영어 말하기, 쓰기 학습자를 위한 문법 오류 검출 시스템

서홍석○, 이성진, 이진식, 이종훈, 이근배
포항공과대학교
{hsseo, junion, palcery, jh21983, gblee}@postech.ac.kr

Grammar Error Detection System for Learners of Spoken and Written English

Hongsuck Seo○, Sungjin Lee, Jinsik Lee, Jonghoon Lee, Gary Geunbae Lee
Pohang University of Science and Technology

요 약

외국어 교육의 필요성이 강조되고 그에 대한 요구가 늘어남에 따라 언어 교육의 기회를 늘리고 비용을 줄이기 위해 컴퓨터 기반의 다양한 기술들의 요구 역시 증가하고 개발되고 있다. 언어 능력 개발의 중요한 요소로서 문법 교육에 대한 컴퓨터 지원 기술 연구가 활발히 진행되고 있다. 본 연구에서는 문법 오류 시뮬레이션을 통해 문법 오류 패턴 데이터베이스를 구축하고 이들 패턴과 사용자 입력의 패턴 매칭으로 생성된 자질 벡터로 기계 학습을 하여 문법성 확인을 했다. 문법성 확인 결과에 따라 오류 종류에 따른 상대 빈도를 고려하여 오류 종류를 분류했다. 또 말하기와 쓰기 작업의 서로 다른 특성을 반영하기 위해 말하기 작업과 쓰기 작업에 대한 두 개의 다른 말뭉치가 학습에 이용 되었다.

주제어: 문법 오류 검출, 영어 교육, 언어 교육

1. 서론

국제화 물결의 확산으로 영어 능력의 필요성이 점차 증대되고 영어 교육에 대한 수요가 늘어남에 따라 상대적으로 적은 비용으로 효과적인 영어 학습의 기회를 제공하기 위한 학습 도구, 교재 및 교육과정의 개발이 활발히 진행 되고 있다. 특히, 컴퓨터의 멀티미디어 컨텐츠 처리 능력과 논리적 처리 능력의 이점을 가진 교육용 소프트웨어 개발에 대한 요구도 증가하여 많은 영어 학습용 소프트웨어가 개발되고 있다.

정확한 문장으로 명확히 의사를 전달하기 위해서는 문법적 감각과 지식이 필수적이다. 그러나 대다수의 기존의 영어 학습용 소프트웨어는 어휘, 발음의 향상에 중점을 두고 있어 상대적으로 문법적 능력을 배양하는데 부족한 면이 있다. 현재 문법 오류를 판단하고 적절한 대안을 제시하는 방법론은 주로 워드 프로세서에 적용되어 쓰기 학습을 위한 교육용 컨텐츠와 말하기 학습을 위한 음성인터페이스와의 결합에 대한 고려가 현재까지는 깊이 이루어 지지 않았다. 따라서 문법적 지식을 학습하기 위한 영어 교육용 소프트웨어 개발할 필요가 있으며 이를 위해서는 문법오류 검출 및 수정 기술을 검토하고 이를 바탕으로 사용자에게 적절한 피드백을 줌으로써 문법적 지식의 학습 효과를 높일 수 있는 방법에 대한 연구가 필요하다.

본 고에서는 이러한 연구의 첫걸음으로서 학습자의 문법 수준 향상을 위해 사용자의 영어 말하기 및 쓰기 작업에서 자동으로 문법 오류를 검출하는 시스템을 소개한다.

2. 관련연구

언어 학습자의 문법 교육을 위한 다양한 컴퓨터 기반의 기술들이 개발되어왔다. 교육 자료 개발을 위해 기계학습 기술을 이용해 문법적으로 정확한 문장에서 문법 오류를 포함하는 문장을 생성하는 기술이 개발되었다[1]. 문법 오류 검출 기술의 경우 초창기에 문법 오류 규칙을 구문 분석기에 포함시켜 출력된 구문 트리가 문법 오류 규칙으로부터 생성되었는지를 판별하여 구조적 오류를 검출 했으며, 이후에는 관사, 전치사 등의 빈번한 오류에 보다 초점이 맞춰져 연구가 되어왔다[2,3,4]. 이때, 관사, 전치사 등의 오류는 구조적으로는 오류를 포함하고 있지 않는 것들로 기존의 방법으로는 검출 할 수 없는 부류들이다. 최근에는 기계 학습 기법을 이용하여 구조적 오류와 전치사, 관사 오류를 포함한 다양한 오류를 검출하는 많은 기술들이 연구되고 있다[5,6]. 이들 연구는 모두 학습자의 쓰기 작업에서의 문법 오류 검출로, 말하기 작업에서의 문법 오류 검출은 비교적 근래에 연구되기 시작되었다. 말하기 작업에서의 문법 오류 검출 기술들은 음성 인식기에 오류 경로를 추가하거나 패턴 매칭을 이용하여 문법 오류를 검출하는 방법 등이 연구되었다[7,8].

본 연구에서는 문법 오류 시뮬레이션으로부터 생성된 문법 오류 패턴 매칭과 기계 학습을 병합하여 학습자의 영어 말하기, 쓰기 작업상의 문법 오류 검출 시스템을 개발하였다.

3. 문법 오류 검출 기법

본 시스템은 문법 오류 시뮬레이션 기법을 이용하여

다양한 문법 오류 패턴을 생성하고 사용자 입력과 이들 패턴 간의 패턴 매칭으로부터 자질 벡터를 생성한다. 마지막으로 기계학습 기법을 이용하여 앞서 생성된 자질 벡터를 분류해 문법 오류를 검출하게 된다. 이번 장에서는 문법 오류 패턴 데이터베이스 구축을 위해 이용되는

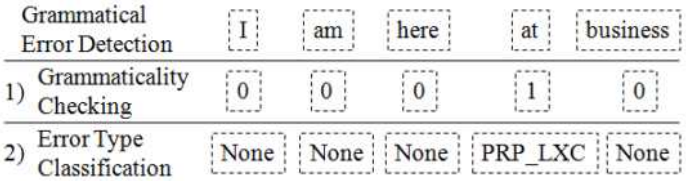


그림 1 문법 오류 검출의 두 하위 단계: 1) 각 위치의 문법성 확인, 2) 문법 오류의 종류 분류

문법 오류 시뮬레이션과 문법 오류 검출 과정인 문법성 확인과정 및 오류 종류 분류 과정을 설명한다.

3.1. 문법 오류 시뮬레이션

문법 오류 시뮬레이션은 상대적으로 새로운 연구 분야로, 문법 오류 검출 기술 개발을 위한 학습용 말뭉치 혹은 실험용 말뭉치 생성의 용도로 사용되어왔다. 그러나 기존에 사용되어 오던 방식들은 문장의 길이를 고려하지 않고 각 문장마다 오직 하나씩의 오류만을 생성하며, 이 과정에서 단지 품사 정보만이 포함되고 문맥 정보와 같은 고차원의 분석 정보가 고려되지 않으므로 생성된 문장이 실제 학습자의 문법 오류 특성을 그대로 반영하지 않는 문제점이 있다. 이러한 문제를 해결하기 위해 사람이 정의한 규칙을 이용하여 오류를 생성하되, markov logic을 이용해 절대적인 규칙이 아닌 문맥을 고려한 규칙으로, 오류가 표지 된 말뭉치로부터 문맥에 따른 규칙의 확률분포를 학습하고, 학습된 규칙에 대한 분포로 부터 오류를 생성해 내는 방법이 개발되었다[1]. 본 연구에서는 보다 실제적인 학습자의 문법 오류 패턴 데이터베이스를 구축하기 위해 문맥 의존 규칙을 이용하는 방법을 이용하나, 효율성의 측면을 고려하여 markov logic의 근사 모델로 maximum entropy 모델을 이용하였다.

3.2. 문법성 확인

본 연구에서 문법 오류 검출 과정은 문법성 확인과 오류 종류 분류 두 단계로 나뉜다.(그림 1) 문법성 확인 과정은 문장 내의 각 단어들에 대해 문법 오류가 있는지를 확인하는 과정으로, 패턴 매칭과 지지 벡터 기계(support vector machine; SVM)를 이용하여 문법성을 확인하며 문법성 확인 과정의 결과는 −1(오류 없음)과 +1(오류 있음) 둘 중 하나로 분류된다.

문법성 확인 과정에서는 먼저 문법 오류를 포함하는 기존의 말뭉치로부터 오류 단어를 중심으로 양쪽 2개씩의 단어를 포함하는 5단어 문법 오류 패턴 데이터베이스와 각 5개 단어의 품사로 구성되는 5품사 문법 오류 패턴 데이터베이스를 구축한다. 문법성을 확인하고자 하는 학습자의 문장이 입력되었을 때 입력 문장의 각 단어를 중심으로 한 5개 단어에 대해 패턴 데이터베이스의 패턴들로부터 7개의 자질 (표 1)을 추출하고 이 중, TS 자질을 기준으로 정렬 한 후 상위 10개의 단어 문법 오류 패턴 벡터와 10개의 품사 문법 오류 패턴 벡터를 합하여 최종적으로 SVM을 이용해 문법성을 확인한다[8].

표 1 문법 오류 패턴으로부터 추출된 자질

자질	설명
S1	패턴과 입력의 첫 번째 단어의 1)일치도
S2	패턴과 입력의 두 번째 단어의 일치도
S3	패턴과 입력의 세 번째 단어의 일치도
S4	패턴과 입력의 네 번째 단어의 일치도
S5	패턴과 입력의 다섯 번째 단어의 일치도
TS	S1, S2, S3, S4, S5의 합
SD	오류 종류 구분자 (치환 오류의 경우 0, 삭제 오류의 경우 1)

3.3. 오류 종류 분류

주어진 입력 문장의 어떤 단어에 대한 문법성 확인의 결과가 +1일 경우, 다음과 같이 오류 종류를 분류하게 된다.

$$Score(e) = TS(e) + a * EF(e)$$

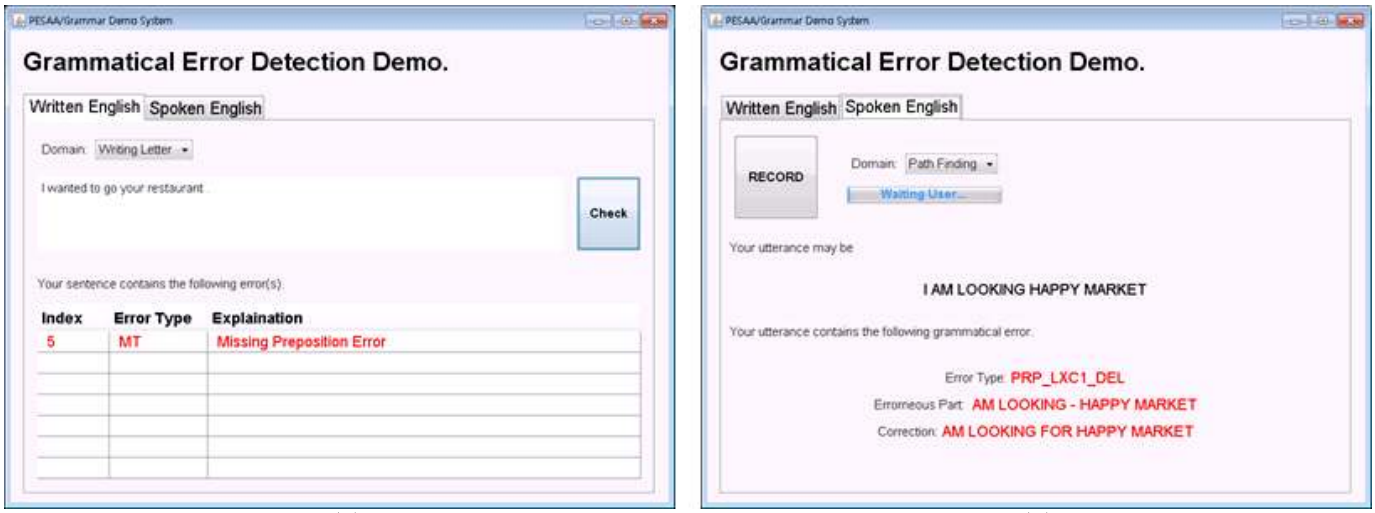
이때, $TS(e)$는 에러 패턴 e의 TS 자질을 의미하며, $EF(e)$는 에러 패턴 e의 오류 종류에 대한 상대빈도수를 의미한다.

오류 종류는 주어진 문법 오류 패턴 중 TS가 가장 높은 것만을 선택하지 않고, 각 패턴의 오류 종류에 따라 다른 추가 점수를 더한 값을 이용하여 오류 종류를 결정한다[8].

4. 말하기, 쓰기 문법 오류 검출 시스템의 구현

본 연구에서는 말하기, 쓰기 작업에 대한 문법 오류 검출 시스템을 개발 했다.(그림 2) 문법 오류 검출 시스템은 두 개의 말하기, 쓰기 작업 각각 대해 독립적으로 처리하는 두 개의 독립된 탭으로 구성되어있다. 쓰기 작업은 입력 상자에 문법 오류 검출을 원하는 문장을 넣고 'Check' 버튼을 누르게 되면 하단에 검출된 오류의 정보를 출력한다.(그림 2−a) 말하기 작업은 'Record' 버튼을 누르면 마이크를 통해 녹음이 시작되며, 녹음된 발화에서 오류가 검출되면 오류 정보를 하단에 출력한다.(그림 2−b) 말하기와 쓰기 작업의 서로 다른 특성을 문법 오류 검출 과정에 반영하기 위해 두 개의 문법 오류가 표지 된 말뭉치가 각각 사용했다.

1) 일치도는 말하기의 경우 입력 발화의 음성인식 결과인 confusion network 내 각 단어의 신뢰도로 하며, 쓰기의 경우 단어가 일치할 경우 1 그렇지 않을 경우 0으로 한다.



(a) (b)

그림 2 쓰기 문법 오류 검출 시스템(a)과 말하기 문법 오류 검출 시스템(b)

4.1. 말하기

학습자의 말하기 작업에 문법 오류 시뮬레이션 학습을 위해 Japanese learner English (JLE) 말뭉치가 이용되었다[9]. JLE 말뭉치는 일본인 학습자들의 말하기 작업을 전사하여 문법 오류를 표시한 말뭉치이다. 학습된 문법 오류 시뮬레이션 기술을 이용해 길 찾기 도메인에서의 문법 오류가 없는 문장들로부터 문법 오류를 포함하는 문장을 생성한다. 생성된 문장들의 일부는 문법 오류 패턴 데이터베이스를 구축하는데 사용하였고 나머지는 SVM 학습에 사용하였다.

말하기 데이터에서의 사용자 입력은 하나의 문장이 아닌 발화 문장에 대한 confusion network (CN)로 자질 벡터를 구성할 때 각 단어에 대한 일치도(S1~S5)는 CN 내 일치하는 단어의 신뢰도([0,1])를 이용한다.

4.2. 쓰기

Cambridge Learner Corpus (CLC) 말뭉치 중 first certificate in English (FCE) 시험을 표지 한 CLC FCE 말뭉치를 이용해 말하기 작업에 대한 오류 시뮬레이션을 학습 학습했다[6]. 학습된 문법 오류 시뮬레이션 기술을 통해 편지 형식의 문법 오류가 없는 문장들로부터 문법 오류를 포함하는 문장을 생성해 문법 오류 패턴 데이터베이스 구축과 SVM 학습에 사용되었다.

쓰기 작업에서 문법성 확인의 자질 벡터를 구성할 때 일치도(S1~S5) 자질은 오류 패턴의 단어와 사용자 입력의 단어가 정확히 일치할 경우 1, 그렇지 않을 경우 0으로 하여 채워졌다.

5. 결론 및 향후 계획

본 연구에서는 문법 오류 시뮬레이션 기술과 패턴 매칭 및 기계학습을 이용하여 영어 학습자의 말하기 및 쓰기 문법 오류 검출 시스템을 개발하였다. 말하기와 쓰기의 다른 특성을 반영하기 위해 말하기 작업에 대한 말뭉치와 쓰기 작업에 대한 말뭉치가 각각의 문법 오류 검출을 위해 사용되었다.

학습자의 모국어에 따라 다른 분포로 문법 오류를 발생 시키기 때문에 앞으로 한국인의 문법 오류 검출을 위해 한국인의 말하기, 쓰기 작업에 대한 문법 오류가 표지 된 말뭉치를 구축하고 한국인 교육에 특화된 문법 오류검출 시스템 개발이 이루어질 예정이다.

Acknowledgement

"이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2011-0027953)."

참고문헌

[1] Ssungjin Lee, Jonghoon Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee, Grammatical error simulation for computer-assisted language learning, Knowledge-Based System, 24, 868-876, 2011.

[2] Heift, T. and Schulze, M., Errors and Intelligence in CALL, Parsers and Pedagogues, 2007.

[3] Ryo Nagata, Tatsuya Iguchi, Kenta Wakidera, Fumito Masui and Atsuo Kawai, Recognizing article errors in the writing of Japanese learners of English, Systems and Computers in Japans, 37, 7, 60-68, 2005a.

[4] Joel Tetreault and Martic Chodorow, The ups and downs of preposition error detection in ESL writing, In proceedings of the 22nd international conference on computational linguistics, 865-872, 2008.

[5] Gamon, M., High-Order Sequence Modeling for

Language Learner Error Detection, Proceedings of the 6th workshop on innovative use of NLP for building educational application, 180-189, 2011.
[6] A new dataset and method for automatically grading ESOL texts, HLT '11 Proceedings of the 49th annual meeting of the association for computational linguistics: Human Language Technologies, 180-189, 2011.
[7] Morton, H. and Jack, M., Scenario-based spoken interaction with virtual agents. Computer Assisted Language Learning, 18, 3, 171-191, 2005.
[8] Sungjin Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee, Grammatical error detection for corrective feedback provision in oral conversations, Proceedings of the 25th AAAI conference on artificial intelligence, 2011.
[9] E. Izumi, K., Uchimoto, H. Isahara, Error annotation for corpus of Japanese Learner English, Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora, 71-80, 2005.