JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 2, pp. 267-272, Feb. 2023

Check for updates

기사체-방송체 텍스트 스타일 변환 연구

김 경 민[1] · 임 상 훈[2] · 김 기 백[3] · 오 흥 선[4*]
[1]한국기술교육대학교 컴퓨터공학부 학부과정 [2]한국기술교육대학교 컴퓨터공학과 박사과정
[3]한국기술교육대학교 컴퓨터공학과 석사과정 [4*]한국기술교육대학교 컴퓨터공학과 교수

Text Style Transfer Study for Article-Broadcast Style

Kyung Min Kim[1] · Sang Hun Im[2] · Gi Baeg Kim[3] · Heung-Seon Oh[4*]

[1]Undergraduate, School of Computer Science and Engineering, KOREATECH, Cheonan 31253, Korea
[2]Doctor's Course, School of Computer Science and Engineering, KOREATECH, Cheonan 31253, Korea
[3]Master's Course, School of Computer Science and Engineering, KOREATECH, Cheonan 31253, Korea
[4*]Professor, School of Computer Science and Engineering, KOREATECH, Cheonan 31253, Korea

[요 약]

뉴스 텍스트-음성 변환 과정에는 기사체-방송체 변환이 요구된다. 기사체-방송체는 content 훼손에 민감하다는 특성을 가지고, 기사체의 괄호와 종결어미를 처리하여 방송체로의 변환이 가능하다. 스타일 토큰 기반 텍스트 스타일 변환 모델은 문장의 일부분만 변형하기 때문에, content 훼손을 최소화할 수 있어 기사체-방송체 변환에 적합하다. 그러나 기존 연구에서는 비병렬 데이터를 사용하여 스타일 토큰 학습이 어렵다는 단점이 있다. 병렬 데이터는 같은 content를 가진 두 문장에서 다른 부분을 명확한 스타일 토큰으로 구분 가능하지만, 구축에 비교적 많은 비용이 요구된다. 프롬프팅을 적용하여 학습에 요구되는 데이터를 줄일 수 있으나, 기존 방식으로는 스타일 토큰의 content 유지가 불가하다는 문제가 발생한다. 본 논문에서는 기사체-방송체 병렬 데이터 2,000건을 구축하였으며, 스타일 토큰의 content를 유지시키는 콘텐트 마커 프롬프팅을 새롭게 제안하였다. 또한 기사체-방송체 데이터셋에서 EM 0.9978의 높은 성능을 달성하였다.

[Abstract]

The news text-to-voice conversion process requires article-broadcast style transfer. The article-broadcast style has the characteristic of being sensitive to content corruption, and can be converted into a broadcast style by processing the parentheses and final suffixes of the article style. Since the style token-based text style transfer model modifies only a portion of the sentence, it is suitable for article-broadcast transfer as it can minimize content corruption. However, there is a disadvantage that learning style tokens are difficult in studies using non-parallel data. In parallel data, different parts of two sentences with the same content can be distinguished by clear style tokens, but it requires much cost to build. Although it is possible to reduce the data required for learning by prompting, there is a problem that the content of the style token cannot be maintained. In this paper, we construct 2,000 parallel article-broadcast data, and newly propose Content Marker Prompting that maintains the content of style tokens. The high performance of EM 0.9778 was achieved on the article-broadcast dataset.

색인어 : 인공지능, 딥러닝, 데이터셋, 자연어처리, 텍스트 스타일 변환
Keyword : Artificial Intelligence, Deep Learning, Dataset, Natural Language Processing, Text Style Transfer

http://dx.doi.org/10.9728/dcs.2023.24.2.267

Received 06 December 2022; **Revised** 10 January 2023
Accepted 17 January 2023

***Corresponding Author; Heung-Seon Oh**
Tel: ※ 개인정보 표시제한
E-mail: heungseon.oh@gmail.com

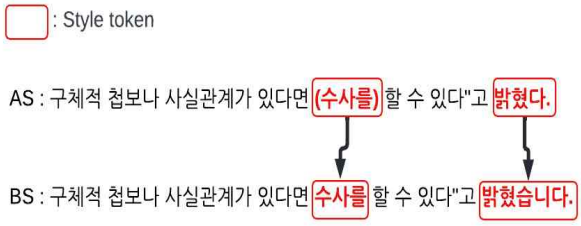
Ⅰ. 서 론

최근 텍스트보다 영상이나 음성 미디어에 익숙한 디지털 세대의 증가로 기존 뉴스 기사를 자동으로 음성 미디어로 변환하려는 시도가 증가하고 있다. 뉴스 텍스트-음성 변환에는 각 형태에 어울리는 어체인 기사체(AS; Article Style)-방송체(BS; Broadcast Style)로의 변환이 요구된다. 뉴스 기사에 사용되는 기사체는 "-하다" 형태의 해라체 종결어미를 가지며 추가적인 정보 제공을 위하여 괄호가 사용된다. 반면 뉴스 방송에 사용되는 방송체는 "-합니다" 형태의 합쇼체 종결어미를 가지며, 괄호는 사용되지 않는다. 따라서 기사체-방송체 변환은 종결어미(Final-Suffix)와 괄호(Parenthesis)를 처리하여 해결할 수 있다. 예를 들어 그림 1의 기사체 문장은 유창성을 고려하여 괄호 "(수사를)" 을 "수사를" 으로 바꾸고, 종결어미 "-다" 를 "-습니다" 로 바꾸어 방송체로 변환한다. 불규칙적인 종결어미 변환과 문맥, 유창성의 파악이 필요한 괄호 변환을 위해서는 언어의 추상적 특징을 활용하는 텍스트 스타일 변환(TST; Text Style Transfer) 모델이 요구된다.

텍스트는 문체, 종결어미 등으로 정의되는 style과 텍스트가 담고 있는 핵심 정보인 content로 구분되며, TST란 content는 유지한 채 style만 변환하는 작업이다. 그림 1의 기사체 문장에서 괄호 "(수사를)" 이나 종결어미 "-다" 와 같이 특정 style에서 나타나는 표현을 스타일 토큰(style token)이라고 하며, 스타일 토큰만을 변환시키는 것으로 TST가 가능하다[1]. 그림 1에서 볼 수 있듯이 스타일 토큰 이외의 부분에서는 content가 온전히 유지된다. 뉴스는 content의 훼손에 굉장히 민감하므로, 스타일 토큰 기반의 TST 모델이 적합하다.

기존 스타일 토큰 기반의 TST 연구는 비병렬 데이터에서 이루어졌으며, 별도의 분류 모델을 사용하여 스타일 토큰을 정의하였다[1],[2]. 그러나 단어 간의 상호작용이 이루어지는 텍스트의 특성상, 명확한 스타일 토큰 학습이 어렵다[3]. 반면 병렬 데이터셋은 같은 content를 가지는 다른 style의 문장 쌍으로 구성되기 때문에 두 문장 간의 다른 부분은 style의 차이에서 오는 것이며, 이는 명확하게 스타일 토큰으로 정의할 수 있다. 따라서 content의 훼손에 민감한 기사체-방송체 변환에서는 병렬 데이터를 사용하는 스타일 토큰 기반 TST가 더 적합하다. 그러나, 병렬 데이터 구축은 비병렬 데이터에 비해 많은 비용이 요구되므로 적은 데이터로 모델을 학습시킬 방법이 필요하다.

프롬프팅(prompting)은 사전학습과 유사하게 미세 조정(fine-tuning)을 진행하는 방식으로, 적은 데이터로 높은 성능 향상을 보였다[4]. 대표적인 사전학습 방법인 span corruption[5]은 문장의 마스크 처리된 부분을 예측하도록 학습된다. 이는 스타일 토큰 기반 TST 방법론과 동일하여 별도의 처리 없이 프롬프팅이 적용 가능하지만 마스크 처리된 스타일 토큰의 content를 유지할 수 없다.



*The actual data applied is Korean, so it is inevitable to insert Korean

그림 1. 스타일 토큰을 통한 텍스트 스타일 변환 과정
Fig. 1. Text style conversion process with style token

본 논문에서는 기사체-방송체 병렬 데이터셋을 구축하고 스타일 토큰 기반 TST 모델에서 content 유지가 가능한 콘텐트 마커 프롬프팅(Content Marker Prompting)을 새롭게 제안한다. 기사체-방송체 병렬 데이터셋은 기사체 문장 2,000건을 수집하고 각각 대응하는 방송체 문장을 직접 작성하여 구축하였다. 콘텐트 마커는 마스크 처리된 원본 스타일 토큰으로 모델의 기존 입력에 추가하여 프롬프팅의 효과를 내면서 content는 유지한다. 콘텐트 마커 프롬프팅은 구축된 기사체-방송체 데이터셋에 EM(Exact Match) 0.9778의 높은 성능을 달성하였다. 또한 콘텐트 마커의 효과를 모델의 실제 출력으로 분석하였다.

Ⅱ. 관련연구

2-1 Prompt

사전학습 언어모델(pretrained language model)은 대량의 데이터를 통해 언어의 특성을 학습하며, 미세조정을 통해 여러 세부 태스크(downstream task)에 적용된다. GPT-2[4]에서는 별도의 세부 태스크 데이터셋 없이 사전학습하는 과정에서 얻은 in-context 정보만으로도 태스크 해결이 가능함을 보였다.

프롬프팅 방식이란 이러한 in-context 정보를 최대로 사용하기 위해 고안된 방식이다. 특히, 사전학습과 유사한 방식으로 미세조정을 진행하는 방식을 사용하여 여러 세부 태스크에서 few-shot learning만으로 높은 성능을 보였다[6]. 감성 분류 태스크의 경우 “문장은 <MASK>적이다”와 같은 프롬프트를 문장에 추가하고, <MASK>에 긍정 혹은 부정의 결과를 생성하게 하여 해당 태스크를 사전학습과 유사한 방식으로 바꾸어 프롬프팅 방식의 적용이 가능하다[7]. 이외에도 질의응답[8], 개체 탐지[9], 관계 추출[10] 등 다양한 세부 태스크에서의 연구가 진행되었으며 적은 데이터를 활용하여 기존 모델 대비 높은 성능을 보였다.

2-2 Text style transfer

텍스트는 문장이 담고 있는 핵심 정보인 content와 문장의 문체나 종결어미로 나타나는 style로 구분된다. TST란 문장의 content를 훼손하지 않으면서 style을 변환하는 태스크를 의미한다.

TST 모델은 사용된 데이터의 형태에 따라 비병렬 데이터 모델과 병렬 데이터 모델로 구분된다. 비병렬 데이터는 content의 구분 없이 해당 style을 가진 문장들로 이루어져 있는 반면 병렬 데이터는 하나의 content를 가진 다른 style의 문장쌍들로 구성된다. 비병렬 데이터 모델의 경우 다양한 방법론들이 제안되었지만, 병렬 데이터 모델은 주로 기계번역에 사용되는 seq2seq 모델을 중심으로 연구되었다[11].

스타일 토큰 기반 TST 모델은 비병렬 데이터에서 연구된 방법론 중 하나이다. 스타일 토큰이란 특정 style에서 특징적으로 나타나는 표현으로, 스타일 토큰만 변환하여 TST가 가능하다. 스타일 토큰 기반 TST 모델로는 검색 기반의 스타일 토큰 분류기를 사용하거나[1] 스타일 토큰 분류기를 학습시키는 방법[2] 등이 연구되었다. 그러나 비병렬 데이터에서는 모든 텍스트가 상호작용하는 특성상, 명확한 스타일 토큰의 학습이 어렵다는 문제가 존재한다[3].

프롬프팅을 통해 TST를 진행하는 연구는 높은 성능의 사전학습 언어모델에 텍스트와 요구 style을 입력하여 zero shot 혹은 few-shot으로 변환된 style의 문장을 생성하였다[12]. 그러나 이러한 방식은 언어모델의 성능에 크게 의존하며 content의 유지가 보장되지 않는다.

비병렬 데이터를 사용할 경우 스타일 토큰 구분이 명확하지 않아 기사체-방송체 변환을 위해서는 병렬 데이터를 사용할 필요가 있다. 그러나 병렬 데이터는 구축에 비교적 많은 시간과 비용이 소요되므로 데이터 수를 줄일 수 있는 방법에 대한 연구가 필요하다.

Ⅲ. 방송체-기사체 병렬 데이터셋

기사체는 텍스트 형태의 뉴스 기사에서 사용되는 style, 방송체는 방송 뉴스와 같은 미디어에서 사용되는 style을 의미한다. 기사체와 방송체는 정확한 정보를 전달해야 하는 뉴스의 특성상 content의 훼손에 굉장히 민감하다. 예를 들어 "어제 밤 범인이 검거되었다." 라는 문장에서 "어제 범인이 잡혔다." 라는 문장으로 변환된다면, "밤"과 "검거"라는 content에서 훼손이 발생한다. 이러한 content의 훼손은 전체적인 문맥은 유지하지만 부정확한 정보가 전달될 수 있다.

기사체와 방송체 모두 문장의 구조나 순서가 높은 수준으로 정제되어 있으므로 단어의 배치나 문장의 구조와 같은 큰 틀은 유지되며, 종결어미와 괄호와 같은 스타일 토큰들이 변환되는 것으로 이루어진다. 변환되는 스타일 토큰은 명확한 규칙성을 가지며 종결어미 변환과 괄호 내 정보(CP; Content in Parenthesis) 유지/삭제 3가지로 구분된다.

종결어미 변환은 기사체의 "-다" 형태의 해라체에서 방송체의 "-입니다" 형태의 합쇼체로의 변환으로 나타난다. 해라체-합쇼체 변환은 다양한 형태를 가지며 여러 불규칙적 변환이 존재한다. 예를 들어 표 1-Change final-suffix에서 등장하는 "공개했다" 는 "-다" 가 "-습니다" 로 변환되지만, 표 1-Preserve C.P 2에서는 "-이다" 가 "-입니다" 로 변환된다.

괄호 내 정보 유지/삭제는 괄호가 사용된 목적에 따라 다른 변환 양식을 보인다. 예를 들어 표 1-Delete C.P의 "(프랑스)", "(UCL)" 과 표 1-Preserve C.P 2의 "(약 13억 8천만원)" 은 추가적인 정보를 제공하기 위하여 사용하였지만, 표 1-Preserve C.P의 "(수사를)" 은 인용한 말의 유창성을 위하여 사용되었다. 방송체에서는 괄호를 사용하지 않기 때문에, 유창성과 모호성을 고려한 변환이 요구된다. "(프랑스)" 와 "(UCL)" 의 경우 삭제되더라도 정보의 모호성이 발생하지 않으므로 삭제한다. 그러나 "(약 13억 8천만원)" 의 경우 환율 등의 이유로 정보의 모호성이 발생할 수 있으므로 "한화 약 13억 8천만원" 으로 변환하여 괄호는 삭제하되 정보는 유지하여 준다. "(수사를)" 또한 삭제 시에 유창성에 문제가 발생하므로, 괄호를 벗긴 "수사를" 으로 변환하여 괄호 내 정보를 유지한다.

해당 규칙들을 바탕으로 데이터셋 2,000건을 구축하였으며, 하나의 인스턴스는 동일한 content를 가지는 기사체-방송체 문장 쌍으로 구성하였다. 기사체 문장은 2020년~2022년 사이 출간된 중앙일보와 연합뉴스의 무작위 기사 321건의 크롤링을 통해 수집하였고, 방송체 문장은 수집된 기사들을 바탕으로 방송사와의 협업을 통해 사람이 변환하여 구축하였다.

크롤링 된 기사의 분류는 원본 기사에서 제공하는 분류를 바탕으로 작성되었고, 각 분류별 기사 수는 표 2와 같다. 문장당 스타일 토큰은 평균 1.37회 등장하며 최대 6개의 스타일 토큰이 한 문장에서 등장하였다. 각 스타일 토큰 종류별 통계는 표 3과 같다.

표 1. 스타일 토큰 별 실제 예시

Table 1. Actual example by style token

*The actual data applied is Korean, so it is inevitable to insert Korean

Change final-suffix	AS: 국민대학교 앞에서 1인 시위를 한 사진을 **공개했다.** BS: 국민대학교 앞에서 1인 시위를 한 사진을 **공개했습니다.**
Delete C.P	AS: 1일 마르세유**(프랑스)**를 상대로 치른 챔피언스리그**(UCL)** 조별리그 경기에서 전반 27분 만에 교체됐다. BS: 1일 마르세유를 상대로 치른 챔피언스리그 조별리그 경기에서 전반 27분 만에 교체됐습니다.
Preserve C.P	AS: "구체적 첩보나 사실관계가 있다면 **(수사를)** 할 수 있다"고 밝혔다. BS: "구체적 첩보나 사실관계가 있다면 **수사를** 할 수 있다"고 밝혔습니다.
Preserve C.P 2	AS: 6개국에서 124만 달러**(약 13억8천만원)**의 수출을 달성한 영어조합법인 '해연'이다. BS: 6개국에서 124만 달러, **한화 약 13억8천만원의** 수출을 달성한 영어조합법인 '해연'입니다.

표 2. 분류 별 문장 수
Table 2. Number of sentences by section

Social	Politics	Economy	Global	Etc	Total
632	501	307	305	255	2,000

표 3. 스타일 토큰 별 등장 통계
Table 3. Appearance statistics by style token

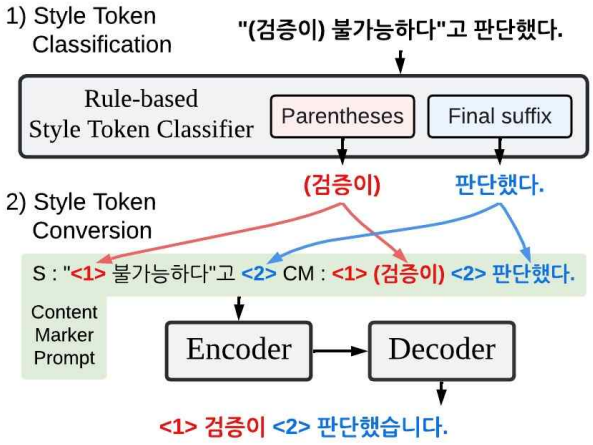
	Sentence (Ratio)	Style token
Change final-suffix	2,000 (1.0000)	2,000
Preserve C.P	291 (0.1455)	400
Delete C.P	352 (0.1760)	422

Ⅳ. 제안 방법

스타일 토큰 기반의 TST 모델은 그림 2와 같이 1) 스타일 토큰 분류(Style Token Classification)와 2) 스타일 토큰 변환(Style Token Conversion) 두 단계로 이루어진다. 기사체-방송체의 스타일 토큰 분류는 표 1에서 볼 수 있듯이 명확한 규칙을 발견하여 규칙 기반 모델을 사용하였다. 종결어미 변환의 경우 문장의 마지막 부분에 있는 단어를 마스크 처리하며, 괄호 내 정보 유지/삭제의 경우 괄호 및 괄호 내 단어를 마스크 처리하였다. 예를 들어 그림 2의 문장에서 나타난 괄호인 "(검증이)" 와 마지막 단어인 "판단했다." 가 스타일 토큰으로 분류되었다. 스타일 토큰 변환은 원본 스타일 토큰을 목표 style에 맞추어 변환하는 과정으로, 인코더-디코더(encoder-decoder) 기반 사전학습 언어모델인 mT5[13]를 사용했다. 모델의 인코더 입력으로 스타일 토큰을 마스크 처리한 문장이 들어가며, 디코더 출력으로 마스크 처리된 부분에 해당하는 스타일 토큰이 목표 style에 맞추어 생성된다.

적은 데이터를 활용한 효과적인 학습을 위하여 프롬프팅 방식을 적용하였다. 마스크 처리된 부분을 예측하는 과정은 mT5의 사전학습 방식인 span corruption과 동일하기 때문에 별도의 처리 없이 프롬프팅이 가능하다. 그러나 이를 적용한 모델은 변환되는 style뿐만 아니라 content의 예측 역시 진행하여야 한다. 기사체-방송체 변환은 온전한 content의 유지가 요구된다. 예를 들어, 그림 2의 "(검증이)" 를 마스크 처리한 "<1>"에서 "확인이"라는 span이 예측된다면, style은 잘 변환되었지만 "검증"에서 "확인"으로 content가 훼손된다.

따라서 프롬프팅 방식에 스타일 토큰의 content를 유지하기 위해 원본 스타일 토큰을 사용하는 콘텐트 마커 프롬프팅을 제안한다. 콘텐트 마커는 마스크 처리된 원본 스타일 토큰으로 기존 content를 유지하는 역할을 한다. 그림 2의 스타일 토큰 변환 부분에서 볼 수 있듯이, 마스크 처리된 문장과 콘텐트 마커가 합쳐져 변환 모델의 입력으로 들어가게 된다.



*The actual data applied is Korean, so it is inevitable to insert Korean

그림 2. 제안 모델 전체 구조
Fig. 2. Proposed model overall structure

mT5는 사전학습 시 두 개 이상의 span을 마스크 처리하기 위해 여러 개의 마스크 토큰이 존재한다. 이러한 점을 활용하여 마스크 토큰으로 마스크 처리된 부분과 원본 스타일 토큰을 매칭하였다.

Ⅴ. 실험

표 4. 스타일 토큰 별 성능
Table 4. Performance by style token

Method	Final-Suffix	Preserve C.P	Delete C.P	Total
CMP	**0.9986**	**0.8833**	**0.9936**	**0.9778**
SC	0.6794	0.1313	0.9395	0.7305

스타일 변환 모델은 mT5-base 모델을 사용하였으며, batch size 8과 5 epoch로 학습하였다. 문장과 콘텐트 마커에 사용된 마스크 토큰은 mT5에 사용된 마스크 토큰과 동일하게 사용하였다. 데이터 부족으로 인한 문제를 최대한 완화하기 위하여 전체 데이터 2,000개를 500개씩으로 나누어 4-fold cross validation을 진행하였다. Metric으로는 content의 훼손에 대해 문자 단위의 측정이 가능한 EM을 사용하였다.

표 4는 SC와 CMP 방식을 적용하여 실험, 비교한 결과표다. 베이스라인으로 사용한 SC는 span corruption 프롬프팅을 적용한 방식이고, CMP는 콘텐트 마커 프롬프팅 방식이다. SC(0.7305)와 CMP(0.9778)에 0.2473의 큰 성능 차이가 존재함을 볼 수 있다. 특히 content의 유지가 필수적인 괄호 내 정보 유지에서 SC는 0.1313의 낮은 성능을 보인 반면, CMP는 0.8833의 성능을 보였다. 종결어미 변환 또한 해라체-합쇼체 변환 이외에도 기존의 content를 유지해야 하므

로 SC에서 0.6794라는 낮은 성능을 보인다.

표 5는 실험된 두 모델에 대한 실제 예측 결과이다. 표 5-(1)로 content 유지의 효과를 확인할 수 있다. SC는 방송체 style의 스타일 토큰을 생성하는 것은 성공하였으나, content를 유지하지 못하여 틀린 예측을 하고 있다. 반면 CMP는 style의 변환과 함께 content 역시 유지하고 있다.

또한, 표 5-(2), (3)에서 볼 수 있듯이 마스크 된 content의 확인 불가로 괄호 내 정보 유지/삭제도 잘 이루어지지 않고 있다. 이러한 분석을 통해 CMP가 기존 프롬프팅의 효과를 내며 content를 유지하고 있음을 알 수 있다. 그러나 표 5-(4), (5)에서 볼 수 있듯이, CMP에서도 괄호 내 정보 유지/삭제에서 오류가 발생한다. "추진위"나 "기저에 있는" 과 같이 잘 쓰이지 않는 표현이 나타날 때 틀린 예측을 하는 경향을 보인다. 이는 두 가지 이유로 분석이 가능하다. 사전학습 언어모델의 희귀한 단어에 대한 문맥화 능력이 떨어지는 단점을 공유했을 가능성이 있다. 혹은 인용문 괄호 정보 유지/삭제 데이터 분포의 다양성 부족으로 인해 충분한 데이터를 학습하지 못하여 발생하였을 수도 있다.

표 5. 실제 모델 출력 결과
Table 5. Actual Model Output Results
*The actual data applied is Korean, so it is inevitable to insert Korean

(1)	Source	:	체험 프로그램도 **다양하다.**
	Target	:	체험 프로그램도 **다양합니다.**
	SC	:	체험 프로그램도 **있습니다**
	CMP	:	체험 프로그램도 **다양합니다.**
(2)	Source	:	"**(문제가)** 끊이지 않는다."
	Target	:	"**문제가** 끊이지 않는다."
	SC	:	"끊이지 않는다."
	CMP	:	"**문제가** 끊이지 않는다."
(3)	Source	:	설문**(오차±3%포인트)**한 결과
	Target	:	설문 한 결과
	SC	:	설문을 한 결과
	CMP	:	설문 한 결과
(4)	Source	:	"추진위원회**(추진위)**가"
	Target	:	"추진위원회가"
	SC	:	"추진위원회가"
	CMP	:	"추진위원회 추진위가"
(5)	Source	:	"**(한·미 동맹 강화와 관련)** 기저에 있는"
	Target	:	"**한·미 동맹 강화와 관련** 기저에 있는"
	SC	:	"기저에 있는"
	CMP	:	"기저에 있는"

Ⅵ. 결론

기사체-방송체 텍스트 스타일 변환은 다른 스타일 변환보다 content의 훼손에 민감하다는 특성을 가지고 있다. 이러한 특성에 맞추어 본 연구에서는 1) 기사체-방송체 병렬 데이터셋을 구축하였고 2) 스타일 토큰 기반 텍스트 스타일 변환 모델에 적용되는 콘텐트 마커 프롬프팅 방식을 제안한다. 콘텐트 마커 프롬프팅으로 content를 유지하면서 프롬프팅 효과를 내었고, 이를 통해 기사체-방송체 데이터셋에서 0.9778의 높은 성능을 달성하였다.

그러나 본 연구는 규칙 기반 방식으로 스타일 토큰을 분류하기 때문에 사람이 직접 규칙을 정립해야 한다. 또한 기사체-방송체와 달리 문장의 구조와 순서 등이 변화하거나 불규칙적인 변환이 이루어지는 경우에는 이러한 규칙 기반 방식을 적용하기 어렵다. 따라서 다양한 변환에 범용적인 적용을 위하여 모델 관점에서의 스타일 토큰 분류기를 적용한 모델이 향후 연구되어야 한다.

감사의 글

본 논문은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0)과 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다. (2021RIS-004)

참고문헌

[1] J. Li, R. Jia, H. He, and P. Liang, “Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1865–1874, Jun. 2018. doi: 10.18653/v1/N18-1169.

[2] J. Lee, “Stable Style Transformer: Delete and Generate Approach with Encoder-Decoder for Text Style Transfer,” Available: https://arxiv.org/abs/2005.12086.

[3] F. Luo et al., “A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer,” in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 5116–5122. Nov. 2019, doi: 10.24963/ijcai.2019/711.

[4] T. B. Brown et al., “Language Models are Few-Shot Learners.” Available: https://arxiv.org/abs/2005.14165.

[5] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” Available : https://arxiv.org/abs/1910.10683.

[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of

Prompting Methods in Natural Language Processing." Available : https://arxiv.org/abs/2107.13586

[7] W. Yin, J. Hay, and D. Roth, "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach," 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), pp. 3914–3923, Nov. 2019. doi: 10.18653/v1/D19-1404.

[8] D. Khashabi et al., "UNIFIEDQA: Crossing Format Boundaries with a Single QA System," in Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1896–1907, Nov. 2020. doi: 10.18653/v1/2020.findings-emnlp.171.

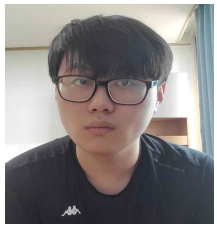
[9] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-Based Named Entity Recognition Using BART." Available : https://arxiv.org/abs/2106.01760.

[10] T. Schick and H. Schütze, "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners." Available : https://arxiv.org/abs/2009.07118.

[11] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep Learning for Text Style Transfer: A Survey." Available : https://arxiv.org/abs/2011.0041.

[12] E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Callison-Burch, and J. Wei, "A Recipe For Arbitrary Text Style Transfer with Large Language Models." Available : https://arxiv.org/abs/2109.03910.

[13] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer." Available : https://arxiv.org/abs/2010.11934



김경민(Kyung Min Kim)

2017년～현 재 : 한국기술교육대학교 컴퓨터공학부 (학사과정)

※관심분야 : 딥러닝(Deep Learning), 자연어처리(Natural Language Process)

임상훈(Sang Hun Im)

2019년 : 한국기술교육대학교 컴퓨터공학부 (학사)
2021년 : 한국기술교육대학교 컴퓨터공학과 (공학석사)
2022년～현 재 : 한국기술교육대학교 컴퓨터공학과 (공학박사과정)

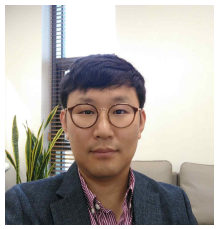
※관심분야 : 자연어처리(Natural Language Process), 계층적 문서 분류(Hierarchical Text Classification)

김기백(Gi Baeg Kim)

2020년 : 한국기술교육대학교 컴퓨터공학부 (학사)
2021년～현 재 : 한국기술교육대학교 컴퓨터공학과 (공학석사과정)

※관심분야 : 딥러닝(Deep Learning), 자연어처리(Natural Language Process)

오흥선(Heung-Seon Oh)



2009년 : 한국과학기술원 전산학 (공학석사)
2014년 : 한국과학기술원 전산학 (공학박사)

2013년～2018년: KISTI
2018년～현 재: 한국기술교육대학교 교수
※관심분야 : 딥러닝(Deep Learning), 자연어처리(Natural Language Process) 등