https://doi.org/10.24303/lakdoi.2021.29.2.1

딥러닝 기반 단어 임베딩을 적용한 사진 자막 영작문 채점 시스템

김동성
(이화여자대학교)

Kim, Dongsung. (2021). Automatic scoring system for picture-based English caption writing test adopting deep learning based word-embedding. *The Linguistic Association of Korea Journal, 29*(2), 1-20. Since human grading of English writing requires substantial resources, many researchers in the area of Computer-Assisted Language Learning (CALL) have been focusing on automatic scoring systems based on natural language processing systems, machine learning, and other automatic processing mechanisms. English Testing Services (ETS) announced several automatic scoring systems for English writing. In this paper, we suggest using a deep learning based automatic scoring system for an English caption writing test. Our method involves using a sentence similarity measurement, which compares different levels of answer sentences with user writing input. We chose different word embedding types (Word2Vec, Word Mover's Distance (WMD), Bidirectional Encoder Representations from Transformers (BERT)) and Abstract Meaning Representation (AMR), a linguistic model for comparing semantic differences between two sentences based on semantic representation. Scoring systems should not only satisfy the requirements of complicated scoring rubrics but also meet the conditions of a language proficiency test. Our results show that BERT outperforms three competitive models in predicting accurate scoring levels and also shows the characteristics of the criterion reference which could theoretically express the standards of a language proficiency test.

주제어(Key Words): 컴퓨터 언어보조학습(computer assisted language learning), 딥러닝(deep learning), 영작문(English writing), 단어 임베딩 (word embedding), 준거참조검사(criterion-referenced test), 채점(scoring)

1. 서론

이 논문은 사진에 기반을 둔 단문 영작문 채점 시스템에 대한 연구이다. 숙련된 영어교육 전문가가 영작문 채점을 담당하면 운영·유지에 대한 많은 비용·재원이 발생하므로 자동 채점 도입이 절실히 요구된다(Wiegle, 2010, 2011, 2013). 이러한 연유로 자동 채점 시스템에 대한 여러 연구들이 있었다. 국외에서는 ETS (Educational Testing Service) (Attali & Burstein, 2006), 국내에서는 한국교육과정 평가원(진경애, 2007)을 포함한 여러 연구들이 발견된다(김지은 & 이공주, 2007; 김동성 외, 2008; 김동성, 2016).

일반적인 영작문 평가는 형태소, 통사, 어휘·의미 등에서 오류를 각각 탐지하고 이에 작문 구성 요소들도 고려해서 이루어진다. 다양한 채점 기준(scoring rubrics)이 다면적으로 복잡하게 얽혀져 있기 때문에 단문 영작문 채점 시스템의 경우에도 복합적인 채점 모델이 구성되고 여러 평가 자질들이 사용된다(Sukkarich & Stoyanchev, 2009). 단문 채점을 하는 ETS c-rater 시스템의 경우에 전통적 방식에 따라 사용자 입력문장에서 형태소, 통사, 어휘·의미 오류를 탐지하고, 이를 작문 구성 요소들과 결합해서 평가한다(Sukkarich & Stoyanchev, 2009). ETS 사진 자막 서술 자동 채점에서는 45개의 많은 평가 자질들을 사용한다 (Somasundaran 외, 2015).

문제는 ETS 영작문 자동 채점 시스템처럼 많은 평가 자질이 활용되면 더 복잡한 구조의 시스템을 만들어야 하는데, 이러한 구성이 효율적인지에 대한 의구심이 발생한다. 복잡한 구조가 더 정확하고 공정한 채점을 보장하는 것은 아니기 때문이다. 연구에서는 더 단순한 처리 방식이면서도 다양한 언어 현상을 포괄적으로 다룰 수 있는 딥러닝 기반의 단어 임베딩을 활용한 효율적인 자동 채점 시스템을 제안하고자 한다. 기존의 영작문 채점 시스템이 형태·통사·의미 분석이라는 언어학에 기초한 복합 처리 방식에 기초한다면 딥러닝 기반의 단어 임베딩 모델은 사용자 입력과 문장 유사도 비교라는 단순한 처리 방식으로도 정확도가 높은 수준의 영작문 채점이 가능하다. 또한 본 연구에서는 전통적인 언어학적 의미·통사 구조에 기초한 AMR (Abstract Meaning Representation)과 비교하여 효용성도 입증하고자 한다.

언어 능력 시험이나 일반적인 시험에서 검사 방식에는 준거참조검사(Criterion-Referenced Test)나 규준참조검사(Norm-Referenced Test)가 있다. 규준참조검사는 개인 검사점수가 전체 집단에서의 상대적 위치로 평가되며, 준거참조검사는 일정 기준에 따라 검사점수를 평가하는 절대평가 검사방식이다. TOEFL, TOEIC과 같은 언어 능력 평가는 준거참조방식으로 일정 기준에 도달했는지가 절대적으로 평가되어져야 한다. 영작문 채점 시스템도 언어 능력 평가의 일환으로 준거참조검사를 입증되어져야 한다. 본 연구에서 이 점에 주목하고 딥러닝 기반 모델 중 특정 모델이 준거참조검사가 가능함을 통계적으로 입증할 것이다.

본 연구에서는 단문 영작문 데이터를 수집하기 위해서 특정 사진을 제시하고 이에 연

관된 대량의 사용자 문장들을 수집했다. 이를 대상으로 채점 기준에 따라 2명의 채점자가 기준이 되는 채점 작업을 했으며, 이것을 토대로 여러 모델들을 적용해서 문장 간 유사도를 측정하고 자동 채점을 수행했다. 활용한 모델들은 딥러닝 기반의 단어 임베딩 모델들인 Word2Vec, WMD (Word Mover's Distance), BERT (Bidirectional Encoder Representations from Transformers)를 활용했다. Word2Vec, WMD는 초기 모델이고 BERT는 최근 각광받고 있는 모델로 기학습된(pre-trained) 딥러닝 모델들을 적용했다. 이러한 통계적 모델에 언어학 기반의 논리 구조 유사성을 측정하는 AMR 모델도 포함했다. 실험에서는 모델들이 예측한 것과 인간 채점과 얼마나 유사한지를 살펴보고 언어 능력 평가 시험이 담보해야 할 준거참조검사를 어떻게 설명할 수 있는지 통계적 방식으로 살펴보았다.

본 논문의 구성은 다음과 같다. 2절은 기존 연구에 대한 논의이다. 3절은 연구 방법으로 연구 자료와 분석 방법에 대한 소개이다. 4절은 연구 결과 및 논의이며 5절은 이 논문의 결론이다.

2. 이론적 배경

2.1. 영작문 자동 채점 시스템

언어학습에서 쓰기 교육은 듣기, 말하기, 읽기 등 다른 언어능력 영역의 발전과 연관되며 효과적인 의사소통 능력과 연결된다(Rivers, 1981). 제2언어 학습에서 학습자의 문법, 어휘, 형태 등의 언어 능력이 작문과 연관되어 쓰기 영역에서 나타난다(Wiegle, 2013). 그런데 제2언어 쓰기 수업에서 수업자가 담당해야 할 채점 분량이 많아서 Wiegle (2010, 2011, 2013)은 작문 교육에 자동 채점 시스템이 요구된다고 주장했다. 이러한 시스템은 교수자의 업무를 덜어주고 학습자에게 빠른 피드백을 전달할 수 있는 방편이 된다.

TOEFL, TOEIC 시험에 영작문이 포함되기 때문에 ETS는 수많은 영작문 데이터를 채점하기 위해 많은 인력·재원·시간을 소모하고 있다. 이러한 문제점을 해결하기 위해서 ETS는 자동 채점 시스템을 오랫동안 연구하고 있으며, 에세이 형식의 작문채점을 위해서 e-rater를, 사진 자막과 같이 짧은 단문 유형의 채점을 위해서 c-rater를 개발·출시했다. 사진자막 영작문을 채점하는 c-rater는 다음과 같은 처리 방식을 갖고 있다. 먼저 사진자막 영작문 채점을 위한 정답 문장들을 생성하고 이를 기준으로 입력된 문장과 비교한다(Sukkarich & Stoyanchev, 2009). 비교 방식은 정답 문장과 학습자나 시험자가 입력한 문장을 토대로 얼마나 유사한 문장을 입력했는지 평가한다(Somasundaran 외, 2015). ETS의 사진자막 영작문 채점기준(scoring rubrics)에 기초해서 문장 간 유사도 측정을 한다(민선식, 2008). C-rater의 채점 기준은 사진과 연관성 측정, 시험자 사용 단어와 필수 답안 단어와의 관련성, 문제에서 주어진 논조와 적합 여부, 시험자 문장내의 논리적 연결성, 서술적 관점과 주어진 지문

과의 연관성 등 총 5가지 영역과 그 외의 채점 9가지 영역을 조합해서 총 45개로 복잡하게 구성된다.

컴퓨터보조 언어학습(Computer-Assisted Language Learning; 이하 CALL)은 단순히 인력·재원·시간 소모를 대체하는 것이 아니라 언어학습의 일환으로 정교한 사용자 오류 피드백을 지원을 포함해서 언어능력 향상 목적으로 활용된다(Nagata, 1996; Heift & Schulze, 2007). 이 분야의 연구자들은 현재 더 효율적인 언어 학습을 위해서 지능적 방법론인 ICALL (Intelligent Computer-Assisted Language Learning, 지능적 컴퓨터 보조 학습)을 언어학습자 피드백·평가·교정에 도입했다(Heift & Schulze, 2007). Condon (2009)은 글쓰기는 일반적인 의사소통 능력을 넘어서 전반적인 교육 체계를 점검하는 방법론이라고 주장하며, 더 나아가서 대학교육 이상에서 지식습득을 평가하기 위해서 주어진 시간 안에 논술형 시험이 적절히 평가되어져야 하므로 평가 방법론으로 자동 채점이 활용되어져야 한다고 주장했다.

언어자질을 활용한 채점은 여러 연구가 있으나 국내 연구에서는 김동성(2016)과 김동성 외 (2008)의 연구가 발견된다. 김동성(2016)에서는 사진 자막 영작문 채점을 위해서 데이터를 모으고, 채점 데이터로 분류해서 언어학적 유사도를 활용한 채점방식을 제안했다. 김동성(2016)에서 활용한 언어학적 방식은 언어학적 의미·통사 해석구조인 AMR로 다면채점 기준들 중에서 가장 주요한 기준이 된다는 것을 통계적으로 입증했다. 김동성 외 (2008)에서는 에세이 형식의 영작문에 문법성, 어휘 응집성을 채점 기준으로 활용한 총괄적 평가 방식을 적용했다.

Settles 외(2020)에서 자연어처리 시스템이나 기계학습의 도입이 기존 언어적 특징이 고려되어져야 하는 언어능력 시험 개발에서 많은 도움을 줄 것이라고 주장했다. 더 나아가서 준거참조검사나 규준참조검사를 구분하면서 준거참조검사가 필요한 언어능력 시험 개발에 자연어처리나 기계학습이 이점으로 작용한다고 주장했다.

영작문 채점을 위해서는 학습자의 오류를 발견하고 오류 수준에 대한 평가가 필요하다. 학습자 오류를 발견하는 방식은 크게 두 가지로 구분된다(Leacock 외, 2014). 하나는 일반적이고 전통적인 언어학에 기초한 절차적 방식이다. 사용자가 입력한 문장을 형태소, 통사, 어휘·의미 등의 언어학적 분석 순서에 따라 각각의 오류를 탐지하고, 이를 작문 구성 요소들과 결합해서 언어학적 분석 순서에 따른 절차적 처리를 활용한다. 다른 하나는 비절차적 방식으로 기계학습 등을 활용한 데이터 처리 기반이다. 학습자 코퍼스 데이터를 수집하고 이를 기반으로 영작문 평가를 위한 통계 기반 모델을 생성하고 사용자가 문장을 입력하면 통계 모델을 활용해서 오류를 탐지한다.

절차적 방식이나 비절차적 방식 모두 여러 채점 기준들이 있으며, 이를 종합적으로 판단해서 채점이 신뢰성을 갖도록 하고 있다. 만약 많은 채점 기준들로 평가해야 한다면 복잡성의 증가로 문제점이 발생한다. ETS c-rater의 경우에 45개의 채점 기준은 45개의 다면적 분석이 필요하다(Leacock & Chodorow, 2003). 수리적인 측면에서 다면적 분석은 다차원 해

석(multi-dimensional analysis)을 요구하므로 차원 왜곡(dimension distortion) 문제점이 발생한다.[1)]

2.2. 단어 임베딩 딥러닝 모델

본 연구는 단어 임베딩 모델을 활용하는데, 단어 임베딩 기술은 최근 각광받고 있는 딥러닝 모델로 이에 대한 자세한 이해를 위해 간략하게 설명하고자 한다. 단어 임베딩은 딥러닝에서 코퍼스에서 단어가 분포하는 특성을 표현하는 모델이다. 모델의 이론적 출발은 언어학적 이론인 분포가설(distributional hypothesis)에서 비롯한다. 언어학자인 Firth (1957)는 단어의 의미는 주변 단어에 의해서 결정된다는 분포가설을 주장했으며, 이 모델을 가장 극대화 시킨 자연어처리 모델이 단어 임베딩 모델이다(Manning, 2017).

전통적인 언어적 분석은 단계적으로 처리된다. 문장은 단어 연쇄로 구성되고 문장 구성의 표층형 구조를 분석하는 방법은 통사 구조를 거쳐서 의미적 명세성(specification)을 통한 모델 해석 단계를 통해서 세계에 대한 완전한 해석을 목표로 한다. 반면에 n-gram 모델로도 알려진 분포가설에 기초한 단어 모델은 단어의 분포적 특성을 고려한 통계적 모델이다. 순차적 방식으로 단어 출현 빈도에 기초한 조건 확률을 통계적으로 활용한다.

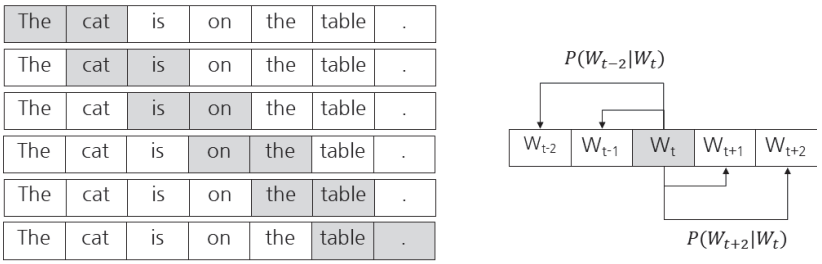

그림 1. N-gram 모델

N-gram 모델은 다차원 연산이기 보다는 구조적으로 입력되는 텍스트에 종속되기 때문에 희소성(sparsity) 문제, n개 선택의 문제인 상충성(trade-off) 문제, 복잡도(perplexity) 문제들이 발생한다. 희소성은 특정 텍스트에서 추출하고자 하는 n-gram이 발견되지 않으면 확률계산이 곱의 합으로 이루어지기 때문에 모든 계산이 0이 되는 문제점이다. 예를 들어서

1) 다차원 해석의 문제는 채점 신뢰성 문제와 연관된다. 여러 채점 요인을 활용한 정교한 채점을 하는 것이 적절한지 아니면 단순한 채점 요인을 활용한 신뢰성이 높은 채점을 할지는 문제점으로 남는다. 복잡한 채점 기준을 갖는 경우가 더 신뢰성이 높거나 또는 단순한 채점 기준을 갖는 경우가 더 신뢰성이 낮은 것은 아니다. 신뢰할 만한 인간 전문가 집단 채점과 가장 유사하며, 여러 글쓰기 능력 평가 항목 평가에 가장 부합하는 방식이 가장 적절한 채점 기준을 갖는다.

‘covid-19’이라는 신조어로 현재 텍스트에 존재하지 않던 단어이면 새로 이 단어가 포함한 문맥은 확률적으로 0이 되기 때문에 확률적으로 중요한 것이 출현하더라도 0이 될 것이다. 상충성과 복잡도는 서로 연관되어 있다(Jurafsky & Martin, 2020)[2].

이러한 n-gram 방식의 한계점을 극복하고 일반적인 문맥과 특정 문맥을 동시에 고려하기 위해서 Mikolov 외(2013)는 딥러닝 기반으로 벡터 공간에서 문맥 이해 기술인 Word2Vec을 발표했다. 간략하게 모델을 설명하면 다음과 같다. 모델은 크게 입력, 계산, 출력으로 이루어진다. 입력은 일반적 텍스트 파일인 코퍼스 형식의 문장들을 무작위로 이루어진다. 자연 언어의 경우 하나의 문장에도 형태, 통사, 의미, 화용 정보 등의 여러 층위 정보가 다면적으로 응축되어 있기 때문에 입력된 정보는 다층위적, 다면적 구조이다. 딥러닝 모델은 코퍼스에서 발견되는 이러한 다층위적, 다면적 구조를 이해하기 위해서 추출된 여러 특징적 정보를 반복적으로 다시 입력하고 계산·처리한다. 소위 ‘역전파(back-propagation)’라는 방식을 활용해서 반복적으로 출력을 다시 입력으로 바꾸어서 연산하는 과정을 통해서 가중치(weight value)를 미세하게 조절해서 정교한 입력 대비 출력 계산이 가능해지게 한다. 이러한 일련의 과정을 통해서 산출된 출력 구조는 일차원이지만 다층위, 다면적 정보를 함축하고 이해하게 된다.

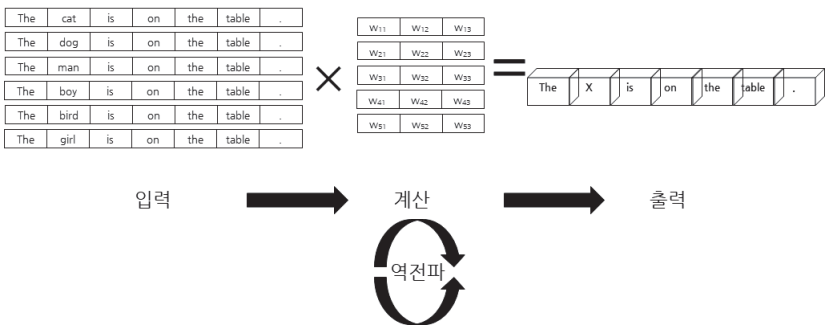

그림 2. 단어 임베딩 모델

이러한 딥러닝 구조는 기존의 자연어처리 시스템을 크게 바꾸어 놓았는데, 가장 중요한

2) N-gram의 n은 문맥의 개수, 즉 고려하는 단어의 총 개수이다. 여기서 n을 몇 개로 하는 것이 처리 복잡도를 낮추고 문맥 이해도를 높이게 되는 것인지는 고려되어져야 한다. N을 높여서 5나 7로 만드는 것이 적절한지 아니면 1이나 2, 3으로 낮추는 것이 좋은지는 문맥과 처리하고자 하는 대상에 달려있다. Jurafsky & Martin (2020)은 Shakespeare에서 n-gram을 추출한 경우를 예로 들고 있다. Shakespeare의 모든 희곡 텍스트를 대상으로 n을 4이상, 5까지 추출한 결과는 Shakespeare 희곡 자체가 추출이 되지만 1로 낮추면 Shakespeare 희곡이 아닌 이해하기 어려운 문맥이 나온다. 2나 3으로 추출한 경우가 문맥을 가장 적절히 보여준다고 주장한다.

측면은 형태, 통사, 의미, 화용 정보 등의 다층위 구조를 시스템이 각각 이해하고 처리해야 하는 부담에서 벗어나 층위적 구분이 필요 없는 평면적 해석 장치로 바꾸었다. 더 자세히 설명하면 형태, 통사, 의미, 화용 정보는 하나의 차원적 정보이며 이러한 여러 차원을 해석해야 전체를 해석할 수가 있다. 그러나 이러한 다면적, 다차원 구조적 해석이 동반하는 다층위적 해석은 차원 왜곡이라는 문제점도 발생하는데, 딥러닝 구조는 다차원을 평면적 해석으로 바꾸어서 이러한 문제점을 해결했다.

Word2Vec, WMD은 그림 2와 같이 유사하게 작동하는 통계 모델이지만 BERT는 이 방식보다 더 복잡하게 구동한다. BERT 모델은 단어 임베딩 모형 중 하나이지만, Word2Vec 모델과 다르게 문맥 이해 능력이 있다. Word2Vec은 단어와 단어 모델에 의한 단일 모델로 연산이 된다면 BERT는 주변 문맥을 더 포괄적으로 연산하고 처리하기 위한 다층형 모델이 가능하다. 그림 3(좌)에서 Word2Vec 모델은 그림 2와 같이 단일 연산 방식으로 다층형 구조를 연산한다. 반면에 BERT는 복잡한 내부 연산 구조를 갖고 있어서 자연언어의 복잡한 구조에 대한 해석이 가능하다. 또한 마스킹(masking)이라는 기법을 통해서 다양한 문맥 요소들도 삽입해서 연산이 가능하다. 결론적으로 BERT 모델은 문맥이라는 요소를 구조화된 단위로 이해하는 것이 가능하다.[3)]

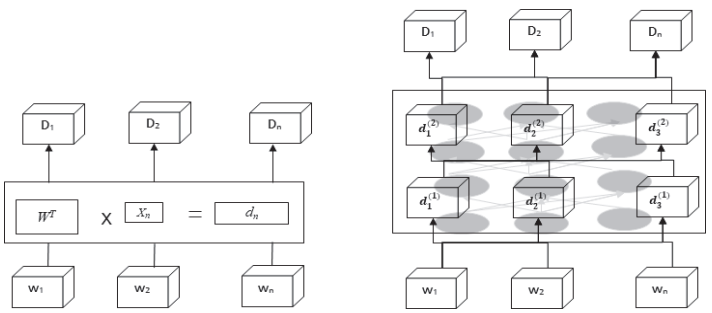

그림 3. (좌) Word2Vec 모델 (우) 다층위 이해가 가능한 문맥 이해 BERT 모델

본 연구는 단문 영작문 자동 채점에서 정답이 되는 문장과 사용자 입력이 얼마나 유사한지를 살펴보는 문제이다. 이를테면 "I saw a snail crawling on the ground."라는 정답 문장에 대해서 "I saw a snail which is crawling on the ground."라는 사용자 입력 문장이 얼마 유사성이 높은지를 채점한다. 이러한 유사도 측정 방식을 다른 자연어처리 시스템과

3) BERT 모델의 경우 인간의 모국어 학습 능력과 유사하다는 연구들이 발견된다. 영어의 경우 주어-동사 일치(Jawahr 외, 2019)를 예측할 수 있으며, 문맥 상 가능한 동사 예측(Goldberg, 2019), 의문문 생성, 재귀사, 부정극어 예측(Warstadt & Bowman, 2020) 등 여러 통사적 현상에 대해서 언어 능력과 유사하거나 동일하다는 것이다. 또한 딥러닝 모델이 제 2언어 학습의 모델로써 L1-L2간 언어 전환 (L1-L2 transfer)에 대한 예측이 가능하다는 연구들이 발견된다(Cohen 외, 2020; Dalad & Manoj, 2018).

비교하면 다음과 같다. 자연어처리는 여러 다양한 문제를 처리하기 위한 여러 분야가 있는데, 본 연구와 비교에서 가장 유사성이 높은 분야는 문장 간 유사도 비교를 하는 텍스트 함의 인식 (Recognizing Textual Entailment, 이하 RTE)이다.

RTE는 다양한 문장 유형들이 무엇을 서로 함의하는지 파악하는 것이다. 다시 말하면 다양한 유형의 텍스트를 무차별적으로 분석하는 경우에 적당한 동일한 주제 또는 분야의 내용을 함의하는지를 판별하는 것이다. 이러한 처리 시스템은 정보처리, 질의분석, 요약시스템과 같이 다양한 자연어처리 시스템의 기초가 된다. RTE 처리 방식은 어떤 언어 정보에 기초하는지에 따라 분류되는데 크게 단어, 문장, 의미정보에 기반을 둔 처리 방식으로 나뉜다(Farouk, 2019). 단어를 활용한 경우는 단어순서 겹침 정보를 활용하고, 문장 정보가 필요한 경우에는 통사적 정보를 활용하며, 의미정보를 활용한 경우는 WordNet과 같은 전자사전을 이용한다. 만약 여러 정보를 통합해서 활용하는 경우에는 복잡한 구조가 된다. 딥러닝을 활용한 경우는 단어 벡터에 의존하기 때문에 복잡한 언어학적 단위들의 연쇄가 필요하지 않다. 이러한 딥러닝 기반 RTE 처리 방식은 크게 기학습된 단어 임베딩 정보를 적용한 방식과 딥러닝 학습 시스템이 적용된 방식으로 크게 나뉜다(Farouk, 2019). 연구에서는 기학습된 단어 임베딩 정보에 기반을 둔 방식을 적용했다.

RTE나 본 연구의 문장 유사도 측정 방식은 두 문장 간의 코사인 유사도(cosine similarity)를 측정하는 것이다(Ranasinghe 외, 2019). 이를 수학적 표현으로 나타내면 (1)과 같다. (1a)는 코사인 유사도 측정 방식을 가리키며 (1b)는 단어 임베딩 모델의 문장 측정 방식이다. 문장의 표현은 단어의 연쇄로 이루어지기 때문에 모델에서 산출되는 각 단어별 수치 연쇄의 평균치이다. (1c)는 단어별 값은 단어 임베딩 모델에 의해서 산출되게 된다.

(1) a. $\cos(sent_1, sent_2) = \frac{\Theta(sent_1) \bullet \Theta(sent_2)}{\|\theta(sent_1)\| \times \|\theta(sent_2)\|}$ $(\theta : embedding\ operator)$

b. $\theta(sent) = \frac{1}{N}\sum p(w_i|w_1,....,w_n)$

c. $p(w_i|w_1,...,w_n) \approx f: X \rightarrow Y\,(f(x)\ is\ one\ of\ W2V, WMD, BERT, ...)$

앞서 논의한 바와 같이 모델이 가지는 성능은 전체적인 성공과 실패를 측정하는 정확도에도 있지만 언어능력 평가에서 준거참조조사를 표현할 수 있는지에 대한 논의도 포함될 것이다. 준거참조조사는 평가에 있어서 성공과 실패로 나타나며 이러한 점은 개별 점수별로 다른 점수와 차이점이 분명하게 나타난다. 이러한 점이 통계적으로 이산적(discrete)으로 표현되며 점수별 차이가 이산적으로 나타나기 때문에 집단 간 구분은 이산적 관계성으로 표현된다. 다시 말하면 모델별 성능은 이러한 이산성을 가장 분명히 하는 경우에 가장 정확성이 높은 것으로 나타날 것이다.

3. 연구 방법

수도권 대학에 재학 중인 대학생들을 대상으로 그림 3과 같이 20여 개의 다양한 사진들을 하나씩 제시하고 영어 자막을 하나의 문장으로 다양하게 작성하도록 했다. 실험자 집단은 초·중·고 과정과 대학 과정에서 적어도 10년 이상을 외국어로 학습한 학습자를 대상으로 하고 영어가 모국어인 경우는 제외했다. 총 100명 학생들을 대상으로 사진 하나당 5개 정도의 다양한 문장을 구성하게 해서 총 1만 여개 문장의 데이터를 획득했다.

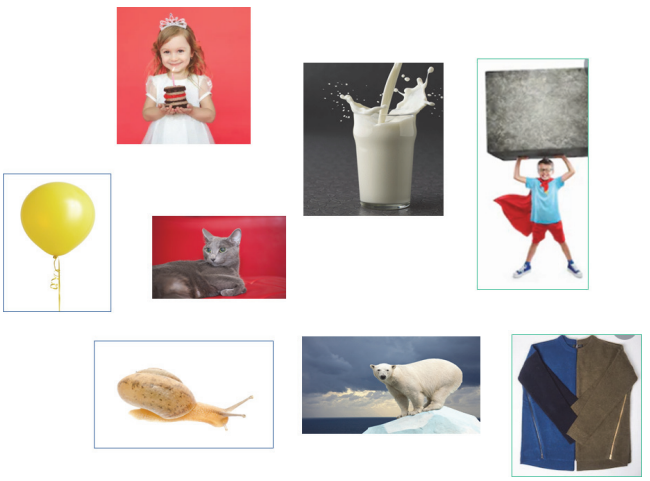

그림 4. 사진 데이터의 예시

본 실험 전에 예비 실험 단계를 거쳐서 실험이 적합한 결과를 구성할 것인가를 검증하는 단계를 거쳤다. 예비 실험 단계에서는 10여 명에게 사진 5개를 주고 각각 5개씩 자막 영작문을 시행했다. ETS TOEIC 사진 자막 영작문 시험 방식에 따라 예비 실험도 유사한 방식으로 진행했다. 주어진 시간을 1분으로 설정하고 자막 영작문을 한 개씩 작성한 경우에 분석한 결과 대부분의 문장들이 유사했다. 본 실험에서는 더 다양한 데이터를 확보하기 위해서 과제로 작문을 제출하게 해서 다양한 문장과 등급의 데이터를 얻을 수 있었다[4]. 이러

4) 심사중에 실험환경과 결과에 대한 영향에 대한 문제를 제기되었다. 본 연구의 예비실험 단계에서 실제 실험환경 및 조건이 실험결과에 영향을 미치는 것이 발견되었다. 예비 실험은 주어진 시간 안에 정해진 장소에서 처리되었는데, 평가해본 결과 평균적으로 2, 3점이라는 특정 점수에 집중되는 경향성을 보였다. 이것은 주어진 시간이 1분으로 짧고 대면형식으로 제시된 결과 피험자인 학생들이 시험적 성격으로 판단하기 때문에 중앙이 되는 평균에 집중되는 경향을 보인다. 본 연구에서 진행한 실험의 목표는 데이터 수집에 있기 때문에 실험방식을 자유스럽게 해서 더 다양한 데이터를 수집하고자 했다. 이를 위해서 과제

한 일련의 예비 단계를 통해서 실험 관련한 다음 사항의 문제점을 발견했다. 제시되는 사진 구성이 복합적이면 자막 영작문이 복잡하게 되므로 채점에서 문제가 발생되는 것을 발견했다. 또한 동일한 주제의 사진이 겹칠 경우에 자막으로 구성한 문장 형식을 다음에도 반복하는 경향이 있어서 사진의 주제가 겹치지 않도록 조절했다[5].

실험자들은 다양한 문장 유형과 어휘를 선택해서 자유롭게 단문 영작문을 구성했다. 그림 3의 상단 왼쪽의 어린 소녀가 촛불이 꼽혀진 케이크를 들고 있는 사진의 경우에 (2)와 같이 다양한 문장들이 수집되었다.

(2) a. A smiling little girls holds a birthday cake.
b. A little girl holds a birthday cake with a smile.
c. A girl is smiling while holding a cake.
d. The little girl is happy to be given the birthday cake.
e. She is smiling at a birthday cake.
f. A pretty girl is smiling.
g. Today is the little girl's birthday.
h. A little girl has a sweet cake.
i. She smiles at me.
j. The girl smiles.
k. She loves cake.
l. I'm happy.
m. I'm happy to see you.

채점 기준은 ETS TOEIC 사진 자막 영작문 채점 기준(민선식, 2008, p. 26)에 따라서 3, 2, 1 채점 등급으로 구분했다. 간략히 설명하면 다음과 같다. 3점은 사진과 연관성이 있으며 오류가 하나도 없는 경우에, 2점은 3점과 같지만 의미를 해치지 않는 오류가 있는 경우에, 1점은 의미를 해치는 문법적 오류가 있으며 사진과 연관성이 없다. (2)에서 필수 단어는 {girl, cake, hold, smile}이며, 3점은 (2a-e)으로 필수 단어와 사진과 연관성, 문법적 무결성 등으로 판단된다. 2점은 (2f-h)로 필수 단어는 하나 정도이고, 사진과 연관성은 있지만 정확한 묘사나 표현이 결여된 경우이다. 1점은 (2i-m)으로 필수 단어가 거의 쓰이지 않았고, 묘사나 표현이 2점보다 덜 정확하게 사용된 경우이다,

---

형식으로 더 많은 작성 시간을 부여하고 비대면 형식으로 실험환경을 변경한 결과는 1점 문장들도 더 많이 수집되는 등 더 다양한 유형의 데이터를 획득할 수 있었다.

5) 또한 제시한 사진은 저작권이 없는 경우로 한정했고, 과제의 경우에도 무기명으로 제출하게 해서 더 자유로운 환경을 만들었다.

채점 과정은 2인이 작업을 했는데, 첫 번째 채점을 1인이 하면 두 번째 채점은 나머지 1인이 채점을 확인하는 형식으로 했다. 각 점수별 분포는 1점은 34%, 2점은 48%, 3점은 18%이다. 1, 2점을 합친 것이 3점보다 더 많다. 등급이 높을수록 더 많은 단어가 사용되었다. 3점은 9.38개, 2점은 7.45개, 1점은 5.15개가 평균적으로 활용되었다. 또한 통사적 분석을 의존 문법 파서(dependency grammar parser)로 문장 내 의존 관계를 추출했다. 의존 문법은 문장 내 구조에 따라 두 단어의 의존성을 추출하는 문법적 방법론으로 더 복합적이고 더 복잡한 문법 구조에서 더 많은 의존 관계가 발견된다. 연구에서는 각 등급별 점수 당 의존 관계의 개수를 측정했는데, 3점은 10.23, 2점은 8.89, 1점은 6.09개의 의존 관계가 평균적으로 추출되었다. 결과적으로 보면 등급이 높을수록 더 많은 수의 단어 사용과 더불어 더 복잡한 구조적 관계가 발견된다.

집단 간 특성 비교를 위해서 문장 간 비교를 했으며, 두 문장 간 얼마나 유사한지를 측정하는 문장 유사도를 활용했다. 유사도 측정은 Word2Vec, WMD, BERT, AMR 모델을 사용했다. Word2Vec, WMD는 30억 영어 단어로 구성된 텍스트로 기학습된 모델을 사용했고, BERT는 Turch 외(2019)의 모델을 사용했다. Word2Vec, WMD, BERT는 딥러닝 기반 단어 임베딩 모델들로 발전의 순서에 따라서 보면 Word2Vec, WMD는 초기 모델로 문맥에 독립적인 임베딩 방식을 사용한다. 반면에 BERT 모델은 최근에 발표되었으며 현재 각광받는 사전학습 모델로 문맥 학습효과가 크다. AMR은 Cai 외 (2013)의 AMR 자연어처리 시스템을 활용해서 문장의 의미·통사적 구조를 원자적 단위로 분해해 차이점을 만들고 각각의 구조적 동질성을 조화평균을 적용한 smatch 자연어처리 도구로 두 문장 간 유사도를 측정한다(김동성, 2016). AMR 방식은 언어학적 분석에 따라서 개별 논항들의 논리적 구조를 분석해 구조화 하는데, 그림 5에서와 같이 두 개의 다른 문장 구조의 차이들이 구조화된다.

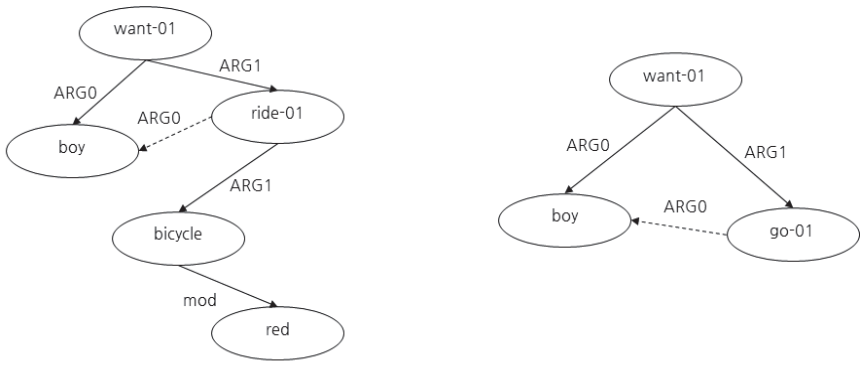

그림 5. AMR 분석

4. 결과 분석 및 논의

결과적으로 가장 최신 단어 임베딩 모형인 BERT 모형이 92.53% 이내의 가장 정확한 분류 결과를 보였다. 그 외의 딥러닝 모델인 Word2Vec이나 WMD 모델이나 의미·통사 구조적 동일성을 관찰하는 AMR은 동일 등급 간에 동질성을 보이지도 않고, 다른 등급 간 차이점도 보이지 않았다[6]. BERT 모델의 분류 능력을 자세히 살펴보면 다음과 같다. 채점 3 문장들을 중심으로 채점 3인 문장들 간의 유사도 점수와 채점 1의 문장들의 문장 유사도 점수를 Bland-Altman 플롯은 그림 6과 같다.

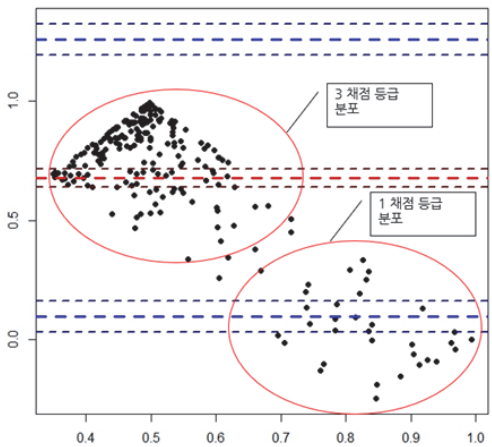

그림 6. Bland-Altman 플롯

Bland-Altman 플롯은 두 변인들 간의 상관성의 산점도 분포로 표현되는 그래프인데, 분포의 상하한선에 각각 선을 그어서 분포 한계를 관찰하는 방법이다. 양의 상관성이 높은 경우는 1에 가깝고, 반대로 음의 상관성이 높은 경우는 0에 가까울 것이다. 양의 상관성은 동질성을 설명하는데 비해서 음의 상관성은 상관성이 반대로 동질성이 희박한 것을 설명한다. 여기에서 채점 등급 3인 문장들 간의 유사도 분포는 위쪽에 분포하고 채점 1 문장들과 유사도는 아래쪽에 분포한다. 다시 말하면 같은 등급 간 분포는 동질적인데 비해서 다른 등급 간 분포는 이질적으로 나타난다.

규준참조조사의 경우 동일 분포 특성을 가진 집단 내의 분포적 특징을 참조하게 되며 백분율이나 표준점수나 표준등급과 같은 정규화된 확률 분포의 특성을 나타낸다. 그림 7 (좌)와 같이 마름모꼴 모양이나 별모양이 집단 내 위치한 분포 특성을 통해서 설명된다. 이

6) Word2Vec은 19.87%, WMD는 25.3%, AMR은 7.59% 정도의 낮은 낮은 분류를 보였다.

러한 분포 특성은 집단 내 상대적 위치로 인해서 발생한다. 반면에 준거참조조사의 경우 서로 다른 분포 특성을 갖는 경우로 설명된다. 일정한 준거에 의해서 이산적 구분이 되면 집단 간 이산화된 특성을 갖게 된다. 그림 7(우)는 서로 다른 분포 특성이 나타나는 경우에 이산성(discreteness)에 따라 준거참조조사의 특징을 나타낸다. 연구에서 살펴본 바에 따르면 BERT 모델은 채점 집단 간 이산성을 명확하게 표현하며 이러한 특성은 준거참조조사의 특징을 나타낸다.

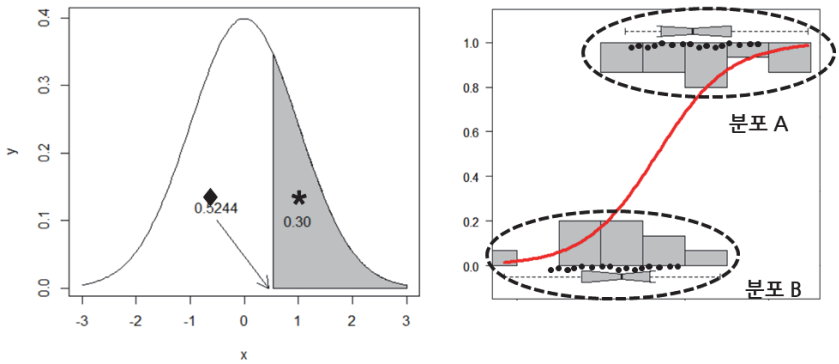

그림 7.（좌）규준참조검사 분포특성（우）준거참조검사 분포특성

그림 7(우)의 이산성은 로지스틱 회귀분석과 같은 통계적 모델이 종속변수와 독립변수 간의 인과성을 포착할 수 있다. 여기서 종속변수를 1 또는 0 (성공 또는 실패)로 간주할 수 있다. 즉 같은 집단인지 아니면 다른 집단인지 구분하는 문제를 통계적으로 살펴볼 수 있다. 다시 말하면 문장 유사도 측정에서 동일 채점 집합 문장일 경우 동일한 분포 특성을 가정하고 다른 채점 집단의 경우에는 다른 분포 특성을 나타내는 것으로 간주한다. 또한 독립변수를 Word2Vec, WMD, BERT, AMR 등으로 측정된 문장 유사도로 간주하면 어떤 방식이 가장 분포를 이산적으로 잘 예측했는지 측정할 수 있다. 측정치는 점수별 집단 마다 각기 다른 값을 보이나 BERT 모델로 측정된 문장 유사도들의 경우 유의수준 0.01이하로 통계적으로 유의미하다. 그 외의 모델로 측정된 문장 유사도는 유의수준이 0.01이상으로 통계적으로 유의미하지 않다.[7]

연구에서 활용한 문장들의 비교 집합은 $10^{4\times6}$의 엄청난 양의 문장들이 있으므로 데이터 분포의 양도 많아서 데이터를 세밀하게 보여줄 수 없다. 유사도 비교 수준을 세밀하게 보여주기 위해서 소량의 데이터를 활용해서 분포의 질적인 면을 논의하면 다음과 같다. (3-4)은 그림 4의 노란 풍선이 떠있는 사전 자막 영작문 문장들의 예들이다. (3, 4)는 각각

7) 활용된 통계툴은 R x64 3.5.1로 일반화 선형 모형(generalized linear model)을 사용했다.

채점 등급 3, 2로 판정된 문장들 중 일부이다.

(3) a. The yellow balloon is tied with a yellow ribbon.
b. There is a yellow balloon floating in the air.
c. This is a yellow balloon tied with a yellow ribbon.
d. There is a yellow balloon floating.
(4) a. This is a yellow balloon.
b. It's a yellow balloon.
c. There is a yellow balloon.
d. This is a balloon.

점수 2-3 문장들 간의 비교를 위해서 점수 3인 (3)의 문장 중 하나를 점수 2인 (4)의 문장들이 포함한 다른 점수 2 문장들 30개와 문장 유사도를 비교해 보았다. 유사도를 활용해서 산포도와 박스플롯을 그리면 그림 8과 같다. 같은 점수 문장들 간에는 문장 유사도가 높고, 다른 점수 문장들 간에는 문장 유사도가 낮게 나타나야 올바른 예측이 가능하다. 즉 분포가 높거나 낮게 밀집되어 나타나야 올바른 점수 문장 집단을 표현하게 된다. 반면에 밀집되지 않고 분포가 퍼져서 나타나면 점수 문장 집단을 올바르게 판정할 수 없다. (3-4)와 같은 점수 2-3 문장들 간에는 유사도가 없기 때문에 수치적으로 낮은 수준의 유사도 분포가 나타나야 한다. 그림 8(좌)와 같이 BERT 모델로 측정된 유사도 수준은 낮게 나타나는데, 반면에 다른 모델들로 측정된 것은 분포적 특성이 넓게 나타나서 점수 2-3 문장 간 유사도 분포를 관찰하기 어렵다. 그림 8(우) 박스플롯에서도 BERT가 밀집한 분포 특성을 보이는 반면에 다른 모델들의 분포는 넓게 나타나는 것을 볼 수 있다. 그러므로 BERT 모델 이외에 다른 모델들은 점수 문장 간 유사도 차이점을 분포적으로 실질적으로 설명하지 못한다.

BERT는 그림 6, 8에서와 같이 다른 점수 문장들 간 유사도의 군집이 서로 다르게 나타나므로 집단 간 이산성을 발견할 수 있다. 같은 점수 문장들 간에는 군집되고 다른 집단 간에는 산개하게 된다. 이러한 면을 채점 시스템에서 활용하면 집단 내 또는 집단 간 구분으로 평가가 가능하다. WMD, Word2Vec, AMR로 측정된 문장 유사도를 활용하면 BERT 모델을 활용한 경우에 보여주는 다른 채점 간 분포 차이를 전혀 찾아 볼 수가 없다. WMD 모델은 문서 당 의미 공간이 동일한지 측정하게 되는데, 하나의 문서에서의 의미 공간을 여러 문서에서 어떻게 측정되는지에 따라 공간을 재조정하게 된다(Kusner 외 2015, Ramasinghe 외 2019). WMD로 측정된 경우는 다른 채점 문장 간 분포가 같은 채점 문장 간을 비교한 것과 큰 차이를 보이지 않는다. 유사도 점수의 분포가 변별성이 없으며, 오히려 다른 점수 문장들 간의 유사도의 분포가 같은 채점 문장들 간의 유사도 분포보다 더 적은 구간에서 나타난다. Word2Vec 모델의 경우도 WMD와 유사하다. 두 모델은 다른 점수 문장들 간 비교가 어렵

다. 같은 점수 문장들 간의 산포도와 다른 점수 문장들 간 유사도 사이에 변별성은 없다.

WMD나 Word2Vec 모델은 단어 임베딩에 기초한 초기 모델로 단어 공간을 일정한 공간으로 변환시키게 된다. 이러한 공간에서 하나의 단어는 주변 어휘들을 중심으로 하는 공기 행렬(co-occurrence matrix)에 다른 어휘들의 상대적 빈도를 측정하게 된다. 문제는 자연언어는 문장, 구, 단락이라는 단위로 의미·통사적으로 다양한 구조를 만들어 내는데 Word2Vec이나 WMD와 같은 측정법은 단순한 단어 연쇄에 의한 공기 행렬 모델로 측정한다. 이러한 방식은 복잡한 자연어 구조를 다 표현하지 못한다. 반면에 현재 각광받는 BERT 모델은 단어와 단어 사이, 구조와 구조 사이, 문장과 문장 사이와 같은 구조적 연관성을 추정하는 장치가 포함되어 있다. 2절에서 논의한 바와 같이 BERT 모델의 마스킹은 문맥을 연산하는 장치로 구조에 대한 이해가 가능하다.[8] 그림 9는 Word2Vec, WMD 모델을 바탕으로 점수 3과 1 문장들 간 유사도의 분포를 보여주는 산포도이다. 두 모델의 경우 모두 분포가 넓게 퍼져 있어서 두 점수 문장들 간 구분이 어렵다.

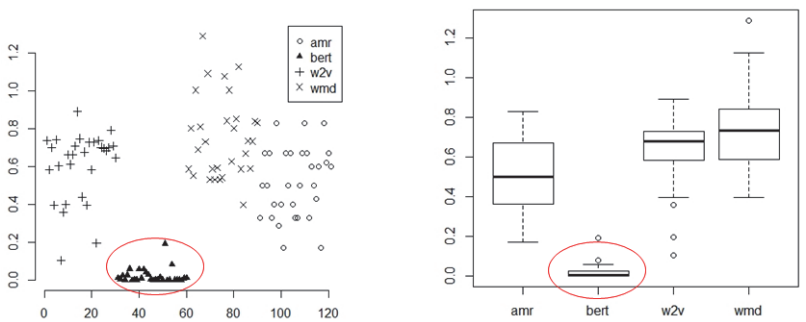

그림 8. (좌) 모델간 산포도 비교 (우) 모델간 박스플롯 비교

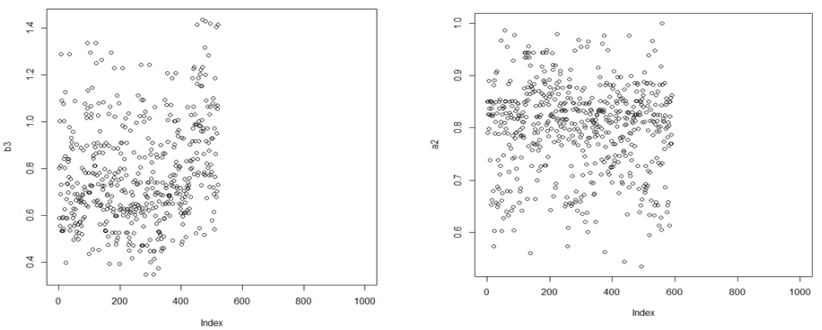

그림 9. 모델별 점수 3-1 문장들 간 유사도 비교 분포 (좌) Word2Vec (우) WMD

8) 주3 참조

AMR 모델을 활용한 문장들 간 유사도 분포는 희박성(sparseness)의 특징을 갖는다. AMR은 문장 간 의미·통사 구조에서 추출되는 이항적 관계 일치도를 간단한 조화평균으로만 계산한다. 따라서 문장 성분으로 만들어지는 관계들의 평균이기 때문에 조밀한 통계적 분포를 표현하지 못한다. 또한 단어 연쇄 통계 모델이 반영되지 않는 구조이기 때문에 다양한 통계 분포가 관찰되지 않고 동일 수치만이 반복된다. 의미·통사 구조적 특성 이외에 단어와 단어간의 연쇄적 특성도 문장 간 차이점으로 나타난다. 예를 들어서 "Tom likes Mary."라는 문장과 "People like holiday."는 문장은 통사적 구조가 동일하며 주어와 목적어가 다른 의미적 구조이다[9]. AMR로 측정한다면 유사성이 높은 구조로 판별이 되지만 단어의 연쇄적 통계 분포를 고려하는 Word2Vec, WMD, BERT는 다른 단어의 연쇄로 판단하고 두 문장의 유사도 수치가 낮게 나타날 것이다. 따라서 AMR이 문장 유사도 측정에서 가장 낮은 정확성을 보인다. 분포도를 살펴보면 그림 10과 같으며 등급 내 분포나 등급 간 분포는 구분하기 어렵다.

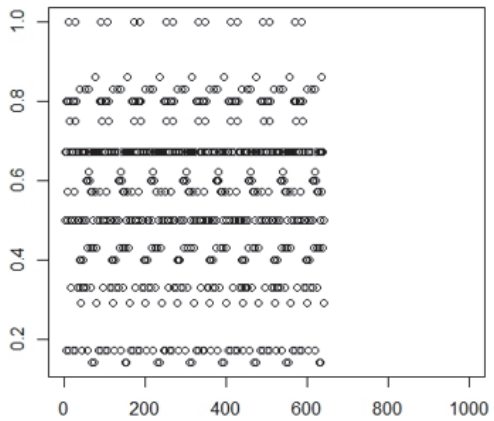

그림 10. AMR로 측정된 점수 3-1 문장들 간 유사도 분포

5. 결론 및 제언

본 연구는 사진 자막 영작문 채점에 대한 ICALL을 다루었다. 영작문 채점이 인간 채점으로 진행되는 경우 엄청나게 많은 인력, 절차, 시간, 재원이 요구되므로, 시스템을 통한 자동 채점이 요구된다. ETS는 자동 채점에 활용하기 위해서 e-rater, c-rater를 개발했으며, 여

9) 실제 구글 검색을 하면 "People like holiday."는 42,500번 출현빈도를 갖고, "Tom likes Mary."는 50, 50번 출현빈도를 갖는다. 이와 같이 각기 다른 단어 연쇄는 각기 다른 용례를 갖으며, 다른 확률을 갖게 된다.

러 연구자들이 다양한 시스템을 제시했다. 연구에서는 자동 채점을 위해서 언어학적 분석에 근거한 의미·통사 구조에 근거한 AMR과 딥러닝 기반의 단어 임베딩 모형에 근거한 Word2Vec, WMD, BERT 모델들을 활용했다. 실험을 위해서 사진들에 대해서 영작문 데이터를 1만 건을 수집하고 이를 작업자들이 등급을 분류했다. 이 데이터를 기준으로 모델들을 적용해서 각 점수별 분류가 이루어지는지 평가했다. 결과적으로 BERT 모델이 정확한 분류를 하는 것에 반해서 다른 모델들의 점수별 분류는 낮은 수준이다.

TOEFL, TOEIC과 같은 영어 능력 평가는 준거참조조사로 일정한 기준에 도달했는지를 평가해야만 한다. 이런 점에서 기계학습 방식을 적용한 시스템이 준거참조조사가 가능한지는 채점 시스템 전체의 신뢰도 측면에서 중요하다. 연구에서 살펴본 바로는 BERT 모델이 적용된 경우 각기 다른 점수별 문장 간 문장 유사도 분포가 이산성을 보이며, 준거참조조사가 가능한 분포적 특성을 나타낸다. 또한 BERT 모델의 경우 인간 직관과 유사성이 높다는 많은 연구들이 발표되고 있다. 이러한 면이 어떠한 연관성이 있는지에 대한 연구는 향후 연구로 미루고자 한다. 본 연구에서는 채점과 연관되어서 모델을 활용한 기계 채점과 인간 채점과의 유사성을 대상으로 하고 BERT 모델이 가장 적합함을 보였다.

이 연구는 채점 모델 개발로, 학습자에 피드백을 주는 방안은 없다. 다시 말하면 오류 검출과 오류에 대한 해결방안 제시, 그리고 학습에 도움이 되는 피드백 콘텐츠 생성 등과 같이 학습을 위한 방안은 없다. 이러한 문제점은 단어 임베딩 모델이 통계형 모델로 언어 구조에 대한 설명이 불가능한 것에 기인한다. 연구에서는 AMR이 통계적 모형으로 가장 희박한 수준의 예측 분포 모형을 제시하였지만 AMR은 의미·통사 구조에 기인하므로 언어학적 구조에 대한 설명력을 가진다. 따라서 어느 논리적 이항구조가 어떻게 다른지 설명이 가능하므로 구조에 대한 해석력을 갖는다. 이러한 점에서 딥러닝 단어 임베딩 모델을 통한 채점과 AMR을 통한 피드백 구조가 서로 순환적으로 이루어진다면 채점과 피드백을 완전하게 연결하는 시스템 구조가 가능할 수도 있을 것이다.

참고문헌

김동성, 채희락, 이상철. (2008). 문법성과 어휘 응집성 기반의 영어 작문 평가 시스템. *인지과학*, *19*(3), 223-255.

김동성. (2016). 추상적 의미 표상을 활용한 사진 자막 영작문 평가. *언어학*, *24*(4), 1-26.

김지은, 이공주. (2007). 중학생 영작문 실력 향상을 위한 자동 문법 채점 시스템 구축. *한국콘텐츠학회논문지*, *7*(5), 36-46.

민선식. (2008). *Toeic Writing Test 공식문제집*. 서울: 시사영어사.

진경애. (2007). 영작문 자동 채점 시스템 개발 연구. *영어어문교육*, *13*(1), 235-259.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rator V.2. *Journal of Technology, Journal of Learning, and Assessment*, *4*(3), 3-30.

Botvin, G., & Sutton, S. (1977). The development of structural complexity in children's fantasy narratives. *Developmental Psychology*, *13*(4), 377–388.

Cai, S., & Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. *In Proceedings of the ACL*, 748-752.

Cohen, C., Higham, C., & Nabi, S. (2020). Deep learnability. *Frontiers in AI*, *3*(43), 1-11.

Condon, W. (2009). Looking beyond judging and ranking. *Assessing Writing, 14*(3), 141-56.

Dalad, N., & Manoj, N. (2018). Transforming second language acquisition modeling. *In Proceedings of NIPS*, 1-9.

Farouk, M. (2019). Measuring sentence similarity. *Indian Journal of Science and Technology, 12*(25), 1-11.

Firth, J. (1957). *Papers in Linguistics 1934–1951* (1957) London: Oxford University Press.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. *arXiv:1901.05287*.

Heift, T., & Schulze M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning*. New York: Routledge.

Jawahr, G., Benoit, S., & Seddah, D. (2019). What does BERT learn about the structure of language? *In Proceedings of ACL*, 3651-3657.

Jurafsky, D., & Martin, J. (2020). *Speech and language processing*. London, UK: Pearson.

Kusner, M., Sun, S., Klkin N., & Weinberger, K. (2015). From word embeddings to document distance. *In Proceedings of International Conference on Machine Learning*, 957-66.

Leacock, C., & Chodorow, M. (2003). C-rater. *Computers and Humanities*, *37*, 389-405.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated Grammatical Error Detection for Language Learners*. San Rafael, CA: Morgan & Claypool Publishers.

Manning, C. (2017). *Representations for Language*. Retrieved January 26, 2021, from http://simons.berkeley.edu/sites/default/files/docs/6449/christo phermanning.pdf.

McKeough, A., & Malcolm, J. (2011). Stories of family, stories of self: Developmental pathways to interpretive thought during adolescence. *New Directions for Child & Adolescent Development*, *2011*(131), 59-71.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In Proceedings of NIPS*, 3111-9.

Nagata, N. (1996). Computer vs. Workbook instruction in second language acquisition. *CALICO Journal*, 14(1), 53–75.

Ramasinghe, T., Orasan, S., & Mitkov, R. (2019). Enhancing unsupervised sentence similarity methods with deep contextualized word representation. *In Proceedings of the Recent Advances in NLP*, 994-1003.

Rivers, W. (1981). *Teaching foreign-language skills*. Chicago: Univ. of Chicago Press.

Somasundaran, S., Lee, C., Chodorow, M., & Wang, X. (2015). Automated scoring of picture-based story narration. *In Proceedings of Innovative Use of NLP for Building Educational Applications*, 42–48.

Settles, B., LaFlair, G., & Hagiwara, M. (2020). Machine learning-driven language asessment. *In Proceedings of the Transactions for Computational Linguistics*, 247-263.

Sukkarieh, J., & Stoyanchev, S. (2009). Automating model building in c-rater. *In Proceedings of Applied Textual Inference*, 61-69.

Turc, I., Chang, M., Lee, K., & Toutanova, K. (2019). *Well-read students learn better: On the importance of pre-training compact models*. Unpublished manuscript.

Warstadt, A., & Bowman, S. (2020). Can neural networks acquire a structural bias from raw linguistic data? *In Proceedings of Cognitive Science Society*, 1737-1943.

Weigle, S. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*(3), 335–353.

Weigle, S. (2011). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *TOEFL iBT Research Report TOEFL iBT-15*. Princeton, NJ: Educational Testing Service.

Wiegle, S. (2013). English as a second language writing and automated essay evaluation. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 36-54). New York: Routledge.

김동성
03760 서울시 서대문구 이화여대길 52
이화여자대학교 인문과학대학 인문테크놀로지 전공 특임교수
전화: (02)3277-4339
이메일: dsk202@ewha.ac.kr

Received on January 31, 2021
Revised version received on April 22, 2021
Accepted on June 30, 2021