음성인식 오류를 포함한 텍스트 대상 언어모델 학습 방법 및 이를 이용한 음성언어 의도 분류 성능 개선에 관한 연구

김준우, 윤혜경, 정호영 *
경북대학교

kaen2891@knu.ac.kr, yhk04150@knu.ac.kr, *hoyjung@knu.ac.kr

A study on training language model of speech recognized text and improvement of spoken language intent classification

Kim June-Woo, Yoon Hyekyung, Jung Ho-Young*
Kyungpook National Univ.

요 약

본 논문은 음성인식 오류로 인해 실제 사용 환경에서 발생하는 사용자 의도 분류 성능 저하를 개선하고자 사전 학습된 BERT 언어모델을 바탕으로 오류를 포함한 음성으로부터 인식된 텍스트와 전사 데이터로 각각 학습을 진행하고 의도 분류 성능을 평가한다. 실제 환경에서 음성인식 오류에 강인한 모델 학습법에 대해 제안하고 실험 결과를 통해 본 방법의 효과에 대해 검증한다.

Ⅰ. 서 론

인공지능 기술이 발전함에 따라 다양한 분야에 이를 적용하는 추세이며, 특히 인공지능 스피커에 대한 수요가 해마다 증가하고 있다. 하지만 시중의 인공지능 스피커들은 사용자의 요청을 이해하는데 여전히 불안정한 성능을 보이고 있다. 사용 환경에서의 주변 소음과 음성인식 시스템을 학습할 때 사용하는 데이터셋에서 희박한 노인과 어린이의 음성, 그리고 불확실한 발음 등이 음성인식 시스템의 결과인 텍스트에 오류를 유발하고, 이 오류들이 그대로 언어모델로 전해지면서 사용자 의도 분류 시 오류가 발생한다.

자연어처리 도메인에서 Transformer [1], 그리고 해당 모델을 기반으로 한 BERT [2] 모델이 널리 사용되고 있다. 특히 BERT 는 대규모 텍스트 데이터셋으로 self-supervised learning 을 통해 representation 추출 및 적은 수의 데이터로 fine-tuning 이 가능한 강력한 모델이다. 하지만 해당 모델을 사용하여 음성인식 오류, 즉 잡음이 포함된 텍스트에 대해서 강력한 효과를 보이지 못하는 단점이 존재한다.

위의 문제를 해결하기 위해, 본 논문에서는 음성인식 오류가 포함된 recognized text 를 추가적으로 얻고, 원본 그대로인 전사 데이터 labeled text 를 함께 활용하는 방법을 제안한다. 이를 위해, Fluent Speech Commands (FSC) [3] 데이터셋과 audio-SNIPS [4] 음성 데이터셋을 open 음성인식 시스템인 wav2vec 2.0 [5] 에 입력으로 사용하여 recognized text 셋을 확보한다. 이후, labeled text 셋과 함께 사전 학습된 언어모델 BERT 2 개를 활용하여 이중 학습을 진행한다.

평가 단계에서는 학습된 target-BERT 모델만 이용하여 recognized, labeled 모두에 대해 평가를 진행한다. 우리의 실험 결과는 recognized text 셋에 개선된 의도 분류 정확도를 보여줌과 동시에 labeled text 셋에도 큰 성능 저하가 없음을 보여준다. 이를 통해 제안하는 방법이 효과적임을 보인다.

Ⅱ. 본론

본 장에서는 제안하는 이중 학습법에 대해 소개한다.

Ⅱ-Ⅰ. 이중 BERT 모델

언어모델 학습을 위해 대규모 텍스트 데이터셋으로 사전 학습된 BERT 를 활용한다. 우리의 방법은 2 개의 BERT 를 이중 공동 학습 방법이며, labeled text 로 학습을 진행할 C-BERT 와 recognized text 로 학습을 진행할 N-BERT 로 구성되어 있다. 이 두 모델을 함께 배치하여 동시에 의도 분류 이중 학습을 진행하며, 학습 중 C-BERT 의 gradient 들이 N-BERT 에 흐르게 하여 N-BERT 가 labeled text 의 정보를 함께 학습할 수 있도록 한다. 이를 통해 정형화된 데이터셋 및 잡음에 강인한 모델이 만들어질 것으로 예상된다. 평가시에는 recognized text 와 labeled text 모두로부터 학습된 N-BERT 만 활용하여 테스트를 진행한다.

Ⅱ-Ⅱ. 실험 세팅

실험을 위해 spoken language understanding 분야의 벤치마크 데이터셋 중 짧은 명령어가 녹음된 FSC 데이터셋과 인공지능 음성 합성기인 Amazon Polly 에 의해 생성된 음성 파일로 구성된 Audio-SNIPS 데이터셋을 활용한다. FSC 데이터셋은 총 30,043 개의 음성파일과 labeled text 로 구성되어 있으며, Audio-SNIPS 데이터셋에는 총 16 개의 서로 다른 AI 음성 합성기가 생성한 각각 2,700 개의 음성 파일과 그에 해당하는 labeled text 가 포함되어 있다.

음성인식을 통해 얻어진 오류가 있는 recognized text 를 확보하기 위하여 self-supervised 기반의 대규모 음성 데이터셋으로 학습된 wav2vec 2.0 모델을 활용한다.

실험은 NVIDIA TITAN RTX 24GB GPU 로 진행되었으며, 학습률은 1e-4, batch 사이즈는 64, AdamW 옵티마이저를 사용하였다. 모든 fine-tuning 은 5 epoch 으로 하였다.

Ⅱ-Ⅲ. 실험 결과

제안하는 모델 성능을 평가하기 위해 사전 학습된 기존 BERT 와 N-BERT 의 성능을 비교한다.

Model	평가 데이터셋: X	
	FSC	A-SNIPS
	의도 분류 성능(%)	
BERT		
X	100	99
$\hat{X}$	98.86	97.97
N-BERT		
$X + \hat{X}$	100	98.86

Table 1: Labeled 테스트셋에 대한 의도 분류 결과

Table 1 은 각각 단일 BERT 를 labeled text X, recognized text $\hat{X}$로 학습시킨 뒤 X의 테스트셋에 대한 평가 결과이다. 단일 BERT 로 학습된 두 결과 모두 비슷한 성능을 나타내고 있으며, 이는 사전 학습된 BERT 가 대규모 text 데이터셋으로부터 언어모델을 충분히 표현하기 때문에 fine-tuning 에 잡음이 추가된 데이터가 오더라도 좋은 퍼포먼스를 보여주는 것을 알 수 있다. 마찬가지로 본 논문에서 제안한 N-BERT 도 두 데이터셋에 대해 높은 성능을 내고 있다.

Model	평가 데이터셋: $\hat{X}$	
	FSC	A-SNIPS
	의도 분류 성능(%)	
BERT		
X (100%)	89.27	90.79
X (100%) + $\hat{X}$(10%)	95.36	97.62
X (100%) + $\hat{X}$(99.9 %)	97.92	98.16
N-BERT		
$X + \hat{X}$	98.49	98.54

Table 2: Recognized text 셋에 대한 각 모델의 실험 결과

Table 2 는 recognized text 의 테스트셋에 대한 실험 결과이다. FSC, Audio-SNIPS 두 데이터셋에 대하여 recognized text 의 비율이 높아질수록 단일 BERT 의 정확도개 개선된다. 하지만, 단일 BERT 로 사용한 의도 분류 성능은 labeled text 셋에 비해 성능 저하가 일어나는 것을 알 수 있다.

반면에 제안한 N-BERT 는 recognized 테스트셋에 대해 각각 98.49%, 98.45%의 결과를 보였으며, Table 1 에서 볼 수 있듯이, N-BERT 는 labeled text 에서도 성능 저하가 크게 일어나지 않는 것을 알 수 있다.

Ⅲ. 결론

본 논문에서는 음성인식 오류로부터 강인한 언어모델 연구에 초점을 맞추었다. 이를 위해 음성인식 시스템을 통과하여 오류를 포함한 recognized text 셋으로 N-BERT 를 학습하고, 원본 그대로인 전사 데이터 labeled text 셋으로 C-BERT 를 학습하는 이중 학습 방법을 제안하였다. N-BERT 와 C-BERT 를 학습한 언어모델의 테스트 성능은 오류가 포함된 입력과 더불어 오류가 포함되지 않은 입력에 대해서도 강인한 성능을 보여주었다. 이를 통해 제안하는 방법이 실제 사용 환경에서 발생하는 음성인식 오류들에 대하여 강인한 언어모델을 제공하여 사용자 음성언어 대상 의도 파악 시스템을 개선할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2021-2020-0-01808*)

참 고 문 헌

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017.

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[3] Lai, Cheng-I., et al. "Towards semi-supervised semantics understanding from speech." arXiv preprint arXiv:2011.06195 (2020).

[4] Lugosch, Loren, et al. "Speech Model Pre-Training for End-to-End Spoken Language Understanding" Interspeech 2019-818.

[5] Baevski, Alexei, et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." Advances in Neural Information Processing Systems 33 (2020).