



(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0074291  
(43) 공개일자 2025년05월27일

(51) 국제특허분류(Int. Cl.)  
G06F 18/214 (2023.01) G06F 40/232 (2020.01)  
G06N 3/08 (2023.01) G10L 13/08 (2006.01)  
G10L 15/26 (2006.01)  
(52) CPC특허분류  
G06F 18/214 (2023.01)  
G06F 40/232 (2020.01)  
(21) 출원번호 10-2023-0161177  
(22) 출원일자 2023년11월20일  
심사청구일자 2023년11월20일

(71) 출원인  
고려대학교 산학협력단  
서울특별시 성북구 안암로 145, 고려대학교 (안암  
동5가)  
(72) 발명자  
임희석  
서울특별시 성북구 안암로 145 자연계캠퍼스 애기  
능생활관 311호 NLP&AI연구실  
구선민  
경기도 의정부시 오목로 171, 204동 1504호  
(74) 대리인  
김동용

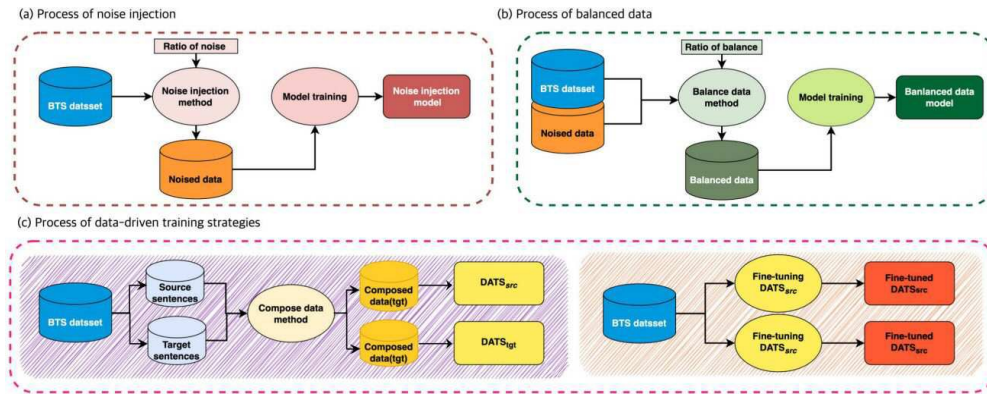
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 합성 데이터 기반 한국어 맞춤법 교정 데이터셋 생성 장치 및 방법

(57) 요약

합성 데이터 기반 한국어 맞춤법 교정 데이터셋 생성 장치 및 방법이 개시된다. 상기 맞춤법 교정 데이터셋 생성 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 미리 정해진 언어로 된 텍스트 형식의 복수의 문장들을 포함하는 단일 말뭉치가 획득되는 단계, 상기 단일 말뭉치로부터 병렬 말뭉치를 생성하는 단계, 상기 병렬 말뭉치에 포함된 소스 문장들에 노이즈를 추가하는 단계, 및 상기 단일 말뭉치에 포함된 복수의 문장들 중 적어도 일부와 노이즈가 추가된 소스 문장을 포함하는 제1 학습 데이터를 포함하는 학습 데이터를 생성하는 단계를 포함한다.

대표도 - 도1



(52) CPC특허분류

*G06N 3/08* (2023.01)

*G10L 13/08* (2013.01)

*G10L 15/26* (2013.01)

이 발명을 지원한 국가연구개발사업

|             |                                    |
|-------------|------------------------------------|
| 과제고유번호      | 1711193711                         |
| 과제번호        | 2018-0-01405-006                   |
| 부처명         | 과학기술정보통신부                          |
| 과제관리(전문)기관명 | 정보통신기획평가원                          |
| 연구사업명       | 정보통신방송혁신인재양성                       |
| 연구과제명       | Human-inspired 복합지능 원천기술 연구 및 인력양성 |
| 기 여 율       | 1/2                                |
| 과제수행기관명     | 고려대학교산학협력단                         |
| 연구기간        | 2023.01.01 ~ 2023.12.31            |

이 발명을 지원한 국가연구개발사업

|             |                         |
|-------------|-------------------------|
| 과제고유번호      | 1345365402              |
| 과제번호        | 2021R1A6A1A03045425     |
| 부처명         | 교육부                     |
| 과제관리(전문)기관명 | 한국연구재단                  |
| 연구사업명       | 이공학학술연구기반구축             |
| 연구과제명       | Human-inspired AI 연구소   |
| 기 여 율       | 1/2                     |
| 과제수행기관명     | 고려대학교                   |
| 연구기간        | 2023.03.01 ~ 2024.02.29 |

---

## 명세서

### 청구범위

#### 청구항 1

적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 맞춤법 교정 데이터셋 생성 방법에 있어서,

미리 정해진 언어로 된 텍스트 형식의 복수의 문장들을 포함하는 단일 말뭉치가 획득되는 단계;

상기 단일 말뭉치로부터 병렬 말뭉치를 생성하는 단계;

상기 병렬 말뭉치에 포함된 소스 문장들에 노이즈를 추가하는 단계; 및

상기 단일 말뭉치에 포함된 복수의 문장들 중 적어도 일부와 노이즈가 추가된 소스 문장을 포함하는 제1 학습 데이터를 포함하는 학습 데이터를 생성하는 단계를 포함하는 맞춤법 교정 데이터셋 생성 방법.

#### 청구항 2

제1항에 있어서,

상기 병렬 말뭉치를 생성하는 단계는,

상기 단일 말뭉치에 포함되는 복수의 문장들 각각을 TTS(Text-to-Speech) 변환 후 다시 STT(Speech-to-Text) 변환하여 상기 병렬 말뭉치를 생성하는,

맞춤법 교정 데이터셋 생성 방법.

#### 청구항 3

제1항에 있어서,

상기 노이즈를 추가하는 단계는, 분리 오류, 모음 변형 오류, 발음 오류, 구두점 부착 오류, 외래어 변환 오류, 및 신조어 오류 중 적어도 하나를 노이즈로 추가하는,

맞춤법 교정 데이터셋 생성 방법.

#### 청구항 4

제1항에 있어서,

상기 노이즈를 추가하는 단계는,

미리 정해진 주입 강도로, 상기 소스 문장들 중 적어도 일부의 문장에 노이즈를 추가하는,

맞춤법 교정 데이터셋 생성 방법.

#### 청구항 5

제4항에 있어서,

상기 미리 정해진 주입 강도는 0.1 내지 0.4이고,

상기 미리 정해진 주입 강도는 노이즈가 삽입된 문장에 포함된 각 단어에 노이즈가 추가되는 빈도를 의미하는,

맞춤법 교정 데이터셋 생성 방법.

#### 청구항 6

제1항에 있어서,

상기 학습 데이터는 상기 병렬 말뭉치에 포함된 복수의 문장 쌍들 중 적어도 일부를 포함하는 제2 학습 데이터

를 포함하는,  
맞춤법 교정 데이터셋 생성 방법.

#### 청구항 7

제6항에 있어서,  
상기 학습 데이터에 포함되는 제2 학습 데이터와 상기 제1 학습 데이터의 비율은 5:5 내지 1:9인,  
맞춤법 교정 데이터셋 생성 방법.

#### 청구항 8

제1항에 있어서,  
상기 학습 데이터를 이용하여 소정의 인공 신경망 모델을 학습시킴으로써 맞춤법 교정 모델을 생성하는 단계를 더 포함하는,  
맞춤법 교정 데이터셋 생성 방법.

#### 청구항 9

제1항에 있어서,  
상기 노이즈를 추가하는 단계는, 분리 오류, 모음 변형 오류, 발음 오류, 구두점 부착 오류, 외래어 변환 오류, 및 신조어 오류를 노이즈로 추가하는,  
맞춤법 교정 데이터셋 생성 방법.

### 발명의 설명

#### 기술 분야

[0001] 본 발명은 합성 데이터 생성 방법에 관한 것이며, 보다 상세하게는 합성 데이터를 기반으로 한국어 맞춤법 교정 분야 데이터셋을 변형 및 생성하는 장치 및 방법에 관한 것이다.

#### 배경 기술

[0002] 합성 데이터셋(synthetic dataset)은 실제 데이터에 대한 대체재로써 실제 데이터의 특성을 반영하여 원본과 유사한 인공 데이터이다. 현실 세계에서는 개인정보가 다수 포함된 주민등록증, 진료비 영수증과 같은 문서 데이터, 피싱 정보를 감지하기 위한 보이스피싱 관련 음성 데이터, 국방 정보가 다수 포함된 데이터셋은 직접 수집하기 쉽지 않다. 그렇기에 이러한 데이터는 실제 데이터 없이 합성 데이터만을 사용해 제작할 수 밖에 없다. 더욱이 최근 자기 지도학습(self-supervised learning)을 기반으로 다양한 합성 데이터 연구가 떠오르고 있고 데이터셋의 양을 기하급수적으로 늘리는 방법으로 모델이 생성한 데이터를 사용하는 것이 굉장히 효과적임이 증명되었다.

[0003] 최근, GPT3, ChatGPT, LaMDA 등 LLM(Large Language Model, 거대 언어 모델)의 등장으로 양질의 합성 데이터를 만들 수 있어, 추후에는 합성 데이터로만 학습 시킨 모델이 등장할 가능성이 높아지고 있다. 기존 연구들은 기본적으로 사람이 구축한 데이터셋 혹은 공공에게 오픈된 데이터를 기반으로 많은 연구들이 이루어져 왔지만, 합성 데이터 독점 환경에서 학습된 모델에 대한 연구는 충분히 이루어지지 않았다. 즉, 다양한 합성 데이터 변형 방법의 부족으로 합성 데이터 독점 환경에서 학습된 모델에 대한 분석이 활발하지 않은 편이다.

[0004] 따라서 이러한 문제점을 완화시키기 위하여, 본 발명에서는 한국어 맞춤법 교정 분야를 대상으로 합성 데이터 독점 환경 연구가 활발히 진행될 수 있도록 다양한 방법의 합성 병렬 데이터 변형 및 생성 방법을 제안하고자 한다. 보다 구체적으로, 단일 말뭉치만 존재하는 환경에서 3가지 방법(데이터 품질 관리, 데이터 균형, 다양한 데이터 기반 학습 전략)과 함께 한국어 맞춤법 교정용 합성 병렬 말뭉치를 생성하는 장치 및 방법을 만들고자 한다. 또한, 데이터의 특성마다 효과적인 합성 비율이 다를 수 있기 때문에 다양한 세팅으로 모델의 성능을 개선할 수 있도록 한다.

## 선행기술문헌

### 특허문헌

- [0005] (특허문헌 0001) 대한민국 공개특허 제2022-0153731호 (2022.11.21. 공개)  
(특허문헌 0002) 대한민국 공개특허 제2022-0060350호 (2022.05.11. 공개)

## 발명의 내용

### 해결하려는 과제

- [0006] 본 발명이 이루고자 하는 기술적인 과제는 합성 데이터 기반 한국어 맞춤법 교정 데이터셋 생성 장치 및 방법을 제공하는 것이다.

### 과제의 해결 수단

- [0007] 본 발명의 일 실시예에 따른, 맞춤법 교정 데이터셋 생성 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 미리 정해진 언어로 된 텍스트 형식의 복수의 문장들을 포함하는 단일 말뭉치가 획득되는 단계, 상기 단일 말뭉치로부터 병렬 말뭉치를 생성하는 단계, 상기 병렬 말뭉치에 포함된 소스 문장들에 노이즈를 추가하는 단계, 및 상기 단일 말뭉치에 포함된 복수의 문장들 중 적어도 일부와 노이즈가 추가된 소스 문장을 포함하는 제1 학습 데이터를 포함하는 학습 데이터를 생성하는 단계를 포함한다.

### 발명의 효과

- [0008] 본 발명의 실시예에 따른, 합성 데이터 기반 한국어 맞춤법 교정 데이터셋 생성 장치 및 방법에 의한 경우, 한국어 맞춤법 교정 분야를 대상으로 다양한 세팅의 합성 병렬 말뭉치 변형 및 생성 방법을 제시함으로써 LLM(Large Language Model, 거대 언어 모델)의 발전으로 앞으로 활발하게 이루어질 합성 데이터 독점 환경에서의 한국어 맞춤법 교정 모델 성능 향상을 유도할 수 있다.

### 도면의 간단한 설명

- [0009] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.  
도 1은 (a) 노이즈 주입 과정, (b) 데이터 균형의 과정, (c) 데이터 기반 학습 전략을 도시한다.  
도 2는 단일 말뭉치를 병렬 말뭉치로 합성하기 위한 아키텍처를 도시한다.  
도 3은 노이즈 주입에 의한 실험 결과로서, (a) AI-HUB의 합성 데이터에 노이즈를 주입했을 때의 결과와, TED의 합성 데이터에 노이즈를 주입했을 때의 결과를 도시하며, x-축은 노이즈 주입의 강도를 나타낸다.  
도 4는 데이터 균형에 의한 실험 결과를 도시하고, x-축은 클린 데이터와 노이즈 데이터의 비율을 나타낸다.  
도 5는 본 발명의 일 실시예에 따른 맞춤법 교정 데이터셋 생성 방법을 설명하기 위한 흐름도이다.

### 발명을 실시하기 위한 구체적인 내용

- [0010] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.
- [0011] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.
- [0012] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로

만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0013] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0014] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0015] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0016] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0017] LLM(Large Language Models, 거대 언어 모델)의 발전으로 합성 데이터(synthetic data)를 용이하게 생성할 수 있게 되었으며, 프라이버시 문제로 인한 합성 데이터의 필요성을 고려하면 앞으로 합성 데이터만을 이용하여 훈련된 모델이 활용될 것으로 예상된다. 본 명세서에서는, 데이터 품질 관리(data quality control)가 합성 데이터만으로 학습되는 모델에 적용되었을 때 긍정적인 영향을 미치는지 알아보려고 한다. 아래와 같이 3가지 실험을 통해 이를 검증한다.

#### [0018] A. 노이즈 주입 강도(Strength of noise injection)가 모델 성능(model performance)에 미치는 영향

[0019] 첫째, 데이터 품질 관리의 접근법으로 노이즈 주입 기법을, 합성 데이터만으로 구성된 모델에, 적용하였을 때의 영향을 조사한다. 도 1의 (a)는 노이즈 주입 과정을 도시한다. 활용되는 데이터셋은 소스 및 타겟 문장 쌍들(source and target sentence pairs)로 구성된 병렬 문장들로 구축된다.

[0020] 합성 데이터에 노이즈를 주입하기 위해, 적어도 하나의(예컨대, 6 가지의) 유형의 노이즈(분리, 모음 변형, 발음, 구두점 부착, 외래어 변환, 신조어 오류(separation, vowel transformation, pronunciation, punctuation, loanword transformation, and neologism errors))를 소스 문장들에 적용할 수 있다. 분리 오류(separation error)는 글자의 자음과 모음이 분리된 경우를, 모음 변형 오류(vowel transformation error)는 한 글자 내에서 모음이 다른 모음으로 교체되는 경우를, 발음 오류(pronunciation error)는 글자가 발음대로 변형된 경우를, 구두점 부착 오류(punctuation error)는 구두점이 문장 내에 불필요한 위치에 부착된 경우를, 외래어 변환 오류(loanword transformation error)는 글자의 일부가 외래어로 변환된 경우를 의미한다. 본 발명에서는 한국어를 대상으로 하므로 외래어는 가장 많이 쓰이는 언어 중 하나인 영어로 정의될 수 있다. 즉, 외래어 변환 오류는 한국어의 일부를 영어로 변환하여 노이즈를 삽입하는 방법이다. 신조어 오류(neologism error)는 기존 문법 체계에 없는 문법을 사용하여 글자를 변형하는 경우를 의미한다.

[0021] 본 발명에서는, 소정의 범위(예컨대, 0.1에서 1.0 까지의 범위)에서 적어도 하나의 노이즈 주입 강도(noise-injection intensities, 예컨대 0.1, 0.2, 0.4, 0.6, 0.8, 및 1.0)에 기초한 다양한 케이스를 고려한다. 노이즈 비율로도 명명될 수 있는 노이즈 주입 강도는 문장 내에 포함되는 각 단어에 노이즈가 부가될 확률로 이해될 수도 있다.

[0022] 노이즈는 각 문장(소스 문장을 의미할 수 있음) 내에서 단어 수준에서 설정된 비율에 따라 삽입될 수 있다. 예컨대, 노이즈 삽입 강도가 1.0이면, 문장 내의 모든 단어에 노이즈가 추가될 수 있다. 생성된 데이터를 이용하

여 모델이 학습됨에 따라 노이즈 주입 모델(noisy injection mode, 노이즈가 주입된 학습 데이터를 이용하여 학습된 모델을 의미할 수 있음)이 획득될 수 있다. 또한, 추가되는 노이즈는 상술한 예시적인 노이즈 중 적어도 하나가 될 수 있다.

[0023] **B. 노이즈 데이터(noisy data)와 클린 데이터(clean data)의 비율이 모델 성능에 미치는 영향**

[0024] 둘째로, 합성 데이터만을 이용하여, 데이터 균형 기법(balanced data method)을 모델에 적용하였을 때의 영향을 조사한다. 데이터 균형은 학습에 이용하기 위해 노이즈 데이터와 클린 데이터의 비율을 설정함으로써, 데이터셋을 구축하는 방법론이다. 클린 데이터는 노이즈 주입과 같은 추가적인 수정(modifications)이 가해지지 않은 합성 데이터를 나타낸다. 반면에, 노이즈 데이터는 노이즈 주입 기법을 통해 생성된 데이터셋을 의미할 수 있다. 예컨대, 설정 비율(노이즈 데이터와 클린 데이터의 비율)이 5:5이면, 전체 데이터의 50%는 노이즈가 부가된 데이터이다. 노이즈를 부가하는 방식은 도 1의 (a)에서의 기법과 동일할 수 있다.

[0025] 도 1의 (b)는 데이터 균형을 획득하는 과정을 도시한다. 클린 데이터와 노이즈 데이터가 설정된 비율에 따라 결합(combined)어 학습 데이터를 구성할 수 있다. 전체 데이터셋에서 노이즈 데이터의 비율에 따른 영향을 관찰하기 위해, 클린 데이터와 노이즈 데이터를 다양한 비율(예컨대, 5가지의 비율(5:5, 4:6, 3:7, 2:8, 및 1:9))로 병합하여 전체 학습 데이터셋을 구축할 수 있다. 각 설정에 대하여, 학습 데이터셋을 이용하여 모델을 학습할 수 있다. 클린 데이터를 이용하여 학습된 베이스라인 모델(baseline model)의 성능과 생성된 학습 데이터셋(클린 데이터와 노이즈 데이터가 병합된 데이터셋을 의미함)을 이용하여 학습된 모델의 성능을 비교함으로써, 클린 데이터와 노이즈 데이터의 비율에 따른 영향을 분석할 수 있다.

[0026] **C. 학습 전략(training strategies)이 모델 성능에 미치는 영향**

[0027] 셋째로, 다양한 데이터 기반 학습 전략(data-driven training strategies)을 적용했을 때의 영향을 살펴본다. 도 1의 (c)는 데이터 기반 학습 전략을 도시한다. 합성 병렬 데이터는 오류를 포함하는 소스 문장들과 노이즈가 존재하지 않은 타겟 문장들로 구성된다.

[0028] 도 1의 (c)에서, 보라색 섹션은 동일한 특징(identical characteristics)을 갖는 데이터만을 이용하여 병렬 말뭉치(parallel corpora)를 구축하는 과정을 도시한다. 이를 위해, 상이한 특징을 가진 병렬 말뭉치를 분리(또는 분할)하고 동일한 학습 특징(identical training characteristics)을 갖는 문장들을 이용하여 쌍들을 재결합한다. 즉, 병렬 문장들로부터의 소스-소스 문장 쌍들을 이용하여 학습을 수행할 수 있다. 반대로, 병렬 문장들로부터 타겟 문장들을 분리하고, 타겟-타겟 문장 쌍들을 이용하여 학습을 수행할 수 있다. 이는, 동일한 특징을 가진 문장 쌍들(예를 포함하는 문장 쌍들 및/또는 예를 포함하지 않는 문장 쌍들)을 이용하여 DATS 모델들(Data-Driven Training Strategy models)을 학습함을 의미할 수 있다. 이에 따라 DATS<sub>src</sub>와 DAT<sub>tgt</sub> 모델들이 각각 생성될 수 있다.

[0029] 도 1의 (c)에서, 주황색 섹션은, 소스-타겟 문장 쌍들로 구성된 합성 데이터를 이용하여, 각 DATS 모델에 추가적인 미세조정(fine-tuning)을 적용하는 과정을 도시한다. 즉, 이러한 전략은 순차적으로 동일한 특징과 상이한 특징을 모두 가진 데이터를 학습하는 것이다. 이러한 과정을 통해, 미세조정된 DATS<sub>src</sub> 모델과 미세조정된 DATS<sub>tgt</sub> 모델을 획득할 수 있다.

[0030] 베이스라인 모델과 두 가지의 데이터 기반 학습 전략을 이용하여 학습된 모델을 비교함으로써, 모델 성능을 분석할 수 있다.

[0031] **실험 결과**

[0032] **A. 합성 데이터(Synthetic data)**

[0033] 학습을 위한 합성 데이터셋을 생성할 수 있다. 도 2는 단일 말뭉치를 병렬 말뭉치로 합성하기 위한 아키텍처를 도시한다. 사후 교정기(post-processor, GEC)를 학습하기 위해서는 인간 주석자(human annotators)에 의해 생성된 병렬 말뭉치가 필요하지만, 본 발명에서는 TTS 기법(text-to-speech technology)과 STT 기법(speech-to-text technology)을 결합하여 병렬 말뭉치를 생성한다. TTS에 의해 생성된 합성된 음성 데이터(synthesized voice data)는 TTS 변환(text-to-speech conversion)을 거친다. 예시적으로, Navers CLOVA Speech Recognition(CSR) API가 이용될 수 있다. 또한, 실시예에 따라, 음성 데이터(voice data)는 WaveNet 모델을 이용하여 합성될 수 있다. 병렬 말뭉치를 구축하는 과정은 다음 단계들을 포함한다. 1) 기존에 널리 사용되는 기법 등을 이용하여 사전에 구축된 단일 말뭉치(pre-built mono corpus)를 크롤링한다. 2) TTS 기법을 이용하여 음성(speech

h)으로 변환한다. 그리고, 3) STT 기법을 이용하여 텍스트로써 변환된 결과를 출력한다. 본 발명에서는, 실험을 위해, 대표적인 한국어 데이터 플랫폼인 AI-HUB(C. Park, M. Shim, S. Eo, S. Lee, J. Seo, H. Moon, and H. Lim, "Empirical analysis of parallel corpora and in-depth analysis using LIWC," Appl. Sci., vol. 12, no. 11, p. 5545, May 2022.)와 TED Korean dataset(<https://www.ted.com/talks?language=ko>)을 이용한다. 또한, 병렬 데이터를 구축하기 위해서는 상술한 기법이 이용될 수 있다.

[0034] 표 1은 실험에 이용된 데이터 통계를 나타낸다. 합성 데이터를 생성하기 위해서, AI-HUB 데이터로부터 92,000 문장들과 TED's Korean Transcript data로부터 119,883 문장들을 이용한다. 모델에 대한 합성 데이터의 영향을 보다 정확하게 분석하기 위해서, 활용 가능한 리소스 내에서 학습 데이터의 양을 최대화하였다. 이러한 데이터는, TTS를 이용한 음성(speech)으로의 변환과 STT를 이용한 텍스트로써의 변환된 결과 도출을 위한 원시 데이터(raw data)로 이용될 수 있다. 다시 말하면, 수도-병렬 말뭉치(pseudo-parallel corpus)를 생성하기 위해 TST 기법(Text-to-Speech-to-Text technology)이 이용될 수 있다.

[0035] [표 1]

| Dataset | Train   | Dev.  | Test  | Total   |
|---------|---------|-------|-------|---------|
| AI-HUB  | 92,000  | 5,000 | 3,000 | 100,000 |
| TED     | 119,883 | 5,000 | 3,000 | 127,883 |

[0036]

## [0037] B. 구현예

[0038] 모델 자체보다는 기법의 성능에 집중하기 위해서, 바닐라 트랜스포머(A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1-11.)를 모델 학습에 이용하였다. 그러나, 본 발명의 권리범위가 학습 모델의 종류나 구조에 제한되는 것은 아니며, 다양한 신경망 모델 기반의 학습 모델이 이용될 수 있다. 실시예에 따라서는, 두 개 이상의 신경망 모델이 결합된 형태의 학습 모델이 이용될 수도 있다. Fairseq(M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," 2019, arXiv:1904.01038.)는 구현을 위해 이용되었다. 서브워드 토큰화(subword tokenization)의 경우, 센텐스피스(SentencePiece, T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," 2018, arXiv:1808.06226.)를 이용하였고, 50,000 워드의 어휘 사전(vocabulary size)을 설정하였다. 이에 따라, BLEU와 GLEU를 평가하였다. BLEU와 GLEU는 높은 값을 가질수록 더 좋은 성능을 나타낸다.

[0039] 이하에서는, 구체적인 실험 결과를 설명한다.

## [0040] A. 노이즈 주입에 의한 결과

[0041] 노이즈 주입에 의한 실험 결과는 도 3에 도시되어 있다. 도 3에서, (a)는 AI-HUB에 기초하여 생성된 합성 데이터의 경우이고, (b)는 TED에 기초하여 생성된 합성 데이터의 경우이다. 노이즈 주입 비율(경우에 따라, 노이즈 주입 강도와 동일한 의미를 가질 수 있음)은 문장 내의 토큰들에 노이즈가 주입될 확률을 나타낼 수 있다. 베이스라인 성능은, AI-HUB와 TED로부터의 합성 데이터로 학습했을 때, 각각 65.69와 56.14이다. 대부분의 경우에서, 성능은 감소하였으며, 노이즈 비율이 0.1과 0.8일 때 가장 낮은 성능 63.83와 54.2을 보였고, 노이즈 비율이 0.4와 0.2일 때 가장 높은 성능 65.37과 56.55를 보였다. 특히, TED의 성능은, 노이즈 비율이 0.2일 때, 베이스라인 보다 약간 높은 성능을 보였다.

[0042] 노이즈 주입 기법을 이용하여 생성된 데이터로 학습된 모델들을 비교하면, 노이즈 비율이 낮을 때 성능은 개선되는 경향을 보인다. 그러나, AI-HUB와 TED에서 강한 부정적인 영향은 노이즈 비율이 각각 0.1과 0.8에서 발생한다.

## [0043] B. 데이터 균형에 의한 결과

[0044] 도 4는 데이터 균형에 의한 실험 결과를 도시한다. 클린 합성 데이터와 노이즈 데이터를 이용한 실험이 진행되었다. 클린 데이터와 노이즈 데이터의 비율은 5 가지로 설정되었다. 노이즈 주입 비율은 도 3에서의 경우와 같이 적용되었다. 실험 결과, 노이즈 주입 비율이 0.6을 초과하는 경우에 성능에 저하가 발생하였다. 특히, 베이스라인과 비교하면, 클린 데이터와 노이즈 데이터의 비율이 1:9이고 노이즈 비율이 1.0일 때 AI-HUB의 결과는 19.07%(53.16) 만큼 저하되었다. 또한, 베이스라인과 비교하면, TED의 결과는 32.04%(38.15) 만큼 저하되었다. 이에 따라, 성능 차이(performance gap)은 노이즈 주입 비율이 증가하고 클린 데이터와 노이즈 데이터의 비율이

노이즈 데이터 쪽으로 편향될 때 발생하는 것을 알 수 있다.

[0045] 한편, 노이즈 주입 비율이 0.1-0.4인 경우에는 데이터 균형의 비율과는 무관하게 성능이 향상되는 것을 볼 수 있다. 이에 따라, 노이즈 비율이 낮을 때, 모델은 데이터 균형의 설정과 무관하게 강건하고 적절한 양의 노이즈가 필요함을 추측할 수 있다.

### [0046] C. 데이터 기반 학습 전략에 의한 결과

[0047] 표 4는 데이터 기반 학습 전략에 따른 실험 결과를 나타낸다. DATS<sub>src</sub> 모델과 DATS<sub>tgt</sub> 모델은 동일한 특성을 갖는 데이터만을 이용하여 학습된 모델이다. DATS<sub>src</sub>의 경우, AI-HUB와 TED는 각각 45.82와 39.27 BLEU를 나타내었다. 이는, 각각 -19.87과 -16.87 포인트의 감소를 의미한다. DATS<sub>tgt</sub>의 경우, AI-HUB는 -34.71 포인트 감소한 30.98을, TED는 -16.87 포인트 감소한 39.27을 나타내었다. DATS<sub>src</sub>의 성능이 DATS<sub>tgt</sub>에 비해 덜 감소하였지만, 두 모델 모두 베이스라인에 비해 성능이 열화됨을 알 수 있다.

[0048] [표 4]

| Model                          | AIHUB          |                | TED            |                |
|--------------------------------|----------------|----------------|----------------|----------------|
|                                | BLEU           | GLEU           | BLEU           | GLEU           |
| Base                           | 39.68          | -              | 46.09          | -              |
| BTS baseline                   | 65.69          | 55.70          | 56.14          | 49.48          |
| DATS <sub>src</sub>            | 45.82 (-19.87) | 17.64 (-38.06) | 39.27 (-16.87) | 11.17 (-38.31) |
| DATS <sub>tgt</sub>            | 30.98 (-34.71) | 27.88 (-27.82) | 39.27 (-16.87) | 34.32 (-15.16) |
| Fine-tuned DATS <sub>src</sub> | 44.90 (-20.79) | 37.94 (-17.76) | 39.23 (-16.91) | 34.20 (-15.28) |
| Fine-tuned DATS <sub>tgt</sub> | 21.66 (-44.03) | 19.89 (-35.81) | 38.83 (-17.31) | 33.86 (-15.62) |

[0049]

[0050] 미세조정된 DATS<sub>src</sub> 모델과 미세조정된 DATS<sub>tgt</sub> 모델은, 동일한 특징과 상이한 특징을 가진 데이터를 이용하여 순차적으로 학습 방법을 적용한 결과를 나타낸다. 미세조정된 DATS<sub>src</sub>의 경우, AI-HUB의 BLEU 스코어는 44.9(베이스라인에 비해 -20.79 포인트 낮음)이고, TED의 BLEU 스코어는 39.23(베이스라인에 비해 -16.91 포인트 낮음)이다. 미세조정된 DATS<sub>tgt</sub> 모델의 경우, AI-HUB의 스코어는 베이스라인에 비해 -44.03 포인트 낮은 21.66이고, TED의 스코어는 베이스라인에 비해 -17.31 포인트 낮은 38.83이다.

[0051] 이는, 합성 데이터로 추가적인 미세조정을 수행하더라도, 성능에 부정적인 영향을 미침을 의미한다. 이를 통해, 초기 모델의 훈련이 중요함을 유추해 볼 수 있다.

[0052] 도 5는 본 발명의 일 실시예에 따른 맞춤법 교정을 위한 교정 데이터셋 생성 방법을 설명하기 위한 흐름도이다.

[0053] 도 5를 참조하면, 맞춤법 교정 모델 생성 방법 등으로도 명명될 수 있는 교정 데이터셋 생성 방법은 적어도 프로세서(Processor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 따라서, 교정 데이터셋 생성 방법을 구성하는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 랩탑 컴퓨터, 태블릿 PC 등을 포함할 수 있다. 또한, 컴퓨팅 장치는 물리적으로 구분되는 복수의 장치들로 구현되는 것도 가능하다.

[0054] 우선, 단일 말뭉치가 획득된다(S110). 단일 말뭉치는 소정의 언어(예컨대, 한국어)로된 텍스트 형식의 문장들을 포함할 수 있다. 단일 말뭉치는, 유무선 통신망을 통하여, 크롤링을 수행함으로써 획득될 수 있다. 실시예에 따라, 단일 말뭉치는 USB 메모리 장치와 같은 외부 장치로부터, 소정의 입출력 인터페이스를 통해, 수신되거나, 컴퓨팅 장치에 미리 저장되어 있을 수도 있다. 획득된 단일 말뭉치는 맞춤법 교정 데이터에서 타겟 데이터(또는 타겟 문장)로 이용될 수 있다.

[0055] 획득된 단일 말뭉치로부터 병렬 말뭉치가 생성될 수 있다(S120). 병렬 말뭉치는 단일 말뭉치에 포함되는 문장을 TTS 변환한 후 다시 STT 변환을 수행함으로써 생성될 수 있다. 이때, 추가적으로 생성된 문장은 맞춤법 교정 데이터에서 소스 데이터(또는 소스 문장)으로 이용될 수 있다.

[0056] 병렬 말뭉치에 포함되는 문장에 노이즈가 주입될 수 있다(S130). 추가되는 노이즈는 앞서 설명한 복수의 노이즈 유형 중 적어도 하나를 포함할 수 있다. 또한, 노이즈는 병렬 말뭉치에 포함된 문장들 중 소스 문장을 대상으로 부가될 수 있다. 노이즈 주입의 강도(또는 노이즈 주입의 비율)은 미리 설정될 수 있고, 실시예에 따라 복수의 노이즈 주입 강도가 적용될 수도 있다. 이에 따라, 설정된 노이즈 주입 강도에 따라 노이즈가 주입된 소스 문장

들이 생성될 수 있다. 이때, 노이즈가 주입되는 강도는 설정에 따라 가변될 수 있으나, 0.1 내지 0.6(또는 0.1 내지 0.4)의 강도로 노이즈를 주입할 경우, 학습 모델의 성능 향상을 기대할 수 있다.

[0057] 학습 데이터가 생성된다(S140). 학습 데이터는 (한국어) 맞춤법 교정을 위한, 즉 맞춤법 교정 모델의 생성 과정에서 학습을 위한 데이터를 의미할 수 있다. 또한, 학습 데이터는 노이즈가 부가된 소스 문장을 포함하는 병렬 문장 쌍들을 의미하는 노이즈 데이터와 노이즈가 부가되지 않은 소스 문장을 포함하는 병렬 문장 쌍들을 의미하는 클린 데이터를 포함할 수 있다. 노이즈 데이터의 소스 문장은 상술한 노이즈 주입 기법을 통해 노이즈가 주입된 문장을 의미할 수 있다. 또한, 클린 데이터는 노이즈가 주입되기 전인 병렬 말뭉치 생성 단계에서 생성된 병렬 문장 쌍들 중 적어도 일부를 포함할 수 있다.

[0058] 또한, 학습 데이터를 이루는 클린 데이터와 노이즈 데이터는 미리 정해진 비율을 가질 수 있다. 예컨대, 클린 데이터와 노이즈 데이터의 비율은 5:5 내지 1:9의 범위를 가질 수 있다.

[0059] 생성된 학습 데이터를 이용하여 소정의 학습 모델을 학습함으로써 맞춤법 교정 모델이 생성될 수 있다(S150). 학습 모델은 임의의 인공 신경망 기반의 모델을 의미할 수 있으며, 실시예에 따라 복수의 신경망 모델이 결합된 형태를 가질 수 있다. 또한, 학습 모델은 바닐라 트랜스포머와 같은 사전 학습된 언어 모델일 수도 있다.

[0060] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 컨트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0061] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

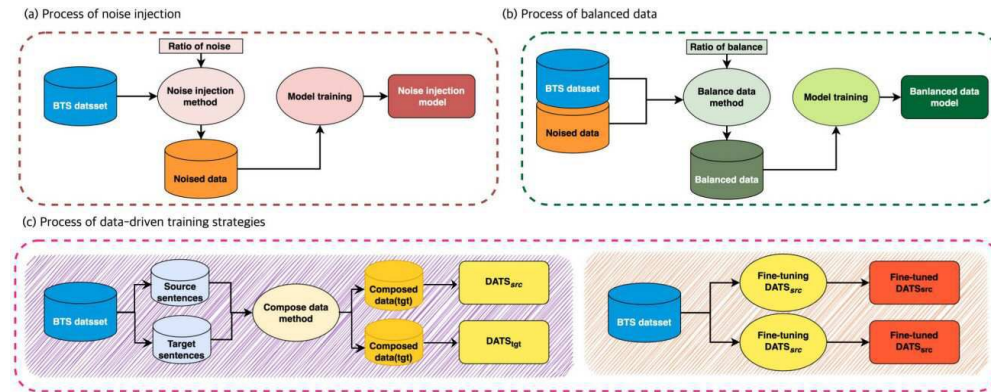
[0062] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0063] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나

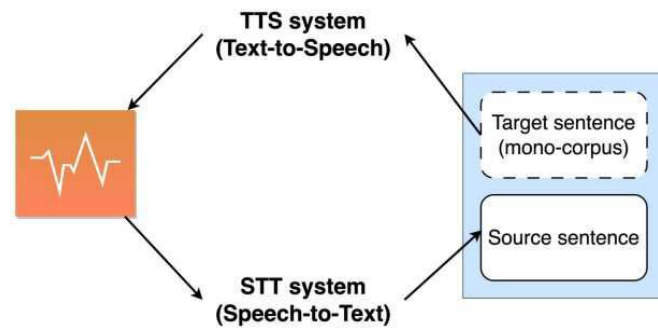
치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

## 도면

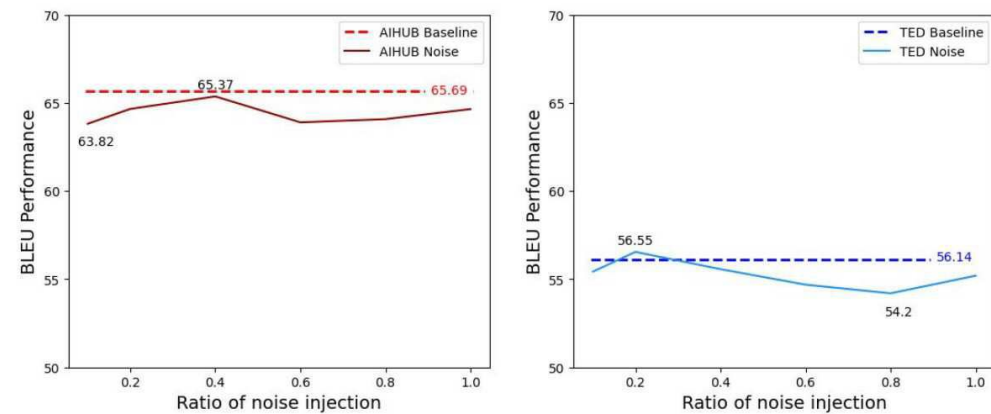
### 도면1



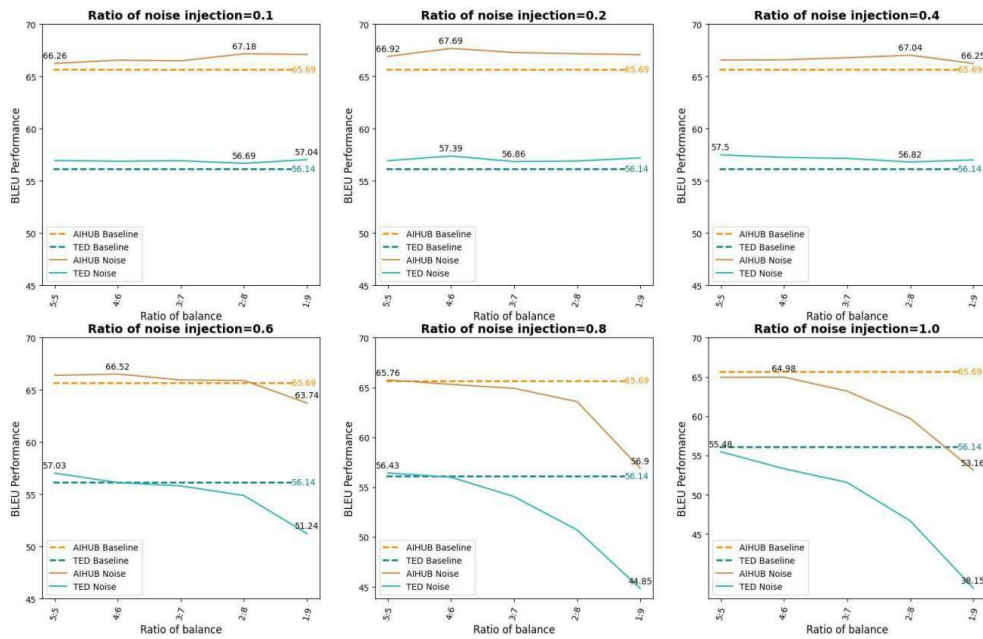
### 도면2



### 도면3



도면4



도면5

