연 구 계 획 서

연구과제명	국문	사전 학습된 언어모델의 fine-tuning을 통한 한국어 문법 오류 교정기
	영문	Grammatical Autocorrection for Korean via Fine-tuning pre-trained Language Models

(*)은 필수 작성부분입니다.

(*) 아이디어 제시 배경	1. 인공지능/기계학습 기반 자연어처리(natural language processing) 모델들의 성능은 최근 2년 새 눈부신 성능 향상을 보였음. 구체적으로, 기계번역(machine translation), 질의응답(question-answering), 과제지향적 대화모델(task-oriented conversation model)을 필두로 하여, 다양한 기계학습 기반 자연어처리 모델들의 성능들은 인간이 이러한 과제를 수행했을 때의 성능에 근접했음. 기계 번역의 경우, 사용자가 많은 영어, 불어, 중국어 등의 주요 언어 간의 번역 모델의 성능은, 이미 다수의 회사가 상용 서비스를 출시하였으며, 이제 병렬 코퍼스가 없거나 적은 상황에서 소수 언어(minor language)에 대한 번역 모델 연구가 활발히 진행되고 있음. 질의응답의 경우, 주어진 지문에서 질문에 대한 답을 찾는 문제(SQUAD 2.0)의 경우 이미 인간 수준을 뛰어넘었으며(Google, BERT), 이제는 일반 대량의 지식 데이터베이스(Wikipedia)를 활용하여 일반 상식(common sense) 질문에 대한 응답을 찾는 모델을 개발하고 있음. 또한, 대화 모델의 경우 과제지향적 대화 모델은 다양한 맥락(고객 응대, 서비스 예약 등)에서 인간과 유사한 수준의 반응을 생성하여 과제를 수행하는 모델을 선보인 바 있음(Google, Duplex) 2. 그런데 한국어를 활용하여 이러한 기계학습 모델을 개발하려 할 경우, 다양한 문제 상황에 직면함. 학습 데이터의 부족뿐만 아니라, 주요 언어(영어)와의 차이에 따른 문제가 발생하고, 이러한 문제는 위와 같은 기계학습 모델의 성능에 큰 영향을 미침. 특히, 언어의 구조의 차이는 모델의 입력(input) 단계부터 큰 차이를 불러옴. 한국어는 영어와 달리 교착어로써 어휘의 다양한 변형이 존재할 뿐만 아니라(morphologically rich), 한국어를 모국어로 사용하는 사용자들조차 맞춤법 오류(Grammatical Error)를 빈번하게 저지르기 때문에, 영어와 비교했을 때 한국어 데이터의 경우 기계학습 모델의

	입력부터 매우 큰 노이즈가 존재함. 3. 입력의 노이즈의 원천으로써의 교착어의 특성(어휘의 다양한 변형 문제)을 해결하기 위한 다양한 기법들은 이미 제안된 바 있음(FastText (Facebook), Byte Pair Encoding (Google)). 이는 어휘의 분산표상을 학습할 때 어휘 수준 정보뿐만 아니라 글자 수준 n-gram 정보까지 활용하여 학습하는 방식으로 해결함. 그러나 **이러한 문제의식에서 출발하여 한국어 맞춤법 오류를 해결하고자 하는 시도는 거의 없었음**. 대부분의 현존하는 한국어 맞춤법 교정 시스템의 경우 국립국어원에서 발간한 한국어 어문규범에 기초한 규칙기반(rule-based) 검사기이며, 이러한 모델로는 한국어 사용자들이 발생시키는 수많은 오류의 종류를 모두 포괄하기 매우 어려울 뿐만 아니라, 새롭게 발생하는 오류를 교정하는 방법에 대한 유연한 대처가 힘듦, 따라서 한국어 맞춤법 오류 교정기 또한 **기계학습 기반 end-to-end 모델로 학습을 해야할 필요성이 있음. 이는 기존 규칙-기반 모델에 비교해 훨씬 더 많은 종류의 맞춤법 오류를 교정할 수 있을 것으로 기대될 뿐만 아니라, 새로운 오류를 포괄하는 교정 모델을 만들기에도 용이해짐**. 규칙-기반 모델의 경우, 어떤 종류의 맞춤법 오류가 새로 발생했는지 오류의 종류를 판별 및 정의하고 이에 대한 어문규범을 찾아서 교정 방식을 구현해야하는 번거로움이 있는 반면, 기계학습 모델의 경우 새로운 오류를 포괄하는 문장이 포함된 데이터셋을 기존 모델에 추가적으로 학습하는 것으로 충분하여 상대적으로 간편함. 4. 따라서 기계학습 기반 한국어 맞춤법 교정기의 개발을 통해, 다양한 맞춤법 오류를 교정하고 이를 **기계 번역, 질의 응답, 대화 모델 등의 입력을 위한 전처리 모델로 활용하여 다양한 한국어 자연어처리 과제의 성능을 높일 수 있을 것으로 기대**함. 5. 더 나아가, 이 모델은 전처리 모델로 기능하는 것 뿐만 아니라, 모델 자체로써 의미가 있음. 한국어 학습자 및 한국어를 모국어로 사용하는 사람들을 위한 **맞춤법 교정 및 교육 모델로 활용될 수 있으며, 이러한 모델을 통해 올바른 맞춤법을 사용하도록 사용자를 돕고, 사용자 간 의사소통의 효율을 높일 수 있을 것으로 기대**함.
(*) 연구목적 및 필요성	**1. 연구의 목표: 기계학습 기반 한국어 맞춤법 교정기의 개발** 위와 같은 배경을 바탕으로, 본 연구는 기계학습 기반 한국어 맞춤법 교정기를 개발하는 것을 목표로 함. 연구 배경에서 기술한 대로, 이를 통해 다양한 한국어 자연어처리 과제의 성능을 높이는 것을 목표로 할 뿐만 아니라, 모델 개발을 통하여 한국어 학습자 및 한국어를 모국어로 사용하는 사람들의 맞춤법 준수 능력을 향상시키는 것에 도움이 되도록 함.

- 연구의 세부 목표:

1.1. 한국어 맞춤법 교정기 개발을 위한 학습 데이터셋 구축

- 연구 목표 달성을 위해, 먼저 기계학습 모델 개발을 위한 데이터셋을 구축함. 현재 공개된 한국어 맞춤법 교정 학습 데이터셋이 전무하기 때문에, 데이터 수집을 진행함. 데이터는 (문법적으로 틀린 문장, 문법적으로 올바른 문장)의 쌍으로 구성될 예정이며, 이러한 문장 쌍이 대량으로 포함된 코퍼스를 구축하여 학습 데이터로 활용함.

1.2. 한국어 맞춤법 교정기 개발을 위한 기계학습 모델 개발

- 첫 번째 세부 목표 달성을 통해 구축한 데이터셋을 기반으로 한국어에 적합한 맞춤법 교정기 모델을 개발함. 최근 맞춤법 교정(Grammatical Error Correction) 분야에서 효과적인 접근법으로 여겨지는 것은 Seq2Seq 구조를 사용하되, 이 구조를 언어 모델(language model)을 학습하는 방식을 이용해 대량의 코퍼스에서 사전학습(pre-train) 한 뒤, 이를 맞춤법 교정 과제에 다시 한 번 학습하는 것임. 본 연구는 이러한 최신 모델을 적용하고, 이러한 모델이 한국어에 더 좋은 성능을 낼 수 있도록 새로운 모델을 개발함.

2. 연구의 필요성

- 본 연구가 달성하고자 하는 (세부)목표는 다음과 같은 필요성에 기반함.

2.1. 기계학습 기반 한국어 맞춤법 교정기 개발을 한국어 데이터셋 부재

기계학습 모델 학습을 위해서는 양질의 데이터가 필수적임. 데이터의 양과 질에 따라 그 모델의 성능이 크게 좌우됨. 대부분의 자연어처리 분야 연구가 공개한 데이터셋은 영어로 구성됨. 문법 오류 교정을 위한 영어 데이터셋은 다양하게 존재하나(CoNLL 2014, BEA 2019), 한국어는 이러한 공개된 데이터셋이 거의 없음. 일부이 다국어 데이터셋의 경우(lang-8) 한국어 데이터셋이 최근 공개되었으나 데이터셋의 오류가 지나치게 많아서, 교정작업이 필요한 수준임. 한국어 데이터셋은 유일하게 한국어 학습자 코퍼스(국립국어원)가 있는데, 이는 교정자들의 오류가 빈번하게 발견되고, 기계학습 모델 학습 용도로 제작된 것이 아니라, 이에 적합하도록 변환하는데 다소 큰 작업이 요구됨. 따라서 본 연구는, 이러한 데이터셋을 모아서 교정작업을 진행하는 한편, 다양한 종류의 문법 오류를 포괄하는 데이터셋을 직접 수집함.

2.2. 한국어 전용 맞춤법 교정(Grammatical Error Correction) 모델의 부재

또한, 한국어를 위한 맞춤법 교정 과제를 위한 전용 모델은 아직 제안된 바 없음. 대부분의 교정 모델은 영어를 기반으로 개발되었음. 최근 이러한 모델들은 사전학습된 언어 모델을 기반으로 하고 있는데, 언어 모델 또한 한국어 전용 모델로 아직 제안된 바 없음,

(*)
연구내용
및 방법

1. 기계학습 기반 한국어 맞춤법 교정기 학습 데이터셋 확보

1.1. lang-8 Corpus

lang-8은 언어 교환을 위한 소셜 네트워크 웹사이트임. 이 곳에서는 모국어와 배우고 싶은 언어를 고르고 글을 쓰면 해당 언어에 대한 모국어 사용자가 내가 쓴 글을 교정을 해주는 서비스를 제공함. 이 사이트에서 한국어를 배우고 싶어하는 외국인들의 글과, 그것을 교정한 글의 한국어 학습자 말뭉치를 수집하고, 교정기 개발을 위한 데이터셋을 구축함.

1 이즘도 비가 많이 와요.	1 요즘도 비가 많이 와요.
2 27 스문일곱 이십칠	2 27 스물일곱 이십칠
3 계를 먹고 싶는 느낌이 들었다.	3 계를 먹고 싶은 느낌이 들었다.
4 이런 작문문제가 있어요.	4 이런 작문문제가 있었어요.
5 몸의 반응에 감사해야죠?	5 몸의 반응에 감사해야겠죠?
6 할 수 없어서 작년 죽었던 때	6 할 수 없어서 작년 죽었을 때
7 65 예순다섯 육십우	7 65 예순다섯 육십오
8 다시 같이 술을 마시고 있는	8 다시 다같이 술을 마시고 있는
9 많이 노력해야 하겠다"고 주장했다.	9 많이 노력해야 겠다"고 주장했다.
10 여러분이 그렇게 생각하지 않아요?	10 여러분도 그렇게 생각하지 않아요?
11 요즘은 고기 많이 먹어요.	11 요즘은 고기를 많이 먹어요.
12 드라마 보며 청취하는 것.	12 드라마를 보며 청취하는 것.
13 바로 내일 출발할대요.	13 바로 내일 출발한대요.
14 26 스문여섯 이십육	14 26 스물여섯 이십육
15 그런데 나는 곧 도작할 거다.	15 그런데 나는 곧 도착할 거다.
16 직원 : 손님,죄송합니다.	16 직원 : 손님, 죄송합니다.
17 나 : 무손 소식이야?	17 나 : 무슨 소식이야?
18 어늘 한국말로 씁니다.	18 오늘 한국말로 씁니다.
19 "가억해 너하나만"	19 "기억해 너하나만"
20 스트레스가 많이 쌓여지요?	20 스트레스가 많이 쌓이지요?
21 에어컨으로 시원하게 된 방안에서	21 에어컨으로시원하게 된 방안에서
22 그렇게 된 이유가 잘 모르겠어요...	22 그렇게 된 이유를 잘 모르겠어요...
23 고등 학교 선생님 이애요.	23 고등 학교 선생님 이예요.
24 온몸에 힘이 없어요.	24 온 몸에 힘이 없어요.
25 입본어로 볼 수 있었어요.	25 일본어로 볼 수 있었어요.
26 옛날의 편이 좋았네요.	26 옛날인 편이 좋았네요.
27 한국말도 말할 수 있었어요.	27 한국말도 할 수 있었어요.
28 손님 안년히 가세요.	28 손님 안녕히 가세요.
29 오일 한국어 시험이 있어요.	29 오늘 한국어 시험이 있어요.
30 35 서른다섯 삼십우	30 35 서른다섯 삼십오
31 화장품을 많은 산 적이 있어요.	31 화장품을 많이 산 적이 있어요.
32 그러니까 우리는 큰 기대를 하지 않았다.	32 그러니까 우리는 큰 기대를 하지 않았다.
33 결과는 엉망 이요.	33 결과는 엉망 이에요.
34 나중에 친구 한 명 만났어요.	34 나중에 친구 한 명을 만났어요.
35 집은 저 못안에 있어요.	35 집은 저 못 안에 있어요.
36 스트레스가 많이 쌓여지요?	36 스트레스가 많이 쌓였지요?

그림1. lang-8 데이터셋에서 추출한 틀린 문장 -> 맞는 문장 pair 예시.

1.2. Wikipedia 수정 히스토리를 이용한 데이터 수집

위와 같은 코퍼스는 일반적으로 용량이 작기 때문에, 보다 자동적인 방법으로 대량의 코퍼스를 수집하는 방법이 제안된 바 있음. 관련 연구("Using wikipedia edits in low resource grammatical error correction") [2]의에서는, 적은 자원의 문법 교정 데이터가 있는 환경에서 Wikipedia edit을 이용해 grammatical error correction을 위한 사전학습용 데이터셋을 만드는방법을 제안함. 즉, 위키피디아 문서에서 수정 히스토리를 수집한 뒤, (수정 전, 수정 후)의 쌍을 구성한 뒤 이를 맞춤법 교정으로 간주하는 것임. 이 아이디어를 차용하여, 한국어 위키피디아 코퍼스를 활용하여 사전학습용 코퍼스를 수집함.

1.3. 일반인이 저지르는 맞춤법 오류를 수집한 데이터셋 확보

1.1.번 데이터셋의 경우, 한국어를 학습하는 외국인이 생성한 문법 오류가 주를 이루게 됨. 1.2.번의 경우, 작성자가 한국어 모국어 화자인지 알 수 없다는 문제가 있음. 그런데 한국어의 특수성을 고려하면, 한국어를 모국어로 사용

하는 화자들 또한 맞춤법 오류를 자주 저지름. 이에 따라, 한국어를 모국어로 하는 사람들이 저지르는 오류를 수정하기 위한 별도의 데이터셋을 구축함.

이 데이터셋의 경우, '문법적으로 오류가 있는 문장'을 수집하고 직접 이를 교정하면 매우 큰 시간적/금전적 비용이 들게 됨. 이러한 비용을 줄이기 위하 보다 간편한 방법을 제안함. 그래서, 먼저 문법적으로 올바른 문장을 국립국어원이 배포한 자료에서 수집하여 이를 제시하고, 이를 "듣고" 받아쓰는 데이터 수집 플랫폼을 개발하여, 이 플랫폼 상에서 데이터를 수집함.

구체적으로, 문법적으로 올바른 문장은 국어교수학습샘터에서 발췌해 오고, 사용자들에게 이 문장을 TTS(text-to-service, 글자를 자동으로 읽어주는 기술) 로 읽어주어 쓰도록 하는 플랫폼을 개발함. 이렇게 하면 자연스럽게 사용자들이 쓰던 습관대로 문법 오류가 있는 문장을 수집할 수 있음. 아래 <그림2>, <그림3>에 플랫폼 예시가 주어짐. 플랫폼 사용자가 사이트에 접속하면 <그림2>와 같은 화면이 나오고, 화면의 play 버튼을 누르게 되면 음성으로 문법적으로 올바른 문장을 읽어주는데, 이를 듣고 자연스럽게 문장을 받아쓰게 됨. 받아쓰는 과정에서 한국어 사용자들이 빈번하게 저지르는 맞춤법 오류가 포함됨.

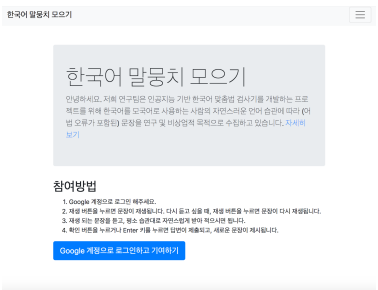

그림2. 한국어 말뭉치 모으기 플랫폼

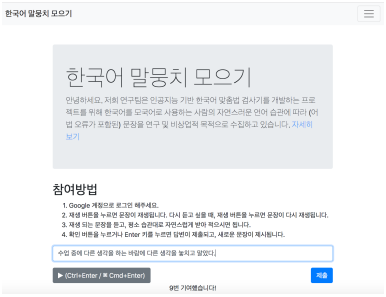

그림3. 한국어 말뭉치 모으기 플랫폼 로그인 후 화면

2. 한국어 맞춤법 교정기 개발을 위한 기계학습 모델 개발

- 모델 개발은 기존 최신 연구에서 출발함. 현재 grammatical error correction을 위한 가장 성능이 좋은 최신 모델(state-of-the-art) model은 copy attention 개념을 사용하는 Transformer 기반 Seq2seq 모델임. 이 모델은 NAACL 2019에서 "Improving Grammatical Error correction via Pre-training a Copy-Augmented Architecture with Unlabeled Data"[3] 라는 제목의 논문으로 발표된 바 있음.

- 이 연구에서 저자들은 문법 교정 과제를 기계번역 과제로 간주함. 따라서, 기존 기계번역 모델들이 자주 사용하는 주의 메커니즘(attention mechanism)에 "copy attention"을 추가해, grammatical error correction 모델의 정확도를 높임. "copy attention"은 Seq2seq 모델의 인코더 입력을 (필요한 경우에) 그대로 복사해올 수 있는 통로를 만들어줌. 인코더를 통해 입력된 문장이

기존 Seq2seq 모델의 디코더를 이용할지, 아니면 copy attention을 기반으로 한 경로를 사용할지를 결정하는 비율은 데이터셋을 통해 학습되며, 이 비율은 아래 그림4의 α_t^{copy} 로 정해짐. 이 매커니즘은 문법 오류 교정 과제 특성 상, 교정된 문장의 결과가 기존 문장과 똑같은 부분이 많다는 것에서 착안함.

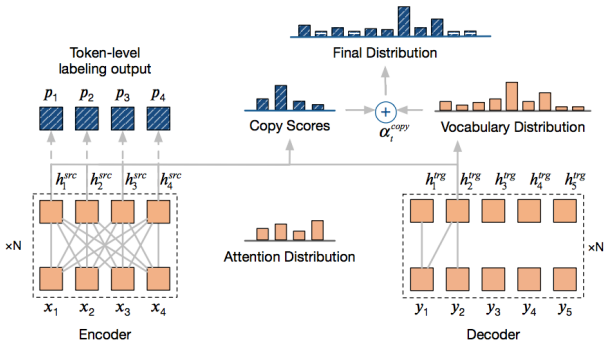

그림4. Copy-augmented Seq2Seq 모델 구조

이 모델은 현재 영어 데이터로 만든 pretrained model이 구현되어 있으며, 띄어쓰기 기반으로만 tokenization이 되어있음. 영어는 character level이나 word level 등으로 tokenization할 수 있지만, 한국어는 영어와 달리 하위 단어 수준(subword-level)으로도 tokenization이 가능함.[4] 하위 단어 수준이란, 한국어에 존재하는 초성과 중성, 그리고 종성의 조합으로 tokenization을 한 것을 말함. 예를 들어, <그림5>의 왼쪽은 subword-level로 lang-8 데이터셋을 tokenization했을 때의 vocabulary를 나타내고, 오른 쪽은 lang-8 데이터의 문장들을 subword-level로 분리했을 때의 결과를 나타냄.

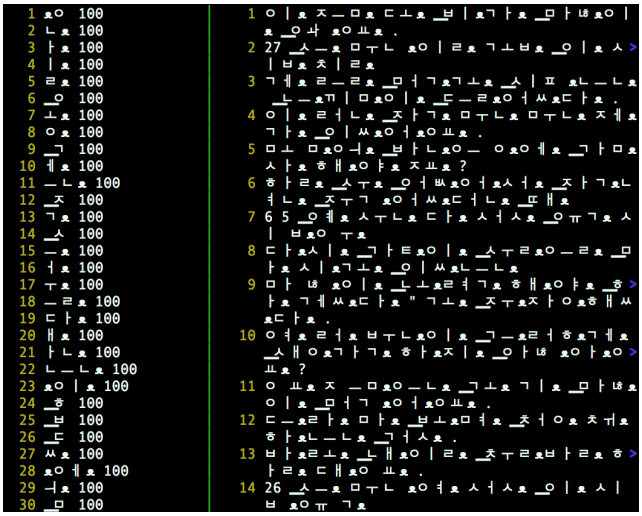

그림5. subword-level로 한국어를 분리했을 때의 결과

	본 연구는 이 모델이 한국어에 적합하도록 변형할 예정임. subword-level이나 character-level tokenization을 진행할 뿐만 아니라, 전용 모델을 개발하여 한국어에 적합한 모델을 맞춤법 오류 교정 모델을 개발함. **3. 학습된 모델 및 연구 결과의 평가** 본 연구에서는 일반적으로 사용되는 문법오류 교정 분야의 성능 측정치를 사용할 계획임. 일반적으로 사용되는 문법오류 교정 과제의 성능 측정치는 GLEU, M2 Score (max-match score), Exact Match, Precision, Recall, F1-score가 있음. 이는 자연어 생성 과제의 평가 측정 방법에서 널리 사용되는 방법을 문법 오류 교정 과제에 맞게 변형한 것으로써, 학술대회에서 열리는 문법 오류 교정 대회(영어, CoNLL 2014, BEA 2019 등)에서 모델의 순위를 가리는데 널리 사용됨.
연구 종료 후 후속조치 및 계획	연구 종료 후, 여러 방법으로 수집한 한국어 교정 데이터셋을 공개하고, 연구 결과는 EMNLP 2020에 컨퍼런스 논문으로 제출할 예정임.
참고문헌	[1] Ross Israel, Markus Dickinson, Sun-Hee Lee, 2013. *Detecting and Correcting Learner Korean Particle Omission Errors.* [2] Adriane Boyd, 2018. *Using Wikipedia Edits in Low Resource Grammatical Error Correction. (*https://www.aclweb.org/anthology/W18-6111/) [3] Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, Jingming Liu, 2019. *Improving Grammatical Error Correction via Pre-training a Copy-Augmented Architecture with Unlabeled Data.* (https://www.aclweb.org/anthology/N19-1014/) [4] Sungjoon Park, Jeongmin Byun, Sion Baek, Yeongseok Cho, Alice Oh. 2018. *Subword-level Word Vector Representations for Korean.* [5] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu, 2019. *MASS: Masked Sequence to Sequence Pre-training for Language Generation.*

본인이 URP 수행학생으로 선정될 경우 위에 작성된 내용에 대하여 연구를 성실히 이행하며 만일 파견, 휴학, 자퇴, 제적 등의 개인적인 사유로 중도에 그만둘 시에는 장학금과 연구비를 반납할 것을 확인합니다.

2019년 11 월 27 일

신청자 윤소영 (인)

위 내용에 대한 2020년 겨울/봄학기 URP 프로그램 참여 학생의 연구 지도를 승인합니다.

2019년 11 월 27 일

지도조교 박성준 (인)

지도교수 오혜연 (인)

연 구 추 진 계 획

일 정	연 구 수 행 내 용
12월 31일 - 1월 24일	Related work 조사, 적용 가능한 GEC 모델 사전조사, evaluation method 연구
1월 25일 - 2월 10일	한국어 말뭉치 수집을 위한 플랫폼 개발
2월 11일 - 2월 17일	한국어 말뭉치 수집을 위한 플랫폼 피드백 반영, 공개화전 Finalize, 홍보 방법과 수집 방법 최종화
2월 18일 - 5월 15일	한국어 말뭉치 수집을 위한 플랫폼 개방, 데이터 수집 및 플랫폼 보수 및 홍보
2월 18일 - 3월 13일	lang8 데이터셋과 wikipedia 데이터셋 추출 및 preprocessing
3월 14일 - 3월 30일	wikipedia 데이터셋으로 copy-attention model을 character level로 pre-training
3월 31일 - 4월 7일	wikipedia 데이터셋으로 copy-attention model을 subword level로 pre-training
4월 8일 - 4월 13일	wikipedia 데이터셋으로 copy-attention model을 word level로 pre-training
4월 14일 - 4월 29일	MASS 사전 조사, 한국어 문법 교정기에 적용할 수 있도록 MASS의 variation model연구
4월 30일 - 5월 15일	lang8 데이터셋으로 character level, subword level, word level pre-trained 된 모델의 fine-tuning
5월 16일 - 5월 30일	플랫폼으로 수집된 데이터셋을 preprocessing 이후 해당 데이터 사용하여 model training
6월 1일 - 6월 7일	wikipedia, lang-8, 플랫폼 수집 데이터셋을 이용한 MASS 의 variation model training
6월 8일 - 6월 12일	사용할 evaluation method 정리 및 개발
6월 13일 - 6월 17일	지금까지 training한 모델 evaluation 진행
6월 16일 - 8월 1일	데이터셋 정리 후 공개화, 논문 작성

※. 종료시까지의 연구계획을 세부일정으로 자세하게 작성