Journal of The Korea Convergence Society
Vol. 12. No. 11, pp. 109-117, 2021
I-SSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.11.109

Back TranScription(BTS)기반 데이터 구축 검증 연구

박찬준[1], 서재형[1], 이설화[1], 문현석[1], 어수경[1], 임희석[2*]
[1]고려대학교 컴퓨터학과 석·박사통합과정, [2]고려대학교 컴퓨터학과 교수

A Study on Verification of Back TranScription(BTS)-based Data Construction

Chanjun Park[1], Jaehyung Seo[1], Seolhwa Lee[1], Hyeonseok Moon[1], Sugyeong Eo[1], Heuiseok Lim[2*]
[1]Master & Ph.D. Combined Student, Department of Computer Science and Engineering, Korea University
[2]Professor, Department of Computer Science and Engineering, Korea University

요 약 최근 인간과 컴퓨터의 상호작용(HCI)을 위한 수단으로 음성기반 인터페이스의 사용률이 높아지고 있다. 이에 음성인식 결과에 오류를 교정하기 위한 후처리기에 대한 관심 또한 높아지고 있다. 그러나 sequence to sequence(S2S)기반의 음성인식 후처리기를 제작하기 위해서는 데이터 구축을 위해 human-labor가 많이 소요된다. 최근 기존의 구축 방법론의 한계를 완화하기 위하여 음성인식 후처리기를 위한 새로운 데이터 구축 방법론인 Back TranScription(BTS)이 제안되었다. BTS란 TTS와 STT 기술을 결합하여 pseudo parallel corpus를 생성하는 기술을 의미한다. 해당 방법론은 전사자(phonetic transcriptor)의 역할을 없애고 방대한 양의 학습 데이터를 자동으로 생성할 수 있기에 데이터 구축에 있어서 시간과 비용을 단축할 수 있다. 본 논문은 기존의 BTS 연구를 확장하여 어떠한 기준 없이 데이터를 구축하는 것보다 어투와 도메인을 고려하여 데이터 구축을 해야함을 실험을 통해 검증을 진행하였다.

주제어 : 기계번역, BackTranScription, 병렬말뭉치, 음성인식, 딥러닝, 언어융합

Abstract Recently, the use of speech-based interfaces is increasing as a means for human-computer interaction (HCI). Accordingly, interest in post-processors for correcting errors in speech recognition results is also increasing. However, a lot of human-labor is required for data construction. in order to manufacture a sequence to sequence (S2S) based speech recognition post-processor. To this end, to alleviate the limitations of the existing construction methodology, a new data construction method called Back TranScription (BTS) was proposed. BTS refers to a technology that combines TTS and STT technology to create a pseudo parallel corpus. This methodology eliminates the role of a phonetic transcriptor and can automatically generate vast amounts of training data, saving the cost. This paper verified through experiments that data should be constructed in consideration of text style and domain rather than constructing data without any criteria by extending the existing BTS research.

Key Words : Machine translation, BackTranscription, Parallel corpus, Speech recognition, Deep learning, Language convergence

*This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)" and this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received August 12, 2021 Revised September 2, 2021
Accepted November 20, 2021 Published November 28, 2021

1. 서론

Automatic Speech Recognition(ASR)이란 사람의 음성을 텍스트로 변환해주는 기술이다. 이 기술은 딥러닝의 등장으로 급격한 인식률 향상을 이루었으며 많은 기업에서 해당 기술을 이용해 비즈니스 모델에 적용하고 있다[1].

Google Cloud Speech API[2], Naver의 CLOVA Speech[3] 등 훌륭한 상용화 API 시스템이 존재하나 대부분의 중소 기업에서는 Kaldi[4]와 같은 오픈소스를 이용하여 자체적으로 ASR software를 구축하고 있다. 이는 기업 내 in-house data에 대한 보안유지, 도메인특화 시스템의 필요성 등을 이유로 이러한 방식을 채택하고 있다. 더불어 실제 서비스를 진행할 시 디코딩 속도를 고려하여 음향모델과 언어모델 기반의 Gaussian Mixture Models(GMM)[5], Hidden Markov Models(HMM)[6] 등과 같은 전통적인 음성인식 아키텍쳐를 기반으로 서비스를 진행하고 있는 기업들이 많이 존재한다.

그러나 이러한 방법론은 음성인식기의 어휘사전에 없는 단어는 여전히 인식할 수 없으며 해당 부분은 다른 어휘로 오인식된다는 문제점이 존재한다. 또한 통계기반 방법이기 때문에 엄청난 양의 음성 DB가 구축이 되어야 만족할 만한 성능이 나온다는 한계점이 존재한다. 또한 음성인식 전문가가 아니면 모델을 다루기 어려워 진입장벽이 높다는 단점이 존재한다.

이러한 한계점을 극복하기 위하여 최근 Pretrained model(PM) 기반의 전이학습을 적용한 ASR 연구가 이루어지고 있다[7-9]. 해당 방법론은 전통적인 ASR 연구들보다 월등한 성능을 보이고 있으나 이를 적용하여 실세계에서 서비스를 진행하기에는 크게 2가지 측면에서 한계점이 존재한다.

첫째로 데이터 측면이다. 해당 방법론을 적용하여 ASR을 서비스하기 위해서는 사전학습을 위해 많은 양의 학습 데이터가 필요하다. 데이터 크기에 크게 의존적인 방법론이기에 한국어와 같은 저자원 언어(low resource language(LRL))에 적용하기에 어려움이 존재한다는 문제가 있다[10]. 더 나아가 최신 연구들은 훈련 데이터가 충분히 존재하는 고자원 언어(high resource language (HRL))를 기반으로 연구를 진행하기 때문에, 이와 동일한 모델을 별다른 처리 없이 저자원 언어에 적용한다면 동일한 성능을 내리라 기대하기 힘들다[11].

두번째로 서비스 환경적 측면이다. 해당 방법론은 대용량 데이터를 처리하기 위하여 충분한 computing power(e.g., GPU)를 갖춘 서비스 환경이 요구된다. Google, Facebook과 같은 대기업이 아니고서야 서비스를 위한 충분한 하드웨어 환경을 갖추기가 어렵다. 즉 모델을 훈련할 때 많은 파라미터들과 데이터를 요구하기 때문에, 서버나 GPU 환경이 충분하지 않은 기업은 최신 모델을 이용한 서비스환경 구성 및 성능 개선에서 많은 어려움이 발생한다. 따라서 저자원 언어에 대해서도 좋은 성능을 내면서 GPU 학습 환경이 부족한 기업도 서비스를 하는데 무리가 없기 위하여, Pretrain model(PM) 기반의 전이학습을 적용한 방법론이 아닌 새로운 방향의 ASR 성능 개선을 이룰 수 있는 방안을 대한 연구가 필요하다.

해당 한계점을 완화하기 위하여 모델의 변경 없이 각종 pre-processing 및 post-processing을 통해 ASR 모델의 성능을 향상시킬 수 있다는 연구의 움직임이 생겨나고 있다[12-14]. 해당 방법론은 ASR 서비스를 진행함에 있어서 데이터 측면에서의 한계점과 서비스 환경적 측면에서의 한계점을 동시에 완화시킬 수 있다.

ASR 모델을 변경하는 방법론이 아니기에 대용량의 학습 데이터를 요구하지 않아 데이터 측면에서의 한계점도 완화가 가능하며 어떠한 모델에도 적용 가능하기에 vanilla Transformer[15]와 같은 CPU로도 충분히 서비스가 가능한 모델에 결합할 수 있어 서비스 환경적 측면에서의 한계점을 완화시킬 수 있다. 따라서 저자원 언어 같은 경우 해당 연구의 중요성이 더 강조된다.

이러한 흐름을 반영하여 음성인식기의 모델 변화 없이 ASR의 성능개선을 이룰 있는 sequence to sequence(S2S) 기반의 후처리기 모델을 위한 데이터 구축 방법론인 Back TranScription(BTS)가 제안되었다[16]. 해당 방법론은 아직 초기단계에 연구가 진행중이며 여러 방면으로 검증이 필요하다.

이에 본 논문은 BTS에 대한 데이터 구축 방향성에 대한 검증을 다양한 실험을 통해 진행하였다. 학습 데이터를 구축할 때 가이드라인과 명확한 기준이 중요하듯, BTS 기반 학습 데이터 구축 시에도 어떠한 기준점이 필요하다고 판단되어 도메인과 어투를 기준으로 데이터 구축 방향성에 대한 검증을 진행하였다. 동일한 도메인을 기반으로 데이터를 구축하는 것이 좋을지, 아니면 이질적인 도메인을 기반으로 데이터를 구축하는 것이 좋을지에 대한 검증을 진행하였고, 구어체 및 문어체 등어투에 통일성을 부여하여 데이터를 구축하는 것이

좋을지에 대한 검증도 진행하였다.

본 논문은 2장에서 기존 음성인식 후처리기 연구들에 대해서 자세히 다루고, 3장과 4장을 통해 BTS 연구에 대한 심도있는 소개와 필요성에 대해서 서술한다. 5장을 통해 실질적인 실험과 검증을 진행 후 6장에서 결론으로 본 논문을 마무리한다.

2. 음성인식 후처리기 연구

음성인식 후처리(ASR post-processing)란 모델의 구조를 변경하여 ASR의 성능을 향상시키는 것이 아닌, 인식된 결과 문장에 대한 교정을 통해 성능을 향상시키는 연구분야이다. 음성인식 분야에서 음성인식 후처리를 적용할 수 있는 방법론은 크게 Conventional Methodology와 Sequence to Sequence(S2S) 방식 2가지가 존재한다.

첫번째로 규칙 및 통계기반 방법론이다[17,18]. 기업에서 ASR 서비스를 진행하면서 사내 자체적인 규칙을 구축하여 ASR의 성능 향상을 위해 노력 중이며 언어학적 규칙을 적용하여 음성인식 결과의 품질을 향상시키고 있다. 이러한 방법론은 충분한 규칙을 제작하는데 엄청난 비용과 시간이 투자되며 규칙 간의 충돌이 발생할 수도 있다는 한계점이 존재한다. 또한 각각의 모듈이 독립적으로 구현이 된다는 단점이 존재한다. 또한 N-gram 언어모델을 이용하여 후처리를 진행하는 방법연구도 존재하나 통계기반 방식은 많은 데이터가 필요하고 문맥을 고려하지 못한다는 한계가 존재한다[19].

두번째로 S2S을 이용하는 방법으로, 기계번역과 같은 방식으로 접근하여 오류를 수정하는 방법론이다. S2S 모델을 기반으로 encoder를 통해 음성인식 결과를 vector화 시킨 후 decoder를 통해 vector를 복호화시켜 사람이 수정한 음성인식 결과 문장을 생성해내는 방법론이다[20,21]. 이러한 방법론은 전통적인 규칙 및 통계기반 방식에 비해서 월등히 좋은 성능을 보이고 있다. 그러나 S2S 방법론 기반의 음성인식 후처리기 역시 크게 데이터 구축적 측면과 서비스적 측면에서 한계가 존재한다.

첫째로 데이터 구축적 측면에서의 한계점이다. 학습을 위한 공개된 데이터가 존재하지 않으며, 사람이 직접 음성인식 후처리기를 위한 parallel corpus를 구축해야 한다. 학습데이터의 형태가 (음성인식 결과(speech recognition sentence), 사람이 수정한 문장(Human post edit sentence))로 구성되며 이를 구축하기 위하여 음성인식 결과를 전사할 많은 인력이 필요하다. 즉 데이터 구축적 측면에서 인건비 및 시간이 많이 든다는 한계가 존재한다. 더불어 전사자에 따라서 품질 차이가 발생할 수도 있다는 한계가 존재한다. 동일한 문장에 대해서 각기 다른 전사자가 다른 인식 결과를 전사할 경우가 존재하며 이는 결국 모델 성능 하락으로 직결된다.

둘째로 서비스적 측면에 대한 한계점이다. 최근 대부분의 NLP 연구들은 Pretrain-Finetuning Approach(PFA)를 기반으로 연구하고 있으나 중소기업 혹은 하드웨어가 부족한 기업에서는 해당 기술을 이용하여 서비스하는 것은 느린 속도, 메모리 부족 등으로 인해 많은 한계점이 존재한다. XLM-E[22], mt5[23]와 같은 연구들이 발표되고 있으며 이는 현재 가장 좋은 성능을 보이고 있으나 해당 모델들은 파라미터 및 모델 크기 등이 너무 커 실제 기업에서 서비스를 진행하기에는 아직 무리가 존재한다. 또한 이러한 방법론은 데이터의 크기에 의존적이라 고자원 언어에는 적용하기 쉬우나 저자원 언어에 적용하기에 어려움이 존재한다는 한계가 있다.

3. 음성인식 서비스의 4대 한계점

첫 번째로 음성인식결과를 살펴보면 띄어쓰기가 제대로 구분되지 않는, 분절(segmentation)에서 한계점이 존재한다. 이를 해결하기 위하여 자동 띄어쓰기(automatic spacing module)에 대한 연구가 다수 이루어지고 있으나 ASR에 특화된 연구는 거의 존재하지 않는다[24]. 띄어쓰기 문제가 해결되지 않고 기업에서 ASR을 기반한 서비스를 진행할 시 end-user들의 만족도가 떨어지고 음성인식결과에 대한 신뢰를 잃을 수도 있다.

두 번째로 외래어 변환 처리를 잘못하는 경우가 존재한다. 예를 들어 “Lotte Tower는 123층입니다”라는 결과를 내보내야하지만 “롯데타워는 123층입니다” 라고 결과를 내보내는 시스템이 존재한다. 이러한 것은 심각한 문제점은 아니나 해결이 된다면 end-user들에게 가독성을 높여주고 만족감을 높여줄 수 있다. 즉 외래어 변환 문제는 서비스 만족감을 위해 필요한 요소이다.

세 번째로 기호부착(쉼표, 마침표, 느낌표, 물음표 등)의 문제이다. “지금 어디가세요?”라고 출력을 해야하나 대부분의 음성인식기는 “지금어디가세요”처럼 기호를 부착하지 않고 출력을 내보낸다. 기호부착이 되지 않을 경우 문장분리 지점을 사용자가 파악하기가 어려우며 화자의 의도를 파악하기 어렵다는 단점이 존재한다. 더불어 상용화 시스템마저도 기호부착을 하지 않는

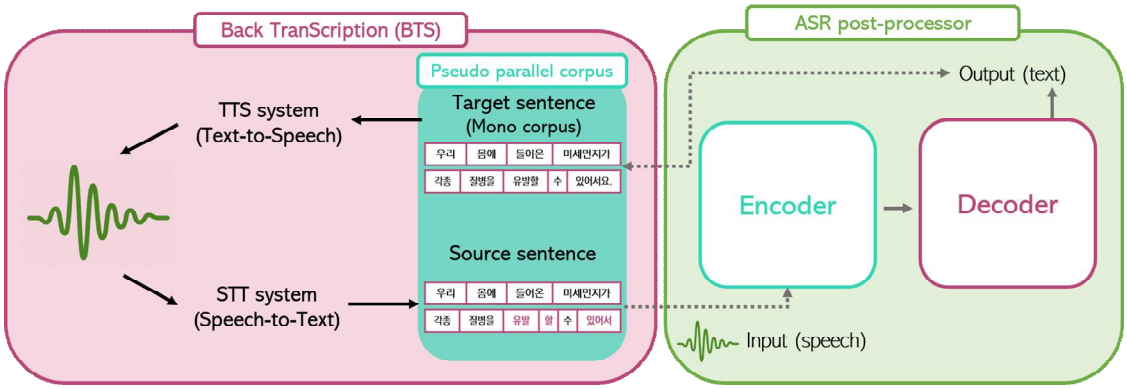


Fig. 1. Overall architecture of BTS [16]

경우가 존재한다. 이러한 문제들을 해결하기 위하여 기호 부착 연구가 독립적으로 이루어지고 있는 추세이다[25].

네 번째로 철자 오류이다. 가장 빈번하게 발견되는 음성인식기의 문제점으로 철자가 오인식된 결과를 출력물로 내보내는 경우가 상당히 많이 존재한다. 이러한 문제점을 해결하기 위하여 철자 오류기 연구가 독립적으로 이루어지고 있다. 그러나 음성인식기에 특화된 연구는 진행된 사례가 거의 없으며 general한 분야를 기반으로 많은 연구들이 진행되었다[20,21].

4. Back TranScription(BTS) 이란?

Back TranScription (BTS)란 음성합성 기술인 TTS(Text To Speech)기술과 STT(Speech To Text) 기술을 결합하여 pseudo parallel corpus를 생성하는 기술을 의미한다[16]. Crawling 등으로 쉽게 구할 수 있는 기 구축된 mono corpus를 TTS 돌려서 음성파일을 만든 후 해당 파일을 STT를 돌려 인식결과를 만든다. 즉 TST(Text-to-Speech-to-Text) 기술을 적용하여 (TTS결과, mono corpus 원문)의 음성인식 후처리기용 pseudo parallel corpus가 생성된다. 기계번역을 기준으로 원시문장(source sentence)에는 TST결과, 목적문장(target sentence)에는 mono corpus 원문이 대입된다.

여기서 TST는 모노 데이터에 대해 음성인식 오류를 생성할 수 있도록 진행하는 프로세스이다. 즉 TST는 TTS와 STT를 결합한 형태로, TTS 모듈은 STT 모듈에서 필요한 데이터를 제공해주며 두 모듈을 거치는 과정에서 음성인식 결과에 대한 오류들이 생성되게 된다. Mono corpus 원문 같은 경우 띄어쓰기가 대부분 완벽하고 외래어 변환 그리고 기호부착이 잘되어 있으며, 철자 오류가 거의 없다. 이러한 특징을 바탕으로 해당 데이터를 이용해 S2S 훈련을 진행하면 고품질의 ASR post-processor를 만들 수 있다.

Fig. 1은 BTS의 구조와 여기에서 파생되는 데이터를 이용해 S2S기반의 음성인식 후처리기 모델 학습 구조를 보여준다. BTS에서 pseudo parallel corpus는 monocorpus로부터 획득된 target sentence와 target sentence를 TTS로 발화한 음성을 STT를 통해 텍스트로 변환한 source sentence로 구성된다. 변환된 source sentence에서는 grammatical error (i.e., spacing, punctuation, 문장의 끝마침의 부재) 문제가 발생한 것을 확인할 수 있다. 최종적으로, 이렇게 생성된 pseudo parallel corpus를 학습 데이터로 사용하여 음성인식 후처리기 모델을 만들 수 있다.

BTS의 장점과 서비스적으로 가치가 있는 이유는 크게 5가지로 분류할 수 있다. 첫째로 학습데이터의 무한한 생성이 가능하다. Parallel corpus를 구축하기 위해서는 많은 시간과 비용이 들고 구하기 힘들다는 단점이 있으나 mono Corpus는 웹을 통해 무한히 구할 수 있다는 장점이 존재한다. 이를 통해 학습데이터의 무한한 구축이 가능하다.

둘째로 기존 음성인식기에 한계점인 띄어쓰기의 문제점, 외래어 변환 문제점, 기호부착 문제점, 철자 오류의 문제점 등을 하나의 모델로 한번에 해결할 수 있다. Mono corpus는 해당 문제에서 자유하기 때문이다. 이를 통해 기존에 독립적으로 연구되었던 각각의 문제들을 본 논문이 제안하는 BTS을 이용해 한번에 해결할 수 있다.

Table 1. Data statistics according to TTS/STT

API	Source	Sentence	Speech token	Speech conversion length (sec)	Processing time (within hour)
TTS					
Google TTS	TED	129,987	7,969,230	2,081,115	36
	AI-HUB	105,000	3,065,086	1,563,990	24
STT					
CLOVA Speech Recognition	TED	-			120
	AI-HUB	-			72

셋째로 구글 음성인식기와 같은 상용화된 음성인식 시스템을 도메인특화 음성인식기로 변환시킬 수 있다. 원하는 도메인의 단일 말뭉치만을 이용해 TTS를 제작하고 구축된 병렬 말뭉치로 후처리기를 제작하면 상용화 시스템을 도메인 특화 음성인식기로 서비스 가능하다. 기업에서 상용화 시스템 보다 자체 음성인식기를 구축하는 이유는 도메인 특화 모델의 니즈 때문이나 BTS을 통해 상용화 시스템의 높은 인식률이라는 장점을 살려 도메인 특화 모델로 발전시킬 수 있다.

넷째로 전사자의 역할이 필요 없는 방법론이며 시간과 비용적인 면에서 엄청난 장점을 보유한 방법론이다. 또한 전사자간이 품질차이에 대한 문제에서 자유하다는 장점이 존재한다.

다섯째로 언어확장이 용이하다. 상용화 시스템은 다양한 TTS와 STT 언어에 대해 API 서비스를 제공한다. 이를 통해 다양한 BTS 데이터가 구축가능하다.

즉 본 논문에서 제안하는 BTS는 ASR을 서비스하는 기업에게 실질적인 도움이 될 수 있다.

5. 실험 및 실험결과

본 논문은 실 서비스 환경을 고려한 데이터 구축의 방향성을 검증하는 2가지 실험을 진행하였다. 첫째로 도메인을 고려한 실험을 진행하였으며 두번째로 어투를 고려한 실험을 진행하였다.

5.1 데이터

실험을 위한 언어는 한국어로 설정하였으며, 데이터 구축을 위하여 크게 두 가지 출처의 mono corpus를 수집하였다. 자세한 데이터 통계치는 Table 1과 같다.

첫 번째로 비즈니스 및 기술 TED에서 한국어 스크립트로 제공된 129,987개의 문장을 웹 크롤링으로 추출했다. 두 번째로 AI-HUB에서 한국어-영어 번역(병렬) 말뭉치 AI 데이터 중 105,000개의 문장을 추출했다[26].

구축한 mono corpus를 기반으로 Google TTS API를 사용해서 mp3 형식의 음성 데이터로 변환했다. TED를 기반한 129,987개 문장은 7,969,230개의 음성 토큰으로 나누어져 2,081,115초의 음성 데이터로 합성하였고 AI-HUB를 기반한 105,000개의 문장은 3,065,086개의 음성 토큰으로 나누어져, 1,563,990초의 음성 데이터로 합성했다. 소요 시간은 TED와 AI-HUB 각각 36시간, 24시간 이내였다. 상용화 시스템을 사용하는 이유로는 기구축된 TTS 시스템이 없는 기업들도 BTS를 사용할 수 있게 하여 진입장벽을 낮추기 위해서이다.

TTS를 통해 구축된 음성 데이터를 바탕으로 네이버 CLOVA Speech Recognition (CSR) API를 사용해서 다시 텍스트 데이터로 변환을 진행한다. 소요 시간은 TED와 AI-HUB 각각 120시간, 72시간 이내였다. 이를 통해 S2S 기반의 ASR post-processor를 위한 229,987 문장 쌍의 pseudo parallel corpus가 구축된다.

Parallel Corpus Filtering(PCF)이란 양질의 병렬 말뭉치를 구축하기 위한 작업이며 좋은 품질의 문장만을 선별하는 작업을 의미한다. 즉 high quality training data만을 training에 이용하여 model의 performance를 올리고자하는 기계번역의 하위분야이다[27,28].

TST를 통해 구축한 pseudo parallel corpus의 경우 STT 및 TTS 시스템의 의도치 않은 오류로 인해 인식이 되어 있지 않아 source 쪽이 비어있거나, 과도하게 길이가 길거나 짧은 이상치 값 등이 존재한다.

이에 [27]에서 제안한 PCF 방법론을 적용하여 고품질의 데이터만을 남기는 작업을 진행하였다. 총 10,669문장이 필터링 되었으며 그 중 대부분은 STT시 인식이 제대로 되지 않아 발생한 저품질의 데이터 때문이었다. 또한 source와 target이 완전히 동일한 문장은 학습을 진행하여도 의미가 없을 수 있기에 제거하였으며 50%이상이 special symbol token으로 구성된 문장도 제거하였다.

PCF를 거치게 되면 mono corpus의 품질이 좋지 못한 문장들과 STT시 인식이 잘못된 문장을 한번에 걸러 줄 수 있다는 장점이 존재한다.

본 논문은 추가적으로 필터링의 효과를 검증하기 위하여 PCF를 진행한 모델과 진행하지 않은 모델의 성능을 비교하였다. 성능 평가는 BLEU[29]와 GLEU[30]를 기반으로 진행한다.

5.2 모델

BTS를 통해 구축된 데이터는 vanilla Transformer 모델을 기반으로 훈련을 진행하였다[15]. 하이퍼 파라미터의 경우의 setting과 동일하게 설정하였다. Vocab size는 32,000개 subword tokenization는 sentencepiece를 이용하였다[31].

5.3 도메인을 고려한 실험

대게 기업에서는 도메인에 특화된 모델을 기반으로 ASR 서비스를 진행한다. 이에 BTS를 후처리기에 적용할 때 도메인에 특화된 데이터로만 학습을 시키는 것이 올바른 방향성일지에 대한 검증이 필요하다. 이에 한국어 TED 도메인과, AI-HUB 도메인을 기반으로 각각 BTS를 적용한 별도의 모델을 제작하였다. 테스트셋은 AI-HUB의 출처로만 구성된 5000개의 테스트셋을 추가적으로 구축하였으며 이를 통해 도메인이 일치 될 때와 일치가 되지 않을 때의 성능변화를 살펴보았다. 실험결과는 Table 2와 같다.

Table 2. Experimental results of considering domain

Model	BLEU	GLEU
Base	36.44	X
TED	40.04 (+3.60)	32.35
TED+PCF	39.71 (+3.27)	31.25
AI-HUB	47.89 (+11.45)	38.84
AI-HUB+PCF	**48.28 (+11.84)**	**39.12**

Base란 원시 문장과 목적 문장간의 BLEU 점수이며 이를 기준으로 성능 향상도를 측정하였다. 실험결과 테스트셋 기준 도메인이 일치된 AI-HUB 모델이 압도적으로 좋은 성능을 보였다. AI-HUB 모델이 11.45 score의 성능 향상을 보인 반면 도메인이 일치하지 않는 TED 모델의 경우 3.60의 성능 향상을 보였다. GLEU를 기준으로도 AI HUB 모델이 38.84 score인 반면 TED 모델은 32.35로 낮은 성능을 보였다. 이를 통해 도메인이 일치하는 mono corpus 만을 이용하여 BTS를 진행하는 것이 더 좋은 후처리기를 만들 수 있는 요인임을 알 수 있었다.

또한 PCF 기술을 적용하니 도메인이 일치하는 AI-HUB 모델에서는 성능 향상을 보였으나 도메인이 일치하지 않은 TED 모델에서는 성능 하락을 보였다. 이는 유의미한 결과이며 진정한 필터링은 도메인의 일관성(consistency)임을 알 수 있다. 즉 BTS를 적용한 음성인식 후처리기를 제작 시 PCF를 적용하는 것도 중요하나 이보다 더 중요한 것은 학습 데이터간 도메인 consistency임을 알 수 있었다.

이러한 현상은 기업에서 서비스를 진행하는 입장에서 의미있는 실험결과이다. 중소기업 혹은 대기업에서 ASR을 서비스 할 시 범용(generic) 모델로 서비스 하기보다는 특정 도메인(콜센터, 비즈니스, 교육 등)에 특화된 ASR을 만들어 서비스한다. 본 실험을 통해서 기업에서 서비스하고 있는 ASR의 도메인의 단일 말뭉치만을 이용해 BTS를 적용하고 후처리기를 제작해 서비스를 해야 함을 알 수 있었다.

5.4 어투를 고려한 실험

두번째로 어투에 따른 성능변화 실험을 진행하였다. 이를 위해 AI-HUB에서 제공하는 구어체 데이터 만을 이용하여 모델을 별도로 제작 후 AI-HUB에서 제공하는 문어체 데이터만으로 구성된 5000개의 테스트셋을 별도로 구축하여 성능을 평가하였다. 즉 구어체만으로 학습된 모델이 문어체에서 후처리 효과가 있는지 검증을 진행하였다. 실험결과는 Table 3와 같다.

Table 3. Experimental results of considering text style

Model	BLEU	GLEU
Base	46.29	X
Colloquial	**48.29 (+2.00)**	**40.81**

실험결과 어투가 달라질 경우 성능의 악영향을 미침을 확인하였다. Base를 기준으로 2점 정도 밖에 성능 향상을 가져오지 못하였다. 즉 Table 2의 도메인이 달랐을 때 Base 기준 3.60의 성능향상 밖에 못가져왔는데 language

style이 달라지니 이보다 더 낮은 2.00의 성능향상을 보였다. 즉 구어체는 구어체끼리, 문어체는 문어체끼리만 BTS를 적용하는 것이 성능 향상에 도움이 됨을 알 수 있었다.

Table 2와 Table 3에서 진행한 실험을 통하여 BTS를 학습시킬 때의 방향성에 대한 지표를 찾을 수 있었으며 이를 통해 양질의 모델을 제작할 수 있다.

6. 결론

본 논문은 S2S 기반의 음성인식 후처리기를 제작하기 위한 새로운 데이터 구축 방법론인 BTS에 대한 데이터 구축 방향성에 대해 논의하였다. 기존 BTS 기반 음성인식 후처리기 연구에서 더 나아가 어떠한 방향성으로 데이터를 구축하는 것이 성능 향상에 더 효과적인지 검증을 진행하였다. 실험결과 도메인과 어투를 고려하여 BTS 기반 음성인식 후처리기 학습 데이터를 구축해야 함을 정량적인 수치로 알 수 있었다. 추후 BTS기반 데이터 구축 시 노이즈 삽입(Noise Injection) 기법을 적용하여 강건한 모델을 훈련할 예정이며 BTS를 다국어로 확장하여 데이터를 추가 배포할 예정이다. 먼저 아시아 언어권의 언어를 바탕으로 추가 구축을 진행 후 유럽어로 확장할 예정이다.

REFERENCES

[1] S. K. Kaya, T. Paksoy & J. A. Garza-Reyes. (2020). The New Challenge of Industry 4.0. *Logistics 4.0: Digital Transformation of Supply Chain Management, 51.*

[2] P. Aleksic, M. Ghodsi, A. Michaely, C. Allauzen, K. Hall, B. Roark & P. Moreno. (2015). *Bringing contextual information to google speech recognition.*

[3] J. W. Ha, K. Nam, J. Kang, S. W. Lee, S. Yang, H. Jung & S. Kim. (2020). ClovaCall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers. *arXiv preprint arXiv:2004.09367.*

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel & K. Vesely. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). *IEEE Signal Processing Society.*

[5] M. N. Stuttle. (2003). A Gaussian mixture model spectral representation for speech recognition *(Doctoral dissertation, University of Cambridge).*

[6] M. Gales & S. Young. (2008). *The application of hidden Markov models in speech recognition.*

[7] A. Baevski, H. Zhou, A. Mohamed & M. Auli. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477.*

[8] C. Wang, J. Pino & J. Gu. (2020). Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation. *arXiv preprint arXiv:2006.05474.*

[9] Z. Q. Zhang, Y. Song, M. H. Wu, X. Fang & L. R. Dai. (2021). XLST: Cross-lingual Self-training to Learn Multilingual Representation for Low Resource Speech Recognition. *arXiv preprint arXiv:2103.08207.*

[10] C. Park, Y. Yang, K. Park & H. Lim. (2020). Decoding strategies for improving low-resource machine translation. *Electronics, 9(10),* 1562.

[11] C. Park, S. Eo, H. Moon & H. S. Lim. (2021, June). Should we find another model?: Improving Neural Machine Translation Performance with ONE-Piece Tokenization Method without Model Modification. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers* (pp. 97-104).

[12] K. Voll, S. Atkins & B. Forster. (2008). Improving the utility of speech recognition through error detection. *Journal of digital imaging, 21(4),* 371.

[13] A. Mani, S. Palaskar, N. V. Meripo, S. Konam & F. Metze. (2020, May). ASR error correction and domain adaptation using machine translation. *In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6344-6348). *IEEE.*

[14] J. Liao, S. E. Eskimez, L. Lu, Y. Shi, M. Gong, L. Shou & M. Zeng. (2020). Improving readability for automatic speech recognition transcription. *arXiv preprint arXiv:2004.04438.*

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez & I. Polosukhin. (2017). Attention is all you need. *In Advances in neural information processing systems* (pp. 5998-6008).

[16] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo & H. Lim. (2021). BTS: Back TranScription for Speech-to-Text Post-Processor using

Text-to-Speech-to-Text. *Proceedings of the 8th Workshop on Asian Translation*, (pp. 106-116).

[17] M. Paulik, S. Rao, I. Lane, S. Vogel & T. Schultz, (2008, March). Sentence segmentation and punctuation recovery for spoken language translation. *In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5105-5108). *IEEE.*

[18] S. Škodová, M. Kuchařová & L. Šeps, (2012, September). Discretion of speech units for the text post-processing phase of automatic transcription (in the czech language). *In International Conference on Text, Speech and Dialogue* (pp. 446-455). *Springer, Berlin, Heidelberg.*

[19] H. Cucu, A. Buzo, L. Besacier & C. Burileanu. (2013, July). Statistical error correction methods for domain-specific ASR systems. *In International Conference on Statistical Language and Speech Processing* (pp. 83-92). *Springer, Berlin, Heidelberg.*

[20] C. Park, K. Kim, Y. Yang, M. Kang & H. Lim. (2020). Neural spelling correction: translating incorrect sentences to correct sentences for multimedia. *Multimedia Tools and Applications*, 1-18.

[21] C. Park, Y. Yang, C. Lee & H. Lim. (2020). Comparison of the evaluation metrics for Neural Grammatical Error Correction with Overcorrection. *IEEE Access, 8,* 106264-106272.

[22] Z. Chi, S. Huang, L. Dong, S. Ma, S. Singhal, P. Bajaj & F. Wei. (2021). XLM-E: Cross-lingual Language Model Pre-training via ELECTRA. *arXiv preprint arXiv:2106.16138.*

[23] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant & C. Raffel. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934.*

[24] C. Lee & H. Kim. (2013). Automatic Korean word spacing using Pegasos algorithm. *Information processing & management, 49(1),* 370-379.

[25] J. Yi, J. Tao, Y. Bai, Z. Tian & C. Fan. (2020). Adversarial transfer learning for punctuation restoration. *arXiv preprint arXiv:2004.00248.*

[26] C. Park & H. Lim. (2020). A Study on the Performance Improvement of Machine Translation Using Public Korean-English Parallel Corpus. *Journal of Digital Convergence, 18(6),* 271-277.
DOI : 10.14400/JDC.2020.18.6.271

[27] C. Park, Y. Lee, C. Lee & H. Lim. (2020). Quality, not quantity?: Effect of parallel corpus quantity and quality on neural machine translation. *In The 32st Annual Conference on Human Cognitive Language Technology* (pp. 363-368).

[28] H. Moon, C. Park, S. Eo, J. Park & H. Lim. (2021). Filter-mBART Based Neural Machine Translation Using Parallel Corpus Filtering. *Journal of the Korea Convergence Society, 12(5),* 1-7.
DOI : /10.15207/JKCS.2021.12.5.001

[29] K. Papineni, S. Roukos, T. Ward & W. J. Zhu. (2002, July). Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

[30] K. Sakaguchi, C. Napoles, M. Post & J. Tetreault. (2016). Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics, 4,* 169-182.

[31] T. Kudo & J. Richardson. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226.*

박 찬 준(Chanjun Park) [학생회원]

· 2019년 2월 : 부산외국어대학교 언어처리창의융합전공(공학사)
· 2018년 6월 ~ 2019년 7월 : SYSTRAN Research Engineer
· 2019년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Data-centric AI, Machine Translation, Grammar Error Correction, Deep Learning
· E-Mail : bcj1210@naver.com

서 재 형(Jaehyung Seo) [학생회원]

· 2020년 8월 : 고려대학교 영어영문학과 및 경영학과(문학사, 경영학사)
· 2020년 9월 ~ 현재: 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Graph Encoder, Commense Reasoning
· E-Mail : seojae777@korea.ac.kr

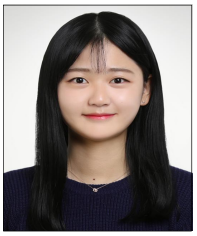
이 설 화(Seolhwa Lee) [학생회원]

· 2015년 2월 : 백석대학교 소프트웨어학과 (공학사)
· 2015년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : 인공지능, 자연어처리, 딥러닝
· E-Mail : whiteldark@korea.ac.kr

문 현 석(Hyeonseok Moon) [학생회원]

· 2021년 2월 : 고려대학교 수학과 및 인공지능학과(이학사, 공학사)
· 2021년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Neural Machine Translation, Natural Language Processing
· E-Mail : glee889@korea.ac.kr

어 수 경(Sugyeong Eo) [학생회원]



· 2020년 8월 : 한국외국어대학교 언어인지과학과, 언어와공학전공 (문학사, 언어공학사)
· 2020년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Neural Machine Translation, Quality Estimation, Deep Learning
· E-Mail : djtnrud@korea.ac.kr

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 : 고려대학교 컴퓨터학과 (이학학사)
· 1994년 : 고려대학교 컴퓨터학과 (이학석사)
· 1997년 : 고려대학교 컴퓨터학과 (이학박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr