**74.82**	**94.66**	**92.91**	**86.42**	**91.32**

Table 2. Evaluation F1 scores of different pre-trained LMs fine-tuned on three downstream tasks. Model *capu restoration* proposed by [9].

Model	en	vi	de
Capu restoration [♣]	89.91	87.47	86.70
Capu restoration [♠]	89.90	88.08	88.03

Table 3. Evaluation F1 score of capu restoration models: [♣] proposed by [9] compares with our [♠] CapuBERT.

Model	en	vi
RoBERTa$_{uncased}$	68.75	63.19 [♣]
RoBERTa$_{uncapu}$	68.80	63.20
RoBERTa$_{cased}$ + capu restoration	71.10	67.13 [♣]
CapuBERT (our)	**73.06**	**68.72**

Table 4. Evaluation F1 scores of different pre-trained LMs for doing NER on spoken text in practice. [♣] reported by [9]

with the RoBERTa$_{uncased}$. Our CapuBERT, however learns spoken text representation through Mask LM and Mask Capu objective functions. The results show our proposed method achieved state-of-the-art performance on all three downstream tasks across all three languages. In particular, the CapuBERT had a significant improvement on the NER task via all three languages (see column NER at row 5), demonstrating the Mask Capu layer's effectiveness in the proposed model.

The capu restoration [9], followed by a RoBERTa$_{cased}$ (row 2) obtained much better results than the RoBERTa$_{cased}$. Table 3 shows another strong ability of our proposed joint pre-trained LM, in which our capu restoration [♠] performed better than the capu model [♣] proposed in [9]. The capu restoration [♠], which is the Mask Capu CRF component in the proposed model, can deal with spoken text from ASR module by providing a correct text conversion. This verify the necessity of the capu recovery layer, which not only brings useful capu features into the learning process but also effectively restores proper formatting text from spoken text.

Table 4 illustrates the result of doing NER on the spoken text in practice. The spoken text is produced from Speech NER dataset using Google Cloud Speech-to-Text API. The word error rate for English is 7.55% and Vietnamese is 6.57%. This table shows that by using CapuBERT, we can overcome other pre-trained LMs in representing spoken text. So the capu information embedded in CapuBERT works well in practice.

5. CONCLUSION

In this paper, we have described a novel CapuBERT language model to enhance the contextual representation of language in the spoken form. We extensively experimented models on three downstream NLU tasks, including PoS tagging, NER, and Chunking in English, German, and Vietnamese languages. Experimental results demonstrated that the proposed CapuBERT model improve the performance and reduce the complexity in performing NLU from ASR output on all evaluation scenarios.

6. ACKNOWLEDGMENT

I would like to thank Van-Khanh Tran, Tran-Thai Dang, Kim-Anh Nguyen and Prof. Alex Waibel for their advice and feedback. This work was partially done during work at Vingroup Big Data Institute. The research has been supported in part by the German Federal Ministry of Education and Research (BMBF) under the project OML (01IS18040A).

7. REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "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)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[2] Sebastian Ruder, Anders Søgaard, and Ivan Vulić, "Unsupervised cross-lingual representation learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*,

Florence, Italy, July 2019, pp. 31–38, Association for Computational Linguistics.

[3] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, 2014, pp. 338–342.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, pp. 5998–6008, Curran Associates, Inc.

[5] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[7] Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel, "Named entity extraction from speech," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*. Citeseer, 1998, pp. 287–292.

[8] Mihai Surdeanu, Jordi Turmo, and Eli Comelles, "Named entity recognition from spontaneous open-domain speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[9] Thai Binh Nguyen, Quang Minh Nguyen, Thi Thu Hien Nguyen, Quoc Truong Do, and Chi Mai Luong, "Improving vietnamese named entity recognition from speech using word capitalization and punctuation recovery models," *Proc. Interspeech 2020*, pp. 4263–4267, 2020.

[10] Eunah Cho, Jan Niehues, and Alex Waibel, "Segmentation and punctuation prediction in speech language translation using a monolingual translation system," in *Proceedings of the 9th International Workshop on Spoken Language Translation: Papers*, Hong Kong, Table of contents, Dec. 6-7 2012, pp. 252–259.

[11] B. Nguyen, V. B. H. Nguyen, H. Nguyen, P. N. Phuong, T. Nguyen, Q. T. Do, and L. C. Mai, "Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging," in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2019, pp. 1–5.

[12] Monica Sunkara, Srikanth Ronanki, Kalpit Dixit, Sravan Bodapati, and Katrin Kirchhoff, "Robust prediction of punctuation and truecasing for medical ASR," in *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, Online, July 2020, pp. 53–62, Association for Computational Linguistics.

[13] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[14] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave, "CCNet: Extracting high quality monolingual datasets from web crawl data," in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4003–4012, European Language Resources Association.

[16] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Aug. 2016, pp. 1715–1725, Association for Computational Linguistics.

[17] Erik F. Tjong Kim Sang and Fien De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147.

[18] Huyen Nguyen, Quyen Ngo, Luong Vu, Vu Tran, and Hien Nguyen, "Vlsp shared task: Named entity recognition," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 283–294, 2019.

[19] Dat Quoc Nguyen and Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 1037–1042, Association for Computational Linguistics.