ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 46, No. 6, pp. 554-562, 2019. 6
https://doi.org/10.5626/JOK.2019.46.6.554

L2 학습자를 위한 주의 기제 기법 기반의 문법 오류 감지

(Grammatical Error Detection for L2 Learners Based on Attention Mechanism)

박 찬 희 † 박 진 욱 †† 조 민 수 †† 박 상 현 †††
(Chanhee Park) (Jinuk Park) (Minsoo Cho) (Sanghyun Park)

요 약 문법 오류 감지는 주어진 문장에서 발생한 문법적인 오류의 존재와 그 위치를 발견하는 작업으로, 새로운 언어를 배우는 L2 학습자의 언어 학습과 평가에 유용하게 활용될 수 있다. 기존에는 문법 오류 교정을 위한 시스템이 활발히 연구되고 있으나, 학습 말뭉치의 부족과 제한된 오류 유형 교정과 같은 한계가 존재한다. 따라서 본 연구에서는 순차 레이블링 문제를 통해 오류의 유형이 사전에 정해지지 않은 일반화된 문법 오류 감지를 위한 모형을 제안한다. 단어와 문자를 동적으로 혼합한 표상을 사용하여 L2 학습자의 쓰기에서 나타나는 예측 불가능한 단어를 다루고, 멀티 태스크 학습을 통해 불균형한 데이터의 학습 과정에서 발생할 수 있는 편향성을 방지하였다. 또한, 주의 기제 기법을 적용하여 오류 예측에 있어 판단의 근거가 될 수 있는 단어에 집중해 효율적으로 오류를 예측하였다. 제안하는 모형의 검증을 위해 3개의 평가 데이터를 사용하였으며 각 구성요소를 제거해 봄으로써 모형의 효용성을 검증하였다.

키워드: 자연어 처리, 문법 오류 감지, 순차 레이블링, 단어 표상, 멀티 태스크 학습, 주의 기제 기법

Abstract Grammar Error Detection refers to the work of discovering the presence and location of grammatical errors in a given sentence, and is considered to be useful for L2 learners to learn and evaluate the language. Systems for grammatical error correction have been actively studied, but there still exist limitations such as lack of training corpus and limited error type correction. Therefore, this paper proposes a model for generalized grammatical error detection through the sequence labeling problem which does not require the determination of error type. The proposed model dynamically decides character-level and word-level representation to deal with unexpected words in L2 learners' writing. Also, based on the proposed model the bias which can occur during the learning process with imbalanced data can be avoided through multi-task learning. Additionally, attention mechanism is applied to efficiently predict errors by concentrating on words for judging errors. To validate the proposed model, three test data were used and the effectiveness of the model was verified through the ablation experiment.

Keywords: natural language processing, grammar error detection, sequence labeling, word representation, multi-task learning, attention mechanism

· 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국 연구재단-차세대 정보·컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(NRF-2015M3C4A7065522)

† 학생회원 : 연세대학교 컴퓨터과학과
channy_12@yonsei.ac.kr
†† 비 회 원 : 연세대학교 컴퓨터과학과
parkju536@yonsei.ac.kr
minsoo0104@yonsei.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수(Yonsei Univ.)
sanghyun@yonsei.ac.kr
(Corresponding author임)

논문접수 : 2018년 12월 31일
(Received 31 December 2018)
논문수정 : 2019년 3월 11일
(Revised 11 March 2019)
심사완료 : 2019년 3월 27일
(Accepted 27 March 2019)

정보과학회논문지 제46권 제6호(2019. 6)

1. 서 론

문법 오류 감지(Grammatical Error Detection)는 주어진 문장에서 문법적인 오류의 존재와 그 위치를 발견하는 작업이다. 모국어를 사용할 때에도 문법적으로 완벽한 표현을 구사하는 경우가 많지 않으므로, 새로운 언어를 배우는 L2 학습자의 경우에는 모국어가 아닌 언어로 작문을 하면서 문법적인 오류를 범하는 것은 자연스러운 일이라고 볼 수 있다. L2 학습자의 작문에 대한 정확하고 빠른 피드백을 제공하는 것은 언어 습득에 있어 중요한 과정이지만, 작문을 평가하기 위해서는 많은 비용과 시간이 필요하다. 따라서 비용과 시간을 절감하기 위해 이러한 과정을 자동화한 오류 감지 시스템은 제2외국어 학습 및 평가에 유용하게 활용될 수 있다.

이와 연관되어 최근 몇 년 동안에는 문법 오류 교정(Grammatical Error Correction)을 위한 시스템이 활발히 연구되고 있다. 문법 오류 교정은 오류 감지와 감지된 오류에 대한 교정 두 과정으로 나눌 수 있으며, 두 과정은 순차적으로 시행되거나 동시에 이루어질 수 있다[1]. 최근에는 기계 번역(Machine translation)을 통한 오류 교정 연구가 활발히 진행되고 있지만, 교정 전과 후의 문장을 병렬적으로 사용하는 학습 말뭉치의 부족으로 인해 오류 교정 시스템의 성능에는 여전히 한계가 존재한다[2]. 또한 지도 학습 기반의 기계 학습 분류기(Machine learning classifier)를 통한 교정 모형은 분류 대상인 오류의 종류가 제한적이므로, 발생 가능한 모든 오류의 경우를 다루지 못한다는 단점이 존재한다[3]. 따라서 오류 교정 시스템의 성능 향상을 위해서는 먼저 일반화된 오류 감지 과정이 정확하게 수행되어야 한다. 또한 교육적인 관점에서도 발생한 오류를 감지하여 학습자에게 피드백을 제공하는 것이 문장에서 발생한 모든 오류를 교정하는 것보다 효율적일 수 있다.

[4]는 처음으로 문법 오류 감지를 위한 연구를 오류 교정 작업과 분리하여 독립된 연구로 진행하였다. 지도 학습 기반의 순차 레이블링(Sequence labeling) 모형을 통해 주어진 문장에 대해 토큰 단위로 문법 오류의 유무를 판단하였다. 본 연구에서도 순차 레이블링 모형을 기반으로 하여 오류 감지 모형의 성능 개선을 위해 단어와 문자를 동적으로 혼합한 표상과 주의 기제(Attention) 기법을 적용하였으며, 순차 레이블링과 언어 모델링(Language modeling)을 동시에 수행하는 멀티 태스크 학습(Multi-task learning)을 진행하였다.

자연 언어 처리 분야에서는 일반적으로 텍스트를 단어 단위로 나누고, 각 단어들의 의미를 갖고 있는 임베딩 벡터를 해당 단어의 표상으로 사용한다. 하지만 임베딩 벡터에 존재하지 않는 미등록(Out-Of-Vocabulary) 단어는 그 의미를 반영한 표상을 사용할 수 없게 된다는 한계가 있다. 특히, L2 학습자의 쓰기에는 잘못된 철자와 같이 일반적으로 사용하지 않는 단어가 등장할 가능성이 높으므로 단어 수준의 임베딩을 적용할 경우에는 그 한계가 더욱 극대화 될 수 있다. 따라서 본 연구에서는 단어 수준 표상의 단점을 보완하기 위해 단어와 문자가 동적으로 혼합된 표상을 사용하였다[5].

문법 오류 감지 모형의 학습을 위해 사용하는 인간 주해 데이터(Human-annotated data)는 문장의 각 토큰에 오류의 유무가 표기되어 있다. 하지만 오류 감지 모형의 학습을 위해 주로 사용하는 FCE(First Certificate in English) 데이터셋의 경우, 전체 학습 데이터에서 실제 오류(Incorrect)에 해당하는 단어는 약 14%로 매우 희소하기 때문에 편향적인 학습이 이루어질 가능성이 있다. 따라서 본 연구에서는 [6]과 같이 가장 기본이 되는 언어 모델링을 동시에 학습시킴으로써 추가적인 자질을 사용하지 않고도 보다 일반화된 학습을 진행하였다. 언어 모델링을 통해 다음에 등장할 단어를 예측함으로써 문장의 의미와 구문 구성에 있어 범용적인 패턴을 학습할 수 있게 된다.

또한, 문장에서 오류가 존재하는지 판단하기 위해서는 문장의 전체 보다는 특정 부분을 주의 깊게 봐야 할 필요성이 있다. 예를 들어, 영어에서 주어와 동사의 수 일치 여부를 판단하기 위해서는 해당 문장의 주어를 보고 판단하는 것이 그 대표적인 예라고 할 수 있다. 따라서 본 연구에서는 문장의 각 단어가 오류 감지에 얼마나 중요한지 파악하고, 중요한 단어에 집중하여 효율적으로 오류를 예측할 수 있도록 주의 기제 기법[7]을 적용하였다.

본 연구의 기여는 다음과 같다. (1) 문법 오류 감지를 위한 새로운 모형을 제안하고, 모형의 적합성과 효용성을 3개의 평가 데이터를 통해 검증하였다. (2) 특히, 문법 오류 감지 작업의 특성을 고려하여, 오류 예측에 있어 판단의 근거가 될 수 있는 단어에 집중하는 주의 기제 기법을 적용하였다. 또한, L2 학습자의 쓰기에서 등장하는 예측 불가능한 단어를 다루기 위해 단어와 문자를 동적으로 혼합한 표상을 사용하여 미등록 단어를 해결하였다. (3) 모형의 구성 요소 제거 실험을 통해 각 요소별 역할과 기여를 확인하고, 추후 연구의 기반이 될 수 있는 영향이 큰 구성 요소들을 파악하였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 소개하고, 3장에서는 기본적인 순차 레이블링 모형과 성능 개선을 위한 방법론을 설명한다. 4장에서는 이에 대한 실험과 그 결과를 분석하고, 마지막 5장에서 결론과 향후 계획에 대해 기술한다.

2. 관련 연구

최근에는 딥러닝을 기반으로 하는 RNN(Recurrent Neural Networks), LSTM(Long Short-Term Memory) 계열의 모형을 통해 순차 레이블링 문제를 위한 다양한 연구가 이루어지고 있다. 순차 레이블링 문제는 주로 개체명 인식(Named Entity Recognition), 품사 태깅(Part-Of-Speech Tagging), 의미역 결정(Semantic Role Labeling)을 위해 쓰일 수 있다[8-10]. [4]는 이러한 순차 레이블링 문제를 문법 오류 감지를 위해 최초로 적용한 바 있다. CNN(Convolutional Neural Networks), RNN, LSTM과 같은 다양한 모형의 비교 실험을 진행하여 결과적으로 가장 성능이 우수한 bi-directional LSTM(bi-LSTM) 기반의 오류 감지 모형을 제안했으며, 본 연구를 포함한 이후 진행되는 문법 오류 감지 연구에서 해당 모형을 기반으로 하여 연구가 진행되었다[11-13].

기존의 문법 오류 감지를 위한 연구는 대부분 잘못된 전치사, 관사, 동사의 형태와 같은 비 원어민 영어 학습자의 쓰기에서 가장 빈번하게 발생하는 유형의 오류를 해결하는 데 초점을 두어 진행되었다[14,15]. 그러나 L2 학습자의 작문에서는 사전에 규정할 수 없는 다양한 오류 유형이 발생할 가능성이 존재하므로, 기존 시스템의 경우 모든 유형의 오류를 다룰 수 없다는 한계가 존재한다. 따라서 순차 레이블링 문제를 적용하여 문법 오류 감지 문제를 해결하는 경우, 정해진 오류의 유형 없이 모든 오류를 다룰 수 있게 된다.

[11]은 L2 학습자의 문법성과 표현을 고려한 단어 임베딩을 적용한 오류 감지 모형을 제안하였다. 원어민에 의해 쓰인 말뭉치로 학습된 일반적인 단어 임베딩을 사용할 경우, 오류가 존재하는 단어와 오류가 존재하지 않는 단어에 대한 표상이 유사한 한계가 존재하기 때문이다. [12]는 기존 데이터에서 추출할 수 있는 오류 유형, 품사, 의존 관계와 같은 정보를 오류 레이블 예측과 더불어 추가적으로 예측하는 학습을 통해 모형의 성능을 향상시킨 바 있다. 또한, 학습 데이터에서 실제 오류에 해당하는 단어가 희소한 한계를 극복하기 위해, Sequence-to-Sequence 모형을 사용하여 실제 발생하는 오류의 분포를 학습하고 오류가 포함된 문장을 생성하여 데이터를 증대(Augmentation)시킬 수 있는 방법론을 제안한 연구가 있다[13].

본 연구에서는 [4]의 순차 레이블링 모형을 확장하여 L2 학습자의 쓰기에서 나타나는 예측 불가능한 단어를 다루고자 단어와 문자를 동적으로 혼합한 표상을 사용하고, 학습 데이터에서 희소한 오류로 인해 발생할 수 있는 편향성을 방지하고자 언어 모델링을 동시에 수행하는 멀티 태스크 학습을 진행하였다. 또한, 이전 연구에서는 다뤄지지 않았던 주의 기제 기법을 통해 오류 예측에 있어 판단의 근거가 될 수 있는 단어에 집중하여 효율적으로 오류를 예측하도록 하였다.

3. 문법 오류 감지 모형

본 장에서는 기본 모형인 순차 레이블링 모형에 대해 설명하고, 본 연구에서 제안하는 모형에 적용한 단어 표상, 멀티 태스크 학습, 주의 기제 기법에 대해 소개한다.

3.1 순차 레이블링 모형

순차 레이블링은 일련의 토큰으로 이루어진 문장을 입력으로 받아 각 토큰에 해당하는 레이블을 예측해 출력한다. 본 연구에서는 [4]에서 제안한 bi-LSTM 구조를 기본 모형으로 사용하여 각 토큰이 문법적으로 옳은지(Correct), 또는 옳지 않은지(Incorrect)를 나타내는 레이블을 예측한다.

그림 1에서의 'Token-level LSTM'은 순차 레이블링 모형의 구조를 나타낸다. 먼저, 문장의 각 토큰은 단어 단위의 임베딩 벡터$[x^1, x^2, x^3, ..., x^T]$로 매핑되고, 매핑된 임베딩 벡터는 bi-LSTM 계층의 입력으로 사용된다. bi-LSTM은 입력에 대해 순방향과 역방향으로 각각 진행하는 2개의 LSTM 계층을 통해 현재 입력 이전의 단어와 현재 입력 이후의 단어를 모두 활용하여 문맥에 의존적인 표현형을 생성하게 된다. LSTM의 각 시점에서는 이전 시점의 은닉 상태(Hidden state)와 현재 시점의 입력을 기반으로 새로운 은닉 상태를 출력한다. 이 과정을 통해 양방향으로 생성된 정보를 모두 고려하기

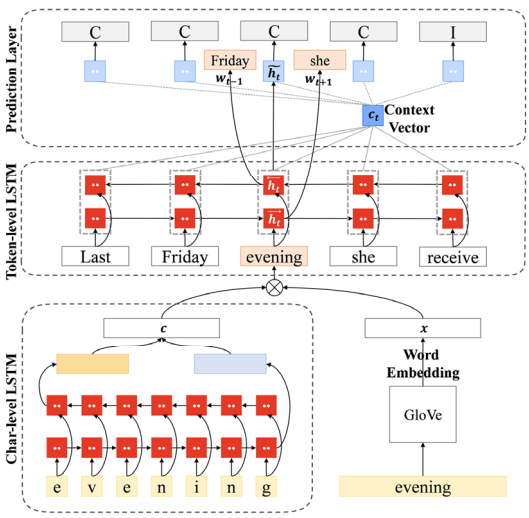


그림 1 문법 오류 감지 모형 구조

Fig. 1 Architecture of the Grammatical Error Detection Model

위해 각각의 은닉 상태를 결합(Concatenate)한 표상을 각 단어의 레이블 예측에 사용한다. 이후 결합된 표상은 완전연결계층(Fully-connected layer)에 통과되어 구성 요소들을 한 공간에 매핑하고 모형이 양 쪽 문맥의 특성을 학습할 수 있도록 한다. 본 구조의 수식은 아래와 같으며, x_t는 t시점에서의 단어 임베딩, $\overrightarrow{h_t}$는 순방향의 은닉 상태, $\overleftarrow{h_t}$는 역방향의 은닉 상태, W_d는 가중치 행렬, $tanh$는 비선형 활성화 함수를 나타낸다.

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}}) \quad (1)$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}}) \quad (2)$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (3)$$

$$d_t = tanh(W_d h_t) \quad (4)$$

마지막으로, 가능한 레이블에 대해 정규화된 분포를 계산하는 소프트맥스(Softmax) 계층을 통해 오류의 유무를 나타내는 레이블을 예측한다.

$$P(y_t|d_t) = softmax(W_o d_t) \quad (5)$$

모형의 학습은 올바른 예측에 대해 음의 로그 확률(Negative log-probability)을 최소화하는 교차 엔트로피(Cross entropy) 함수를 최소화시키는 방향으로 진행된다.

$$E = -\sum_{t=1}^{T} log(P(y_t|d_t)) \quad (6)$$

3.2 문자와 단어를 동적으로 혼합한 표상

단어 임베딩이란 문맥에서 단어들의 유사성을 고려하기 위해 단어를 다차원 공간에 벡터화한 것을 말한다[16]. 이렇게 단어의 의미를 반영하여 벡터 공간상에 분산 표현(Distributed representations)된 임베딩 벡터는 텍스트 데이터의 표현형으로서 문서 분류, 기계 번역과 같은 다양한 자연 언어 처리 분야에서 널리 사용되고 있다[17,18]. 그러나 단어 임베딩은 각 단어들 사이의 표면적 또는 형태론적 유사점을 고려하지 않고, 임베딩 벡터를 만들기 위한 학습 말뭉치에 등장한 단어만을 표현한다는 단점이 존재한다. 따라서 미등록된 단어는 그 의미를 반영한 표현형을 사용할 수 없게 된다.

이러한 한계를 극복하기 위해 문자 수준의 표현형을 고려한 연구가 진행되고 있다[19,20]. 문자 수준의 표현형은 유사한 형태소 패턴을 포착할 수 있으므로 등록 단어 뿐 아니라 미등록 단어에 대한 표현형을 개선할 수 있다. 또한, 이미 학습된 단어의 의미와 문자 수준 표현형을 모두 고려하기 위해 LSTM 계층을 통과한 문자 수준 표현형과 단어 임베딩을 결합(Concatenate)한 표상이 제안된 바 있다[21].

본 연구에서는 각 문자를 입력으로 하는 bi-LSTM 계층을 통과한 마지막 은닉 상태(Final hidden states)를 단어의 문자 수준 표현형으로 사용하였다. 문자와 단어 수준의 표현형을 단순히 결합하는 대신, 최종 표상에 단어 임베딩과 문자 수준 표현형이 각각 얼마나 많은 정보를 제공하는지 학습을 통해 동적으로 결정하게 된다. 본 과정의 수식은 아래와 같다. x는 단어 임베딩, c는 문자 수준 표현형, $W_g^{(1)}$, $W_g^{(2)}$, $W_g^{(3)}$은 가중치 행렬, σ는 [0,1] 범위의 로지스틱(Logistic) 함수를 나타낸다. 그리고 g와의 원소 곱(Element-wise multiplication)을 통해 두 표현형을 각각 단어의 최종 표상($\tilde{x}$)에 어느 정도 반영할지 결정하고, $\tilde{x}$을 3.1의 순차 레이블링 모형의 입력으로 사용하였다.

$$g = \sigma(W_g^{(3)} tanh(W_g^{(1)} x + W_g^{(2)} c)) \quad (7)$$

$$\tilde{x} = (1-g) \cdot x + g \cdot c \quad (8)$$

3.3 멀티 태스크 학습

멀티 태스크 학습은 여러 학습 과제를 동시에 해결하여 예측 성능을 향상시키는 방법이다. 본 연구에서는 불균형한 데이터셋으로 인한 편향된 학습을 방지하기 위해, 다음에 등장할 단어를 예측하는 언어 모델링을 오류 예측과 동시에 수행하였다. 언어 모델링을 통해 문장의 의미와 구문 구성에 있어 범용적인 패턴을 학습함으로써 보다 일반화된 학습을 진행하여 예측 성능을 향상시킬 수 있다.

그림 1에서의 'Prediction Layer'는 순차 레이블링과 언어 모델링을 동시에 수행하는 멀티 태스크 학습 과정을 나타낸다. 순방향 LSTM의 은닉 상태 $\overrightarrow{h}$를 통해 순방향의 다음 단어를 예측하고, 역방향 LSTM의 은닉 상태 $\overleftarrow{h}$를 통해 역방향의 다음 단어를 예측한다. 또한 3.1과 같이 각 방향의 은닉 상태를 결합하여 본래의 목적인 오류 예측을 위해서도 사용하므로, bi-LSTM 모형이 두 과제를 모두 수행하도록 최적화된다. 본 과정은 아래 수식과 같이, 각 방향 모두 은닉 상태로부터 언어 모델링을 위한 자질 $\overrightarrow{m_t}$, $\overleftarrow{m_t}$를 비선형 계층을 통해 추출하고, 해당 자질을 통해 다음에 등장할 단어(w_{t+1}, w_{t-1})을 예측한다.

$$\overrightarrow{m_t} = tanh(\overrightarrow{W_m}\overrightarrow{h_t}) \quad (9)$$

$$\overleftarrow{m_t} = tanh(\overleftarrow{W_m}\overleftarrow{h_t}) \quad (10)$$

$$P(w_{t+1}|\overrightarrow{m_t}) = softmax(\overrightarrow{W_l}\overrightarrow{m_t}) \quad (11)$$

$$P(w_{t-1}|\overleftarrow{m_t}) = softmax(\overleftarrow{W_l}\overleftarrow{m_t}) \quad (12)$$

언어 모델링의 목적 함수(Objective function)는 식 (13), (14)와 같이 올바른 단어 예측에 대해 음의 로그 확률을 최소화하는 하는 방향으로 진행된다. 그리고 식 (15)와 같이, 언어 모델링의 목적 함수와 순차 레이블링의 목적 함수인 식 (6)을 결합해 최종 목적 함수로 사

용한다. γ는 언어 모델링 목적 함수의 기여도를 제어하기 위해 사용되며, 본 연구에서는 0.1로 지정하였다.

$$\overrightarrow{E}=-\sum_{t=1}^{T-1} log(P(w_{t+1}|\overrightarrow{m_t})) \quad (13)$$

$$\overleftarrow{E}=-\sum_{t=2}^{T} log(P(w_{t-1}|\overleftarrow{m_t})) \quad (14)$$

$$\tilde{E}=E+\gamma(\overrightarrow{E}+\overleftarrow{E}) \quad (15)$$

3.4 주의 기제 기법

주의 기제 기법은 딥러닝 모형이 결정을 내려야 하는 특정 시점에서 관련성이 높은 중요한 맥락에 강조를 주는 기계학습의 한 기법이다. 최근 기계 번역, 이미지 캡셔닝, 감성 분석과 같은 다양한 분야에서 널리 사용되어 정확도 향상에 기여한 바 있다. 본 연구에서는 문장에서 오류가 존재하는지 예측할 때, 문장의 전체 보다는 예측에서 판단의 근거가 되는 특정 단어에 집중하여 효율적으로 오류를 예측할 수 있도록 주의 기제 기법[7]을 적용하였다.

본 과정의 수식은 아래와 같다. bi-LSTM 계층에서 t번째 단어의 은닉 상태인 h_t와 문장 내의 각 단어에 대한 은닉 상태 $\bar{h}_j$에 대해 식 (17)을 통해 유사도를 계산하고, 문장의 모든 k개의 단어에 대해 주의 집중 가중치 $\alpha_{t,j}$를 계산한다. 그리고 주의 집중 가중치와 은닉 상태의 가중 합(Weighted sum)을 통해 문맥 벡터 c_t를 얻는다. 여기서 문맥 벡터는 오류 예측에 필요한 중요도 정보를 저장하고, 문맥 벡터를 은닉 상태와 결합해 완전 연결계층에 통과시킴으로써 중요도 정보를 반영한 새로운 은닉 상태 $\tilde{h}_t$를 얻을 수 있다. 최종적으로, 중요도 정보가 가중되어있는 은닉 상태는 식 (4)~(5)와 같이 오류 예측을 위한 소프트맥스 계층에서 이용된다.

$$\alpha_{t,j}=\frac{\exp(score(h_t,\bar{h}_j))}{\sum_k \exp(score(h_t,\bar{h}_k))} \quad (16)$$

$$score(h_t,\bar{h}_j)=tanh(W_a[h_t;\bar{h}_j]) \quad (17)$$

$$c_t=\sum_{j=1}^{N}\alpha_{t,j}h_j \quad (18)$$

$$\tilde{h}_t=tanh(W_c[c_t;h_t]) \quad (19)$$

4. 실험 및 평가

4.1 실험 데이터

모형의 학습을 위해서 사용한 데이터는 FCE로, 본래는 오류 교정을 위해 영어 L2 학습자가 작성한 문장과 그에 대한 문법 오류 유형 및 범위가 표기되어 있다. 본 연구에서는 [4]에서 오류 감지 연구를 위해 FCE 데이터를 각 토큰에 대한 이진 분류(Correct/Incorrect) 형식으로 변환하여 공개한 데이터를 사용하였다. 총 33,673개의 문장 중에서 28,731개는 학습 데이터로, 2,222개는 검증 데이터로, 2,720개는 평가 데이터로 구성되어 있다.

I	was	trully	disappointing	by	it	.
C	C	I	I	C	C	C

그림 2 학습 데이터 예시
Fig. 2 Example of the training data

또한, 제안하는 모형의 추가적인 평가를 위해 CoNLL-14 Shared Task 평가 데이터[22]를 사용하였다. CoNLL-14는 본래 문법 오류 교정을 위한 평가 데이터로, 보다 길고 능숙하게 쓰인 1,300개의 문장으로 이루어져 있으며 2명의 전문가(Annotator)에 의해 2개의 평가 데이터로 구축되었다. 본 연구에서는 해당 데이터를 NLTK(Natural Language Toolkit)[23]를 통한 토큰화(Tokenize) 과정을 거쳐, 오류 교정 전후의 문장을 비교하여 그림 2와 같은 오류 감지 형식으로 변환해 사용하였다. 실험에 사용된 데이터의 전처리를 위해 대문자를 모두 소문자로, 등장하는 숫자는 모두 동일하게 '0'으로 변경하였다.

4.2 실험 환경 및 모수 설정

모든 실험은 Intel(R) Core(TM) i7-6700K CPU와 NVIDIA GeForce GTX 1080 GPU가 장착된 PC에서 Tensorflow[24] 프레임워크를 통해 진행되었다. 단어 임베딩 벡터는 Wikipedia 2014와 Gigaword 5를 통해 사전 학습된(Pre-trained) 300차원의 GloVe[25]를 사용하였다. 문자 수준의 임베딩 벡터는 100차원으로 Xavier 초기화[26] 방식을 통해 초기화한 후 학습되도록 설정하였다. 단어의 문자 수준 표현형을 얻기 위한 LSTM 은닉층의 크기는 100으로, 단어 수준의 표현형을 얻기 위한 LSTM 은닉층의 크기는 300으로, 멀티 태스크 학습에서 언어 모델링을 진행하는 계층의 은닉층 크기는 50으로 지정하였다.

학습에 사용되는 모든 가중치는 Xavier 초기화 방식으로 초기화 되며, 과적합(Over-fitting)을 방지하기 위한 드롭아웃(Drop out)의 비율을 0.5로 설정하였다. 학습은 총 50번의 에폭(Epoch)으로 Adam 최적화 기법[27]을 통해 진행하였다. 초기 학습률(Learning rate)은 0.001로 설정한 후, 점진적으로 학습률을 감소(Learning rate decay)시키는 방법을 이용하였다. 또한, 각 에폭마다 검증 데이터를 통해 성능을 측정하여, 7번의 에폭 동안 $F_{0.5}$ 점수를 기준으로 성능 개선이 없는 경우 과적합을 막기 위해 학습을 초기 종료(Early stopping) 하였다. 실험 결과는 $F_{0.5}$ 점수를 기준으로 최고의 성능을 나타내는 에폭에 대해 4.3의 평가 척도를 이용하여 평가하였다.

표 1 FCE 데이터를 사용한 모형의 성능 평가 및 구성 요소 제거 실험 결과
Table 1 Experimental results of the proposed model and ablation experiments of each component using FCE dataset

	FCE-DEV					FCE-TEST				
	predicted	correct	P	R	$F_{0.5}$	predicted	correct	P	R	$F_{0.5}$
Proposed model	**2334**	**1660**	**68.55**	**34.48**	**57.24**	**2977**	**1977**	**66.41**	**31.36**	**54.27**
(-)char attention	1921	1314	68.40	28.32	53.31	2569	1675	65.20	26.57	50.51
(-)multitask learning	1933	1233	63.78	26.57	49.83	2535	1567	61.81	24.85	47.64
(-)attention	2126	1436	67.54	30.95	54.63	2964	1852	62.48	29.37	50.99
(+)CRF	2160	1500	69.44	32.32	56.47	2685	1787	66.55	28.34	52.42

표 2 CoNLL 평가 데이터를 사용한 모형의 성능 평가 및 구성 요소 제거 실험 결과
Table 2 Experimental results of the proposed model and ablation experiments of each component using CoNLL-14 test data

		CoNLL-14 TEST1				CoNLL-14 TEST2			
	predicted	correct	P	R	$F_{0.5}$	correct	P	R	$F_{0.5}$
Proposed model	**2152**	**540**	**25.09**	**17.83**	**23.20**	**803**	**37.31**	**19.03**	**31.30**
(-)char attention	1910	466	24.40	15.38	21.84	683	35.76	16.18	28.79
(-)multitask learning	1699	356	20.95	11.17	18.11	553	32.54	13.10	25.09
(-)attention	2146	467	21.76	15.41	20.10	715	33.31	16.94	27.92
(+)CRF	1693	452	26.69	14.92	23.05	677	39.98	16.04	30.79

4.3 평가 척도

모든 실험에 대한 성능 평가는 정밀도(Precision), 재현율(Recall), $F_{0.5}$ 점수로 측정된다. 정밀도는 시스템의 예측 중에서 올바른 예측인 비율을, 재현율은 실제 정답 중에서 시스템이 올바른 예측을 한 비율을 나타낸다. $F_{0.5}$ 점수는 문법 오류 교정을 위한 CoNLL-14 Shared Task에서 채택된 평가 척도이다. 정확한 피드백이 중요한 오류 감지 시스템의 특성을 반영하기 위해, 식 (22)와 같이 β의 값을 0.5로 조절하여 재현율보다 정밀도에 2배의 가중치를 주어 계산한다[28]. 또한, [29]에 따라 예측한 토큰과 실제 올바른 예측에 해당하는 토큰의 개수를 측정하였다.

$$precision = \frac{TP}{TP+FP} \tag{20}$$

$$recall = \frac{TP}{TP+FN} \tag{21}$$

$$F_{\beta} = \frac{(\beta^2+1)precision \cdot recall}{\beta^2 precision + recall} \tag{22}$$

4.4 문자와 단어를 동적으로 혼합한 표상 실험

표 3은 3.1의 순차 레이블링 모형에 각각 단어 임베딩(word-level), 단어 임베딩과 문자 수준의 표현형을 단순 결합한 표상(char concat), 단어와 문자의 혼합 비율을 학습을 통해 동적으로 결정하는 표상(char attention)을 적용한 실험 결과를 나타낸다. 실험은 FCE 데이터를 통해 학습을 진행한 모형에 평가 데이터 3개의 $F_{0.5}$ 점수를 측정해 진행하였다. 본 실험을 통해 3가지 표상 방법을 비교한 결과, 3개의 평가 데이터 모두에서 단어 임베딩과 문자 수준 표현형의 혼합 비율을 학습을 통해 동적으로 결정하는 표상이 가장 우수한 성능을 나타냄을 확인하였다. 따라서 본 실험 결과를 바탕으로 문자와 단어 수준을 동적으로 혼합하는 표상을 제안하는 모형의 표상으로 사용하였다.

표 3 단어 표상 방법론에 따른 실험 결과($F_{0.5}$)
Table 3 Experimental results of word representation methodology($F_{0.5}$)

	FCE TEST	CoNLL-14 TEST1	CoNLL-14 TEST2
word-level	44.27	18.73	24.51
char concat	45.27	18.78	24.98
char attention	45.97	18.96	25.30

4.5 구성요소 제거 실험

표 1과 표 2는 본 연구에서 제안하는 모형을 FCE 학습 데이터를 통해 학습하고, 검증 데이터와 평가 데이터, 그리고 CoNLL-14 평가 데이터 2개를 통해 성능을 측정한 결과를 나타낸다. 또한, 모형에 적용된 구성요소의 영향력을 평가하기 위해 각 구성요소를 제거하는 실험을 진행하였다.

먼저, 단어와 문자의 혼합 비율을 동적으로 결정하는 표상을 제거한 경우, FCE 평가 데이터의 $F_{0.5}$ 점수를 기준으로 3.76%p 하락하였다. 단어와 문자의 혼합 비율을 결정하는 값(g)은 미등록 단어와 같이 등장 빈도가 낮은 단어에 대해 높게 나타나는 경향이 있다[21]. 단어에 대한 표상을 제대로 추정할 수 없을 때 단어의 철자를 이용하는 것을 학습한 것이라 할 수 있다. 예를 들어, 'dogs'라는 단어가 임베딩 벡터에 존재하지 않는다면 일반적으로 무작위로 초기화된 벡터를 해당 단어의 표상

표 4 선행 연구와 성능 비교 결과
Table 4 Performance comparison of previously studied models

	FCE-TEST			CoNLL-14 TEST1			CoNLL-14 TEST2		
	P	R	$F_{0.5}$	P	R	F0.5	P	R	$F_{0.5}$
Kaneko et al. [11]	46.70	28.60	41.40	-	-	-	-	-	-
Rei&Yannakoudakis [12]	57.50	28.30	47.70	16.70	22.00	17.50	26.50	24.90	26.20
Rei et al. [6]	58.88	28.92	48.48	17.68	19.07	17.86	27.62	21.18	25.88
Kasewa et al. [13]	-	-	50.40	-	-	22.10	-	-	30.80
Proposed model	66.41	31.36	**54.27**	25.09	17.83	**23.20**	37.31	19.03	**31.30**

으로 사용하게 된다. 하지만 본 표상을 통해서는 'dog'와 접미사 's'가 결합된 것을 유추하고, 그 결합의 정도 또한 동적으로 결정한다. 즉, 본 표상을 통해 단어 임베딩에 존재하지 않는 미등록 단어 뿐 아니라 L2 학습자의 작문에서 나타날 수 있는 다양한 단어까지도 의미있는 표상으로 나타내는 것이 가능한 것으로 해석할 수 있다. 따라서 본 표상이 제안하는 모형에 적합한 표상임을 입증하였다.

언어 모델링을 통한 멀티 태스크 학습을 제거한 경우에는 성능이 6.63%p 하락하였다. 멀티 태스크 학습을 진행하여 불균형한 학습 데이터의 한계를 보완할 수 있을 뿐 아니라, 언어 모델링을 통해 문장의 의미와 구문 구성에 있어 범용적인 패턴의 학습이 이루어진 것으로 분석된다. 특히, 가장 큰 폭으로 성능이 하락함에 따라 학습 데이터의 편향성을 보완할 수 있는 구성 요소가 모형의 성능을 좌우한다는 것을 알 수 있다.

또한, 주의 기제 기법을 제거한 경우에는 성능이 3.28%p 하락하였다. 주의 기제 기법을 적용하여 각 단어의 오류를 예측할 때, 문장의 각 단어가 오류 감지에 얼마나 중요한지를 학습하고 중요한 단어에 보다 집중하여 예측하므로 모형의 오류 예측 성능 향상에 기여함을 확인하였다.

추가로, 순차 레이블링 모형에서 최종적으로 오류를 예측할 때 소프트맥스 계층 대신 주어진 레이블 간의 인접성을 통해 최적의 레이블을 예측하는 CRF(Conditional Random Fields)를 사용한 실험을 진행하였다. CRF는 일반적으로 순차 레이블링 모형의 출력 계층에서 많이 사용하지만, 본 실험에서는 CRF를 사용한 경우 1.85%p의 성능 저하가 확인되었다. 문법 오류 감지 모형의 경우에는 오류(Incorrect) 레이블이 희소할 뿐 아니라, 다른 순차 레이블링 모형과 달리 예측 레이블이 2개로 그 개수가 적기 때문인 것으로 분석된다.

CoNLL-14 평가 데이터의 경우에도 FCE 평가 데이터의 결과와 유사한 양상을 보였다. 따라서 문자와 단어를 동적으로 혼합한 표상을 기반으로, 언어 모델링을 통한 멀티 태스크 학습과 주의 기제 기법을 적용한 본 연구에서 제안하는 모형이 문법 오류 감지 작업에 효과적임을 본 실험을 통해 입증하였다.

4.6 선행 연구 결과 비교

표 4에서는 FCE와 CoNLL-14[1]평가 데이터를 사용한 선행 연구 결과와 본 연구에서 제안하는 모형의 성능을 비교하였다. 데이터에 따라 선행 연구 결과가 없는 경우에는 '-'으로 표기하였다. [11]은 학습자의 문법성과 표현을 고려한 단어 임베딩을 적용한 모형이고, [12]는 오류 레이블 예측과 기존 데이터에서 추출할 수 있는 오류 유형, 품사, 의존 관계와 같은 정보를 추가적으로 예측하는 학습을 진행하였다. 또한 [6]은 언어 모델링을 통한 멀티 태스크 학습만을 진행한 모형이고, [13]은 FCE 데이터와 추가 데이터를 통해 실제 발생하는 오류의 분포를 학습하여 데이터를 증대시켜 학습한 모형이다. 선행 연구 결과와 비교한 결과, 본 연구에서 제안하는 모형이 추가적인 자질을 사용하지 않고도 $F_{0.5}$ 점수를 기준으로 가장 높은 성능을 보임을 확인하였다.

5. 결 론

본 연구에서는 문자 수준 표현형과 단어 임베딩을 동적으로 혼합한 표상을 기반으로, 멀티 태스크 학습과 주의 기제 기법을 적용한 문법 오류 감지 모형을 제안하였다. 제안하는 모형을 검증하기 위해 3개의 평가 데이터를 사용하였으며, 추가적인 자질을 사용하지 않음에도 선행 연구들을 $F_{0.5}$ 점수를 기준으로 3.87%p 상회하며 우수한 모형임을 입증하였다. 또한, 모형의 각 구성 요소들을 제거해 봄으로써 각 구성 요소들의 역할과 기여를 확인하였다.

하지만 개체명 인식, 품사 태깅과 같은 다른 순차 레이블링 문제에 비해 상대적으로 낮은 성능을 보이고 있다. 데이터에서 한 문장 내에 문법 오류에 해당하는 레이블이 다른 작업에 비해 희소하기 때문인 것으로 분석된다. 따라서 오류의 희소성 문제를 해결하기 위해 실제로 존재할 수 있는 다양한 오류를 보강하는 방법을 모색하여 성능 개선을 위한 향후 연구를 진행할 것이다.

1) 각 선행 연구에서 진행한 오류 감지 데이터로의 변환 과정에 따라 데이터 차이가 발생할 수 있음.

References

[1] C. Napoles, and C. Callison-Burch, "Systematically Adapting Machine Translation for Grammatical Error Correction," *Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 345-356, 2017.

[2] M. Rei, M. Felice, Z. Yuan, and T. Briscoe, "Artificial Error Generation with Machine Translation and Syntactic Patterns," *Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 287-292, 2017.

[3] A. Rozovskaya, KW. Chang, M. Sammons, D. Roth, and N. Habash, "The Illinois-Columbia System in the CoNLL-2014 Shared Task," *Proc. of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 34-42, 2014.

[4] M. Rei, and H. Yannakoudakis, "Compositional Sequence Labeling Models for Error Detection in Learner Writing," *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1181-1191, 2016.

[5] Y. Miyamoto, and K. Cho, "Gated Word-Character Recurrent Language Model," *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1992-1997, 2016.

[6] M. Rei, "Semi-supervised Multitask Learning for Sequence Labeling," P*roc. of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 2121-2130, 2017.

[7] M.T. Luong, H. Pham, and C.D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412-1421, 2015.

[8] H. Yu, and Y. Ko, "Expansion of Word Representation for Named Entity Recognition Based on Bidirectional LSTM CRFs," *Journal of KIISE*, Vol. 44, No. 3, pp. 306-313, Mar. 2017. (in Korean)

[9] S.W. Kim, and S.P. Ko, "Research on Joint Models for Korean Word Spacing and POS (Part-Of-Speech) Tagging based on Bidirectional LSTM-CRF," *Journal of KIISE*, Vol. 45, No. 8, pp. 792-800, Aug. 2018. (in Korean)

[10] J. Bae, and C. Lee, "Korean Semantic Role Labeling using Stacked Bidirectional LSTM-CRFs," *Journal of KIISE*, Vol. 44, No. 1, pp. 36-43, Jan. 2017. (in Korean)

[11] M. Kaneko, Y Sakaizawa, and M Komachi, "Grammatical Error Detection Using Error- and Grammaticality-Specific Word Embeddings," *Proc. of the 8th International Joint Conference on Natural Language Processing*, pp. 40-48, 2017.

[12] M. Rei, and H. Yannakoudakis, "Auxiliary Objectives for Neural Error Detection Models," *Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 33-43, 2017.

[13] S. Kasewa, P. Stenetorp, and S. Riedel, "Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection," *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4977-4983, 2018.

[14] E. Kochmar, and T. Briscoe, "Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics," *Proc. of COLING 2014, the 25th International Conference on Computational Linguistics*, pp. 1740-1751, 2014.

[15] M. Gamon, "Using Mostly Native Data to Correct Errors in Learners' Writing: A Meta-Classifier Approach," *Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 163-171, 2010.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Workshop at International Conference on Learning Representations, arXiv preprint arXiv:1301.3781*, 2013.

[17] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746-1751, 2014.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Feature," *Proc. of the 2015 International Conference on Learning Representations, arXiv preprint arXiv:1409.0473*, 2014.

[19] W. Ling, I. Trancoso, C. Dyer, and A.W. Black, "Character-based Neural Machine Translation," a*rXiv preprint arXiv:1511.04586*, 2015.

[20] Y. Kim, Y. Jernite, D. Sontag, and A.M. Rush, "Character-Aware Neural Language Models," *Proc. of the 30th AAAI Conference on Artificial Intelligence (AAAI16)*, pp. 2741-2749, 2016.

[21] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," *Proc. of the NAACL-HLT 2016*, pp. 260-270, 2016.

[22] H.T. Ng et al., "The CoNLL-2014 Shared Task on Grammatical Error Correction," *Proc. of the 18th Conference on Computational Natural Language Learning: Shared Task*, pp. 1-14, 2014.

[23] E. Loper, and S. Bird, "NLTK: The Natural Language Toolkit," *Proc. of the Association for Computational Linguistics-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics-volume 1,* pp. 63-70, 2002.

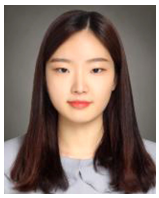
[24] M. Abadi et al., "TensorFlow: A System for Large-scale Machine Learning," *Proc. of the 12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265-283, 2016.

[25] J. Pennington, R. Socher, and C.D. Manning, "GloVe:

Global Vectors for Word Representation," *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532-1543, 2014.
[26] X. Glorot, and Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," *Proc. of the International Conference on Artificial Intelligence and Statistics*, pp. 249-256, 2010.
[27] D.P. Kingma, and J. Ba, "Adam: A Method for Stochastic Optimization," *Proc. of the 2015 International Conference on Learning Representations, arXiv preprint arXiv:1412.6980*, 2014.
[28] R. Nagata, and K. Nakatani, "Evaluating Performance of Grammatical Error Detection to Maximize Learning Effect," *Proc. of COLING 2010, the 23rd International Conference on Computational Linguistics: Posters*, pp. 894-900, 2010.
[29] M. Chodorow, M. Dickinson, R. Israel, and J. Tetreault, "Problems in Evaluating Grammatical Error Detection Systems," *Proc. of COLING 2012, the 24th International Conference on Computational Linguistics*, pp. 611-628, 2012.

박 찬 희
2018년 서울여자대학교 컴퓨터학과(학사) 2018년~현재 연세대학교 컴퓨터과학과 석사과정. 관심분야는 빅데이터마이닝 & 기계 학습

박 진 욱
2016년 서울시립대학교 통계학과(학사) 2017년~현재 연세대학교 컴퓨터과학과 석박사통합과정. 관심분야는 빅데이터마이닝 & 기계 학습



조 민 수
2017년 동국대학교 정보통신공학과(학사) 2017년~현재 연세대학교 컴퓨터과학과 석사과정. 관심분야는 빅데이터마이닝 & 기계 학습

박 상 현
1989년 서울대학교 컴퓨터공학과 졸업(학사). 1991년 서울대학교 대학원 컴퓨터공학과(공학석사). 2001년 UCLA 대학원 컴퓨터과학과(공학박사). 1991년~1996년 대우통신 연구원, 2001년~2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow. 2002년~2003년 포항공과대학교 컴퓨터공학과 조교수. 2003년~2006년 연세대학교 컴퓨터과학과 조교수. 2006년~2011년 연세대학교 컴퓨터과학과 부교수. 2011년~현재 연세대학교 컴퓨터과학과 교수. 관심분야는 데이터베이스, 데이터마이닝, 바이오인포매틱스, 빅데이터마이닝 & 기계 학습