
인공지능 학습용 데이터 구축 사업

- 인터페이스(자판/음성)별 고빈도 오류 교정 데이터-

(주) 유핏 컨소시엄

I 사업개요

1. 추진배경

가. 추진배경 및 필요성

<p>VOICE_01 >>></p> <p>Benchmark Dataset 필요</p> <ul style="list-style-type: none">한국어 맞춤법 교정 임무의 성능 평가 기준 데이터셋 필요새로운 언어 교정 모델, 기존 교정 모델의 새 버전 등이 개발된다면 기존 모델과 비교를 통해 성능 검증과 개발 기준으로 활용	<p>VOICE_02 >>></p> <p>한국어 맞춤법 교정 관련 연구의 부재</p> <ul style="list-style-type: none">한국어 오류에 대한 효율성있는 분류 체계 적립한국어 문법적 오류에 대한 공통 연구 촉진한국어 문법적 오류 연구에 대한 기반 데이터 확보
<p>VOICE_03 >>></p> <p>다양한 인터페이스에 사용 맞는 교정 도구</p> <ul style="list-style-type: none">각 인터페이스별 고빈도 오류 통계 확보오류 통계에 따른 오탈자 교정 도구 필요	<p>VOICE_04 >>></p> <p>AI 알고리즘 맞춤법 교정 모델 제작 및 활용</p> <ul style="list-style-type: none">AI 알고리즘을 활용한 맞춤법 교정 모델 개발디지털/비대면 교육 확대에 따른 맞춤법 교정 기술 도입한국어 학습 중인 외국인이 늘어나면서 외국인을 대상으로 한국어 맞춤법 교육 서비스 고도화 필요

1) 인터페이스(자판/음성)별 고빈도 오류 교정 데이터

○ 언어 교정 모델의 성능 기준 데이터셋(Benchmark Dataset) 필요

- 한국어 맞춤법 교정 모델은 많은 연구가 진행되지 못한 분야로 일부 부산대 맞춤법 검사 API가 있지만, 관련 분야를 Benchmarking 하기에는 부족하며, 이러한 언어 교정 성능을 검증할 기준 데이터셋이 요구됨
- 새로운 언어 교정 모델이 개발되거나 기존 교정 모델이 향상 된다면 기존 교정 모델과의 비교를 통해 성능 검증과 개발 기준에 중요한 참고자료로 활용할 수 있음

○ 한국어 맞춤법 교정 관련 연구의 부재

- 한국어 맞춤법 교정 임무는 많은 연구가 진행되지 못한 분야로 일부 연구가 진행되고 있지만 공통 연구는 진행되고 있지 않아 해당 연구를 공통 연구 분야로 확장시킬 필요가 있음
- 연구가 부족한 만큼 데이터셋의 카테고리 분류의 연구도 부족한 상태이며, 이 분야에 있어서 영어 문법적 교정 임무(English Grammatical Error Correction)를 다룬 CoNLL-2014 논문(The CoNLL-2014 Shared Task on Grammatical

Error Correction)에서는 28종류의 문법적 오류를 정의하여 데이터셋을 분류하고 성능 측정을 하고 있음

- 해당 데이터셋은 입력 인터페이스에 따른 오류에 대해 적절한 맞춤법으로 교정하기 위한 데이터 구축을 위해 필요함
 - PC, 모바일, 천지인 키보드, 음성 입력을 포함한 입력 인터페이스에서 특징적으로 나타나는 오타자 및 띄어쓰기 등 오류를 탐지하고 적절한 맞춤법으로 교정하기 위한 데이터 구축
- AI 언어 모델을 활용한 맞춤법 교정 모델 제작 기반 데이터 필요
 - AI 언어 모델 중 뛰어난 성능을 보인 BERT, Transformer 등을 활용하여 통계 기반의 맞춤법 교정 모델 제작의 기반이 될 수 있음
- AI 언어 모델을 활용한 맞춤법 교정 모델 제작 및 활용
 - 개발된 한국어 교정 모델을 통해 청소년 문법 교육 및 외국인의 한국어 교육에 인공지능 활용의 신호탄이 될 것을 기대

나. 추진목적 및 기대효과

1) 추진목적 및 활용대상

본 과제 수행 단계	다음 단계	세분화 단계	다음 단계	고도화 단계
- 한국어 오류 분류 체계 정의 - 공통 연구 기반이 되는 생태계 조성 - 한국어 문법적 오류 교정 임무의 실험 및 연구 - 인터페이스별 통계기반 오타자 교정 AI 모델 개발		- 오타자, 맞춤법 오류의 세분화 연구 - 문법적 오류 분류 체계 세분화 정의 - 문법적 오류 문제 세분화 정의 - 한국어 문법적 오류 교정 임무의 새로운 실험 및 연구 이론 등장		- 고도화된 오타자 교정 이론, 모델 등장 - 세분화 된 맞춤법 오류에 대한 순차적 해결 방안 등장 - 고도화된 한국어 문법적 오류 교정 모델 등장

본 컨소시엄에서 제안하는 연구 청사진

- 한국어 오류 분야 연구 촉진
 - 본 사업에서 진행하는 실험 결과 및 연구 자료를 공유, 한국어 문법적 오류 교정 연구에 있어 중요한 참고 자료로 활용
 - 한국어 오류 분류에 대한 새로운 정의 제시 및 세분화 연구에 활용
- 한국어 맞춤법 교정 임무에 대한 성능 평가 데이터셋 확보 및 고도화
 - 현재 존재하는 한국어 맞춤법 검사기 혹은 개발 중인 맞춤법 검사기의 성능을 확인하고 비교할 수 있는 기준 데이터셋(Benchmark Dataset)으로 활용
 - 한국어 오류 데이터 품질 기준 연구에 활용

○ AI 언어 모델을 활용한 맞춤법 교정 모델 제작 기반 데이터 구축 및 서비스 활용

논술 준비 과정의 어려운 점



오탈자, 맞춤법 교정에 대한 통계(출처: 메가스터디교육 내부 통계자료)

- 위 통계자료를 보면 논술 준비과정에서 약 50% 학생들이 맞춤법 어려움을 겪고있음
- 맞춤법 교정 모델을 이용하면 맞춤법이 부족한 청소년, 한국어를 학습 중인 외국인이 한국어 맞춤법 연습 및 작문 시 교정을 해줄 수 있는 언어 교정 어플리케이션 개발용 데이터 구축

2) 기대효과

○ 기술적 관점

- 한국어 문법적 오류 교정 임무에 대한 공통 연구가 활발히 진행될 것을 기대
- 한국어 언어 모델에 대한 연구가 다양한 분야에서 진행될 것을 기대
- 한국어의 오류 분류에 대한 새로운 제안 및 연구가 될 것을 기대

○ 산업적 관점

- 본 과제를 통해 개발한 기술과 데이터셋을 통해 학생들이 글을 가장 많이 작성하는 선생님 질문 게시판, QUBE 문제 질문 어플리케이션에 맞춤법 교정 기능을 추가하여 자연스러운 맞춤법 교육 효과를 볼 계획
- 출판업, 뉴스 기사 등 문법적 오류, 맞춤법 오류 교정이 필요한 분야 활용

○ 사회적 관점

- 사회적 문제가 되고 있는 청소년 맞춤법 교육 부실에 대한 대응책이 본 과제를 통해 생성된 데이터와 AI 기술로 연구 개발 될 것을 기대
- 한류에 의해서 한국어 학습을 시작하는 외국인, 문법에 어려움을 겪는 외국인을 대상으로 한국어 교육, 맞춤법 연습 서비스가 개발 될 것을 기대

2. 과제 개요

가. 과제 개요

1) 기본개념

○ 오류 분류 체계 정의

- 수집시점과 인터페이스에 따른 오탈자, 맞춤법 오류, 음성오류 3가지 분류로 정의

오류 구분	정의	예시
오탈자	입력시점에 발생하는 실수 -누르고자 하는 키의 근접한 키를 잘못 누르는 것 -같은 키가 한번 더 눌리는 것 -눌러야할 키가 눌리지 못한 것 등을 포함한 오류	"커피거 맛이 좋더ㅏ."
맞춤법 오류	작성자가 작성을 완료한 후 재차 확인하였음에도 인지하지 못한 오류	"어른이 돼고나면"
음성오류	음성 인식 인터페이스를 통한 입력을 통해 입력된 문장에서 발생한 음성 발화 인식 오류 - 음성인식 모델: 구글, 네이버 등의 범용 모델과 메가스터디교육의 자동 자막 생성에 사용되는 특정 용도 모델	발화문장: 20일이 남아가지고 음성인식 : 29일이 남아가지구

- 본 사업은 디지털 뉴딜 사업의 핵심인 디지털담 구축사업의 일환으로, 본 사업을 통해 인공지능 학습을 위한 최적의 한국어 오탈자 포함 문장 데이터셋 구축 및 이를 공개하는 것을 최종 목표로 함
- [오탈자 데이터]** 다양한 크라우드 워커를 통해 읽으며 입력하기, 들으며 입력하기를 통해 다양한 입력 환경에서 발생한 오류 문장 데이터
- [맞춤법 오류 데이터]** 공개된 SNS 데이터와 메가스터디교육에서 제공하는 서비스 중 질의 응답 게시판, 자유 게시판 등에 등록된 글 중 메타 맞춤법 오류를 포함한 데이터
- [음성 오류 데이터]** 범용 음성인식모델(구글, 네이버) 및 메가스터디교육에서 제공하는 자동 자막 서비스(STT)에 보유 중인 강의 음성 및 다양한 음성 데이터 등을 인식시켜 발생한 발화 인식 오류 문장 데이터
- [자동 생성 오류 데이터]** 오류 자동 생성 알고리즘을 제작하여 뉴스 기사, 대화체 등의 기구축 데이터에 적용하여 생성한 오류 문장 데이터
- [띄어쓰기/문장부호 오류 데이터]** 맞춤법 오류 중 띄어쓰기 오류, 문장부호 오류를 포함한 문장 데이터
- [자주 틀리는 맞춤법 오류 데이터]** 국립국어원, 맞춤법 교재 등에서 제시된 자주 틀리는 맞춤법 오류를 포함한 문장 데이터
-

2) 구성내용

- 인터페이스(자판/음성)별 고빈도 오류 교정 데이터

○ 데이터셋 구성 (인터페이스(자판/음성)별 고빈도 오류 교정 데이터)

- **[오탈자 데이터]** : 초중고 학생 및 20대~40대, 40대 이상을 포함한 다양한 연령대에서 PC 또는 모바일 등의 각기 다른 입력 방법으로 입력한 문장 중 오탈자를 포함한 문장 약 27만 개 이상을 원천 데이터로 구축하고, 이 중 2개 이상의 오류를 포함하고 교정 문장을 포함시킨 데이터 11만 개 이상을 가공 데이터로 구축, 공개하는 것을 목표로 함
- **[맞춤법 오류 데이터]** : SNS와 메가스터디교육에서 서비스 중인 질문 게시판, 자유 게시판 등에 등록된 글 중 띄어쓰기, 문장부호 오류를 제외한 오류를 포함한 문장 15만 개 이상을 원천 데이터로 구축하고, 이 중 2개 이상의 오류를 포함하고 교정 문장을 포함시킨 데이터 6만 개 이상을 가공 데이터로 구축, 공개하는 것을 목표로 함
- **[음성 오류 데이터]** : 메가스터디교육이 보유한 강의 및 서비스 상담, 설명회 영상 등을 활용하여 구글 STT, 네이버 STT, 메가스터디교육 STT 모델로 인식 시킨 문자 데이터 중 오류를 포함한 문장 7.5만 개를 원천 데이터로 구축하고, 이 중 2개 이상의 오류를 포함하고 교정 문장을 포함시킨 데이터 3만 개 이상을 가공 데이터로 구축, 공개하는 것을 목표로 함
- **[자동 생성 오류 데이터]** 오류 자동 생성 알고리즘을 제작하여 뉴스 기사, 대화체 등의 기구축 데이터에 적용하여 생성한 오류 문장 데이터로 수집 목표 20만 개 외에 추가로 구성하여 구축, 공개하는 것을 목표로 함
- **[띄어쓰기/문장부호 오류 데이터]** : 맞춤법 오류 데이터 중 띄어쓰기 오류, 문장부호 오류를 포함한 문장을 수집하며, 해당 데이터는 수집 목표 20만 개 외에 추가로 구성하여 구축, 공개하는 것을 목표로 함
- **[자주 틀리는 맞춤법 오류 데이터]** : 국립국어원, 맞춤법 교재 등에 제시된 자주 틀리는 맞춤법 오류를 포함한 문장 500개를 한국어 전문가를 통해 예문을 만들고 해당 교정 문장을 추가하여 구축, 공개하는 것을 목표로 함

구분	내용
수집 데이터	<ul style="list-style-type: none"> ▪ 오탈자 데이터 : 20대~40대, 40대 이상을 포함한 다양한 연령대의 클라우드 워커가 PC, 모바일, 천지인 키보드를 사용하여 빠르게 입력한 문장 중 오류가 존재하는 문장 ▪ 맞춤법 오류 데이터 : SNS, 질문 게시판, 자유 게시판에 작성한 게시 글 중 오류 문장 ▪ 음성 오류 데이터 : 음성 인식 서비스에 음성을 인식시켜 획득한 문장 중 오류가 발생한 문장
대상 데이터	<ul style="list-style-type: none"> ▪ 오탈자 데이터 : 20대~40대를 포함한 다양한 연령대의 클라우드 워커가 PC, 모바일, 천지인 키보드를 사용하여 입력한 문장 중 오류가 2개 이상 포함되고 편향성 없이 각 오류가 고르게 분포한 데이터 ▪ 맞춤법 오류 데이터 : 초중고 학생 및 성인이 직접 질문게시판에 작성한

	<p>게시 글에 포함된 문장 중 2개 이상의 오류가 포함되고 편향성 없이 각 오류가 고르게 분포한 데이터</p> <ul style="list-style-type: none"> ▪ 음성 오류 데이터 : 범용 음성인식모델(구글, 네이버) 및 메가스터디교육에서 제공하는 자동 자막 서비스(STT)에 보유 중인 강의 음성, 뉴스 음성 등을 인식시켜 발생된 발화 인식 오류 텍스트 중 문장으로 구성되어 있으며 입력 오류가 2개 이상 발생한 문장 ▪ 자동 생성 오류 데이터 : 기구축 데이터에 오류 생성 알고리즘을 적용하여 생성된 오류 문장 데이터 ▪ 띄어쓰기/문장부호 오류 데이터 : 맞춤법 오류 문장 중 띄어쓰기, 문장부호 오류에 해당하는 문장 ▪ 자주 틀리는 맞춤법 오류 : 한국어 전문가가 자주 틀리는 맞춤법에 맞춰 제작한 문장
데이터 구조	<ul style="list-style-type: none"> ▪ 원천 데이터 : 오류 문장으로 구성 ▪ 가공 데이터 : 오류 문장과 교정된 문장(라벨링 데이터)을 1쌍으로 구성
데이터 정의	<ul style="list-style-type: none"> ▪ 데이터 수집에 따른 구분 정의 <ul style="list-style-type: none"> - 오타자 : 순간적인 입력 실수로 생기는 오류 - 맞춤법 오류 : 작성자가 재차 확인하였음에도 인지하지 못한 맞춤법 오류 - 음성 오류 : STT모델이 음성 인식을 할 때에 발생하는 인식 오류 - 자동 생성 오류 : 오류 생성 알고리즘을 적용하여 생성된 오류 - 띄어쓰기/문장부호 오류 : 띄어쓰기 오류, 문장부호 오류 - 자주 틀리는 맞춤법 오류 : 국립국어원에 제시된 오류 ▪ 오타자 데이터의 상세 분류 구분 <ul style="list-style-type: none"> - 연령에 따라 초/중등, 고등, 20~40대, 40~대 분류 - 입력매체에 따라 PC, 모바일 분류 - 키보드에 따라 쿼티, 천지인 분류 - 성별에 따라 분류 - 입력 방식에 따라 보고 입력, 듣고 입력으로 분류 ▪ 맞춤법 오류 데이터의 수집 출처에 따른 분류 구분 <ul style="list-style-type: none"> - SNS, 질문 게시판, 자유 주제 게시판으로 분류 ▪ 입력 매체에 따른 구분 정의 <ul style="list-style-type: none"> - PC : QWERTY 키보드를 사용한 PC에서의 입력 - 모바일 : QWERTY 키보드를 사용한 모바일 환경에서의 입력 - 천지인 : 천지인 키보드를 사용한 모바일 환경에서의 입력 - ASR(음성 입력) : 범용 음성인식모델(구글, 네이버) 및 메가스터디교육에서 제공하는 자동 자막 서비스(STT)에 보유 중인 강의 및 서비스 상담, 설명회 영상 등을 입력
데이터 포맷	<ul style="list-style-type: none"> ▪ JSON
수집 데이터 양	<ul style="list-style-type: none"> ▪ 원천 데이터 : 500,000 문장(100%) <ul style="list-style-type: none"> - 오타자 : 275,000 문장(55%) - 맞춤법 오류 : 150,000 문장(30%) - 음성 오류 : 75,000 문장(15%) ▪ 가공 데이터 : 290,500 문장(100%) <ul style="list-style-type: none"> - 오타자 : 110,000 문장(55%) - 맞춤법 오류 : 60,000 문장(30%) - 음성 오류 : 30,000 문장(15%) - 자동 생성 오류 : 30,000 문장(별도) - 띄어쓰기/문장부호 오류 : 60,000 문장(별도) - 자주 틀리는 맞춤법 오류 : 500 문장(별도)
체크사항	<ul style="list-style-type: none"> ▪ 원천 데이터는 오류를 포함하고 있는 문장이며, 가공 데이터는 원천 데이터 중 오류가 2개 이상 포함되고 라벨링한 데이터를 의미 ▪ 자동 생성 오류, 띄어쓰기와 문장부호 오류, 자주 틀리는 맞춤법 오류 데이터는 별도 수집

나. 국가 AI정책과의 연관성 및 중요성

1) 국가 AI 정책

○ 디지털 뉴딜(20.06)

- 한국형 뉴딜은 경제, 사회 전반의 국가적 도약을 위한 미래정책으로, 정보통신(ICT) 산업 기반 '디지털 뉴딜'과 친환경, 에너지 산업 기반 '그린 뉴딜'을 두 축으로 진행됨
- 디지털 뉴딜은 D.N.A(Data, Network, AI) 생태계 강화, 교육인프라 디지털 전환, 비대면 산업 육성, 사회간접자본(SOC) 디지털화 등을 추진함
- 이 중 데이터 전(全)주기 생태계 강화 및 데이터 컨트롤타워 마련의 일환으로 공공데이터 개방, 인공지능 학습용 데이터 구축 등의 사업을 진행

한국형 뉴딜	
디지털 뉴딜	그린 뉴딜
1) D.N.A. 생태계 강화 ① 국민생활과 밀접한 분야의 데이터 구축·개방·활용 ② 5G 국가망 확산 및 클라우드 전환 ③ 1·2·3차 산업 5G-AI 융합 확산 ④ AI·SW 핵심인재 10만명 양성	1) 도시·공간·생활 인프라 녹색 전환 ① 국민생활과 밀접한 공공시설의 제로에너지화 전면 전환 ② 스마트 그린도시 조성을 위한 선도프로젝트 100개 추진 ③ 취수원부터 가정까지 ICT 기반 스마트 상수도 관리체계 구축
2) 디지털 포용 및 안전망 구축 ① 농어촌 초고속 인터넷망 및 공공시설 WiFi 구축 ② K-사이버 보안체계 구축	2) 녹색산업 혁신 생태계 구축 ① 그린뉴딜 선도 100대 유망기업 및 5대 선도 녹색산업 육성 ② 주력 제조업 녹색전환을 위한 저탄소·녹색산단 조성
3) 비대면 산업 육성 ① 모든 초중고에 디지털 기반 교육 인프라 구축 ② 전국 대학 및 직업훈련기관 온라인 교육 강화 ③ 감염병 안심 비대면인프라 및 건강취약계층 디지털 돌봄 구축 ④ 중소기업 16만개 대상 원격근무 인프라 보급	3) 저탄소산업 에너지 확산 ① 에너지관리 효율화 지능형 스마트 그리드 구축 ② 태양광·풍력·수소 등 3대 신재생에너지 확산 기반 구축 ③ 온실가스 저장효과가 큰 친환경 차량·선박으로 조기 전환
4) SOC 디지털화 ① 4대 핵심시설 디지털 관리체계 구축 ② 도시·산단 디지털 혁신 및 스마트 물류 체계 구축	

* 자료: 과학기술정보통신부, 디지털 뉴딜 로드맵(2020-2025), 2021.02.22.

○ 과제 추진방향과의 연관성

- 본 사업은 디지털 뉴딜 중 데이터 댐 구축사업의 일환으로 데이터 전(全)주기 및 D.N.A 생태계 강화 및 일자리 창출 등에 직접적으로 기여함
- 본 사업을 통해 한국어 오류 교정에 대한 기준을 확립하고, 성능 평가의 기준이 되는 데이터셋(Benchmark Dataset)을 제공
- 본 사업을 통해 한국어에 특화된 오탃자 및 문법 오류 문장 데이터를 구축함에 따라 한국어 자연어 처리 분야에서의 응용 기술 개발에 기여함

○ 인공지능(AI) 국가 전략(19.12)

- '인공지능 국가 전략'은 인공지능이 단순한 신기술이 아닌 각국 경제 성장에 비약적인 파급효과를 불러일으키는 경제사회 대변혁의 핵심 동력인 바, AI 기술력 확보의 시급성과 4차 산업혁명 대응계획(I-Korea 4.0)의 차질 없는 실현을 위해 마련



* 자료: 과학기술정보통신부, AI 국가전략 보도자료, 2019.12.17.

○ 과제 추진방향과의 연관성

- 글로벌 AI 시장에서의 영어의 자연어 처리 연구와 발맞추어 한국어의 연구를 위하고자 함
- 추진전략 1.3) 과감한 규제 혁신 및 법제도 준비를 통해 AI 분야의 규제패러다임을 포괄적 네거티브로 전환하고, AI 시대의 미래지향적인 법제도를 마련하고자 함

3. 관련 동향

■ Grammatical Error Correction(GEC : 문법적 오류 수정)

- 기술 개요: 키보드, 모바일 입력을 통해 단어 입력이나 작문 등을 할 때 생기는 오타자, 맞춤법 오류, 문법적 오류를 통계 기반, 규칙 기반으로 교정

1) 국내 동향

- 한국어 맞춤법 검사기는 부산대 맞춤법 검사기가 있으며, 해당 맞춤법 검사기는 1990년도부터 연구가 시작되어 현재 버전에 이르기까지 많은 연구가 이루어졌지만 관련된 자료, 벤치마크 데이터셋(Benchmark Dataset) 등의 연구 자료가 공유되지 않은 실정, 현재 한국어 맞춤법 검사기로써는 유일무이한 모델
- 카카오에서 맞춤법 검사 API를 오픈한 적이 있으나 현재는 서비스 되고 있지 않으며, 네이버 또한 맞춤법 검사 API를 오픈했었으나 현재는 검색을 통한 베타버전 사용이 전부이며 공식적인 API는 미제공 상태
- 한국어의 맞춤법이나 문법은 언어 중에서는 어려운 편에 속하며, 관련 연구 또한 영어에 비해 부족한 상태로, 이와 같은 흐름은 관련 데이터 수집의 어려움, 언어에 대한 전문성 필요 등의 이유가 있음
- 한국어 문법적 오류 수정과 관련된 **벤치마크 데이터셋(Benchmark Dataset)의 부재가 관련 연구 수행에 어려움으로 지적되고 있음**

2) 국외 동향

- 영어에서는 Grammatical Error Correction(GEC) Task로 연구가 진행되고 있으며 관련 논문과 벤치마크 데이터가 많은 상황이며, 품질 지표 연구를 통해 F0.5 Measure 지표를 사용
- 2019년도 BEA-2019에서 Grammatical Error Correction Task의 Restricted Track과 Low Resource Track에서 카카오 브레인이 각 2위를 기록
- 다른 분야도 마찬가지겠지만, Grammatical Error Correction 분야는 특이나 기준 데이터나 연구가 없기 때문에 본 데이터 구축은 한국어 오류 교정 연구에 신호탄이 될 것을 기대

II 과제 수행 계획

1. 인공지능 학습용 데이터 구축

가. 구축 배경 및 목적

인터넷 이용자수(이용률)
만 3세 이상 국민 5,098만명 중 91% 메시지 이용



연령별 인터넷 이용률
60대 이상 고령층의 급증



네이버 맞춤법 검사기 *Beta* dnm 맞춤법 검사기

원문	교정 결과	검사하기날. ✖
감사하바니다.	감사하나니다.	감사하비날.
감사합니다나	감사합니다나	감삼합니다. ✖
	감사합니다나	감삼합니다.

오타자 맞춤법 교정의 한계
네이버, 다음, 부산대 맞춤법 검사기



발달 장애, 시각 장애
사회적 약자와 함께 공존하는 DT

○ 언어 교정 모델의 성능 기준 데이터셋(Benchmark Dataset) 필요

- 새로운 언어 교정 모델이 개발되거나 기존 교정 모델이 향상 된다면 기존 교정 모델과의 비교를 통해 성능 검증과 개발 기준에 중요한 참고자료로 활용할 수 있음

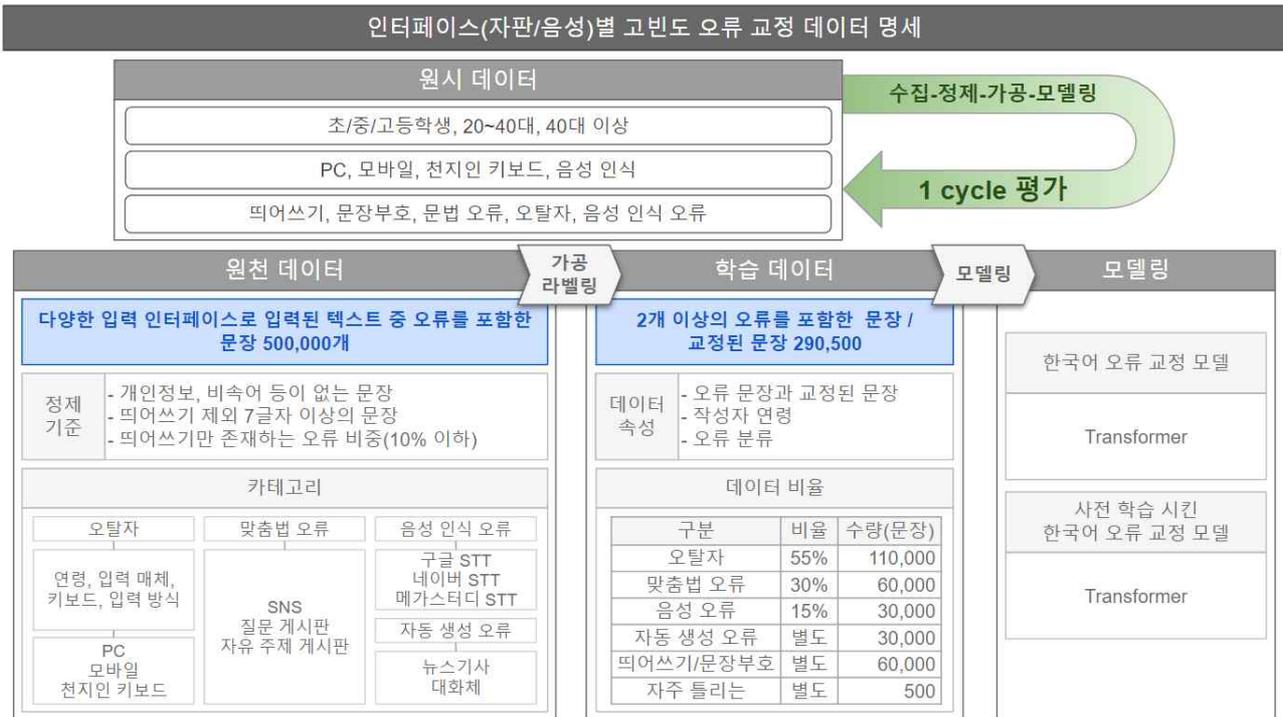
○ 한국어 맞춤법 교정 관련 공통 연구의 목적

- 문법적 오류 교정(Grammatical Error Correction) 임무에 있어서 영어는 많은 연구와 연구대회 등이 있는 반면 한국어는 모든 연구가 폐쇄적으로 진행
- 본 과제를 통해 한국어 맞춤법 오류 교정 임무를 공통 연구 분야로 확장시켜 연구 속도를 높이는 것을 목적으로 두고 있음

○ 한국어 맞춤법 교정 모델의 기반 데이터 필요

- 한국어 맞춤법 오류 교정 서비스 제작, 모델 제작에 활용 될 기반 데이터 필요

나. 구축데이터 명세



다. 인공지능 학습용 데이터 구축 환경

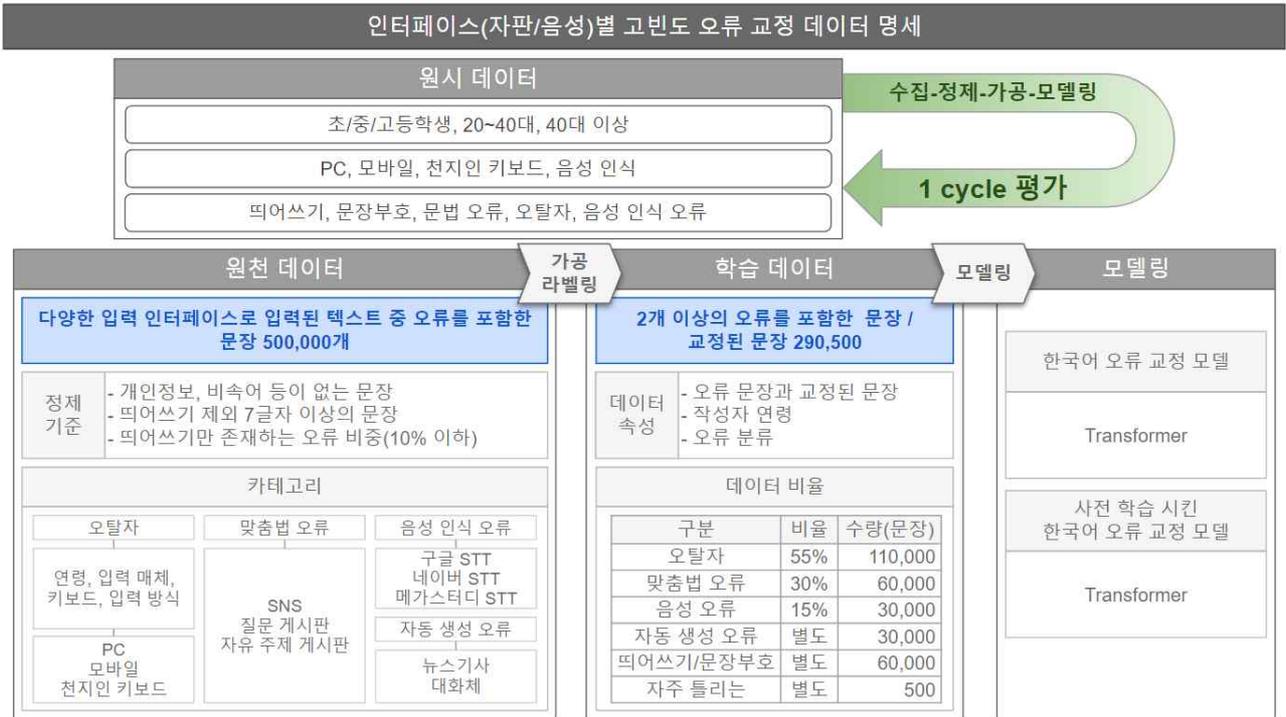
○ 수집 환경

- 본 컨소시엄에서는 인터페이스별 고빈도 오류 데이터를 데이터 성격에 따라 오타자, 맞춤법 오류, 음성 오류, 띄어쓰기/문장부호 오류, 자주 틀리는 맞춤법 오류 5가지로 정의하여 분류했으며, 각 특징을 살려 수집계획을 수립하였음
- **[오타자 데이터]** : 오타자 수집은 입력한 순간 오류가 발생하면 수집을 해야 하기 때문에 직접 현장에서 저작도구를 통해 주어진 문장을 입력하도록 하여 수집할 계획
- **[맞춤법 오류 데이터]** : 맞춤법 오류 데이터 수집은 SNS, 질문 게시판, 자유 주제 게시판에서 수집한 데이터를 워커들이 원활하게 작업하도록 맞춤법 교정 오픈 소스인 Hunspell을 활용해 텍스트 데이터의 오류 부분을 강조 표시한 뒤 해당 전체 텍스트를 워커들이 확인하여 오류를 포함한 부분을 문장 단위로 수집
- **[음성 오류 데이터]** : 메가스터디교육이 보유한 강의 및 서비스 상담, 설명회 영상 등을 활용하여 구글 STT, 네이버 STT, 메가스터디교육 STT 모델로 인식시킨 문자 데이터 중 오류를 포함한 데이터를 수집할 계획
- **[자동 생성 오류 데이터]** : 오류 자동 생성 알고리즘을 제작하여 뉴스 기사, 대화체 등의 기구축 데이터에 적용하여 생성한 오류 문장 데이터로 수집 목표 20만 개 외에 추가로 구성하여 구축할 계획
- **[띄어쓰기/문장부호 오류 데이터]** : 맞춤법 오류 데이터 중 띄어쓰기 오류, 문

장부호 오류를 포함한 문장을 수집하며, 목표 수량 외에 추가로 구축할 계획

- [자주 틀리는 맞춤법 오류 데이터] : 국립국어원의 상담 사례 모음 게시판에 게시된 자주 틀리는 맞춤법 오류를 포함한 문장을 한국어 전문가를 통해 예문을 만들고 해당 교정 문장을 추가하여 구축할 계획

라. 인공지능 학습용 데이터 구축 방법

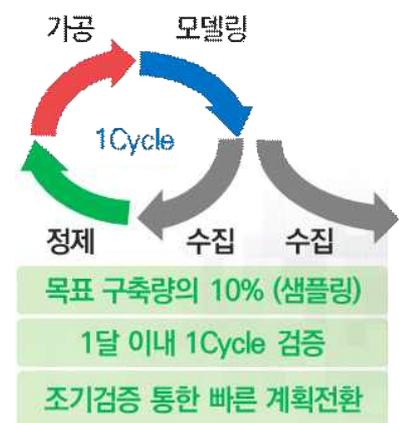


마. 인공지능 학습용 데이터 검사 방법

㉠ 검사조직 및 절차



㉡ 애자일 데이터셋 구축방법론



2. 인공지능 데이터 활용 모델 개발

■ 인공지능 데이터 활용 모델 개발 요약표

데이터명	AI 모델	모델 품질 지표	기준 지표	응용서비스(예시)
인터페이스(자판/음성)별 고빈도 오류 교정 데이터	한국어 오류 교정 모델 [Transformer]	F0.5 Measure	50	맞춤법 검사 서비스, 오타자 교정 서비스
	한국어로 사전 학습시킨 한국어 오류 교정 모델 [Transformer]	F0.5 Measure	50	

■ 인터페이스(자판/음성)별 고빈도 오류 교정 데이터

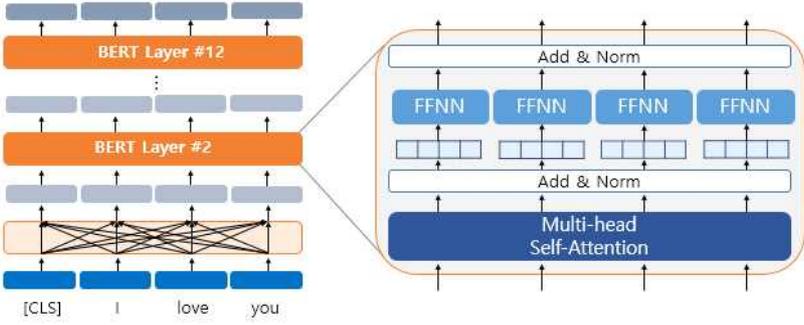
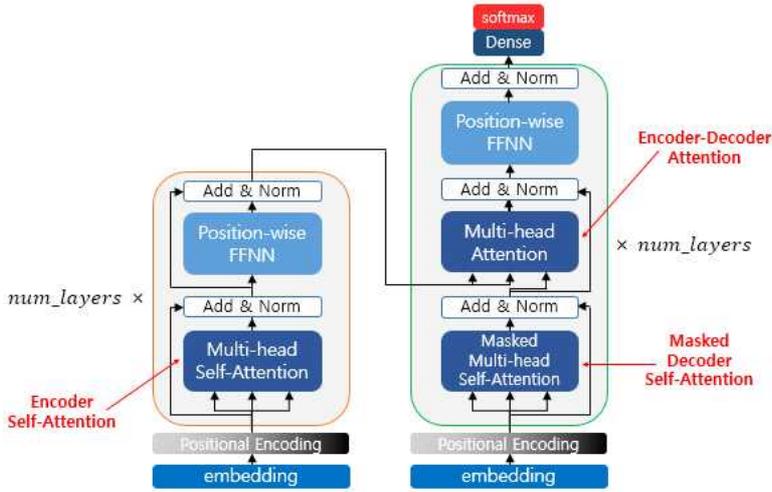
가. 인공지능 학습용 데이터 활용모델 개발 방안

1) 모델/알고리즘의 연구개발 방안 및 개발 목표

○ 인공지능 모델 알고리즘 방법론

- Grammatical Error Correction(GEC: 문법적 오류 교정) 임무에서 영어는 데이터도 풍부하며 연구된 결과와 성능 지표가 다수 있어, 해당 방식을 적용하고자 하지만, 한국어는 영어와 다르게 조합 문자로서 초성, 중성, 종성이 모여 한글자를 이룸
- 한국어 문장에서 글자를 초성, 중성, 종성으로 분리하여 알파벳과 마찬가지로 문자열로 만들어 학습시킨다면 영어에서처럼 GEC 임무를 해결 가능할 것이라고 판단하여 해당 방법론을 제시
- 자음, 모음으로 이루어진 Tokenizer를 활용할 계획이며, Tokenizer를 새롭게 제작하기 때문에 한국어 사전 학습 모델은 적용 불가능, 그래서 직접 위키피디아 한국어 데이터셋을 활용해 새 Tokenizer를 이용한 한국어 사전 학습 모델을 연구할 계획
- 자세한 연구 및 실험 내용은 구축계획서 2장 5.4 항목에 제시

○ 학습모델 후보군

구분	내 용
후보1	<ul style="list-style-type: none"> ▪ 학습모델: BERT ▪ 개요 <ul style="list-style-type: none"> - 'Grammatical Error Correction' 분야 'BEA-2019 (test)' 데이터셋에서 'Sequence tagging + token-level transformations + two-stage fine-tuning (+BERT, RoBERTa, XLNet)' 모델이 2위를 기록 중 ▪ Architecture  <p>The diagram illustrates the BERT architecture. On the left, a stack of layers is shown, starting with an input layer containing tokens '[CLS]', 'I', 'love', and 'you'. This is followed by a 'Multi-head Self-Attention' layer, then a 'BERT Layer #2', and finally a 'BERT Layer #12'. A callout box on the right provides a detailed view of a single BERT layer. It shows a 'Multi-head Self-Attention' block at the bottom, followed by an 'Add & Norm' layer. Above this is a 'FFNN' (Feed-Forward Network) block, which is composed of four parallel FFNN units. This is followed by another 'Add & Norm' layer. The entire layer structure is repeated for multiple layers.</p>
후보2	<ul style="list-style-type: none"> ▪ 학습모델: Transformer ▪ 개요 <ul style="list-style-type: none"> - 'Grammatical Error Correction' 분야 'CoNLL-2014' 데이터셋에서 'Transformer' 모델이 F0.5 Measure 55.8점을 기록 ▪ Architecture  <p>The diagram shows the Transformer architecture, divided into an encoder and a decoder. The encoder (left) starts with an 'embedding' layer, followed by 'Positional Encoding'. It then consists of $num_layers \times$ layers, each containing a 'Multi-head Self-Attention' block, an 'Add & Norm' layer, and a 'Position-wise FFNN' block, followed by another 'Add & Norm' layer. This is labeled as 'Encoder Self-Attention'. The decoder (right) starts with an 'embedding' layer and 'Positional Encoding'. It consists of $num_layers \times$ layers, each containing a 'Masked Multi-head Self-Attention' block, an 'Add & Norm' layer, a 'Multi-head Attention' block (receiving input from the encoder), an 'Add & Norm' layer, a 'Position-wise FFNN' block, and a final 'Add & Norm' layer. This is labeled as 'Masked Decoder Self-Attention'. The output of the decoder goes through a 'Dense' layer and a 'softmax' layer.</p>

○ 학습모델 품질지표 선정

구분	내 용				
Transformer 오류 문장 교정	품질지표	F0.5 Measure			
	선행연구	'Grammatical Error Correction' 분야 'CoNLL-2014' 데이터셋에서 'Transformer' 모델이 F0.5 Measure 55.8점을 기록 [2018] 출처: https://paperswithcode.com/sota/grammatical-error-correction-on-conll-2014			
	지표기준	<p style="text-align: center;">✓ F0.5 Measure 50 이상 [TTA 품질 지표 기준]</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr style="background-color: #d9d9d9;"> <th style="width: 30%;">품질특성</th> <th>검증 방법</th> </tr> </thead> <tbody> <tr> <td>유효성</td> <td> ① 훈련/검증/평가용 데이터셋 준비 ② 알고리즘을 훈련/검증 데이터셋으로 학습하여 인공지능 모델 준비 ③ 평가용 데이터셋을 인공지능 모델에 입력하여 결과를 기록 </td> </tr> </tbody> </table>	품질특성	검증 방법	유효성
품질특성	검증 방법				
유효성	① 훈련/검증/평가용 데이터셋 준비 ② 알고리즘을 훈련/검증 데이터셋으로 학습하여 인공지능 모델 준비 ③ 평가용 데이터셋을 인공지능 모델에 입력하여 결과를 기록				
Transformer 사전 학습 오류 문장 교정	품질지표	F0.5 Measure			
	선행연구	'Grammatical Error Correction' 분야 'CoNLL-2014' 데이터셋에서 'Transformer' 모델이 F0.5 Measure 55.8점을 기록 [2018] 출처: https://paperswithcode.com/sota/grammatical-error-correction-on-conll-2014			
	지표기준	✓ F0.5 Measure 50 이상			

○ 품질 지표 기준 근거

- 영어 문법적 오류 탐지, 문법적 오류 교정 임무의 선행 연구를 참고하여, 지표 수치를 책정하였으며, 다년간의 연구 노하우가 들어가 있는 점, 한국어와 영어의 언어 구성의 차이가 존재하는 점을 감안하여 책정

임무	Grammatical Error Detection	Grammatical Error Correction																																										
데이터셋	CoNLL-2014 A2	CoNLL-2014 Shared Task																																										
성능 순위표	<table border="1"> <thead> <tr> <th>Rank</th> <th>Model</th> <th>F0.5 ↑</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>VERNet</td> <td>63.1</td> </tr> <tr> <td>2</td> <td>Bi-LSTM + POS (unrestricted data)</td> <td>45.1</td> </tr> <tr> <td>3</td> <td>Bi-LSTM (unrestricted data)</td> <td>44.0</td> </tr> <tr> <td>4</td> <td>Ann+PAT+MT</td> <td>30.13</td> </tr> <tr> <td>5</td> <td>BiLSTM-JOINT (trained on FCE)</td> <td>29.65</td> </tr> <tr> <td>6</td> <td>Bi-LSTM + POS (trained on FCE)</td> <td>26.2</td> </tr> </tbody> </table>	Rank	Model	F0.5 ↑	1	VERNet	63.1	2	Bi-LSTM + POS (unrestricted data)	45.1	3	Bi-LSTM (unrestricted data)	44.0	4	Ann+PAT+MT	30.13	5	BiLSTM-JOINT (trained on FCE)	29.65	6	Bi-LSTM + POS (trained on FCE)	26.2	<table border="1"> <thead> <tr> <th>Rank</th> <th>Model</th> <th>F0.5 ↑</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>Sequence tagging + token-level transformations + two-stage fine-tuning (+BERT, RoBERTa, XLNet)</td> <td>66.5</td> </tr> <tr> <td>2</td> <td>LM-Critic</td> <td>65.8</td> </tr> <tr> <td>3</td> <td>Sequence tagging + token-level transformations + two-stage fine-tuning (+XLNet)</td> <td>65.3</td> </tr> <tr> <td>4</td> <td>Transformer + Pre-train with Pseudo Data (+BERT)</td> <td>65.2</td> </tr> <tr> <td>5</td> <td>Transformer + Pre-train with Pseudo Data</td> <td>65.0</td> </tr> <tr> <td>6</td> <td>VERNet</td> <td>63.7</td> </tr> </tbody> </table>	Rank	Model	F0.5 ↑	1	Sequence tagging + token-level transformations + two-stage fine-tuning (+BERT, RoBERTa, XLNet)	66.5	2	LM-Critic	65.8	3	Sequence tagging + token-level transformations + two-stage fine-tuning (+XLNet)	65.3	4	Transformer + Pre-train with Pseudo Data (+BERT)	65.2	5	Transformer + Pre-train with Pseudo Data	65.0	6	VERNet	63.7
	Rank	Model	F0.5 ↑																																									
	1	VERNet	63.1																																									
	2	Bi-LSTM + POS (unrestricted data)	45.1																																									
	3	Bi-LSTM (unrestricted data)	44.0																																									
	4	Ann+PAT+MT	30.13																																									
5	BiLSTM-JOINT (trained on FCE)	29.65																																										
6	Bi-LSTM + POS (trained on FCE)	26.2																																										
Rank	Model	F0.5 ↑																																										
1	Sequence tagging + token-level transformations + two-stage fine-tuning (+BERT, RoBERTa, XLNet)	66.5																																										
2	LM-Critic	65.8																																										
3	Sequence tagging + token-level transformations + two-stage fine-tuning (+XLNet)	65.3																																										
4	Transformer + Pre-train with Pseudo Data (+BERT)	65.2																																										
5	Transformer + Pre-train with Pseudo Data	65.0																																										
6	VERNet	63.7																																										

출처 : <https://paperswithcode.com/>, 성능 순위표

2) 인공지능 서비스의 개발 및 사업화, 제품화 방안

○ 인공지능 서비스 적용방안

구분	적용내용
응용 서비스	<p>① 청소년을 대상으로한 한국어 맞춤법 교정 서비스</p> <ul style="list-style-type: none"> 맞춤법에 약한 청소년을 대상으로 한국어 맞춤법, 오타자 교정 서비스 제공 <p>② 한국어를 공부 중인 외국인을 대상으로한 문법 교정 서비스</p> <ul style="list-style-type: none"> 작문 후 발생하는 시제, 조사 및 단어, 띄어쓰기를 교정해주는 서비스 제공 <p>③ 출판, 뉴스 등을 대상으로한 맞춤법 오류 체크 서비스</p> <ul style="list-style-type: none"> 맞춤법과 오타자에 대해 고밀도의 분석을 해야 하는 산업을 대상으로 맞춤법 오류, 오타자를 찾아주고 교정까지 해주는 서비스

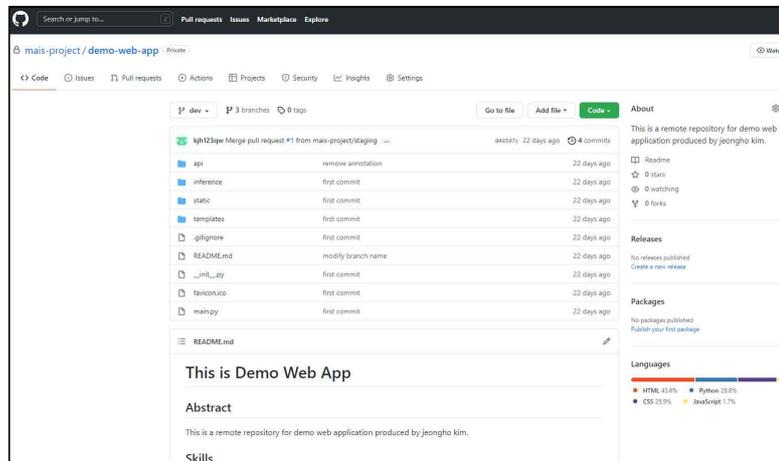
나. 인공지능 모델을 적용한 기술혁신 지원 방안

○ 한국어 오탃자, 문법적 오류 문장 교정 모델의 학습을 위한 기초자료로 활용

- 한국어에 대한 문법적 오류 교정 임무와 관련된 데이터는 매우 부족하기 때문에 본 사업의 결과물을 통해 얻어진 데이터를 문법적 오류 교정 임무의 초기 연구에 활용
- 본 과제에서 진행한 연구 자료를 공유하여 공동 연구 진행을 촉진
- 본 과제에서 개발한 Tokenizer의 연구 결과, 기록 등을 공개하여 한국어 문법적 오류 교정 임무에 사용 될 수 있는 자료로써 활용

○ 연구 결과, 코드 공유

- 본 사업 결과물 중 소스 코드와 알고리즘, 모델 연구 결과, 모델 관련 코드를 Github를 통해 오픈소스로 활용할 수 있도록 공개



○ 맞춤법 교정 데모페이지 공개

- 본 사업에서 수행되는 모델 연구를 통해 확보되는 모델 중 3~4개의 모델을 API로 서빙하여, 웹페이지에서 해당 교정 모델을 테스트하고 비교해 볼 수 있도록 제공할 예정



3. 인공지능 학습용 데이터 품질관리 및 검증

■ 인터페이스(자판/음성)별 고빈도 오류 교정 데이터

가. 품질요구사항

품질 요구사항 구분	구축공정	구축데이터	AI학습모델
품질 요구사항 수	84개	20개	4개

나. 품질 목표

품질관리 영역	품질지표		품질목표	품질 목표 달성 기준
구축공정 품질	준비성		95% 이상	체크리스트 목록*의 95% 이상 '적정'
	완전성		95% 이상	체크리스트 목록의 95% 이상 '적정'
	유용성		95% 이상	체크리스트 목록의 95% 이상 '적정'
구축데이터 품질	적합성	기준적합성	95% 이상	체크리스트 목록의 95% 이상 '적정'
		포괄성	99% 이상	세부항목별 준수율 평균 99% 만족
		통계적 다양성	99% 이상	클래스 분포, 인스턴스 분포 평균 99% 만족
	구문적 정확성	구조 정확성	정확도	99% 이상
		형식 정확성	정확도	99% 이상
	의미적 정확성	오류 유형 분류 정확도	정확도	95% 이상
		오류 문장 교정 정확도	정확도	90% 이상
AI학습모델 품질	알고리즘 적정성		Pass	알고리즘 적정성 과제심의 통과
	유효성	50 이상		Transformer 오류 문장 교정 F0.5 Measure
		50 이상		Transformer 사전 학습 오류 문장 교정 F0.5 Measure

다. 품질관리방안 개요 및 기준

○ 내부(자체) 품질관리 방안

- 고품질 인공지능 학습용 데이터 구축을 위해 적합성, 정확성, 유효성, 준비성, 완전성, 유용성의 6가지* 품질관리 기준을 기반으로 품질 검증 및 관리를 수행하고 체계적이고 효율적인 품질관리를 위해 품질관리 전담 조직 및 전문가 자문단을 구성하여 운영함

* 인공지능 학습용 데이터 품질관리 가이드라인(2022, NIA) 품질관리 지표를 준용하여 본 사업 수행 예정

○ 인공지능 학습용 데이터 구축공정 및 데이터로 구분하여 품질관리 기준 관리

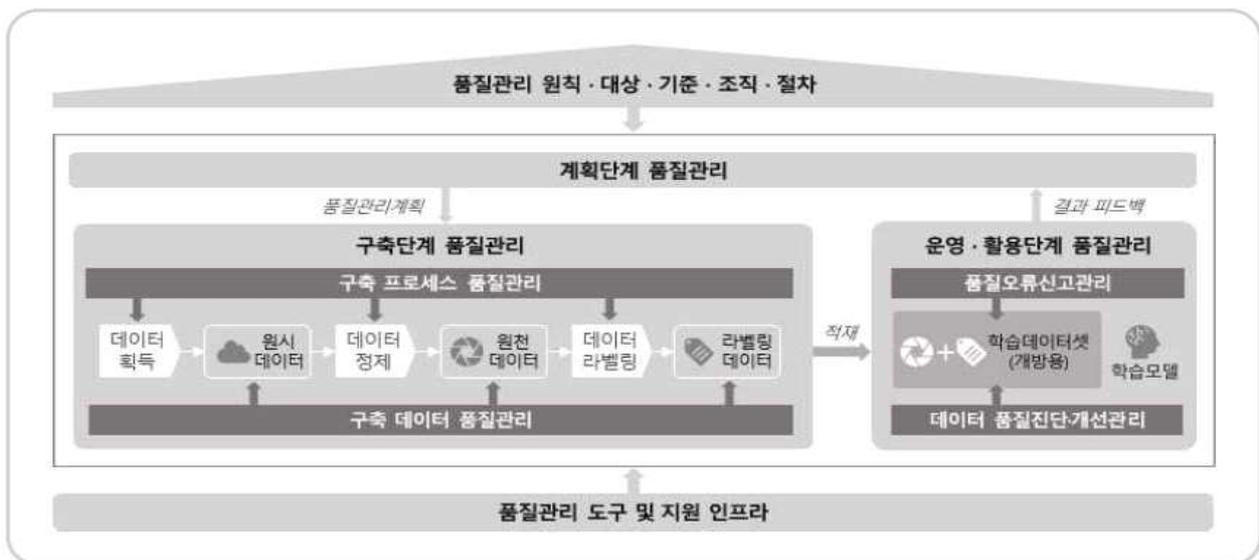
<인공지능 학습용 데이터셋 구축품질관리 지표>

품질관리 영역	품질지표	설명
구축공정 품질	준비성	<ul style="list-style-type: none"> • 인공지능 학습용 데이터 품질관리를 위해 기본적으로 관리해야 하는 정책, 규정(저작권, 초상권, 개인정보보호 및 정보보호 등에 대한 검토 결과를 포함), 조직, 절차 등을 마련하고, 최신의 내용으로 충실하게 관리되는지를 검사하는 지표
	완전성	<ul style="list-style-type: none"> • 인공지능 학습용 데이터를 구축함에 있어 물리적인 구조를 갖추고, 정의한 데이터 형식 및 입력값 범위에 맞게 데이터가 저장되도록 설계·구축 되었는지를 검사하는 지표
	유용성	<ul style="list-style-type: none"> • 발주기관(수요자)의 요구사항이 충분히 반영되었는지, 임무정의에 적합한 인공지능 학습용 데이터의 범위와 상세화 정도를 충족시키는지를 검사하는 지표
구축데이터 품질	적합성	<ul style="list-style-type: none"> • 학습용도 적합성을 측정하는 지표로 <ul style="list-style-type: none"> - (기준적합성) 다양성, 신뢰성, 충분성, 사실성 - (기술적합성) 파일포맷, 해상도, 선명도, 컬러, 크기, 길이, 음질 등통계적 다양성 : 클래스 분포도, 인스턴스 분포도, 문장길이, 어휘개수 등
	정확성	<ul style="list-style-type: none"> • 라벨링 정확성을 측정하는 지표로 <ul style="list-style-type: none"> - (의미 정확성) 정확도, 정밀도, 재현율을 측정하는 지표 - (구문 정확성) 어노테이션 데이터를 구성하는 속성 값들과 원래 정의한 데이터 형식 및 입력 값 범위와의 일치성을 측정하는 지표
	유효성	<ul style="list-style-type: none"> • 학습용 데이터로 훈련시키는데 적합한 인공지능 알고리즘의 유효성을 측정하는 지표

라. 품질관리 절차

○ 품질관리 원칙

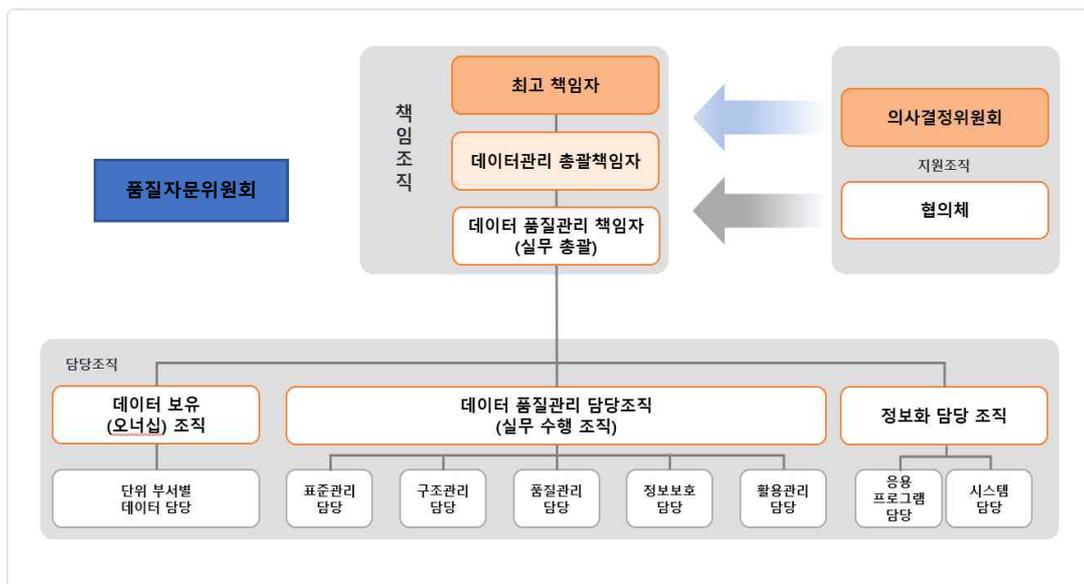
- 인공지능 학습용 데이터 품질관리의 기본 방향과 품질관리 시 참조해야 할 기준과 절차 수립하고, 품질관리 활동을 지원하는 도구나 기술, 인프라도 품질관리 모델에 포함
 - 품질관리 원칙은 데이터 품질관리 데이터 측면과 품질관리 측면에서 총 9개의 원칙을 통한 방향성 제시
 - 품질관리 대상은 AI Hub를 통해 민간에 개방하는 인공지능 학습용 데이터셋을 우선
 - 품질관리 기준은 인공지능 학습용 데이터의 자체적인 품질 및 인공지능 학습용 데이터를 구축하는 과정의 품질을 측정하고 검증하는데 필요한 지표
 - 품질관리 조직은 인공지능 학습용 데이터의 품질확보 및 품질관리 활동을 수행하는 조직
 - 품질관리 절차는 인공지능 학습용 데이터 품질을 검사하고, 원인을 분석해서, 개선을 조치하는 일련의 활동
 - 생애주기별 품질관리 활동은 계획, 구축, 운영, 활용의 각 영역에서 수행해야 할 품질관리 활동을 정의
 - 품질관리 도구 및 지원 인프라는 인공지능 학습용 데이터의 품질검사나 품질관리 활동을 수행하는데 사용하는 도구(Tool)나 기술(Technology), 플랫폼(Platform) 등을 의미(데이터의 값이나 구조, 형식 등의 오류를 찾아내고 개선하는 데 있어 품질검사자를 지원하거나, 자동으로 품질검사 등을 수행하는 기능을 갖춘 시스템이나 솔루션)



[인공지능 학습용 데이터셋 품질관리 모델]

○ 품질관리 조직의 구성

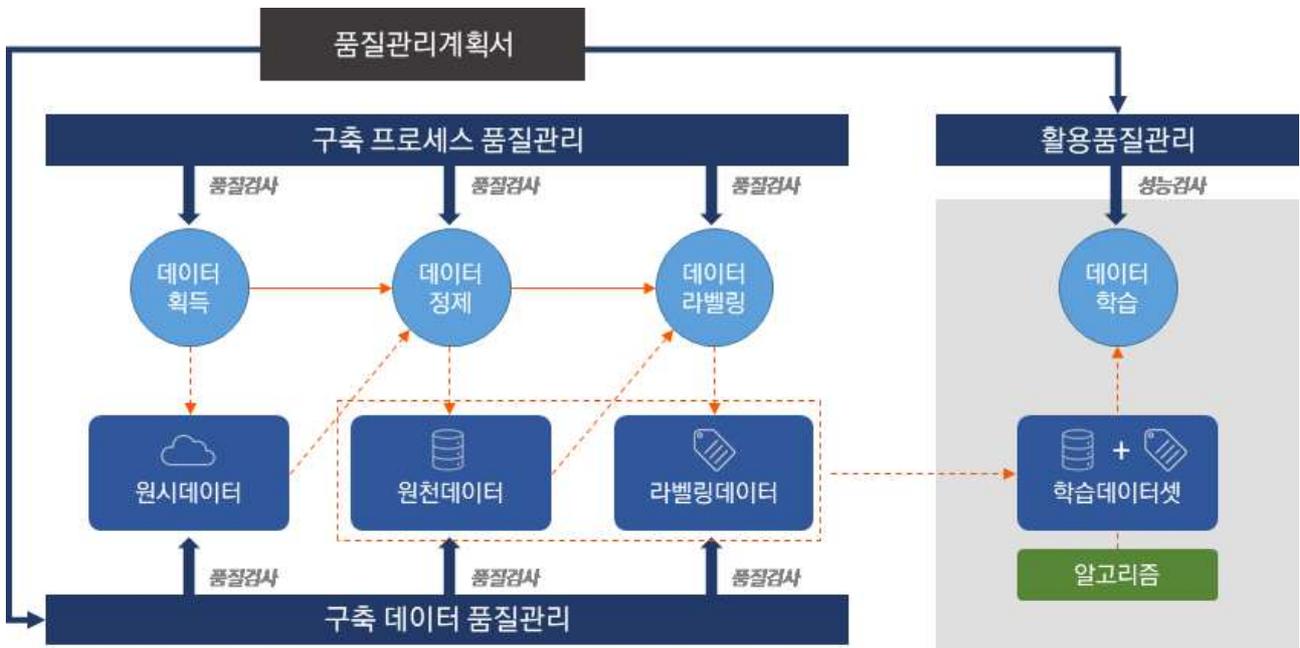
- 인공지능 학습용 데이터의 품질관리를 위해 인공지능 학습용 데이터 구축사업에 참여하는 수행기관 및 참여기관은 품질관리 총괄책임자 및 실무책임자를 지정하여 해당 학습용 데이터의 품질관리 활동을 계획하고, 품질관리 업무 수행을 위한 품질관리 조직을 구성하여 역할과 책임을 부여
- 인공지능 학습용 데이터 구축사업 수행기관 및 참여기관의 품질관리 조직은 학습용 데이터 구축의 목적, 구축과정, 구축하는 데이터의 규모 등에 따라 다양한 형태로 품질관리 조직을 구성



[인공지능 학습용 데이터 구축사업 품질관리체계]

○ 품질관리 체계

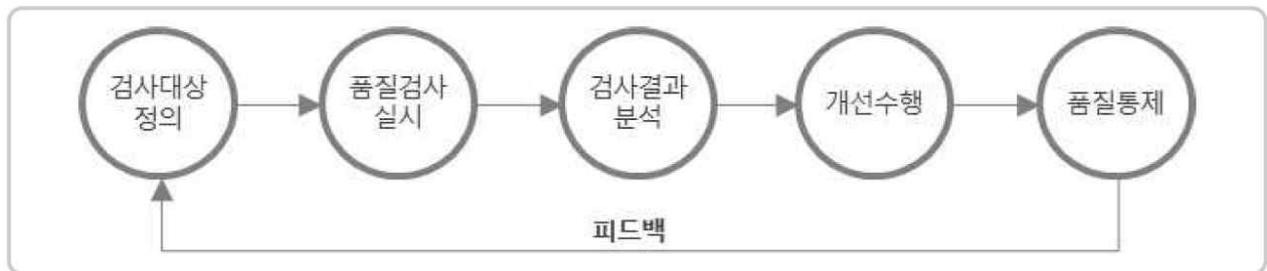
- 인공지능 학습용 데이터의 구축 시 생산되는 원시데이터, 원천데이터, 라벨링데이터의 품질을 확보하기 위한 사전 활동을 의미하며, 원시데이터와 라벨링데이터의 품질관리가 학습데이터의 품질 측면에서 중요
- 인공지능 학습용 데이터 구축에 필요한 데이터를 대상으로 획득단계 품질관리, 정제단계 품질관리, 라벨링단계 품질관리, 운영단계 품질관리, 활용단계 품질관리를 진행 함



[인공지능 학습용 데이터 구축사업 품질관리 체계]

○ 운영 및 활용단계 품질관리 체계

- 운영 및 활용단계 품질관리는 인공지능 학습용 데이터 구축사업을 통해 AI Hub에 적재한 학습데이터셋을 대상으로 품질을 검사하고 개선하는 활동을 의미하며, 이를 위해서는 품질검사 및 개선 활동의 수행을 위한 절차와 품질검사를 위한 기준, 품질검사 방법 진행
- 품질검사는 '검사대상정의(Define)', '품질검사실시(Measure)', '검사결과분석(Analyze)', '개선수행(Implementation)', '품질통제(Control)'의 절차로 수행



[인공지능 학습용 데이터 구축사업 품질관리 체계]

마. 외부 전문 검증기관(TTA)과의 협력 방안

- 고품질 인공지능 학습 데이터 확보 및 품질관리 측면의 효과적 대응을 위해 외부 검증기관(TTA)과의 체계적이고 효율적인 의사소통 체계를 운영하고 수행기관 내 신속한 의사결정 지원을 위해 대내·외 협업체계를 구성 및 운영함

○ 협업 목적

- 과제 총괄 품질 관리 관련 역할 정의
 - 주관기관, 수행 사업단, 외부 검증기관 (TTA) 간 협업체계 구축 및 운영
 - 품질영역 전반의 평가 항목 및 평가 방안 조율 및 품질관리 오너쉽 역할 정의
 - 학습 데이터 품질 평가 및 개선 조치 이행을 통한 품질 확보

- 학습 데이터 공정 및 구축 데이터 품질진단 결과 오류 여부, 원인 분석
 - 오류 데이터 개선 방안에 따른 검토 회의 (의사결정)
 - 외부 검증 기관 품질 검증 지원 및 담당자간 협의

- 인공지능 학습용 데이터의 구축 공정과 결과물의 품질을 종합적으로 관리 및 검증하기 위해 품질 검증 협력 체계 구성

구분	TTA 인공지능 데이터 품질검증 체계의 내용								
검증범위	<div style="border: 1px solid black; padding: 10px;"> <p style="text-align: center; background-color: #4a69bd; color: white; padding: 5px;">인공지능 학습용 데이터 품질관리 및 검증 범위</p> <table border="1" style="width: 100%; border-collapse: collapse; margin: 10px 0;"> <tr> <th style="width: 33%;">공정정의</th> <th style="width: 33%;">데이터 정의</th> <th style="width: 33%;">학습모델 정의</th> </tr> <tr> <td style="text-align: center; vertical-align: top;"> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">획득 및 정제</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">어노테이션</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">검수</div> </td> <td style="text-align: center; vertical-align: top;"> <div style="border: 2px solid gray; padding: 5px; margin-bottom: 5px; background-color: #d3d3d3;">데이터셋 구축</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">데이터 생성</div> <div style="border: 2px solid gray; padding: 5px; margin-bottom: 5px; background-color: #d3d3d3;">활용</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">데이터셋 제공</div> </td> <td style="text-align: center; vertical-align: top;"> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">알고리즘 구현</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">학습모델 생성</div> </td> </tr> </table> </div>			공정정의	데이터 정의	학습모델 정의	<div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">획득 및 정제</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">어노테이션</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">검수</div>	<div style="border: 2px solid gray; padding: 5px; margin-bottom: 5px; background-color: #d3d3d3;">데이터셋 구축</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">데이터 생성</div> <div style="border: 2px solid gray; padding: 5px; margin-bottom: 5px; background-color: #d3d3d3;">활용</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">데이터셋 제공</div>	<div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">알고리즘 구현</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">학습모델 생성</div>
	공정정의	데이터 정의	학습모델 정의						
<div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">획득 및 정제</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">어노테이션</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">검수</div>	<div style="border: 2px solid gray; padding: 5px; margin-bottom: 5px; background-color: #d3d3d3;">데이터셋 구축</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">데이터 생성</div> <div style="border: 2px solid gray; padding: 5px; margin-bottom: 5px; background-color: #d3d3d3;">활용</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">데이터셋 제공</div>	<div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">알고리즘 구현</div> <div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;">학습모델 생성</div>							
검증대상	공정 전주기	데이터 및 저장소	학습 모델						
검증기준	사업수행계획서(계획, 목표), 데이터구축,활용 가이드라인(구축지침, 명세정보)								
	- 체크리스트 이행 여부	- 통계정보 - 데이터, 어노테이션, 저장소 구조 - 참값(Ground Truth)	- 학습성능						
검증방법	검증범위, 항목선정 → 검증기준, 절차, 방법 수립 → 검증결과 확인 및 분석								
	- 문서검토	- 전수 또는 샘플링 검사	- 학습 조건 설정 및 수행						

	- 수행그룹 인터뷰 - 현장 점검 및 근거자료 확인	- 자동화 검수 도구 - 검증 데이터 분석	(데이터 구분, 반복 횟수 등)
검증결과	품질검증 결과서 (구축 공정, 정확도, 유효성)		

○ 외부 전문 검증기관(TTA) 협력방안 내용

- (초기 검증 진행) 초기 인공지능 학습 데이터 설계 시, 분류체계의 초안을 공유하여 보완 의견 수렴
- (가이드라인 제공) 과제 진행 시, 월 단위로 기간을 나누어 일정량의 데이터를 구축했을 때마다 가이드라인을 제동하여 TTA로부터 지속적으로 데이터 품질을 검증받을 수 있도록 함
- (체크리스트 제공) 과제의 진행상황을 확인할 수 있는 체크리스트를 작성하여 일정 기간 별로 TTA와 검증하는 작업을 진행
- 데이터 가공 및 구축에 필요한 항목을 작성하여, 과제 진행 중 수시로 체크리스트를 통한 과제 진척도 확인 및 일정 점검 협의 수행

4. 인공지능 학습용 데이터 활용 지원

가. 인공지능 학습용 데이터 활용 지원 방안

1) 해당 데이터를 활용한 실험 및 연구 자료 공개

- 연구 자료를 통해 해당 데이터의 학습 방법, 활용 방법에 대해 공개
- 연구 자료를 기준으로 해당 데이터의 특징, 카테고리별 사용처에 대한 제시
- 연구 자료 공개를 통하여 공통 연구로서의 진행 기반 마련

2) 모델링 관련 실습 및 학습용 가이드 자료 공개

- 인공지능을 학습 중인 학생, 비전공자들도 쉽게 따라해 보고 이해할 수 있도록 가이드라인을 상세하게 만들어 코드와 함께 공개

3) 인공지능 학습용 데이터 구축도구 공개

- AI Hub에서 허가받은 사람 누구나 사용할 수 있도록 웹 기반의 구축도구를 공개
- 레이블링, 추가데이터 업로드, 데이터검색 등 주요기능 모두 공개
- 구축 도구의 구현 방법, 동작 순서, 지원사항 등에 대해 상세하게 기술
- 레이블링, 검증, 검색작업 등에 대한 상세한 설명을 동영상 촬영하여 공개

4) 데이터셋 고도화 및 재구축

- 새로운 레이블링이 필요한 기업 또는 기관의 요청 시 자체 구축 도구를 활용하여 새로운 인공지능 학습용 데이터를 구축 및 정제하여 제공 (일부 유상판매)
- 정기적인 알고리즘 테스트를 통해 알고리즘의 개선과 동시에 이슈가 있는 데이터에 대한 추가 수정으로 데이터셋 고도화 진행
- 오류 분류 카테고리 세분화 연구를 진행하여 새로운 오류 분류 카테고리 적용, 새 버전 구축 방안 마련

5) 분야별 데이터셋 활용 지원

○ 데이터셋 활용 및 자체 구축을 위한 경험과 전문기술 지원

- 문법적 오류 교정 임무에 대한 연구를 진행하는 학계와 연구계를 대상으로 학습데이터의 효율적인 학습방법 및 오류검증, 편향성 제거 등의 전문경험 지원
- 문법적 오류 교정 임무에 사용될 새로운 기준 데이터셋(Benchmark Dataset)을 자체 개발하고자 하는 기관을 대상으로 새로운 오류 분류 체계, 노하우 및 고도화 경험 지원

○ 학습데이터를 활용한 알고리즘 및 관련 인공지능 제품 개발 지원

- 문법적 오류 교정 임무 모델을 활용한 알고리즘 적용 및 고도화를 진행하는 산업계를 대상으로 데이터 학습과 검증, 품질관리 등 단계별 관련 경험 지원
- 학습 데이터를 활용한 응용제품 및 서비스 모델의 사업화 과정에서 필요한 관련 경험과 전문 지식 공유
- 인공지능 제품 간 통합, 융합 서비스 모델 개발 등 기술간, 서비스 간의 협업 모델을 발굴하고, 공통 연구를 진행

나. 인공지능 학습용 데이터 확산 생태계 구현 방안



○ 데이터 카테고리의 세분화 연구

- 문법적 오류 문장 데이터에 대한 연구가 부족하다보니 해당 데이터의 카테고리 분류 또한 명확하게 연구된 바가 없음
- 본 과제 제안에서 카테고리는 크게 오타자, 맞춤법 오류, 음성 오류이며 상세 분류로 오타자, 맞춤법 오류, 띄어쓰기 오류, 문장부호 오류를 제시
- 데이터를 활용해 학습, 추론, 분류 모델, 교정 모델, 사전 학습 모델 등을 연구하며 오류 문장의 카테고리를 한 층 더 세분화 시킬 수 있는 연구 자료 공개
- 영어에 있어서 오류 문장의 세분화는 CoNLL-2014 논문에 따르면 28개가 제시되었으며 해당 수준까지 도달하기 위해 단계적으로 연구, 본 연구는 그 첫 번째 연구로써 4개의 세분화 카테고리를 사용