병렬 말뭉치를 이용한 CEFR 기반 문장 작문 평가

최승권 [1], 권오욱 [1]
[1] 한국전자통신연구원 언어지능연구실 책임연구원

choisk@etri.re.kr, ohwoog@etri.re.kr

CEFR-based Sentence Writing Assessment using Bilingual Corpus

Sung-Kwon Choi[1], Oh-Woog Kwon[1]
[1]Language Intelligence Research Section, ETRI

요 약

CEFR(Common European Framework of Reference for Language)는 유럽 전역의 교육기관에서 언어 구사 능력을 평가하는 평가 기준이다. 본 논문은 학습자가 문장 작문한 것을 CEFR 에 기반하여 평가하는 모델을 기술하는 것을 목표로 한다. CEFR 기반 문장 작문 평가는 크게 전처리 단계, 작문 단계, 평가 단계로 구성된다. CEFR 기반 문장 작문 평가 모델의 평가는 CEFR 수준별로 분류한 문장들이 전문가의 수동 분류와 일치하는 지의 정확도와 학습자가 작문한 결과의 자동 평가로 측정되었다. 실험은 독일어를 대상으로 하였으며 독일어 전공 41 명의 대학생에게 CEFR 6 등급별로 5 문장씩 총 30 문장의 2 세트를 만들어 실험을 실시하였다. 그 결과 CEFR 등급별 자동 분류는 전문가의 수동 분류와 61.67%로 일치하는 정확도를 보였다.

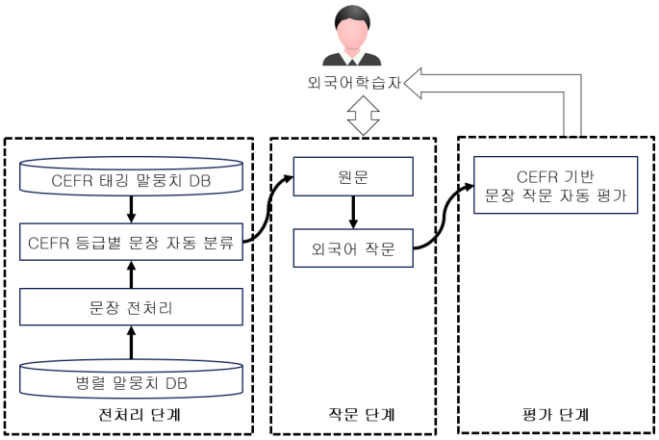
1. 서론

작문 자동 평가 방법은 작문의 크기에 따라 에세이(Essay) 평가와 문장 평가로 나눌 수 있다. 에세이 작문 자동 평가는 언어 사용의 정확성이나 글의 전반적인 내용과 구성 등을 종합적으로 평가하는 방법인 반면, 문장 작문 자동 평가는 내용의 정확성에 초점을 맞추어 특정 단어나 내용에 대한 유무에 초점을 두고 평가하는 방법이다. [1] 에세이 작문 자동 평가는 인공지능 기술을 연계시켜 구문, 의미, 수사학적 특징을 추출하고 딥러닝(deep leaning)에 의해 점수와 피드백을 생성하는 수준으로 발전하고 있으며[2] 문장 작문 자동 평가는 문장의 단어 길이, 문법 오류와 같은 기계적인 자질을 바탕으로 규칙 기반의 시스템이 연구되었다. [3]

작문 자동 평가는 딥러닝 기술이 등장하면서 문장 작문 자동 평가에서 에세이 작문 자동 평가로 확장되고 있다. 그러나 문장의 복잡한 의미를 확인하는 기술적 한계로 에세이를 전적으로 자동 평가하는 것은 아직도 연구 중이다. [4] 게다가 딥러닝 기술에 의한 문장 작문 자동 평가를 위해서는 문장마다 평가 점수가 태깅된 대량의 문장 작문용 학습 데이터가 구축되어 있어야 하나 대부분은 문서에 적용된 평가 점수를 문서의 모든 문장에 동일하게 적용하여 사용함으로써 올바른 문장 작문 자동 평가가 이루어지지 않았다. [5]
이런 점에서 본 논문에서는 CEFR 점수가 부여된 소량의 문서를 이용하여 CEFR 등급 자질을 추출한 후 대량의 병렬 말뭉치에 적용하여 문장 작문 평가를 하는 모델을 기술하는 것을 목표로 한다.

2. CEFR 기반 문장 작문 평가 모델 구성도

CEFR 기반 문장 작문 평가는 크게 세단계로 구성된다. 첫 번째 단계는 전처리 단계로 CEFR 등급이 태깅된 소량의 말뭉치를 토대로 CEFR 등급 언어 자질을 추출하여 다국어 병렬 말뭉치의 문장에 CEFR 등급을 부여하는 단계이다. 두 번째 단계는 작문 단계로 학습자가 병렬 말뭉치의 원문을 외국어로 작문하는 단계이다. 세 번째 단계는 평가 단계로 작문 결과에 대해 학습자의 CEFR 실력을 자동으로 평가하는 단계이다. 전체적인 구성도는 다음과 같다.



(그림 1) CEFR 기반 문장 작문 평가 모델 구성도

3. 전처리 단계

3.1. CEFR 태깅 말뭉치 DB

CEFR 는 6 등급으로 구분되며 각 등급의 의미와 TOEIC 점수와의 상호 관계성을 기술하면 다음과 같다. [6]

<표 1> CEFR 6 등급

CEFR	수준	설명	TOEIC
A1	입문	일상 생활에서 사용되는 간단한 영어표현을 이해하고 구사할 수 있다.	
A2	초급	자주 사용되거나 직접적으로 연계된 언어적 표현을 이해하고 구사할 수 있다.	500
B1	중급	친숙한 주제들을 표준어로 분명하게 표현해주면 대화의 핵심을 이해할 수 있다.	600
B2	중상급	세부적이고 추상적인 상당히 복잡한 글을 이해하고 일부 분야에 대해서는 전문적인 토론이 가능하다.	800
C1	상급	긴 담화 내용을 들으면서 그 텍스트 구조가 복잡하고 명확하지 않아도 내재된 의미를 잘 이해하고 정리할 수 있다.	900
C2	고급	해당 언어를 사용하는 원어민과 어려움 없이 실질적인 대화를 할 수 있다.	

CEFR 등급이 태깅되어 있는 독일어 말뭉치로 MERLIN[7]이 있다. MERLIN 은 교사의 작문 지시 내용에 따라 외국인 학생들이 독일어 편지를 작문하고 교사가 그 독일어 작문 편지를 CEFR 등급으로 평가한 말뭉치이다.

3.2. CEFR 등급별 문장 자동 분류

CEFR 태깅 말뭉치 DB 를 이용하여 CEFR 등급에 영향을 끼치는 문장 난이도를 측정하고자 하였다. 이 문장 난이도를 측정하기 위해 본 논문에서는 언어에 관계없이 기계적으로 추출이 가능한 문장 난이도를 결정할 수 있는 언어적 자질을 얻었으며 문장 난이도를 결정하는 언어 자질로 다음과 같은 자질들을 사용하였다.

<표 2> 문장 난이도 결정 자질

문장 난이도 결정 자질	작문 난이도 예측
문장의 문자 수	문자 수가 많은 경우, 다양한 어휘가 나타날 확률이 높으므로 작문 난이도가 높을 것이다.
문장의 단어 수	단어 수가 많은 경우, 구문론적 복잡성이 커서 작문 난이도가 높을 것이다.
문장의 저빈도 어휘 수	저빈도 어휘 수가 높을 경우, 어려운 어휘가 발생한 것이므로 작문 난이도가 높을 것이다.

자질들의 값의 단위와 분포가 다르기 때문에 상호 비

교하기 쉽게 각 자질들을 정규화하였다. 문장 난이도 측정 식은 다음과 같았다:

$$\text{문장의\_문자수\_정규화} = (x - x_{min}) / (x_{max} - x_{min})$$
$$\text{문장의\_단어수\_정규화} = (y - y_{min}) / (y_{max} - y_{min})$$
$$\text{문장의\_저빈도어휘수\_정규화} = (z - z_{min}) / (z_{max} - z_{min})$$

$$\text{문장 난이도} = (w1 * \text{문장의\_문자수\_정규화} + w2 * \text{문장의\_단어수\_정규화} + w3 * \text{문장의\_저빈도어휘수\_정규화})/3$$

위의 식에서 x 는 해당 문장의 문자수, y 는 해당 문장의 단어수, z 는 해당 문장의 저빈소 어휘수를 의미하며 min 은 해당 데이터 중 최소값을 max 는 해당 데이터 중 최대값을 의미한다. 문장 난이도 결정 자질은 작문에 미치는 영향을 모두 동일한 것으로 간주하여 각 자질들의 w1, w2, w3 가중치를 모두 동일하게 1 로 부여하였다. MERLIN 말뭉치로부터 CEFR 등급별로 문장 난이도를 측정한 결과는 다음과 같았다.

<표 3> MERLIN 독일어 작문 말뭉치의 CEFR 등급별 문장 난이도

CEFR	작문 문장수	문장 난이도	문장의 평균 단어수
A1	329	0.000000~0.162595	6.705521
A2	2,531	0.162596~0.199564	7.871515
B1	3,966	0.199565~0.283905	10.115661
B2	4,299	0.283906~0.399620	12.567256
C1	629	0.399620~0.463466	13.778589
C2	54	0.463466~	16.793103
계	11,808		

3.3. 병렬 말뭉치 DB

2018 년 평창 동계올림픽의 다국어 음성통역 서비스를 위해 한국정보화진흥원에서 구축한 한국어-독일어 50,000 문장과 웹으로부터 수집한 112 문장을 병렬 말뭉치로 활용하였다.

3.4. 문장 전처리

병렬 말뭉치와 CEFR 태깅 말뭉치의 결측치 제거, 통계적 이상치 제거, Umlaut 와 같은 특수문자 처리가 이루어졌다.

4. 작문 단계

학습자가 한국어-독일어 병렬 말뭉치 DB 로부터 각 등급별로 작문하려는 문장수를 정하면, 임의로 추출된 한국어 문장이 학습자에게 제공되며 학습자는 한국어 문장만 보고 독일어 작문을 하고 독일어 정답문은 보이지 않도록 하였다.

5. 평가 단계

5.1. CEFR 기반 문장 작문 자동 평가

작문 문장과 정답문의 정확도를 측정하는 자동 평가 방법 중에 가장 많이 사용하는 방법이 BLEU[8]와 ROUGE[9]이다. BLEU 는 작문 문장에 촛점을 맞추어 작문이 정답 문장과 N-gram 일치하는 정확률(precision)을 측정하는 평가 지표로 BLEU 점수가 높을수록 정답 문장과 유사한 것으로 간주한다. 이에 반해, ROUGE 는 정답 문장에 초점을 맞추어 정답 문장이 작문과 N-gram 일치하는 재현률(recall)을 측정하는 평가 지표이다. ROUGE 는 일반적으로 텍스트 요약 모델의 성능을 평가하는데 사용되는데 ROUGE 의 다양한 평가 지표 중에 ROUGE-L 은 가장 길게 겹치는 것을 평가할 수 있기 때문에 BLEU 에는 없는 어순 평가를 반영할 수 있다. 독일어 학습자의 작문 실력을 자동으로 평가하기 위한 식으로 BLEU 값과 ROUGE-L 값의 평균을 사용하였다.

$$\text{문장작문자동평가} = ((\text{BLEU}(\text{작문문장}, \text{정답문}) + \text{ROUGE-L}(\text{작문문장}, \text{정답문}))/2) \times 100$$

6. 실험

6.1. 실험 세트와 실험자

실험을 위해 CEFR 등급을 알고 있는 독일어 전공 대학생들을 실험자로 활용하였다. 실험자에게는 실험

의 목적에 대해 설명하였으며 개인의 동의 하에 실험을 실시하였다. 실험자는 총 41 명이었으며 A1 23 명, A2 4 명, B1 8 명, B2 6 명이었다. C1 이상의 실력을 가진 학생의 경우는 찾을 수가 없어서 배제하였다. 실험자에게는 각 등급별로 임의로 5 문장씩을 추출하여 총 30 문장의 2 세트를 제공하고 독일어로 수동 작문을 하도록 요청하였다. 수동 작문 시간은 제약을 두지 않았다.

6.2. 문장 난이도 기반 CEFR 분류 정확도

작문 실험을 하기 위해 학생들에게 주었던 30 문장의 2 세트 실험 데이터를 임의로 섞어서 독일어능력시험의 평가 경력이 있는 독일어 전문가에게 제공하고 CEFR 분류를 요청하였다. 그 결과 전문가의 분류와 자동 분류가 일치한 정확도는 다음과 같이 측정되었다.

<표 4> 문장 난이도 기반 자동 분류의 정확도

CEFR	문장수	일치한 문장수	정확도
A1	10	7	70%
A2	10	8	80%
B1	10	5	50%
B2	10	7	70%
C1	10	5	50%
C2	10	5	50%
계	60	37	61.67%

전문가의 수를 더 확보하여 수동 분류를 하였다면 신뢰도가 높은 정확도가 측정되었을 것이다. 하지만 전문가 확보의 어려움을 감안하여 전문가 1 인이 측정한 문장 난이도 기반 CEFR 수준 자동 분류 정확도는 61.67%가 나왔다.

6.3. 문장 난이도 기반 문장 작문 정확도 평가

실험자 41 명이 30 문장 2 세트의 실험 문장을 독일어로 작문한 것을 자동으로 평가한 결과는 다음과 같았다.

<표 5> 학습자 수준과 작문 자동 평가의 상관 관계

CEFR	A1	A2	B1	B2	C1	C2	전체 평균	실력 평균
A1	**25.13**	21.59	20.70	12.04	12.88	15.46	17.97	25.13
A2	**45.15**	**35.80**	37.03	21.85	20.33	20.90	30.18	40.48
B1	**38.28**	**28.96**	**33.44**	19.29	21.51	21.40	27.15	33.56
B2	**40.53**	**31.65**	**33.25**	**19.38**	21.92	21.70	28.07	31.20
계	37.27	29.50	31.10	18.14	19.16	19.86	25.84	

도표에서 첫 번째 칼럼 ‘CEFR’은 학습자 본인의 CEFR 등급을 말한다. 두 번째 칼럼 ‘A1’부터 ‘C2’까지는 학습자의 작문을 자동 평가한 점수이다. ‘전체평균’은 A1 부터 C2 까지의 전체 평균을 의미한다. ‘실력평균’은 학습자의 CEFR 실력까지의 평균을 말하며 도표에서는 굵게 표시되어 있다. 예를 들어 B1 실력 학습자의 ‘실력 평균’은 A1, A2, B1 점수의 평균인 것이다. 따라서 A1, A2, B1, B2 학습자의 ‘실력 평균’은 각각 25.13%, 40.48%, 33.56%, 31.20%였으며 ‘전체 평균’은 25.84%였다. 전체적으로 학습자들의 작문 실력이 낮다는 것을 알 수 있다. 그 이유로 추측할 수 있는 것은 학생들이 제시한 자신의 CEFR 실력이 잘못 기재되었을 수 있다. 그럼에도 불구하고 표로부터 관찰할 수 있는 점은 ‘실력 평균’로부터 자신의 실력을 파악할 수 있다는 것이다. 즉 자신의 실력이라고 알고 있는 CEFR 수준보다 ‘실력 평균’이 낮다면 자신의 CEFR 수준을 하향 조정해야 한다는 것이다.

7. 결론

본 논문은 병렬 말뭉치를 이용하여 CEFR 등급별 문장 작문 실력을 자동으로 평가할 수 있는 모델을 제안했다는 점에서 그 의의가 있다고 할 수 있다. 하지만 앞으로 기술적으로나 교육적으로 개선할 부분이 많다. 기술적으로 개선할 점으로는 BLEU, ROUGE-L 의 측정에 사용되는 병렬 말뭉치의 정답수를 Paraphrase 생성 기술[10]이나 ChatGPT 를 이용하여 확장하는 것이다. 왜냐하면 정답이 현재 1 개이기 때문에 학습자의 작문 결과가 의미적으로 맞지만 정답과

일치하지 않을 경우, 잘못 평가하는 결과가 나올 수 있기 때문이다. 교육적으로 개선할 점으로는 CEFR 등급에 따라 작문 점수는 제공했지만, 학습자의 작문 오류에 대한 언어학적 피드백을 제공하지 못했다. 이 언어학적 피드백에 관한 연구는 추후 더 연구되어야 할 부분이다.

Acknowledgement

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

참고문헌

[1] 장지현. “머신 러닝 기법을 활용한 영어 에세이 자동채점 방안 연구”, 서울대학교 대학원 교육학과 교육학전공 박사학위논문, 2021.

[2] Wilson, J. and Roscoe, R. “Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy”, Journal of Educational Computing Research, 58(1), 87-125, 2019.

[3] 엄진희, 곽동민. “기계학습기법을 이용한 영어작문 문장 수준평가 시스템”, 제 40 회 한국정보처리학회 추계학술발표대회 논문집, 제 20 권 2, 1290-1293, 2013.

[4] 이경건, 하민수. “인공지능 기반 자동평가의 현재와 미래: 서술형 문항에 관한 문헌 고찰과 그 너머”. 교육공학연구, 36 권 2 호, 353-382, 2020.

[5] Arase, Y., Uchida, S., and Kajiwara T. “CEFR-BASED Sentence Difficulty Annotation and Assessment”. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 6206-6219, 2022.

[6] https://www.cefr.co.kr/cefr

[7] Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schone, K., Stindlova, B., and Vettori, C. “The MERLIN corpus: Learner language and the CEFR”, LREC, 1281−1288, 2014.

[8] Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. “BLEU: a method for automatic evaluation of machine translation”, Proceedings of the 40th annual meeting on association for computational linguistics, 311-318, 2002.

[9] Lin, C. “ROUGE: A Package for Automatic Evaluation of Summaries”. Association for Computational Linguistics, 74-81, 2004.

[10] Zhou, J. and Bhat, S. “Paraphrase Generation: A Survey of the State of the Art”, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 5075-5086, 2021.