(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0122146
(43) 공개일자 2024년08월12일

(51) 국제특허분류(Int. Cl.)
G06F 40/151 (2020.01) *G06F 16/951* (2019.01)
G06F 40/205 (2020.01) *G06F 40/284* (2020.01)
G06F 40/58 (2020.01) *G10L 13/08* (2006.01)
(52) CPC특허분류
G06F 40/151 (2020.01)
G06F 16/951 (2019.01)
(21) 출원번호 10-2023-0014961
(22) 출원일자 2023년02월03일
심사청구일자 2023년02월03일

(71) 출원인
한국기술교육대학교 산학협력단
충청남도 천안시 동남구 병천면 충절로 1600 (한국기술교육대학교내)
(72) 발명자
오홍선
충청남도 천안시 동남구 병천면 충절로 1600, 434호(공학2관)
김경민
충청남도 천안시 서북구 불당11로 82, 610동 604호(불당동, 대원칸타빌아파트)
(뒷면에 계속)
(74) 대리인
특허법인오암

전체 청구항 수 : 총 7 항

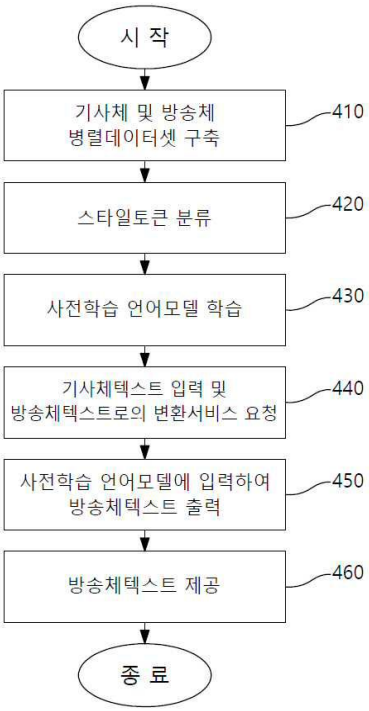
(54) 발명의 명칭 **뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법**

(57) 요 약

본 발명은 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법에 관한 것으로, 변환서비스제공서버에서 기사체 및 방송체 병렬 데이터셋을 구축하는 단계; 상기 변환서비스제공서버에서 상기 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류하는 단계; 상기 변환서비스제공서버에서 상기 분류된

(뒷면에 계속)

대 표 도 - 도1



스타일토큰을 이용하여 사전학습 언어모델을 학습시키는 단계; 사용자단말기에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 상기 변환서비스제공서버에 요청하는 단계; 상기 사용자단말기로부터의 요청에 따라 상기 변환서비스제공서버에서 상기 사전학습 언어모델에 상기 기사체텍스트를 입력시켜 상기 방송체텍스트를 출력하는 단계; 및 상기 변환서비스제공서버에서 상기 방송체텍스트에 대한 상기 변환서비스를 상기 사용자 단말기에 제공하는 단계;를 포함함으로써, 뉴스기사의 음성변환 시 문장의 유창성이 훼손되지 않을 뿐만 아니라 문장의 정보가 정확하게 제공될 수 있도록 테스트스타일을 효과적으로 변환할 수 있다.

(52) CPC특허분류
G06F 40/205 (2020.01)
G06F 40/284 (2020.01)
G06F 40/58 (2020.01)
G10L 13/08 (2013.01)

(72) 발명자
임상훈
충청남도 천안시 동남구 병천면 충절로 1634-29, 302호
김기백
충청남도 천안시 동남구 병천면 충절로 1634-13, 1동 304호
안태훈
서울특별시 마포구 상암산로 48-6(상암동, 중앙일보빌딩(JTBC))
김민
서울특별시 마포구 상암산로 48-6(상암동, 중앙일보빌딩(JTBC))

이 발명을 지원한 국가연구개발사업

과제고유번호	1345356194
과제번호	LINC 3.0-2022-31
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	산학연협력고도화지원
연구과제명	3단계 산학연 협력 선도대학 육성사업(LINC3.0)(0.5)
기 여 율	1/1
과제수행기관명	한국기술교육대학교
연구기간	2022.03.01 ~ 2023.02.28

명 세 서

청구범위

청구항 1

변환서비스제공서버에서 기사체 및 방송체 병렬 데이터셋을 구축하는 단계;

상기 변환서비스제공서버에서 상기 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류하는 단계;

상기 변환서비스제공서버에서 상기 분류된 스타일토큰을 이용하여 사전학습 언어모델을 학습시키는 단계;

사용자단말기에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 상기 변환서비스제공서버에 요청하는 단계;

상기 사용자단말기로부터의 요청에 따라 상기 변환서비스제공서버에서 상기 사전학습 언어모델에 상기 기사체텍스트를 입력시켜 상기 방송체텍스트를 출력하는 단계; 및

상기 변환서비스제공서버에서 상기 기사체텍스트에 대응하는 상기 변환서비스로 상기 방송체텍스트를 상기 사용자 단말기에 제공하는 단계;

를 포함하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법.

청구항 2

청구항 1에 있어서,

상기 기사체 및 방송체 병렬 데이터셋을 구축하는 단계는,

상기 변환서비스제공서버의 데이터셋구축부에서 뉴스기사를 크롤링하여 문장단위로 구분하고, 기사체문장과 방송체문장을 함께 포함하는 상기 기사체 및 방송체 병렬 데이터셋을 구축하는

뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법.

청구항 3

청구항 2에 있어서,

상기 기사체 및 방송체 병렬 데이터셋을 구축하는 단계는,

상기 데이터셋구축부에서 URL과 섹션이 추가로 포함되어 기 설정된 파일포맷으로 변환하여 상기 기사체 및 방송체 병렬 데이터셋을 구축하는

뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법.

청구항 4

청구항 3에 있어서,

상기 스타일토큰을 분류하는 단계는,

상기 변환서비스제공서버의 스타일토큰분류부에서 구축된 상기 기사체 및 방송체 병렬 데이터셋에 포함된 상기 기사체문장에서 괄호 및 종결어미를 상기 스타일토큰을 분류하는

뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법.

청구항 5

청구항 4에 있어서,

상기 스타일토큰을 분류하는 단계는,

상기 스타일토큰분류부에서 상기 스타일토큰 중에서 문맥파악이 요구되지 않은 스타일토큰의 경우 상기 기사체 문장에서 마스킹처리하는

뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법.

청구항 6

청구항 5에 있어서,

상기 사전학습 언어모델을 학습시키는 단계는,

상기 변환서비스제공서버의 스타일토큰변환부에서 상기 스타일토큰을 프롬프팅방식으로 전처리하고, 전처리된 상기 스타일토큰을 이용하여 상기 방송체텍스트를 예측 및 출력할 수 있도록 상기 사전학습 언어모델을 학습 및 평가하는

뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법.

청구항 7

청구항 6에 있어서,

상기 사전학습 언어모델을 학습시키는 단계는,

상기 스타일토큰변환부에서 상기 스타일토큰을 상기 프롬프팅방식과 함께 콘텐트마커방식으로 전처리하는

뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 본 발명은 변환서비스제공서버에서 기사체 및 방송체 병렬 데이터셋을 구축하고, 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류하며, 분류된 스타일토큰을 이용하여 사전학습 언어모델을 학습시킨 후에, 사용자단말기에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 요청할 경우 변환서비스제공서버에서 사전학습 언어모델에 기사체텍스트를 입력시켜 방송체텍스트를 출력하고, 출력된 방송체텍스트를 기사체텍스트에 대한 변환서비스로 제공함으로써, 뉴스기사의 음성변환 시 문장의 유창성이 훼손되지 않을 뿐만 아니라 문장의 정보가 정확하게 제공될 수 있도록 테스트스타일을 효과적으로 변환할 수 있는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법에 관한 것이다.

배 경 기 술

[0002] 잘 알려진 바와 같이, TTS(Text-To-Speech)라고도 불리는 음성합성기술은, 가상의 목소리를 이용하여 텍스트를 음성으로 변환하는 기술로서, 안내방송, 내비게이션, 인공지능 비서 등에서 활용되고 있다.

[0003] 이러한 음성합성기술은 예를 들면, 음성을 음소 등과 같이 아주 짧은 단위로 미리 잘라서 저장해두고, 합성할 문장을 구성하는 음소들을 결합하여 음성을 합성하는 연결합성(concatenative) 방식, 음성의 특징을 파라미터(parameter)로 표현하고 합성할 문장을 구성하는 음성 특징들을 나타내는 파라미터들을 보코더(vocoder)를 이용해 문장에 대응하는 음성으로 합성하는 파라미터(parametric) 방식 등으로 제공될 수 있다.

[0004] 한편, 텍스트형태의 뉴스기사를 상술한 바와 같은 음성합성기술을 통해 음성형태로 변환하여 사용자에게 제공하고자 하는 기술이 제안되고 있는데, 음성화 과정에서는 독해에 적합한 기사체(article style) 문장을 음성화되었을 때의 유창성과 정보의 정확성을 위해 방송체(broadcasting style) 문장으로 먼저 변환한 후에 음성으로 변

환해야만 한다.

[0005] 하지만, 종래의 음성변환에서는 기사체 문장에서 단순히 괄호만을 삭제하는 변환만이 수행되기 때문에, 텍스트 형태의 뉴스기사를 음성변환할 경우 음성변환에 따른 출력결과물에서 문장의 유창성이 훼손될 뿐만 아니라 문장의 정보가 부정확하게 제공되는 문제가 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 1. 한국공개특허 제10-2021-0095010호(2021.07.30.공개)

발명의 내용

해결하려는 과제

[0007] 본 발명은 본 발명은 변환서비스제공서버에서 기사체 및 방송체 병렬 데이터셋을 구축하고, 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류하며, 분류된 스타일토큰을 이용하여 사전학습 언어모델을 학습시킨 후에, 사용자단말기에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 요청할 경우 변환서비스제공서버에서 사전학습 언어모델에 기사체텍스트를 입력시켜 방송체텍스트를 출력하고, 출력된 방송체텍스트를 기사체텍스트에 대한 변환서비스로 제공함으로써, 뉴스기사의 음성변환 시 문장의 유창성이 훼손되지 않을 뿐만 아니라 문장의 정보가 정확하게 제공될 수 있도록 테스트스타일을 효과적으로 변환할 수 있는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법을 제공하고자 한다.

[0008] 본 발명의 실시예들의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0009] 본 발명의 실시예에 따르면, 변환서비스제공서버에서 기사체 및 방송체 병렬 데이터셋을 구축하는 단계; 상기 변환서비스제공서버에서 상기 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류하는 단계; 상기 변환서비스제공서버에서 상기 분류된 스타일토큰을 이용하여 사전학습 언어모델을 학습시키는 단계; 사용자단말기에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 상기 변환서비스제공서버에 요청하는 단계; 상기 사용자단말기로부터의 요청에 따라 상기 변환서비스제공서버에서 상기 사전학습 언어모델에 상기 기사체텍스트를 입력시켜 상기 방송체텍스트를 출력하는 단계; 및 상기 변환서비스제공서버에서 상기 기사체텍스트에 대응하는 상기 변환서비스로 상기 방송체텍스트를 상기 사용자 단말기에 제공하는 단계;를 포함하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법이 제공될 수 있다.

[0010] 또한, 본 발명의 실시예에 따르면, 상기 기사체 및 방송체 병렬 데이터셋을 구축하는 단계는, 상기 변환서비스제공서버의 데이터셋구축부에서 뉴스기사를 크롤링하여 문장단위로 구분하고, 기사체문장과 방송체문장을 함께 포함하는 상기 기사체 및 방송체 병렬 데이터셋을 구축하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법이 제공될 수 있다.

[0011] 또한, 본 발명의 실시예에 따르면, 상기 기사체 및 방송체 병렬 데이터셋을 구축하는 단계는, 상기 데이터셋구축부에서 URL과 섹션이 추가로 포함되어 기 설정된 파일포맷으로 변환하여 상기 기사체 및 방송체 병렬 데이터셋을 구축하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법이 제공될 수 있다.

[0012] 또한, 본 발명의 실시예에 따르면, 상기 스타일토큰을 분류하는 단계는, 상기 변환서비스제공서버의 스타일토큰분류부에서 구축된 상기 기사체 및 방송체 병렬 데이터셋에 포함된 상기 기사체문장에서 괄호 및 종결어미를 상기 스타일토큰을 분류하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법이 제공될 수 있다.

[0013] 또한, 본 발명의 실시예에 따르면, 상기 스타일토큰을 분류하는 단계는, 상기 스타일토큰분류부에서 상기 스타일토큰 중에서 문맥파악이 요구되지 않은 스타일토큰의 경우 상기 기사체문장에서 마스킹처리하는 뉴스텍스트의

음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법이 제공될 수 있다.

[0014] 또한, 본 발명의 실시예에 따르면, 상기 사전학습 언어모델을 학습시키는 단계는, 상기 변환서비스제공서버의 스타일토큰변환부에서 상기 스타일토큰을 프롬프팅방식으로 전처리하고, 전처리된 상기 스타일토큰을 이용하여 상기 방송체텍스트를 예측 및 출력할 수 있도록 상기 사전학습 언어모델을 학습 및 평가하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법이 제공될 수 있다.

[0015] 또한, 본 발명의 실시예에 따르면, 상기 사전학습 언어모델을 학습시키는 단계는, 상기 스타일토큰변환부에서 상기 스타일토큰을 상기 프롬프팅방식과 함께 콘텐트마커방식으로 전처리하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법이 제공될 수 있다.

발명의 효과

[0016] 본 발명은 본 발명은 변환서비스제공서버에서 기사체 및 방송체 병렬 데이터셋을구축하고, 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류하며, 분류된 스타일토큰을 이용하여 사전학습 언어모델을 학습시킨 후에, 사용자단말기에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 요청할 경우 변환서비스제공서버에서 사전학습 언어모델에 기사체텍스트를 입력시켜 방송체텍스트를 출력하고, 출력된 방송체텍스트를 기사체텍스트에 대한 변환서비스로 제공함으로써, 뉴스기사의 음성변환 시 문장의 유창성이 훼손되지 않을 뿐만 아니라 문장의 정보가 정확하게 제공될 수 있도록 테스트스타일을 효과적으로 변환할 수 있다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 실시예에 따라 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템에서 변환서비스를 제공하는 과정을 나타낸 플로우차트이고,

도 2는 본 발명의 실시예에 따라 변환서비스를 제공하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 블록구성도이며,

도 3 내지 도 14는 본 발명의 실시예에 따라 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템에서 변환서비스를 제공하는 상세 동작을 설명하기 위한 도면이고,

도 15는 본 발명의 실시예에 따라 시뮬레이션된 스타일토큰 기반의 텍스트스타일 변환(TST) 모델을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0018] 본 발명의 실시예들에 대한 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성요소를 지칭한다.

[0019] 본 발명의 실시예들을 설명함에 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다. 그리고 후술되는 용어들은 본 발명의 실시예에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0020] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명하기로 한다.

[0021] 먼저, 본 발명의 실시예에서 사용되는 용어에 대해 간략하게 설명하면, 스타일(style)은 문장의 어체, 종결어미 등으로 나타나는 문장의 특징을 의미하고, 콘텐트(content)는 텍스트가 가지고 있는 핵심적인 정보를 의미한다.

[0022] 그리고, 텍스트 스타일 변환(text style transfer)은 텍스트가 가지고 있는 정보인 콘텐트(content)는 유지하면서 스타일(style)을 변환시키는 작업을 의미하고, 스타일 토큰(style token)은 특정 스타일(style)에서 특징적으로 나타나는 표현을 담은 단어를 의미한다.

[0023] 또한, 기사체(article style)는 독해에 적합한 텍스트 형태의 뉴스기사에서 나타나는 스타일을 의미하고, 방송체(broadcasting style)는 청해에 적합한 음성화된 형태의 뉴스기사에서 나타나는 스타일을 의미한다.

[0024] 예를 들어, "한국이 국제축구연맹(FIFA) 월드컵 본선에서 세계 랭킹 1위와 가장 많이 만난 나라 대열에 동참했다."라는 뉴스기사를 참조하면, 콘텐트는 "한국이 국제축구연맹 월드컵 본선에서 세계 랭킹 1위와 가장 많이 만난 나라 대열에 동참함"이고, 스타일은 기사체문장이며, 스타일 토큰은 "(FIFA)", "동참했다"이다.

[0025] 또한, 기사체를 방송체로 변환하는 규칙에 대해 설명하면, 정보의 모호성 및 유창성을 고려하여 변환 규칙을 정립할 수 있는데, 첫째, 괄호는 일반적으로 문장의 유창성을 크게 해치기 때문에 괄호와 내용전체를 삭제할 수 있다. 예를 들면, "마르세유(프랑스) 상대로"를 "마르세유 상대로"로 변환할 수 있다.

[0026] 둘째, 통화변환 괄호에서는 환율의 변환에 따라 정보의 모호성이 발생할 수 있으므로 괄호만을 삭제하면서 유창성을 고려한 변환을 통해 내용을 유지할 수 있다. 예를 들면, "124만 달러(약 13억 8천만원)의"를 "124만달러, 환화 약 13억 8천만원의"로 변환할 수 있다.

[0027] 셋째, 현지시간 괄호에서는 시간의 경우 삭제 시에 정보의 모호성이 발생할 수 있으므로 괄호만을 삭제하면서 유창성을 고려한 변환을 통해 내용을 유지할 수 있다. 예를 들면, "13일(현지시간)"을 "현지시간으로 13일"로 변환할 수 있다.

[0028] 넷째, 인용문 내 괄호에서는 인용문 내에 미처 끝내지 못한 말을 표현하는 괄호의 경우 삭제 시에 유창성에 큰 문제가 발생할 수 있기 때문에 괄호만을 삭제하고 내용을 유지할 수 있다. 예를 들면, "(수사를) 진행해야 한다고"를 "수사를 진행해야 한다고"로 변환할 수 있다.

[0029] 다섯째, 인용문 내 괄호에서는 인용문 내에 추가 정보를 나타내는 괄호의 경우 유지 시에 유창성에 문제가 발생할 수 있으므로, 괄호와 내용전체를 삭제할 수 있다. 예를 들면, "추진위원회(추진위)"를 "추진위원회"로 변환할 수 있다.

[0030] 여섯째, 종결어미에서는 텍스트형태에 적합한 해라체 종결어미에서 발화에 적합한 합쇼체 종결어미로 변환할 수 있다. 예를 들면, "공개했다"를 "공개했습니다"로 변환할 수 있다.

[0031] 도 1은 본 발명의 실시예에 따라 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템에서 변환서비스를 제공하는 과정을 나타낸 플로우차트이고, 도 2는 본 발명의 실시예에 따라 변환서비스를 제공하는 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 블록구성도이며, 도 3 내지 도 14는 본 발명의 실시예에 따라 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템에서 변환서비스를 제공하는 상세 동작을 설명하기 위한 도면이다.

[0032] 도 1 내지 도 14를 참조하면, 변환서비스제공서버(300)에서 기사체 및 방송체 병렬 데이터셋을 구축할 수 있다(단계410).

[0033] 여기에서, 상기 기사체 및 방송체 병렬 데이터셋을 구축하는 단계(410)에서는, 변환서비스제공서버(300)의 데이터셋구축부(320)에서 뉴스기사를 크롤링하여 문장단위로 구분하고, 기사체문장과 방송체문장을 함께 포함하는 기사체 및 방송체 병렬 데이터셋을 구축할 수 있다.

[0034] 또한, 상기 기사체 및 방송체 병렬 데이터셋을 구축하는 단계(410)에서는, 데이터셋구축부(320)에서 URL과 섹션이 추가로 포함되어 기 설정된 파일포맷으로 변환하여 기사체 및 방송체 병렬 데이터셋을 구축할 수 있다.

[0035] 예를 들면, 데이터셋구축부(320)에서는 뉴스기사를 크롤링할 수 있는데, 도 3에 도시한 바와 같이 뉴스기사를 크롤링할 수 있는데, 특정신문사, 특정방송국 등을 지정하면서 기간을 특정하여 무작위로 크롤링하여 뉴스기사를 수집할 수 있고, 그 과정에서 특정신문사 또는 특정방송국에서 제공하는 섹션(즉, 카테고리), 기사의 URL, 기사원문 등을 수집할 수 있다.

[0036] 그리고, 데이터셋구축부(320)에서는 크롤링된 뉴스기사를 전처리할 수 있는데, 도 4에 도시한 바와 같이 크롤링된 뉴스기사를 문장단위로 구분하고, 기사정보와 같은 불필요한 텍스트의 경우에는 제거하며, 문장단위로 기사의 URL과 섹션을 유지할 수 있다.

[0037] 또한, 데이터셋구축부(320)에서는 전처리된 뉴스기사를 텍스트포맷파일(예를 들면, 워드포맷파일 등)로 변환한 후에 방송체문장을 작성하고, 기 설정된 파일포맷으로 변환할 수 있는데, 도 5에 도시한 바와 같이 방송체문장이 작성된 워드파일을 사전학습 언어모델에서 사용될 수 있도록 기 설정된 파일포맷(예를 들면, jsonl포맷 등)으로 변환할 수 있으며, 하나의 인스턴스(instance)에는 원본기사체문장, 방송체문장, URL 및 섹션이 포함될 수 있다.

[0038] 다음에, 변환서비스제공서버(300)에서 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류할 수 있

다(단계420).

[0039] 여기에서, 상기 스타일토큰을 분류하는 단계(420)에서는, 변환서비스제공서버(300)의 스타일토큰분류부(330)에서 구축된 기사체 및 방송체 병렬 데이터셋에 포함된 기사체문장에서 괄호 및 종결어미를 스타일토큰을 분류할 수 있다.

[0040] 또한, 상기 스타일토큰을 분류하는 단계(420)에서는, 스타일토큰분류부(330)에서 스타일토큰 중에서 문맥파악이 요구되지 않은 스타일토큰의 경우 기사체문장에서 마스킹처리할 수 있다.

[0041] 예를 들면, 스타일토큰분류부(330)에서는 알고리즘을 이용하여 괄호 및 종결어미를 스타일토큰을 분류할 수 있는데, 도 6에 도시한 바와 같이 기사체-방송체 변환은 규칙적인 변환형태를 가지고 있으며, 변환되는 부분은 괄호와 종결어미에 한정될 수 있고, 이에 따라 괄호와 종결어미가 명확한 스타일토큰으로 분류될 수 있다.

[0042] 여기에서, 괄호의 경우 정규표현식을 통해 검색 및 검출하여 스타일토큰으로 분류할 수 있고, 종결어미의 경우 문장의 마지막 단어를 선택하는 방식으로 검출하여 스타일토큰으로 분류할 수 있다.

[0043] 또한, 스타일토큰분류부(330)에서는 분류된 스타일토큰을 전처리 및 마스킹처리할 수 있는데, 도 7에 도시한 바와 같이 분류된 스타일토큰을 마스킹 기사체문장에서 마스킹처리할 수 있고, 기사체와 방송체 특성 상 문맥적인 파악이 요구되지 않은 요소들의 경우 사전합습 언어모델을 사용하지 않고 방송체로의 변환이 가능하기 때문에 사전학습 언어모델을 이용한 텍스트스타일 변환이 이루어지지 않도록 마스킹처리할 수 있다.

[0044] 예를 들면, "서훈(68.구속기소)"에서 "(68.구속기소)"의 경우 기사체 및 방송체 변환규칙 상 문맥의 파악 없이 제거될 수 있는데, 이러한 문맥 파악이 요구되지 않은 스타일토큰들을 기사체 및 방송체 변환규칙에 따라 처리할 경우 문맥 파악이 요구되는 스타일토큰(예를 들면, 인용문 내 괄호 등)에서도 더 높은 정확도로 변환될 수 있다.

[0045] 다음에, 변환서비스제공서버(300)에서 분류된 스타일토큰을 이용하여 사전학습 언어모델을 학습시킬 수 있다(단계430).

[0046] 여기에서, 상기 사전학습 언어모델을 학습시키는 단계(430)에서는, 변환서비스제공서버(300)의 스타일토큰변환부(340)에서 스타일토큰을 프롬프팅방식으로 전처리하고, 전처리된 스타일토큰을 이용하여 방송체텍스트를 예측 및 출력할 수 있도록 사전학습 언어모델을 학습 및 평가할 수 있다.

[0047] 또한, 상기 사전학습 언어모델을 학습시키는 단계(430)에서는, 스타일토큰변환부(340)에서 스타일토큰을 프롬프팅방식과 함께 콘텐트마커방식으로 전처리할 수 있다.

[0048] 예를 들면, 스타일토큰변환부(340)에서는 스타일토큰을 프로프팅방식으로 전처리할 수 있는데, 도 8에 도시한 바와 같이 기사체문장을 프롬프팅방식으로 전처리하여 문장으로 변환하면서, 스타일토큰을 콘텐트마커(CM : content marker)를 이용하여 지정하는 방식으로 전처리할 수 있다.

[0049] 여기에서, 프롬프팅방식은 사전학습 언어모델의 학습 과정과 유사한 방식으로 세부 태스크를 적용하는 것으로, mT5알고리즘의 경우 사전학습 시 마스킹된 부분을 예측하는 학습을 하며, 도 9에 도시(빨간색)한 바와 같이 모든 마스크는 "<extra_id_{마스크 번호}>"의 형태로 사용될 수 있다.

[0050] 또한, 콘텐트마커 프롬프팅방식은 마스크처리된 원본 콘텐트를 알 수 있도록 도 10에 도시(빨간색)한 바와 같이 콘텐트마커(CM)를 이용하여 원본문장의 콘텐트를 사전학습 언어모델에 함께 입력시켜 원본문장의 콘텐트를 유지할 수 있으며, 이를 통해 이러한 콘텐트마커가 없을 경우 사전학습 언어모델에서 원본문장의 콘텐트없이 완전히 새롭게 예측함에 따른 콘텐트 훼손을 미연에 방지할 수 있다.

[0051] 한편, 스타일토큰변환부(340)에서는 스타일토큰을 이용하여 방송체텍스트를 예측 및 출력할 수 있도록 사전학습 언어모델을 학습 및 평가할 수 있는데, 도 11의 위쪽에 도시한 바와 같은 문장 및 콘텐트마커(CM)가 사전학습 언어모델에 입력될 경우 도 11의 아래쪽에 도시한 바와 같은 출력값으로 출력될 수 있으며, 각 스타일토큰이 기사체형태로 변환되었을 때의 값을 예측할 수 있고, 괄호가 삭제되는 경우 공백 문자만을 예측할 수 있다.

[0052] 여기에서, "<extra_id_{마스크 번호}>"를 통해 변환된 스타일토큰의 위치를 파악할 수 있다.

[0053] 또한, 도 12에 도시한 바와 같이 학습에 사용되는 정답은 미리 구축한 기사체 및 방송체 병렬데이터셋을 이용하여 구축할 수 있으며, 정답값에 해당하는 방송체텍스트를 예측할 수 있도록 사전학습 언어모델을 학습시킬 수 있다.

[0054] 다음에, 사용자단말기(100)에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 변환서비스제공서버(300)에 요청할 수 있다(단계440).

[0055] 여기에서, 사용자단말기(100)는 예를 들면, 스마트폰(smart phone), 휴대폰, 내비게이션, 컴퓨터, 노트북, 디지털방송용 단말, PDA(Personal Digital Assistants), PMP(Portable Multimedia Player), 태블릿 PC, 게임 콘솔(game console), 웨어러블 디바이스(wearable device), IoT(internet of things) 디바이스, VR(virtual reality) 디바이스, AR(augmented reality) 디바이스 등을 포함할 수 있다.

[0056] 이러한 사용자단말기(100)에서는 변환서비스서버(300)에서 운영하는 웹페이지에 접속하거나, 별도의 변환서비스 어플리케이션을 설치한 후 그 어플리케이션을 통해 변환서비스서버(300)에 접속한 상태에서 텍스트입력화면을 통해 기사체텍스트를 입력하여 방송체텍스트로의 변환을 요청할 수 있다.

[0057] 또한, 통신망(200)은 예를 들면, 이동통신망, 유선인터넷통신망, 무선인터넷통신망, 방송망 등을 포함할 수 있으며, 구체적으로 PAN(personal area network), LAN(local area network), CAN(campus area network), MAN(metropolitan area network), WAN(wide area network), BBN(broadband network), 인터넷 등의 네트워크 중 하나 이상의 네트워크를 포함할 수 있고, 버스 네트워크, 스타 네트워크, 링 네트워크, 메쉬 네트워크, 스타-버스 네트워크, 트리 네트워크, 계층적(hierarchical) 네트워크 등을 포함하는 네트워크 토폴로지 중 하나 이상을 포함할 수 있다.

[0058] 이러한 통신망(400)을 통해 사용자단말기(100) 및 변환서비스서버(300) 상호 간의 유무선데이터 송수신을 위한 통신환경을 제공할 수 있다.

[0059] 아울러, 변환서비스제공서버(300)에 구비된 인터페이스부(310)에서는 웹페이지 기반으로 사용자단말기(100)가 접속하여 변환서비스를 이용할 수 있도록 통신인터페이스를 제공하거나, 혹은 변환서비스어플리케이션을 사용자단말기(100)에 제공하여 변환서비스어플리케이션을 통해 접속되는 사용자단말기(100)의 요청에 따른 변환서비스를 이용할 수 있도록 통신인터페이스를 제공할 수 있다.

[0060] 다음에, 사용자단말기(100)로부터의 요청에 따라 변환서비스제공서버(300)에서 사전학습 언어모델에 기사체텍스트를 입력시켜 방송체텍스트를 출력할 수 있다(단계450).

[0061] 여기에서, 스타일토큰변환부(340)에서는 사전학습 언어모델에 기사체텍스트를 입력시켜 방송체텍스트를 출력할 수 있는데, 도 13에 도시한 바와 같이 입력값과 예측된 출력값을 바탕으로 최종적인 방송체문장을 완성하여 출력할 수 있다.

[0062] 그리고, 변환서비스제공서버(300)에서 기사체텍스트에 대응하는 변환서비스로 방송체텍스트를 사용자단말기(100)에 제공할 수 있다(단계460).

[0063] 여기에서, 인터페이스부(310)에서는 사용자단말기(100)와의 접속 및 통신을 담당하여 스타일토큰변환부(340)를 통해 출력되는 방송체텍스트를 사용자단말기(100)로 전송할 수 있다.

[0064] 이 후, 사용자단말기(100)에서는 변환서비스서버(300)로부터 제공되는 방송체텍스트를 서비스제공화면을 통해 디스플레이할 수 있다.

[0065] 상술한 바와 같은 본 발명의 실시예에 따른 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템의 변환서비스 제공방법에서 제공되는 서비스제공화면은 도 14에 도시한 바와 같이 위쪽에 기사체텍스트를 입력할 경우 그 아래쪽에 방송체텍스트로 변환하여 디스플레이하는 방식으로 제공될 수 있다.

[0066] 따라서, 본 발명의 실시예에 따르면, 본 발명은 변환서비스제공서버에서 기사체 및 방송체 병렬 데이터셋을 구축하고, 구축된 기사체 및 방송체 병렬 데이터셋에서 스타일토큰을 분류하며, 분류된 스타일토큰을 이용하여 사전학습 언어모델을 학습시킨 후에, 사용자단말기에서 기사체텍스트를 입력하여 방송체텍스트로의 변환서비스를 요청할 경우 변환서비스제공서버에서 사전학습 언어모델에 기사체텍스트를 입력시켜 방송체텍스트를 출력하고, 출력된 방송체텍스트를 기사체텍스트에 대한 변환서비스로 제공함으로써, 뉴스기사의 음성변환 시 문장의 유창성이 훼손되지 않을 뿐만 아니라 문장의 정보가 정확하게 제공될 수 있도록 테스트스타일을 효과적으로 변환할 수 있다.

[0067] 한편, 도 15는 본 발명의 실시예에 따라 시뮬레이션된 스타일토큰 기반의 텍스트스타일 변환(TST) 모델을 설명하기 위한 도면으로, 상술한 바와 같은 본 발명의 실시예에 따른 뉴스텍스트의 음성변환을 위한 텍스트스타일 변환시스템을 구현하여 변환서비스를 제공하는 시뮬레이션을 수행한 후 그 시뮬레이션 결과에 대해 설명하기로

한다.

[0068] 먼저, 기사체 및 방송체 데이터셋에 대해 설명하면, 기사체(AS : article style)와 방송체(BS : broadcasting style)는 정확한 정보를 전달해야 하는 뉴스의 특성상 콘텐트(content)의 훼손에 굉장히 민감한데, 예를 들어 "어제 밤 범인이 검거되었다." 라는 문장에서 "어제 범인이 잡혔다." 라는 문장으로 변환될 경우, "밤"과 "검거"라는 콘텐트에서 훼손이 발생할 수 있으며, 이러한 콘텐트의 훼손은 전체적인 문맥은 유지하지만 부정확한 정보가 전달될 수 있는 문제점이 있다.

[0069] 그리고, 기사체(AS)와 방송체(BS) 모두 문장의 구조나 순서가 높은 수준으로 정제되어 있으므로 단어의 배치나 문장의 구조와 같은 큰 틀은 유지되며, 종결어미와 괄호와 같은 스타일토큰들이 변환되는 것으로 이루어지는데, 변환되는 스타일토큰은 명확한 규칙성을 가지며, 그 기사체 및 방송체 변환규칙은 종결어미 변환과 괄호 내 정보(CP; Content in Parenthesis) 유지 또는 삭제 등 3가지로 구분될 수 있다.

[0070] 여기에서, 종결어미 변환에서는 기사체의 "-다" 형태의 해라체에서 방송체의 "-입니다" 형태의 합쇼체로의 변환으로 나타나는데, 해라체-합쇼체 변환은 다양한 형태를 가지며 여러 불규칙적 변환이 존재한다. 예를 들어 아래의 표 1에서 Change final-suffix에서 등장하는 "공개했다" 는 "-다" 가 "-습니다" 로 변환되지만, 표 1-Preserve C.P 2에서는 "-이다" 가 "-입니다" 로 변환된다.

표 1

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.

[0071]

[0072] 또한, 괄호 내 정보 유지 또는 삭제에서는 괄호가 사용된 목적에 따라 다른 변환 양식을 보이는데, 예를 들어 상기 표 1에서 Delete C.P의 "(프랑스)", "(UCL)" 과 Preserve C.P 2의 "(약 13억 8천만원)" 은 추가적인 정보를 제공하기 위하여 사용하였지만, Preserve C.P의 "(수사를)" 은 인용한 말의 유창성을 위하여 사용되었으며, 방송체에서는 괄호를 사용하지 않기 때문에, 유창성과 모호성을 고려한 변환이 요구된다.

[0073] 예를 들면, "(프랑스)" 와 "(UCL)" 의 경우 삭제되더라도 정보의 모호성이 발생하지 않으므로 삭제하지만, "(약 13억 8천만원)" 의 경우 환율 등의 이유로 정보의 모호성이 발생할 수 있으므로 "한화 약 13억 8천만원" 으로 변환하여 괄호는 삭제하되 정보는 유지할 수 있으며, "(수사를)" 또한 삭제 시에 유창성에 문제가 발생하므로, 괄호를 벗긴 "수사를" 으로 변환하여 괄호 내 정보를 유지할 수 있다.

[0074] 상술한 바와 같은 기사체 및 방송체 변환 규칙을 바탕으로 데이터셋 2,000건을 구축하였으며, 하나의 인스턴스는 동일한 콘텐트를 가지는 기사체-방송체 문장 쌍으로 구성하였고, 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.

[0075] 그리고, 크롤링된 기사의 분류는 원본기사에서 제공하는 분류를 바탕으로 작성되었고, 각 분류별 기사 수는 아래의 표 2에 나타낸 바와 같은데, 문장 당 스타일 토큰은 평균 1.37회 등장하며, 최대 6개의 스타일 토큰이 한 문장에서 등장하였고, 각 스타일 토큰 종류별 통계는 아래의 표 3과 같다.

표 2

[0076]

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

표 3

[0077]

	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

[0078] 한편, 시뮬레이션을 위한 스타일토큰 기반의 텍스트스타일 변환(TST) 모델은 도 15에 도시한 바와 같이 1) 스타일 토큰 분류(Style Token Classification)와 2) 스타일 토큰 변환(Style Token Conversion) 두 단계로 이루어질 수 있는데, 기사체-방송체의 스타일 토큰 분류는 상기 표 1에 나타낸 바와 같이 명확한 규칙을 발견하여 규칙기반모델을 사용하였다.

[0079] 예를 들면, 종결어미 변환의 경우 문장의 마지막 부분에 있는 단어를 마스크 처리하며, 괄호 내 정보 유지 또는 삭제의 경우 괄호 및 괄호 내 단어를 마스크 처리하였는데, 예를 들어 도 15에 도시한 바와 같은 문장에서 나타난 괄호인 "(검증이)" 와 마지막 단어인 "판단했다." 가 스타일 토큰으로 분류되었다.

[0080] 이러한 스타일토큰 변환은 원본 스타일토큰을 목표스타일에 맞추어 변환하는 과정으로, 인코더-디코더(encoder-decoder) 기반 사전학습 언어모델인 mT5를 사용하였고, 모델의 인코더입력으로 스타일토큰을 마스크 처리한 문장이 들어가며, 디코더출력으로 마스크 처리된 부분에 해당하는 스타일토큰이 목표스타일에 맞추어 생성될 수 있다.

[0081] 또한, 적은 데이터를 활용한 효과적인 학습을 위하여 프롬프팅 방식을 적용하였는데, 마스크 처리된 부분을 예측하는 과정은 mT5의 사전학습 방식인 스팬커럽션(span corruption)과 동일하기 때문에, 별도의 처리 없이 프롬프팅을 수행할 수 있지만, 이를 적용한 모델은 변환되는 스타일뿐만 아니라 콘텐트의 예측 역시 진행하여야 하기 때문에, 기사체-방송체 변환은 온전한 콘텐트의 유지가 요구된다. 예를 들어, 도 14에 도시한 바와 같이 "(검증이)" 를 마스크 처리한 "<1>"에서 "확인이"라는 스팬(span)이 예측된다면, 스타일은 잘 변환되었지만 "검증"에서 "확인"으로 콘텐트가 훼손될 수 있다.

[0082] 따라서, 프롬프팅 방식에 스타일 토큰의 콘텐트를 유지하기 위해 원본 스타일토큰을 사용하는 콘텐트마커 프롬프팅(CMP)을 적용할 수 있는데, 콘텐트마커는 마스크 처리된 원본 스타일토큰으로 기존 콘텐트를 유지하는 역할을 하며, 도 13에 도시한 바와 같은 스타일토큰 변환부분에서 볼 수 있듯이, 마스크 처리된 문장과 콘텐트마커가 합쳐져 변환모델의 입력으로 들어갈 수 있고, mT5는 사전학습 시 두 개 이상의 스팬(span)을 마스크 처리하기 위해 여러 개의 마스크 토큰이 존재하며,이러한 점을 활용하여 마스크토큰으로 마스크 처리된 부분과 원본 스타일토큰을 매칭하였다.

[0083] 여기에서, 스타일 변환모델은 mT5-base 모델을 사용하였으며, batch size 8과 5 epoch로 학습하였고, 문장과 콘텐트마커에 사용된 마스크토큰은 mT5에 사용된 마스크토큰과 동일하게 사용하였으며, 데이터셋 2,000개의 문장에 대해 500개씩으로 나누어 검증(4-fold validation)을 진행하였고, 모델 성능 측정을 위한 지표(metric)는 EM을 사용하였다.

[0084] 아래의 표 4는 SC와 CMP 방식을 적용하여 실험, 비교한 결과표를 나타내는데, 베이스라인으로 사용한 SC는 스팬커럽션(span corruption) 프롬프팅을 적용하였고, SC(0.7305)와 CMP(0.9778)에 0.2473의 큰 성능 차이가 존재함을 볼 수 있으며, 특히 콘텐트의 유지가 필수적인 괄호 내 정보 유지에서 SC는 0.1313의 낮은 성능을 보인 반

면, CMP는 0.8833의 성능을 보였고, 종결어미 변환 또한 해라체-합쇼체 변환 이외에도 기존의 콘텐트를 유지해야 하므로 SC에서 0.6794라는 낮은 성능을 보였음을 알 수 있다.

표 4

[0085]

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

[0086] 아래의 표 5는 실험된 두 모델에 대한 실제예측결과를 나타내는 것으로, (1)에서 콘텐트 유지의 효과를 확인할 수 있는데, SC는 방송체스타일의 스타일토큰을 생성하는 것은 성공하였으나, 콘텐트를 유지하지 못하여 틀린 예측을 하고 있는 반면에, CMP는 스타일의 변환과 함께 콘텐트 역시 유지하고 있음을 알 수 있다.

[0087] 또한, (2)와 (3)에서 볼 수 있듯이 마스크 된 콘텐트의 확인 불가로 괄호 내 정보 유지 또는 삭제도 잘 이루어지지 않고 있는데, 이러한 분석을 통해 CMP가 기존 프롬프팅의 효과를 내며 콘텐트를 유지하고 있음을 알 수 있다.

[0088] 그러나, (4)와 (5)에서 볼 수 있듯이, CMP에서도 괄호 내 정보 유지 또는 삭제에서 오류가 발생하는데, "추진위"나 "기저에 있는" 과 같이 잘 쓰이지 않는 표현이 나타날 때 틀린 예측을 하는 경향을 보이며, 이는 데이터셋이 포함한 표현의 다양화를 통하여 해결 가능한 문제로 예상된다.

표 5

[0089]

(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"

[0090] 상술한 바와 같이 기사체-방송체 텍스트스타일 변환은 다른 스타일 변환보다 콘텐트의 훼손에 민감하다는 특성

을 가지고 있으며, 이러한 특성에 맞추어 본 발명의 실시예에서는 기사체-방송체 병렬 데이터셋을 구축하였고, 스타일토큰 기반 텍스트스타일 변환모델에 적용되는 콘텐트마커 프롬프팅방식을 통해 콘텐트를 유지하면서 프롬프팅 효과를 내었으며, 이를 통해 기사체-방송체 데이터셋에서 0.9778의 높은 변환 성능을 달성하였다.

[0091] 이상의 설명에서는 본 발명의 다양한 실시예들을 제시하여 설명하였으나 본 발명이 반드시 이에 한정되는 것은 아니며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 여러 가지 치환, 변형 및 변경이 가능함을 쉽게 알 수 있을 것이다.

부호의 설명

[0092] 100 : 사용자단말기

200 : 통신망

300 : 변환서비스제공서버

310 : 인터페이스부

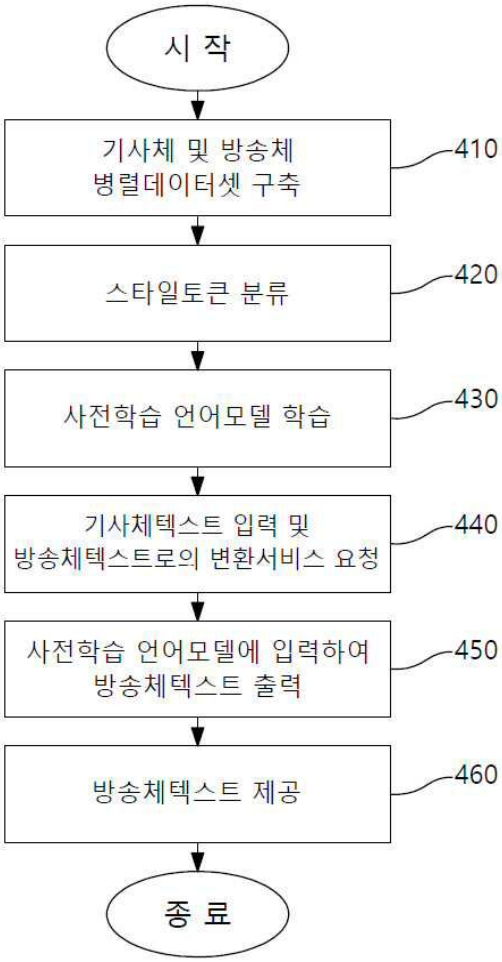
320 : 데이터셋구축부

330 : 스타일토큰분류부

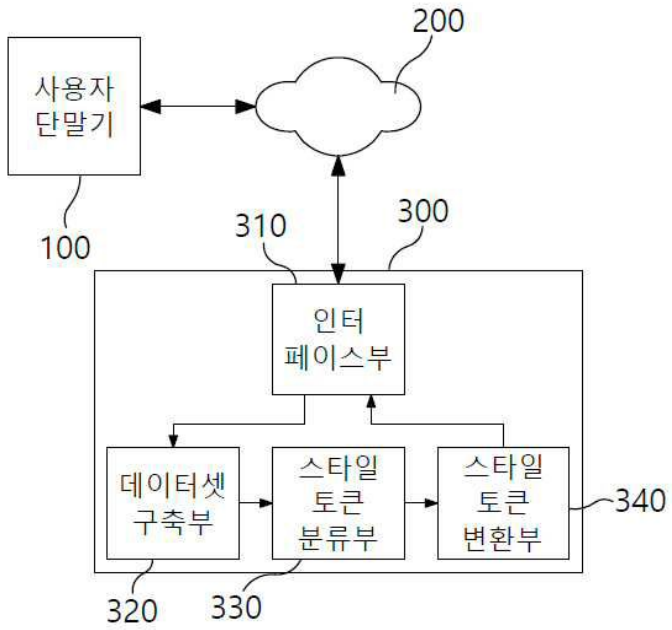
340 : 스타일토큰변환부

도면

도면1



도면2



도면3

기사체 : (서울=연합뉴스) 조다운 기자 = ' 서해 공무원 피격 사건 ' 을 수사하는 검찰이 문재인 전 대통령에 대한 조사 없이 다음 주 중 수사를 마무리할 전망이다. '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈(68·구속기소) 전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는 분위기다.

url : https://n.news.naver.com/mnews/article/001/0013659709?rc=N&ntype=RANKING

섹션: 정치

도면4

기사체 : 서해 공무원 피격 사건 ' 을 수사하는 검찰이 문재인 전 대통령에 대한 조사 없이 다음 주 중 수사를 마무리할 전망이다.
url :https://n.news.naver.com/mnews/article/001/0013659709?rc=N&ntype=RANKING **섹션:** 정치

기사체 : '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈(68·구속기소) 전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는 분위기다.
url : https://n.news.naver.com/mnews/article/001/0013659709?rc=N&ntype=RANKING **섹션:** 정치

도면5

{**기사체** : ' 서해 공무원 피격 사건 ' 을 수사하는 검찰이 문재인 전 대통령에 대한 조사 없이 다음 주 중 수사를 마무리할 전망이다. , **방송체** : ' 서해 공무원 피격 사건 ' 을 수사하는 검찰이 문재인 전 대통령에 대한 조사 없이 다음 주 중 수사를 마무리할 전망입니다.'
url : https://n.news.naver.com/mnews/article/001/0013659709?rc=N&ntype=RANKING, **섹션:** 정치}
{**기사체** : '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈(68·구속기소) 전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는 분위기다. , **방송체** : '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈 전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는 분위기입니다. url : https://n.news.naver.com/mnews/article/001/0013659709?rc=N&ntype=RANKING, **섹션:** 정치}

도면6

기사체 : '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈(68·**구속기소**) 전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는 **분위기다**.

괄호 : (68·**구속기소**) **종결어미** : **분위기다**.

도면7

기사체 : '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈(68·**구속기소**) 전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는 **분위기다**.
마스킹, 전처리된 기사체 : '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈 전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는<mask>
마스킹된 기사체 : '사건 은폐'·'첩보 삭제' 의혹의 최종 책임은 서훈<mask>전 국가안보실장에게 있다고 보고 그와 그의 지시를 받아 사건 은폐에 가담한 이들을 사법처리하는 선에서 정리하는 <mask>

도면8

기사체 문장 : "(검증이) 불가능하다고" 판단했다.

⇩

문장 : "<extra_id_1> 불가능하다고" <extra_id_2>
CM : <extra_id_1> (검증이) <extra_id_2> 판단했다.

도면9

문장 : "<extra_id_1> 불가능하다고" <extra_id_2>
CM : <extra_id_1> (검증이) <extra_id_2> 판단했다.

도면10

문장 : "<extra_id_1> 불가능하다고" <extra_id_2>
CM : <extra_id_1> (검증이) <extra_id_2> 판단했다.

도면11

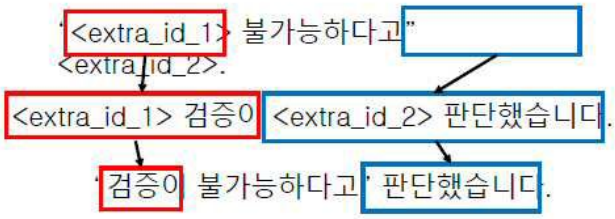
문장 : “<extra_id_1> 불가능하다고” <extra_id_2>
CM : <extra_id_1> (검증이) <extra_id_2> 판단했다.

<extra_id_1> 검증이 <extra_id_2> 판단했습니다.

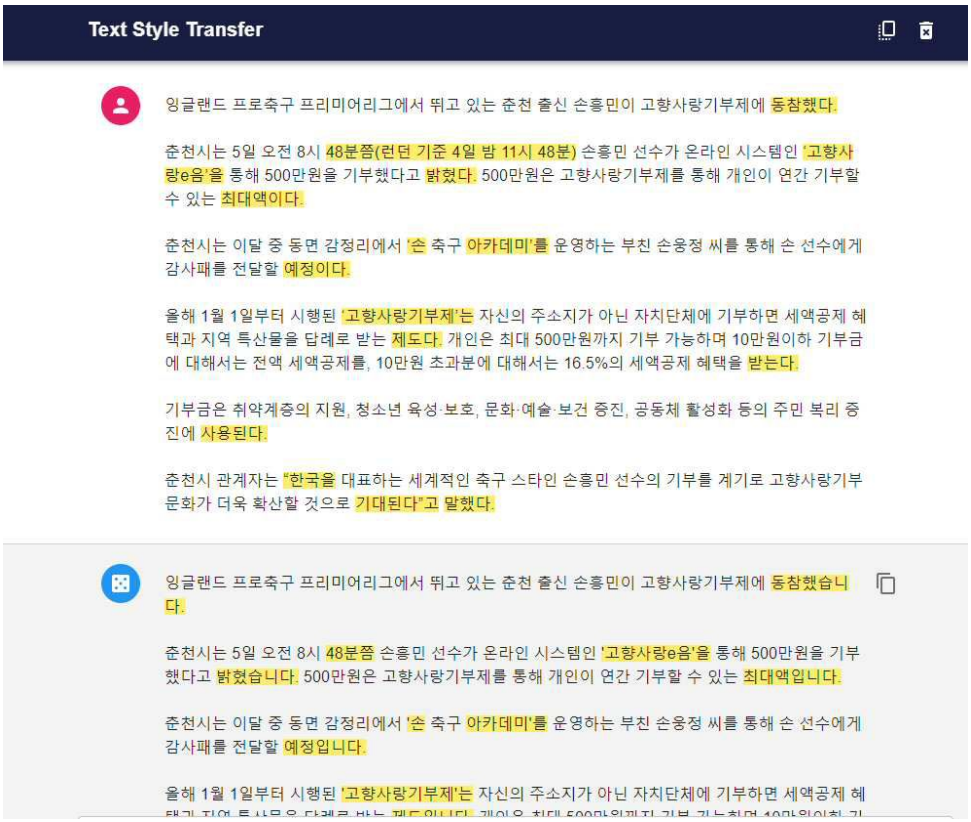
도면12

<extra_id_1> 검증이 <extra_id_2> 판단했습니다.

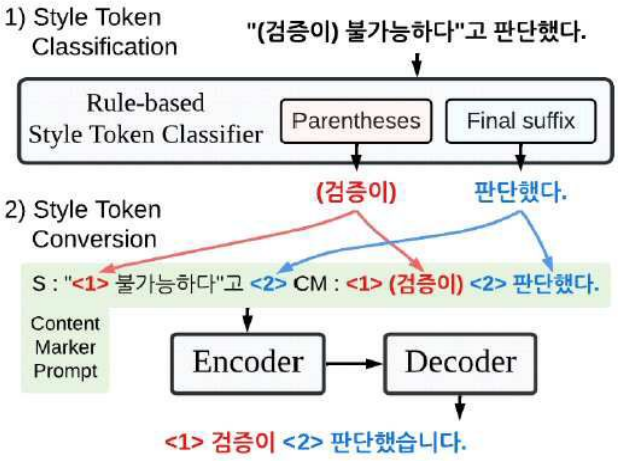
도면13



도면14



도면15


1) Style Token Classification
"(검증이) 불가능하다"고 판단했다.
Rule-based Style Token Classifier
Parentheses
Final suffix
(검증이)
판단했다.
2) Style Token Conversion
S : "<1> 불가능하다"고 <2> CM : <1> (검증이) <2> 판단했다.
Content Marker Prompt
Encoder
Decoder
<1> 검증이 <2> 판단했습니다.