https://doi.org/10.24303/lakdoi.2022.30.1.45

적대적 사례에 기반한 언어 모형의 한국어 격 교체 이해 능력 평가*

송상헌** · **노강산** · **박권식** · **신운섭** · **황동진**
(고려대학교)

Song, Sanghoun; Noh, Kang San; Park, Kwonsik; Shin, Un-sub & Hwang, Dongjin. (2022). Adversarial example-based evaluation of how language models understand Korean case alternation. *The Linguistic Association of Korea Journal*, *30*(1), 45-72. In the field of deep learning-based language understanding, adversarial examples refer to deliberately constructed examples of data, slightly different from original examples. The contrasts between the original and adversarial examples are less perceivable to human readers, but the disruption has a notorious effect on the performance of machines. Thus, adversarial examples facilitate assessing whether and how a specific deep learning architecture (e.g., a language model) robustly works. Out of the multiple layers of linguistic structures, this study lays focus on a morpho-syntactic phenomenon in Korean, namely, case alternation. We created a set of adversarial examples regarding case alternation, and then tested the morpho-syntactic ability of neural language models. We extracted the instances of case alternation from the Sejong Electronic Dictionary, and made use of mBERT and KR-BERT as the language models. The results (measured by means of surprisal) indicate that the language models are unexpectedly good at discerning case alternation in Korean. In addition, it turns out that the Korean-specific language model performs better than the multilingual model. These imply that an in-depth knowledge of linguistics is essential for creating adversarial examples in Korean.

주제어(Key Words): 적대적 사례(adversarial examples), 격 교체(case alternation), 딥러닝(deep learning), 고의적 잡음(intended noise), 견고성(robustness), 언어 모형(language model), 평가(evaluation)

* 이 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020S1A5A2A03042760).
** 제1저자 및 교신저자

1. 서론

본고는 '적대적 사례(adversarial examples)'를 활용하여 딥러닝 언어 모형이 한국어의 격 교체 현상을 어느 정도로 다룰 수 있는지를 측정하는 데 목적이 있다. 최근 주목을 받는 인공지능 성능 평가 기법은 마치 속임수를 쓰듯 인공지능의 학습 방식을 우회하여 그 모형의 견고성(robustness)을 검증하는 특성을 가진다. 이 방식을 차용하여 본 연구는 대표적인 한국어 형태통사적 제약을 언어 모형이 충분히 포괄하고 있는지를 검토하고자 한다.

적대적 사례 기법은 평가 데이터에 의도적인 잡음을 삽입하여 특정 딥러닝 모형이 그러한 잡음에 유연하게 대처할 수 있는지를 공격적으로 평가한다. 이를 간단한 도식으로 개념화하면 아래 그림 1과 같다.

그림 1. 적대적 사례의 기본 개념

위에서 제일 좌측의 A는 정사각형의 형태를 지니고 있다. 여기에서 의도적으로 한쪽 귀퉁이를 작게 깎으면 두 번째 A′ 의 형태가 나온다. 비록 한쪽에 약간의 잡음(noise)가 발생하였다고는 하나 A′ 는 원형으로 표시된 B와 달리 A와 거의 합치하는 면적과 구성을 보인다. 다시 말해, A/B는 본질적으로 다른 특성을 보이지만 A/A′ 는 변이형(variant)의 일종으로 간주할 수 있다.

여기에서 적대적 사례의 중요한 두 특징을 볼 수 있다. 첫째, A가 B보다 A′ 와 더 가깝다는 것은 사람의 직관에 의한 것이다. 시각적 인지작용 등을 갖추지 못한 컴퓨터는 이 양자의 차이를 별도로 입력받고 계산하기 전까지는 쉽게 알 수 없다. 즉, 사람에게는 쉽고 컴퓨터에는 어려운 과제를 고안하는 것이 적대적 사례의 특이점이다. 이는 인공지능 평가 과제를 난도 높게 만들 수 있는 바탕이 된다. 둘째, 그림 1의 오른쪽 두 도형이 나타내는 바와 같이 잡음은 다양한 위상에 포함될 수 있다. 예컨대, 각 귀퉁이를 돌아가면서 같은 흠집을 내는 것만으로도 최소한 4개의 사례가 생산된다(A′ , A′ ′ , A′ ′ ′ , A′ ′ ′ ′). 즉, 하나의 적대적 사례가 확립되면 이를 확장하는 것은 상대적으로 용이하다. 이를 통해 딥러닝 연구자 또는 개발자는 성능 평가를 위한 데이터를 지속적으로 늘려갈 수 있다.

이상의 개념을 언어에서 단어 단위를 대상으로 적용해 보면 아래 (1)과 같다.

(1) a. 서울대학교 vs. 서을대학교
    b. 텔레비전 vs. 텔비레전

위 (1a-b)는 각기 정규형과 잡음을 삽입한 오류형으로 대립된다. 사람의 경우 '서을대학교'라고 오기가 되어 있다고 하더라도 그냥 '서울대학교'라고 읽거나 혹은 '서울대학교'를 잘못 입력한 것이라고 바로 알 수 있다. 글자의 순서를 뒤바꾼 (1b)의 경우도 마찬가지이다. 위에서 언급한 적대적 사례의 특성이 여기에도 적용된다. 첫째, 사람은 오류형(A′)을 보고도 정규형(A)과 쉽게 연결시킬 수 있지만, 컴퓨터는 이러한 유연한 인지능력을 발휘할 수 없다. 둘째, '서을대학교' 다음으로 '서울데학교', '서울대힉교' 등으로 사례를 쉽게 늘려갈 수 있다. 언어자료와 관련하여 한 가지 더 특색이 드러나는데, 문맥적 정보의 사용여부이다. 사람이 '서을대학교'를 '서울대학교'와 쉽게 연결할 수 있는 것은 '대학교'라는 뒷 문맥이 힌트 역할을 하기 때문이다. '텔비레전'도 마찬가지로 그 형태가 방송 등에 관련된 글 안에서 출현한 것이라면 사람은 더 쉽게 '텔레비전'을 연상할 수 있다.

단어 단위를 넘어 문장 단위에서 형태통사적 특징을 포착하면 다음 예시와 같다.

(2) A: 철수가 설악산에 올랐다.
A′ : 철수가 설악산을 올랐다.
B: ?철수가 결승을 올랐다.

표준국어대사전에 따르면, 용언 '오르다'가 '사람이나 동물 따위가 아래에서 위쪽으로 움직여 가다'는 뜻으로 쓰이면 보충어의 격틀을 【…에】 또는 【 …을】 형태로 취할 수 있다. 즉, 격 교체가 위 A 및 A′ 와 같이 가능하고 이들 사이에 별다른 수용성의 차이는 크게 없다. 그러나 B와 같이 의미가 미세하게 달라지면, 격 교체는 수용성을 떨어뜨린다. 한국어 형태통사 단계에서 이러한 격 교체 현상은 적대적 사례의 일반적 특성을 가진다. 사람은 직관이 있어 A′ 와 B의 차이를 쉽게 인지하지만, 인공지능 언어 모형이 이 관계를 얼마나 잘 포착할 수 있는지는 검증된 바 없기 때문이다.

이상과 같은 맥락에서 본 연구는 한국어 격 교체 구문이 실제 인공지능의 적대적 사례로 기능할 수 있는지를 확인한다. 언어 모형이 격 교체를 제대로 처리할 수 없다면, 언어 모형 평가에서 형태통사적 지식부터 반영하여야 한다. 반대로 언어 모형이 이미 격 교체를 다룰 수 있다면, 한국어 인공지능을 위한 적대적 사례는 더 심층으로 구현되어야 한다.

본고의 구성은 다음과 같다. 2절은 적대적 사례의 언어학적 예시를 더 다층적으로 다루고 다른 한편으로 격 교체 구문의 일반적 제약을 기술한다. 3절은 본 연구의 실험 데이터 구성과 실험 진행 과정에 관해 기술한다. 4절은 실험 결과를 제시하고, 5절은 그 결과가 가지는 함의에 대해 논한다. 끝으로 6절은 본 연구의 결론이다.

2. 배경

2.1. 전산적 배경: 적대적 사례

딥러닝 인공지능의 기반이 되는 신경망은 인간의 뉴런을 모방한 함수들의 네트워크로 구성된다. 이러한 네트워크의 복잡한 조합은 대량의 원시 텍스트 자료로부터 스스로 규칙성을 학습할 수 있게끔 한다. 원시 텍스트 자료 다시 말해 코퍼스를 이용하되, 현대 신경망은 어휘적 빈도에만 의존하는 단순 휴리스틱(heuristic) 모형이 아니다. 즉, 동사나 명사의 통계적 빈도와 무관한 언어 자질(인칭, 수) 사이의 규칙성을 포착하는 것이 딥러닝 인공지능이 보여주는 성능이다(Wei et al. 2021). 이러한 언어적 지식을 새로운 데이터에 적용하여 규칙성을 판별할 수 있다는 점이 딥러닝 인공지능의 놀라운 점이다. 이러한 특성은 자연언어 탐침에서 신경망 모형이 타 처리모형보다 대체로 우수한 성능을 보이는 주요 원인이 된다. 그러나, 딥러닝 인공지능의 언어적 지식은 아직 완벽하지 않기 때문에, 구체적인 언어 현상의 복잡성에 따라 매우 낮은 성능을 보이며 실패하기도 한다.

예를 들어, 단순한 주어-동사 수 일치나 재귀사 결속은 인간(96%)과 영어를 학습한 LSTM 언어 모형(83%~94%)의 차이가 크지 않다고 보고되고 있다(Marvin & Linzen, 2018; Hu et al., 2020). 반면에, 밀접한 두 사건의 인과 관계를 포착하는 추론 과제(자동자의 보닛을 열었다 → 자동차의 엔진을 검사하였다)에서는 인간(88%)과 LSTM 언어 모형(50% 미만)의 성능 차이가 크게 나타난다(Zellers et al., 2018). 이러한 편차는 신경망이 다양한 도메인에 존재하는 언어적 복잡성을 아직 유연하게 처리하지 못한다는 점을 보여준다.

후자의 예와 같이 신경망이 처리에 취약할 것으로 예상되는 항목을 평가 목적으로 수집한 데이터를 흔히 적대적 사례라고 한다(Szegedy et al., 2014; Goodfellow et al., 2015). 즉, 넓은 의미에서 문법적 오류는 적대적 사례의 하나라고 볼 수 있다. 공학적으로 말하였을 때, 적대적 사례는 원본 데이터에 의도적 잡음(intended noise)을 삽입하여 신경망의 예측 오류율을 높이는 데 목적이 있다. 언어 모형에 대한 적대적 사례는 다양한 언어 층위에서 동작할 수 있다. 예를 들어, 신경망 기계번역의 성능을 평가하고자 원문에 고의로 오탈자를 삽입할 수 있다. (3a)와 (4a)는 독일어이고 (3b)와 (4b)는 각각 대응하는 영어 신경망 기계번역이다. 독일어 원문 (3a)에 독일어 단어 'Psychoterapeut'를 오탈자 'Psy6hoterapeut'로 교체한 적대적 사례가 (4a)이다.

(3) a. Das ist Dr. Bob Childs – er ist Geigenbauer und Psychotherapeut.
b. This is Dr. Bob Childs – he's a wizard maker and a therapist's therapist.
(4) a. Das ist Dr. Bob Childs – er ist Geigenbauer und Psy6hothearpeiut.
b. This is Dr. Bob Childs – he's a brick maker and a psychopath.

(Ebrahimi et al., 2018)

정상적인 독일어 화자라면 ‘c’ 글자 하나를 ‘6’으로 바꾼 글귀를 보고 오탈자임을 바로 인지함은 물론이거니와, ‘Psy6hoterapeut’가 주어진 맥락에서 무엇을 의미하는가도 쉽게 파악할 수 있을 것이다. 이와 달리 기계번역 결과인 (3b)와 (4b)를 비교해보자. ‘therapist’s therapist’는 비교적 원문의 의미와 유사한 맥락을 공유하고 있으나, ‘psychopath’는 독일어 원문과 상충한다. 이처럼, 적대적 사례 평가는 데이터에 미미한 교체만 가해도 부정적인 예측 실패를 초래할 수 있음을 보여준다. 인간과 달리 컴퓨터는 유연한 인지작용을 수행하기 어렵기 때문이다.

최근에는 문장 완성 또는 클로즈 테스트(cloze test)에 착안하여 신경망을 훈련시킨 BERT 언어 모형이 제안되었다(Taylor, 1957; Devlin et al, 2019). BERT는 단어를 순차적으로 학습하는 LSTM과 달리, 문장 전체를 입력받아 스스로 완성하는 학습을 한다. 즉, 주변 문맥에 비추어 문장을 가장 자연스럽게 완성하는 단어가 무엇인지 스스로 유추하는 학습을 인공지능 모형으로 구현한 것이 BERT이다. 이러한 모델 훈련은 인간에 가까운 언어적 지식을 학습하는 과정에서 매우 효과적이었다. 그 결과, BERT는 단순한 오탈자 삽입과 같은 적대적 사례에 대하여 강건한 성능을 보였고, 더 복잡한 적대적 사례를 필요로 하게 되었다.

최근에는 BERT에 특화된 적대적 사례 평가 또한 개발되었다(Garg & Ramakrishnan, 2020; Jin et al., 2020; Sinha et al., 2021; Yanaka & Mineshima, 2021). 예를 들어, 긍정적인 또는 부정적인 영화 리뷰의 일부 어휘를 교체하여 적대적 사례 평가를 구성할 수 있다. (5a)는 부정적인 영화 리뷰 원문이고, (5b)는 원문의 부정적인 어휘 몇 개를 긍정적인 어휘로 교체했지만 문장 자체의 의미는 유사하여 부정적인 감성을 드러내는 적대적 사례이다. 즉, 원문의 부정적인 어휘들을 긍정적 어휘로 교체해도 인간은 문장에 나타난 감성을 여전히 부정적인 감성으로 인식한다. 그러나, 해당 연구에서 보고된 바에 따르면 이러한 단순 교체만으로도 BERT의 예측 정확도는 크게 떨어졌다(86.0% → 68.3%).

(5) a. The characters, cast in impossibly contrived situations, are totally estranged from reality. (원문, 데이터 라벨: 부정적인 감성)
b. The characters, cast in impossibly engineered circumstances, are fully estranged from reality. (적대적 사례, 데이터 라벨: 긍정적인 감성)
(Jin et al., 2020)

BERT는 통사구조에 그 이전 모형보다 민감하게 반응한다는 점에서(Goldberg, 2019), 어휘적 교체를 넘어 형태-통사 단계에서 적대적 사례 평가를 시도한 선행 연구도 존재한다. Sinha et al. (2021)은 영어 텍스트의 어순을 무작위로 뒤섞은 적대적 사례 말뭉치로 BERT를 사전 학습시켜 성능을 평가하였다. 놀랍게도, BERT의 성능에 통계적으로 유의미한 차이가 없었다. 이는 어순을 뒤섞어 문장 이해를 평가하는 방법이 예상과 달리 BERT에게 어렵

지 않았음을 의미한다.

그러나 이러한 결과와 상충하는 분석 사례도 존재한다. 일본어를 대상으로 Yanaka & Mineshima(2021)은 격 표지의 순서를 바꾸어 적대적 사례를 구성하였다. 아래 제시된 대립쌍 (6a)와 (6b)는 주격과 대격의 순서가 바뀌어 출현한 예이다.

(6) a. ライダー が サーファー を 助け出した
라이더가 서퍼를 도왔다.
b. ライダー を サーファー が 助け出した
라이더를 서퍼가 도왔다.
(Yanaka & Mineshima, 2021)

영어의 어순과 달리, 다국어 BERT와 일본어 BERT는 격 표지의 순서 변경에 대하여 매우 성능이 떨어졌고(49.9%~53.8%), 일본인 화자는 정확도는 매우 높았다(93.3%). 이는 유사한 언어 현상이라고 하더라도, 매개 언어의 구조적 특성에 따라 적대적 사례 평가의 결과가 달라질 수 있음을 시사한다.

이상과 같이, 인간의 언어직관에는 아주 쉬운 과제인데도 신경망 언어 모형은 실패를 겪을 것으로 예상되는 사례를 통해 인공지능의 고급 평가(견고성 또는 취약성)를 도모하는 것이 최근 인공지능 평가의 주요한 동향이다. 적대적 사례도 언어 층위에 따라 다양하게 구성할 수 있다. 언어학적 지식에 의존하지 않는 방법으로는 단순히 문장 일부를 오탈자로 교체하여 적대적 사례를 구성할 수 있다. 그러나 최근의 언어 모형은 표층적 층위의 단순 교체 정도에는 비교적 견고하게 작동하도록 개량되었다고 평가된다. 언어학적 지식을 근소하게 사용하는 방식으로 문장의 어순이나 격 표지를 변경하여 형태통사적 특성에 대한 이해를 묻는 적대적 사례 공격도 제안된 바 있다. 언어학적 지식을 더 깊이 있게 사용하자면, 문장의 진리치와 명제에 대한 화자 확신성 등을 바탕으로 한 언어 추론 과제를 적대적으로 운용할 수도 있다.

이러한 맥락에서 본 연구는 한국어의 격 교체가 BERT의 성능을 크게 떨어뜨리는 적대적 사례에 해당하는지 판별하는 실험을 수행한다. 개괄한 바와 같이 BERT 모형은 영어의 어순 뒤섞기에 견고하지만, 일본어의 격 표지 순서 변경에는 취약하다고 알려져 있다. 유형론적 관점에서 보면, 한국어의 격 표지는 영어의 어순 구조와 일본어의 격 표지와 유사한 기능을 수행한다. 그러나 BERT가 한국어의 격 교체 현상을 정확하게 이해하는지에 대하여는 알려진 바가 없다. 다른 한편으로, 본 실험은 신경망 언어 모형을 평가하기 위하여 언어학적으로 얼마나 복잡하고 어려운 적대적 사례가 필요한지를 경험적으로 뒷받침한다.

2.2. 언어학적 배경: 한국어 격 교체

2.2.1. 격 교체 현상

격 교체(case alternation) 현상은 둘 이상의 격조사가 같은 자리에 나타날 때 발생한다. 예컨대, 다음의 예문들에서 격 교체가 이루어지고 있음을 볼 수 있다.

(7) a. 철수가 포항{에/을} 갔다.
b. 암탉이 마당{에서/을} 나왔다.
c. 영희가 동생{에게/을} 선물을 주었다.

(7a)의 경우에는 위치를 나타내는 부사격 조사 '에'와 목적격 조사 '을'이 같은 자리에 나타날 수 있다. 마찬가지로 (7b)의 경우에도 위치를 나타내는 부사격 조사 '에서'와 목적격 조사 '을'이 같은 자리에 나타날 수 있다. 그리고 (7c)의 경우에는 수혜자(beneficiary)를 나타내는 부사격 조사 '에게'와 목적격 조사 '을'이 같은 자리에 나타날 수 있다(Park & Yi, 2021).

격 교체는 교체되는 조사의 유형과 문형에 따라 여러 가지 양상을 보인다. 송창선(2019)는 한국어의 격 교체 현상에 네 가지 양상을 정리하였다. 다음의 예문들에서 격 교체 현상의 네 가지 양상을 확인할 수 있다.

(8) a. 철수가 학교{에/를} 간다.
b. 물이 얼음{이/으로} 되었다.
c. 우리 어머니{가/의} 손이 크다.
d. 철수가 영희{를/에게} 선물을 주었다.

(8a)는 부사격 조사와 '을/를'의 교체가 가능한 경우이다. 부사격 조사 '에'가 '를'로 교체될 수 있음을 볼 수 있다. (8b)는 '되다' 구문의 격 교체에 해당한다. 보격 조사 '이'가 '으로'로 교체될 수 있음을 볼 수 있다. (8c)는 이중 주어 구문의 격 교체에 해당한다. 주격 조사가 중복되어 나타나는 이중 주어 구문에서 첫 번째 주격 조사 '가'가 '의'로 교체될 수 있음을 볼 수 있다. 마지막으로 (8d)는 이중 목적어 구문의 격 교체이다. 목적격 조사가 중복되어 나타나는 중복 목적어 구문에서 첫 번째 목적격 조사 '를'이 '에게'로 교체될 수 있는 경우이다. (8b-d)에서 예시된 격 교체의 나머지 유형들에 대해서도 격 교체가 허용되지 않는 경우가 있으나, 본 연구에서는 분석의 범위를 (8a)에서 예시된 부사격 조사와 '을/를'의 교체에 한정하고자 한다.

물론 격 교체가 항상 허용되는 것은 아니다. 아래의 예문들은 부사격 조사와 '을/를'의 교체가 항상 허용되는 것은 아님을 보여준다.

(9) a. 철수가 수학여행을 갔다.
b. ??철수가 수학여행에 갔다.
(10) a. 철수가 영희와 결혼했다.
b. *철수가 영희를 결혼했다.

(9)의 경우에는 목적격 조사에서 부사격 조사로의 격 교체가 허용되지 않는 경우이다. 목적격 조사 '을'이 쓰인 (9a)와 달리 부사격 조사 '에'가 쓰인 (9b)는 수용성이 높지 않다. 이와 반대로 (10)의 경우에는 부사격 조사에서 목적격 조사로의 격 교체가 허용되지 않는 경우이다. 부사격 조사가 쓰인 (10a)와 달리 목적격 조사가 쓰인 (10b)는 비문이다. 다음은 부사격 조사와 '을/를'의 교체가 허용되는 경우와 그렇지 않은 경우를 함께 분류한 표이다.

표 1. 부사격 조사와 '을/를' 간의 격 교체 가능/불가능한 경우

목적격 조사 '을/를'	부사격 조사	예문
Y	N	철수가 수학여행{을/*에} 갔다.
N	Y	철수가 영희{*를/와} 결혼했다.

(i)은 목적격 조사 '을/를'만이 허용되며 부사격 조사로의 격 교체가 허용되지 않는 경우이다. 그리고 (ii)는 반대로 부사격 조사만이 허용되며 목적격 조사 '을/를'로의 격 교체가 허용되지 않는 경우이다. 본 연구에서는 (ii)를 대상으로 실험 문장을 구성하였다.

2.2.2. 격 교체 현상과 조사

앞서 밝혔듯이, 본고는 한국어 격 교체 양상에 있어 분석의 범위를 조사 '을/를'과 부사격 조사 간의 교체로 국한한다. 가장 큰 이유는 격 교체 현상 가운데서도 한국어에서 특징적인 형태통사적 분포를 보일 것으로 예상되는 문법 요소에 분석을 집중하기 때문이다. 즉, 인간에게는 쉽고 컴퓨터에는 어려운 적대적 사례의 측면에서 조사 '을/를'의 특수성을 고려하였다.

이홍식(2004)은 조사 '을/를'의 특수성을 강조하며 목적격 조사 '을/를'을 (격 표지도 아니고 의미역 표지도 아닌) 독자적 의미를 지닌 조사로 설명하고 있다.[1] 여기서 조사 '을/를'은 동작의 대상이라는 관계를 표현하는 것으로 파악되며 이에 더해 그 의미와 사용이 확장될 수도 있다. 동작의 대상에서 대상이 아닌 성분으로 확장될 수 있으며 동작을 상정할 수 없는 문장에 사용되어 대상만이 강조될 수도 있다. 아래의 예문들은 조사 '을/를'의 의

1) 조사 '을/를'의 격을 지칭하는 용어에 있어 기존의 연구에서는 목적격과 대격의 두 용어가 혼용되었으나 이홍식(2004)을 따라 목적격으로 지칭하였음을 밝힌다. 이와 함께 혼용되었던 '부사격'과 '사격'의 경우도 일관되게 부사격으로 지칭하였다.

미가 확장된 경우를 보여주고 있다.

(11) a. 철수는 학교에 갔다.
b. 철수는 학교를 갔다.
c. 철수가 세 시간을 잤다.
d. 영희가 예쁘지를 않다.
e. 영희가 생김새가 예쁘지를 않다.
(이홍식 2004)

(11a-b)의 경우, 조사 '을/를'과 부사격 조사 '에' 사이의 교체가 이루어지고 있다. 이 경우에는 대상의 확장이 이루어진 것인데 두 문장 (11a-b) 사이에는 의미역의 차이가 있는 것으로 이해된다. (11c)에서 '세 시간을'은 단순한 대상이 아니라 자는 행위의 시간적 범위를 나타내며 대상의 의미가 확장된 경우로 설명된다. (11e)의 경우에는 (11d)와 비교했을 때 동작이 축소되고 대상이 강조되는 경우를 보여주고 있다. (11d)는 영희가 예쁜 짓을 하지 않는다고 해석될 수 있으나 (11e)에서는 영희의 생김새라는 특징이나 상태에 대한 부정이 이루어질 뿐이며 동작이 개입될 여지는 보이지 않는다. 위의 예시 가운데 (11a-b)에 집중하고, 대응 조사가 불명확한 (11c-e)의 예시는 후행 연구에서 다루기로 한다.

2.2.3. 격 교체 현상과 동사

조사 '을/를'의 성질에 이어 동사에 초점을 맞추어 격 교체 현상의 가능성에 대해 접근해 볼 수 있다. 우형식(1996)에서는 목적격 조사 '을/를'과 부사격 조사 '에' 간의 교체가 가능한 경우를 도달성 이동 동사, 태도 동사, 결과-상황 동사의 세 가지 유형으로 분류하여 설명하고 있다. 이에 더해 김미령(2004)에서도 우형식(1996)의 분류와 유사하게 목적격 조사 '을/를'과 부사격 조사 '에' 간의 교체가 가능한 경우를 이동 동사, 태도 동사, 그리고 결과-상황 동사의 세 가지 유형으로 분류하고 있다. 격 교체를 허용하는 이 세 가지 유형을 더 자세히 살펴보면 표 2와 같다.

표 2. 부사격 조사 '에'와 '을/를'의 교체가 허용되는 경우

유형	세부 유형
이동 동사	[NP(행위자)+이/가] [NP(물리적 공간)+에/를] V
	[NP(행위주)+이/가] [NP(추상적 공간)+에/를] V
태도 동사	[NP(행위주1)+이/가] [NP1(행위주2)의 NP2(서술성 명사)]+에/를] V
	[NP(행위주)+이/가] [NP(추상적 대상이나 사태)에/를] V
결과-상황 동사	NP(THEME)+이/가 NP(처소성)+에/를 V
	NP(THEME)+이/가 NP(사태지시성)+에/를 V

부사격 조사와 '을/를' 간의 교체가 가능한 첫 번째 경우는 이동 동사의 경우이다. [NP(행위자)+이/가] [NP(물리적 공간)+에/를] V에 해당하는 이동 동사의 경우, 복합동사가 기본 의미를 가지는 동사와 결합하여 격 교체가 가능해지는 경우가 많다는 것이 특징이다.

(12) a. 아기가 방바닥{*에/을} 기었다.
b. 아기가 방바닥{에/을} 기어간다.
c. 그는 운동장{*에/을} 달렸다.
d. 그는 운동장{에/을} 달려나갔다.
(김미령 2004)

이렇게 복합동사와 기본 의미를 가지는 동사 간의 결합으로 격 교체가 가능해지는 현상은 두 동사의 논항 구조 간에도 결합이 발생하여 격 교체의 허용으로 이어지는 것으로 이해된다. 이어서 [NP(행위주)+이/가] [NP(추상적 공간)+에/를] V의 경우에 해당하는 예시는 (13)과 같다.

(13) a. 그는 우리 동아리{에/를} 가입했다.
b. 부모들이 아이들의 싸움{에/을} 개입해서는 안된다.
(김미령 2004)

부사격 조사와 '을/를' 간의 교체가 가능한 두 번째 경우는 태도 동사의 경우이다. 태도 동사의 경우는 [NP(행위주1)+이/가] [NP1(행위주2)의 NP2(서술성 명사)]+에/를] V와 [NP(행위주)+이/가] [NP(추상적 대상이나 사태)에/를] V의 두 가지 경우로 나뉠 수 있다. 우선 [NP(행위주1)+이/가] [NP1(행위주2)의 NP2(서술성 명사)]+에/를] V에 해당하는 예시는 (14)와 같다. 다음으로 [NP(행위주)+이/가] [NP(추상적 대상이나 사태)에/를] V에 해당하는 예시는 (15)와 같다.

(14) a. 그의 부모는 그의 행동{에/을} 자주 간섭한다.
b. 그는 상사의 명령{에/을} 거역했다.
(김미령 2004)
(15) a. 우리는 항상 전쟁{에/을} 대비해야 한다.
b. 그는 이번 입시{에/를} 철저히 준비했다.
(김미령 2004)

마지막으로 부사격 조사와 '을/를' 간의 교체가 가능한 세 번째 경우는 결과-상황 동사

의 경우이다. NP(THEME)+이/가 NP(처소성)+에/를 V의 경우와 NP(THEME)+이/가 NP(사태지시성)+에/를 V의 경우 두 가지로 나뉠 수 있다. NP(THEME)+이/가 NP(처소성)+에/를 V의 경우는 (16)과 같이 어떠한 장소나 위치를 나타내는 처소성을 가지고 있다. NP(THEME)+이/가 NP(사태지시성)+에/를 V의 경우에는 (17)과 같이 다소 추상적인 상황을 나타내는 사태지시성을 가지고 있다.

(16) a. 우리 마을은 강가{에/를} 못 미쳐 있다.
b. 그녀의 자리는 방 출구{에/를} 접해 있다.
(김미령 2004)

(17) a. 우리 회사는 파산 위기{에/를} 직면했다.
b. 그의 재능은 아버지의 재능{에/을} 앞선다.
(김미령 2004)

김미령(2004)에 이어 이종근(2006)에서는 이동 동사, 태도 동사, 그리고 결과-상황 동사의 세 가지 유형 외에 사역동사 구문에서도 목적격 조사 '을/를'과 부사격 조사간의 격 교체가 허용됨을 보이고 있다. (18)의 예문들은 모두 사역동사 구문에 해당하는데 '을/를'과 부사격 조사 간의 격 교체가 허용됨을 볼 수 있다.

(18) a. 그는 상관한테 욕을 먹으면 하급자{에게/를} 못살게 군다.
b. 선생님의 지시대로 반장이 지각한 아이들{에게/을} 청소를 시켰다.
b. 부모님{께/을} 욕 먹이는 행동을 해서는 안 된다.
(이종근 2006)

이종근(2006)은 사역동사 구문에서의 '-에(게)/를' 격 교체가 발생하는 것에 대해 사역동사의 특성에 초점을 맞추어 설명하고 있다. 사역동사가 영향력이 강한 동사이며 격 교체의 허용은 사역동사의 내부 논항이 Patient-Proto-Role의 의미 속성을 갖춘다는 것에 기인한다는 것이다.

격 교체는 단순한 격조사의 교체가 아니라 동사의 유형과 성질에 의해서도 영향을 받는다는 현상이라는 것은 한국어 외의 언어에서도 찾아볼 수 있다. Fukuda(2020)에서는 일본어의 목적격과 여격 사이의 격 교체 현상에 대해 다루고 있는데 격 교체를 허용하는 동사를 기준으로 하여 일본어의 격 교체 현상을 목적격-원천(source) 동사의 경우와 목적격-목적(goal) 동사의 경우의 두 가지로 나누어 설명하고 있다.

(19) a. Taroo-ga yama-o/ni nobot-ta
T-NOM mountain-ACC/GOAL climb-PST
'타루가 산을 올랐다.'
b. Taroo-ga ie-o/kara de-ta
T-NOM home-ACC/SOURCE come.out-PST
'타루가 집을/집으로부터 떠났다.'

(Fukuda 2020)

(19a)는 목적격-원천 동사의 경우에 해당하고 (19b)는 목적격-목적 동사의 경우에 해당한다. 목적격-원천(source) 동사의 경우에는 타동사 구조와 비목적격(unaccusative) 구조와 연관되는 반면, 목적격-목적(goal) 동사의 경우에는 두 개의 서로 다른 타동사 구조와 연관된다.

종합적으로 고려해 볼 때, 한국어와 일본어의 격 교체를 다룬 기존의 선행연구에서 볼 수 있듯이 격 교체 현상은 격 표지 역할을 하는 조사뿐만이 아니라 해당 조사와 함께 사용되는 동사의 유형과 성질에 의해서도 영향을 받는 복합적인 언어 현상이라고 할 수 있다. 즉, 격 교체에 있어 단순히 조사에 국한된 분석을 하기보다 조사와 함께 사용된 동사, 더 나아가 격 교체가 발생하는 해당 문장의 격틀 구조를 대상으로 한 분석이 요구된다.

2.2.4. 세종 전자사전

이상과 같은 이유에서 본 연구는 격 교체 현상에 대해 동사 중심의 격틀 구조 차원에서의 분석을 시도하였다. 한국어에서 격틀의 구조를 중심으로 격 교체 현상을 동사의 특성에 근거하여 잘 정리한 대표적인 언어자원이 '세종 전자사전'이다. 본 실험에서는 세종 전자사전의 동사별 격틀을 준용하고 그에 따라 정제된 예문을 활용함을 원칙으로 한다. 이를 통해 격틀 구조 차원에서 분석이 요구되는 복합적인 언어 현상에 해당하는 격 교체 현상에 대한 딥러닝 모델의 이해 능력을 시험하였다.

세종 전자사전은 언어 정보 처리를 위해 최적화된 자원으로 현대 한국어 사용자의 어휘 지식을 폭넓게 반영하고 있다(홍재성 2007). 따라서 이론과 실증의 균형을 이루기에 적합한 바탕이 된다. 김미령(2004) 및 Fukuda(2020) 등의 이론적 선행 연구는 대부분 특정 부사격 조사 '에'와 '을/를'의 교체 가능성에 대해서만 다루었다. 이에 비해, 실증적 연구를 지향하는 본 연구는 부사격 조사의 범위를 '에'에 국한하지 않고 부사격 조사 일반과 '을/를' 간의 격 교체에 대해 다루고자 한다. 세종 전자사전은 이와 같은 본 연구의 분석 취지에도 잘 부합하는 예문들을 다수 포함하고 있다. 한국어의 여러 부사격 조사의 분포를 격틀 안에서 표상한 연구 결과물이기 때문이다.

세종 전자사전에 기술된 예문을 사용하는 것은 실험 구성의 측면에서도 이점을 가진다.

실험을 위한 예문을 구성하여 한국어 화자들의 검증 과정을 거치거나, 일반적인 코퍼스 예문을 사용하는 경우, 개인어의 차이에 따른 격틀의 차이나 언어 사용에서의 오류 등을 고려하여 데이터를 수정해야 하는 수고가 발생한다. 거기에 더해, 각 문장에 사용된 술어가 이번 실험을 위해 알맞은 격틀을 갖추고 있는지를 검증해야 한다. 이에 반해, 세종 전자사전의 경우 기술 체계와 지침에서 문형정보와 격조사에 대한 정보를 기술하도록 규정하여 격틀에 대한 정보를 이미 포함하고 있다. 따라서, 실험 절차를 간소화하여 시간과 비용을 절약하고, 이번 실험의 핵심이라 할 수 있는 격틀 정보에 맞게 구성된 예문들을 활용하기 위해 전문가들에 의해 구성된 세종 전자사전의 예문을 사용했다.

3. 방법

3.1. 데이터

데이터는 크게 두 종류로 나뉜다. 부사격과 목적격 사이의 격 교체가 가능한 acc-obl 데이터와 부사격과 목적격 사이의 격 교체가 불가능한 non-acc 데이터이다. 각 데이터는 사전에 등재된 동사의 격틀을 기준으로 선정되었다. 격틀의 논항에 부사격 조사와 목적격 조사가 동시에 등록되어 서로 교체가 가능한 경우에는 acc-obl에 포함되었고, 격틀 논항에 부사격 조사만 존재하는 경우에는 non-acc에 포함시켰다.

표 3. 실험 예문 구성

	전체 예문	제외 예문	실험 예문
acc-obl (alterable)	1,014	158	856
non-acc (inalterable)	1,500	771	729

acc-obl 데이터의 예문의 경우 총 1,014개의 예문 중 157개의 예문을 기준에 따라 제외하고 총 856개, non-acc 데이터의 예문의 경우 총 1,500개의 예문 중 실험에 부적절한 예문 771개를 제외한 729개로 각각의 경우 최소 700개 이상의 충분한 예문을 사용하여 실험을 진행했다. 예문을 구성하는 데 있어 부사격과 목적격 사이의 격 교체가 가능한 예문들과 가능하지 않은 예문들에 동일한 기준이 적용되었고, 격 교체가 불가한 예문 목록을 고려하여 따로 추가된 기준들이 존재한다.

non-acc 데이터의 경우 별도의 검증 과정을 거쳤다. 먼저 1차 작업자가 1,500개의 제외 기준에 부합하거나 격 교체가 가능해 보이는 예문들을 제외했다. 이후 2차 작업자들이 예문을 최종적으로 검증하는 과정을 거쳐 non-acc 데이터의 예문들의 경우 격 교체가 불가능해 보이는 예문들만을 포함하였다. 2차 검증에서는 총 8명의 주석 작업자들이 예문 검증 과정

에 참여하였으며, 작업자마다 약 300개씩의 문장을 검증했다. 그 과정에서 한 문장을 두 작업자가 교차로 검증하여 양자 모두 격 교체가 가능한 문장이라고 동의를 하는 경우에만 실험 예문에서 제외했다.

먼저, 다음의 기준들은 격 교체 가능 목록과 격 교체 불가능 목록에 같게 적용된 기준들이다.

(20) 매도되다(X=N0-이 Y=N1-에게|에의해 Z=S2-다고 V): 기존 간부들을 모두 새 지도부에 의해 축출 대상으로 매도되었다.
(21) 입히다(X=N0-이 Z=N2-에|에게|을 Y=N1-을 V): 엄마는 발버둥을 치는 동생을 붙잡고 옷을 입혔다.
(22) 부속하다(X=N0-이 Y=N1-에 V):
a. 정부 기관에 부속한 연구 기관들은 대개 일반 직장보다 편안한 직장 생활을 한다.
b. 정부 기관을 부속한 연구 기관들은 대개 일반 직장보다 편안한 직장 생활을 한다.
(23) 놓다(X=N0-이 Z=N2-에|을 Npr1-을 V): 그들은 우리가 들어가려는 데 계속 훼방을 놓았다.
(24) 거들다(X=N0-이 Y=N1-에|을 V): 이번 일에는 좀 거들지 말아라.

첫째, 예문 원본이 비문이거나 오타가 존재하는 경우 실험용 예문에 포함하지 않았다. (20)의 경우, 주어인 '기존 간부들'에 주격 조사인 '-은'이 아닌 목적격 조사 '-을'이 붙는 오타가 포함되어 비문이기 때문에 제외되었다. 그러나, 띄어쓰기가 오류가 있는 경우는 데이터 토큰화 과정을 거치면서 띄어쓰기 오류의 영향이 없어질 것으로 보고 실험용 예문에 포함시켰다. 둘째, 사전에 등록된 격틀의 논항이 모두 명시적으로 실현되지 않으면 예문 목록에 포함시키지 않았다. (21)이 설명하는 '입히다'에 해당하는 격틀은 'X=N0-이 Z=N2-에|에게|을 Y=N1-을 V'이다. 이 경우 술어 '입혔다'의 격틀 중 'Z=N2-에|에게|을'에 해당하는 논항이 명시적으로 실현되지 않았기 때문에 제외했다. 셋째, 부사격 조사에서 목적격 조사로 격조사를 교체하는 것은 가능하지만 두 문장의 의미가 서로 다르다면 실험 예문에서 제외했다. 위 (22a)는 연구 기관이 정부 기관에 속해 있다는 것을 함의하는 반면, (22b)에서는 정부 기관이 연구 기관에 속해 있다는 것을 함의하기 때문에 문장 전체의 의미가 달라진다. 이런 경우는 격 교체 현상을 살펴보기 위한 예문으로 적절하지 않기에 데이터에서 제외되었다. 넷째, 격틀 논항의 격조사가 명시적으로 실현되지 않으면 실험 예문에서 제외했다. (23)는 술어 '놓았다'의 논항 중 하나인 의존명사 '데'에 격조사가 부착되지 않았기 때문에 제외되었다. 끝으로, 격조사에 보조사가 부착되어 사용된 예문들도 예문 목록에 포함시키지 않았다. (24)에서는 '거들지'의 격틀 논항인 '일'에 격조사인 '에'가 실현되어 있지만, 그 뒤에 보조사 '는'이 붙어있기 때문에 예문 목록에 포함되지 못했다.

다음은 부사격과 목적격 사이의 격 교체가 불가능한 예문 목록에 추가로 적용한 기준들이다. 부사격 논항들이 목적격으로의 격 교체가 불가능한, 즉, 목적격이 사용되면 비문이 되는 예문들만 포함했다는 점에서 격 교체 가능 예문 목록의 예문들의 성격과 근본적인 차이가 존재한다.

(25) 신뢰하다(X=N0-이 Y=N1-와 서로 V):
  a. 여당은 야당과 서로 신뢰한다.
  b. 여당은 야당을 신뢰한다.
  c. ?여당은 야당을 서로 신뢰한다.
(26) 경주하다(X=N0-이 Y=N1-와 (서로) V):
  a. 선생님은 운동회에서 철수가 영희와 서로 경주하는 모습을 비디오 카메라에 담았다.
  b. *선생님은 운동회에서 철수가 영희를 (서로) 경주하는 모습을 비디오 카메라에 담았다.
(27) 가시다(X=N0-이 Y=N1-이 V): 그 약을 먹었더니 변비가 싹 가셨다.

첫째, 격틀의 논항 중 격조사 '-와/과 (서로)'가 붙어 실현되는 논항이 존재할 때 별도의 선별 기준을 마련하였다. '와/과'를 목적격 조사 '을/를'로 교체했을 때 부사 '서로'의 존재 여부와 상관없이 비문을 형성한다면 실험 예문으로 선택하였다. 예컨대, '신뢰하다'를 사용한 예문의 경우, (25c)와 같이 '서로'가 존재하면 그 수용성이 떨어지지만 (25b)에서와 같이 '서로'가 존재하지 않을 때는 정문이기 때문에 제외했다. 반면, (26)는 (26b)에서 볼 수 있듯이 '서로'의 존재 여부와 상관없이 비문이기 때문에 예문에 포함하였다. 둘째, 서술어의 격틀에 부사격이 존재하지 않으면 예문으로 사용하지 않았다. 예컨대, (27)의 '가시다'와 같은 동사는 격틀이 'X=N0-이 Y=N1-이 V'로 예문에 부사격이 존재하지 않는다. 이러한 예문들은 이 논문의 대상인 부사격과 목적격 사이의 격 교체 현상으로 보지 않고 제외했다.

위의 기준들을 통해 선정한 예문들에서 서술어의 격틀에 해당하는 부사격 조사의 자리에 [MASK]를 집어넣어 부사격과 목적격 사이의 격 교체 현상을 딥러닝 모델이 구분할 수 있는지 실험을 진행하였다. [MASK] 적용에서는 다음의 예외적인 경우들이 존재했다.

(28) 분패하다(X=N0-이 Z=N2-에서 Y=N1-에|에게 V): 김선수는 세계 유도 선수권 대회에서 프랑스 선수에게 분패했다.
(29) 곁눈질하다(X=N0-이 Y=N1-에게|에게로|을 V): 철수는 그 시험 시간에 나를 곁눈질하다가 걸렸다.

격틀 논항 중 논항이 세 개 이상이고, 그중 두 개 이상이 부사격 논항인 경우, Y 논항

을 [MASK]로 선택했다. (28)의 경우, 격틀에 부사격 논항이 Z와 Y, 두 개가 존재한다. 이 때, Y 논항이 '-에|에게'이기 때문에, '프랑스 선수' 뒤의 격조사에 [MASK]를 씌워 주었다. 다음으로, 사전에 기재된 격틀에는 논항이 2개만 존재하지만, 예문에 격틀이 3개로 실현된 경우에는 사전에 등재된 격틀 논항을 [MASK]로 선택했다. (29)의 경우 원래 격틀에는 존재하지 않는 부사격인 '시험 기간에'가 문장에 사용되었다. 이러한 경우, 격틀에 포함되지 않는 부사격은 고려하지 않고, 격틀에 포함된 논항이 실현된 '나' 뒤의 격조사에 [MASK]를 입력했다.

실험에 사용된 최종 데이터의 형태는 다음과 같다.

표 4. 최종 데이터 형태

동사	승압되다
격틀	X=N0-이 Z=N2-에서 Y=N1-로 V
예문	이 지역은 전부 220 볼트로 승압되었다.
실험 문장	이 지역은 전부 220 볼트[MASK] 승압되었다.
목적격 MASK	##를
부사격 MASK	##로

표 4와 같은 방식은 이 실험에 사용되는 BERT의 특성에 기인한다. 첫째, [MASK]에 각 항목이 대입되었을 때 통사론적 관점에서의 최소대립쌍이 아닌 [MASK]모델에서의 최소대립쌍이 될 수 있도록 구성했다. 예시는 (30)과 같다.

(30) a. 철수가 책을 안 읽었다.
b. 철수가 책을 읽지 않았다.
(이규민 외, 2021)

(30)의 경우, 통사론적 관점에서는 단형 부정과 장형 부정의 차이로 최소대립쌍을 이루는 문장이지만, [MASK] 모형의 관점에서는 두 개 이상의 토큰에서 차이가 발생하기 때문에 최소대립쌍이라고 볼 수 없다. 따라서, 이 실험 예문들은 모두 하나의 토큰에서만 차이가 발생하여 최소대립쌍 문장이 만들어질 수 있도록 [MASK]와 [MASK]에 들어갈 항목들을 구성했다. 둘째, 다음으로 BERT의 토큰화 방식을 고려했다. BERT의 경우 워드피스 방식을 통해 문장을 토큰화한다(Wu et al., 2016). 워드피스 토큰화는 단어를 서브워드(subword)로 분절한 이후 다시 조합하는 과정을 통해 토큰화를 하므로 어휘 목록에 존재하지 않는 단어(Out of Vocabulary)를 처리하는 부분에 이점이 있지만, 그 과정에서 언어의 형태통사적인 규칙이 아닌 모델의 가능도를 높이는 방향에 맞추어 수학적으로 이루어지기 때문에 한국어

와 같이 어절 단위가 많은 언어에서는 토큰화가 잘못 이루어질 수 있다는 문제점이 있다(박권식 외, 2021). 이러한 문제가 [MASK]에 들어갈 항목의 토큰화 과정에 영향을 미치지 않도록, 서브워드 토큰의 경계를 나타내기 위해 사용하는 기호인 '##'를 [MASK]에 들어갈 항목들의 앞에 부착하여 나타냈다.

3.2. 실험

본 연구는 딥러닝 언어 모형 중 가장 널리 쓰이는 동시에 가장 높은 성능을 보이는 모델 중 하나인 BERT(Bidirectional Encoder Representations from Transformers; Devlin et al., 2019)를 기반으로 한 한국어 사전학습 모델을 활용한다. BERT 기반 여러 한국어 사전학습 모델을 비교한 박권식 외(2021)의 결과에 따라 한국어 사전학습 모델 중 높은 성능을 보이는 KR-BERT(KoRean based BERT pre-trained; Lee 외, 2020)를 채택하였고, KR-BERT의 결과와 비교할 수 있는 베이스라인(baseline) 모델로 가장 낮은 성능을 보이는 mBERT (Multilingual BERT; Devlin 외, 2019)를 채택하였다.[2] KR-BERT와 mBERT의 단어장 및 파라미터 크기는 아래와 같다.

표 5. KR-BERT와 mBERT

모델	단어장 크기	파라미터 크기	특징
KR-BERT	16,424	99,265,066	한국어 특정적 모델
mBERT	119,547	167,346,416	104개 Wikipedia를 학습데이터로 활용

본 연구에서 활용하는 평가 지표는 surprisal이다(Levy, 2008; Hale, 2001). Surprisal은 모델이 어떤 단어를 만났을 때 '얼마나 놀라는지'에 대한 측정값으로, BERT의 surprisal 값은 특정 단어가 그 단어를 좌우로 둘러싼 주변 단어들을 환경으로 나타날 수 있는 로그 확률의 역수로 계산된다. 따라서 특정 단어가 출현할 확률이 높을수록 surprisal 값은 낮아지며, surprisal 값이 낮아질수록 모형 입장에서 해당 단어에 대한 수용성이 높아진다고 해석할 수 있다. 만약 인간의 입장에서 수용성이 높은 단어에 대해 모델이 출력한 surprisal 값이 낮게 나타나면, 이는 모델이 해당 구문을 적절히 학습했음을 암시한다. Meister et al.(2021)는 surprisal에 근거한 BERT의 확률적 추론이 인간의 수용성 판단과 높은 상관관계를 가진다는 결과를 보고하였다.[3] 또한, 이외 여러 선행 연구에서 surprisal이 활용된 바 있

2) 박권식 외 (2021)에 따르면 KorBERT(Korean Bidirectional Encoder Representations from Transformers; 한국전자통신연구원, 2019)도 KR-BERT와 같이 높은 성능을 보이지만 세부 하위 과업(task) 중에서 KR-BERT가 KorBERT보다 더 높은 성능을 보인 과업이 더 많았다. 본 연구는 모델 간 비교가 주된 목적이 아니므로 KR-BERT를 대표 모델로 선정하였다.

3) 문장 전체의 음의 로그 확률의 평균을 뜻하며, 해당 논문에서는 pseudo-surprisal 값이라고 표현하였다

으므로(예, Park 외, 2021; 이규민 외, 2021; Da Costa & Chaves, 2020; Wilcox 외, 2019), 본 연구는 surprisal을 평가 지표로 활용한다.

구체적으로 본 연구에서 활용하는 surprisal 측정 예시는 다음과 같다.

(31) a. 문장: 나는 할머니[MASK] 콩을 먹지 않는다고 타박맞았다.
b. surprisal(목적격 ##를/부사격 ##에게): 9.698370934/6.859820843

(31a)는 문장에서 목적격 혹은 부사격 조사가 들어갈 자리에 마스크 토큰을 집어넣은 형태이다. (31b)는 마스크 토큰에 목적격 혹은 부사격 조사를 입력했을 때 KR-BERT 모델이 출력한 각 조사의 surprisal 값이다. (31a) 문장은 부사격의 목적격으로의 격 교체가 허용되지 않는 예문이기 때문에, 모델이 인간의 수용성 판단을 잘 학습하였다면 목적격 조사의 surprisal 값이 부사격 조사의 surprisal 값보다 높아야 한다. 따라서 (31b)의 결과는 모델이 격 교체가 불가능함을 학습하였다는 것을 나타내는 지표가 된다.

4. 결과

격 교체를 허용하지 않는 데이터에 대한 mBERT와 KR-BERT의 정확도는 각각 81.76%와 92.46%이다. 그림 2는 격 교체를 허용하지 않는 구문에 대한 mBERT와 KR-BERT의 surprisal 분포를 나타낸다.

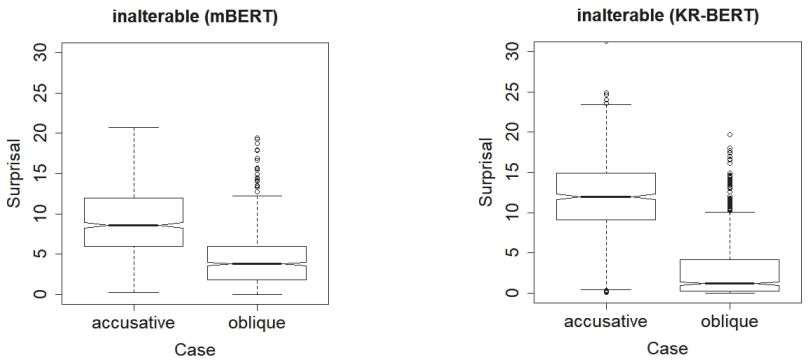

그림 2. 격 교체 비허용 구문에 대한 surprisal 분포

대응표본 t-검정 결과 격 교체를 허용하지 않는 구문에 대해서 mBERT와 KR-BERT 모두 부사격과 목적격의 surprisal 분포가 통계적으로 유의미한 차이를 보였다(mBERT:

t(728)=23.816, p<0.001 '***', KR-BERT: t(728)=43.056, p<0.001 '***'). 이는 mBERT와 KR-BERT 모두 격 교체 비허용 구문에 대해 목적격보다 부사격을 사용하는 것이 더 수용 가능하다고 적절히 판단한다는 것을 가리킨다. 따라서 mBERT와 KR-BERT 모두 한국어에서 격 교체가 허용되지 않는 용언의 보충어 격틀 구조를 잘 학습하였다고 할 수 있다.

그림 3은 격 교체를 허용하는 구문에 대한 mBERT와 KR-BERT의 surprisal 분포를 나타낸다.

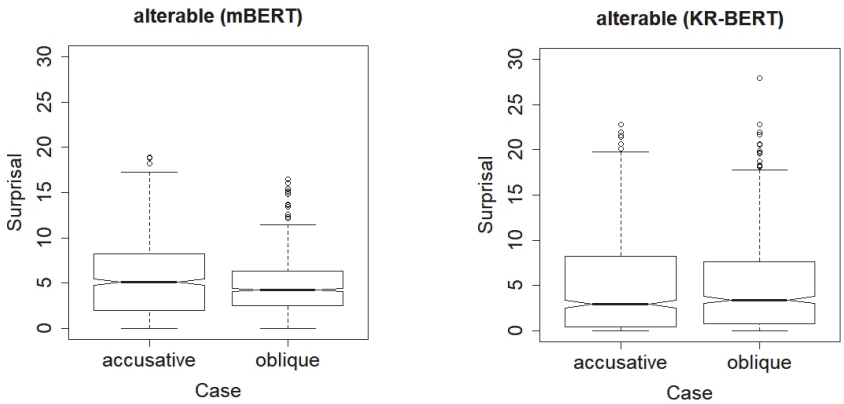

그림 3. 격 교체 허용 구문에 대한 surprisal 분포

대응표본 t-검정 결과 격 교체가 허용되는 구문에 대해서는 KR-BERT의 경우 surprisal 분포와 목적격의 surprisal 분포가 통계적으로 유의미한 차이를 보이지 않았지만 (t(855)=-0.27453, p<1 'ns'), mBERT는 유의미한 차이를 보였다(t(855)=6.0558, p<0.001 '***'). 격 교체가 허용되는 구문에 대해서는 목적격 조사와 부사격 조사의 분포가 통계적으로 유의미한 차이가 없어야 모델이 적절히 학습하였다고 할 수 있다. 따라서 그림 3의 결과는 격 교체가 허용되는 구문에 대해서 KR-BERT는 잘 학습하였지만 mBERT는 적절히 학습하지 않았다는 것을 의미한다.

그림 3의 결과는 KR-BERT가 한국어에서 격 교체가 가능한 용언의 격틀 구조까지도 학습했다는 것을 의미한다. 반면 mBERT의 경우 그림 2의 결과보다 목적격과 부사격의 분포가 서로 가까워진 것은 사실이지만, 두 분포가 통계적으로 유의미한 차이를 보이기 때문에 해당 현상을 적절히 학습했다고 할 수 없다. 이는 한국어에 특정적이지 않은 다국어용 사전학습모델보단 한국어에 특화된 사전학습모델이 격 교체 현상에 대한 이해 능력이 뛰어나다는 것을 보여준다.

그림 4는 mBERT와 KR-BERT의 격 교체 허용/비허용 구문과 목적격/부사격 간의 상호작용 그래프이다.

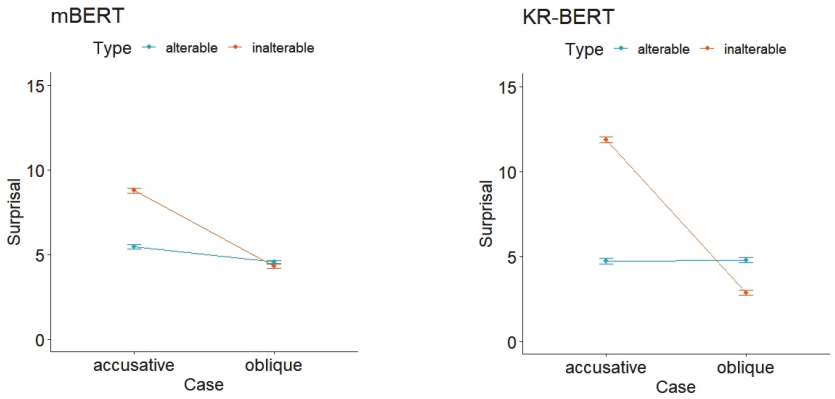

그림 4. 격 교체 허용/비허용 구문과 목적격/부사격 간의 상호작용 그래프

이원분산분석 결과 mBERT와 KR-BERT 모두 격 교체 허용/비허용 구문의 효과와 목적격/부사격의 효과 간에 통계적으로 유의미한 상호작용을 보였다(mBERT: $F(1, 3166)=753.5$, $p<0.001$ '***', KR-BERT: $F(1, 3166)=753.5$, $p<0.001$ '***'). Tukey의 HSD 사후 검정 결과 mBERT는 격 교체 허용 구문의 목적격과 부사격 조사의 surprisal이 유의미한 차이를 보이지만(부사격:목적격=-0.8830140, $p<0.001$ '***'), KR-BERT는 유의미한 차이를 보이지 않았다(부사격:목적격=0.066334, $p<1$ 'ns'). 이는 전술한 격 교체 허용/비허용 구문에 대한 목적격/부사격 조사의 surprisal 분포에 관해 기술한 내용과 일치한다. 즉, KR-BERT와 같이 한국어에 특화된 모델이 용언별 격 교체 허용 여부를 더 정확하게 처리하고 있다. 박권식 외(2021)에 따르면, mBERT가 KR-BERT에 비해 낮은 성능을 보이는 이유는 mBERT는 104개 언어의 데이터를 학습하였기 때문에 전체 학습 데이터에서 한국어 데이터가 차지하는 비중이 작을 것이며, KR-BERT에 비해 매우 불규칙적인 토큰화를 하여 언어학적 관점에 맞지 않는 분절을 하기 때문으로 볼 수 있다.

본 실험에서 사용한 격 교체 허용 구문의 예문은 856문장, 격 교체 비허용 구문의 예문은 729문장으로 충분한 수의 문장을 사용하였다. 즉, 산술적으로 충분한 신뢰성을 갖추었다고 판단한다. 그러므로 본 연구의 결과는 한국어 사전학습모델에 대하여 격 교체와 같이 표층형에서 조사 하나 바꾸는 식의 단순한 방식으로는 적대적 사례를 구축하기 어렵다는 것을 보인 구체적 사례로 볼 수 있다.

5. 논의

형태통사 층위에서 신경망 언어 모형의 자연어 이해 능력에 대한 평가는 복수의 언어

를 대상으로 시도되었다. 영어 화자는 동사의 논항을 인코딩하기 위하여 어순을 참조하기 때문에 어순에 비교적 민감하다고 알려져 있다. Sinha et al. (2021)은 BERT가 논항 구조와 어순의 규칙성을 학습하지 않고도, 어휘의 분포적 정보에 의존하여 매우 견고한 학습이 가능함을 밝혔다. 따라서 단순히 어순을 뒤섞어 인공지능의 언어 지식을 평가하는 것은 인간과 인공지능의 차이를 보여주는 유의미한 평가라고 보기 어렵다. 이와 달리, 일본어의 격 표지 변경은 적대적 사례로써 BERT의 성능을 크게 떨어뜨렸다(Yanaka & Mineshima, 2021). 이는 어순이 상대적으로 자유롭고 격 표지로 동사의 논항을 인코딩하는 일본어의 구조를 BERT가 정확하게 포착하지 못한다는 점을 의미한다.

이와 같은 선행 연구의 흐름에 비추어, 지금까지 한국어 격 표지의 교체 현상이 적대적 사례 평가로 타당한지 평가하였다. 평가 결과, mBERT와 KR-BERT는 격 표지 현상을 포착하고 있으며, 더 많은 한국어 말뭉치를 학습한 KR-BERT가 mBERT에 비하여 더 정확하게 격 표지 현상을 포착하고 있다. 이러한 평가 결과가 함의하는 바를 다음과 같이 논의한다.

5.1. 한국어에 부합하는 적대적 사례

본 실험에 사용된 언어 모형 mBERT와 KR-BERT는 모두 격 표지 교체에 대한 취약성을 보이지 않았다. 이는 격 교체 현상이 BERT가 한국어의 언어적 자질을 심층적으로 이해하고 있는지 평가하는 적대적 사례가 아님을 의미한다. 따라서 Sinha et al. (2021)의 제안과 같은 맥락에서, 한국어 인공지능 모델의 견고성을 평가하기 위하여 더 어렵고 복잡한 적대적 사례 평가가 필요하다고 제안한다.

본고가 제안하는 견고성 평가는 더 심층적인 언어 층위에서 적대적 사례를 생성하여 신경망 언어 모형을 평가하는 것을 말한다. 예를 들어, 의미화용적 층위에서 사건의 사실성을 BERT가 이해하는지 평가할 수 있다(Jeretic et al., 2020; Jiang & de Marneffe, 2021). 사건의 사실성은 술어의 의미에 의하여 결정되지만, 화용적 맥락을 구성하는 요소에 의하여 사건의 사실성이 변동한다. 따라서, 문장이 내포하는 사건의 사실성을 이해하는지 묻는 적대적 사례는 어휘 의미에 대한 추론뿐만 아니라, 맥락적 요소를 병합하여 추론할 수 있는지 묻는 복잡한 평가 과제이다. 영어를 학습한 BERT는 사건의 사실성을 가리키는 술어의 어휘적 패턴을 비교적 정확하게 추론하는 것으로 알려졌으나, 화용적 맥락에 의하여 사건의 사실성이 취소되거나 변동하는 사례는 BERT가 포착하지 못하고 있다(Jiang & de Marneffe, 2021). 따라서 향후 한국어 언어 모형을 평가하기 위한 적대적 사례를 생성하기 위하여 사건의 사실성을 묻는 과제를 구성하여 심층적인 자연어 이해를 확인할 수 있다.

5.2. 한국어 고유의 언어 모형

평가 결과에 따르면 mBERT에 비하여 KR-BERT가 격 교체에 더 견고한 것으로 나타났

다. 다국어 언어 모형인 mBERT는 동일한 신경망 구조와 학습방법으로 104개 언어의 유형론적 특성을 어느 정도 포착한다(Pires et al., 2019). 그러나 통합 모델인 mBERT는 매개 언어인 한국어를 특정하여 학습한 모델인 KR-BERT와 비교하면 격 교체에 취약하다. 이는 언어 독립적인 알고리즘이 격 표지 체계와 같은 언어 의존적인 특성과 구조적 변형 또는 형태통사적 변이형을 학습하는 데 한계가 있음을 의미한다(Bender, 2009). 예를 들어, 문장을 구성하는 단어의 연쇄를 순차적으로 입력받아 학습하는 모형은 어휘의 순서에 매우 민감하며, 입력되는 단어가 동일한 층위의 정보라고 인식한다. 즉, 단어의 형태적 굴절과 어휘소가 언어 구조적으로 다른 층위의 정보라는 점을 모형이 인식하기 어렵다.

이러한 문제는 낮은 형태적 굴절 빈도와 상대적으로 고착된 어순을 가진 영어를 학습할 때는 모형의 성능을 크게 떨어뜨리지 않을 수 있다. 그러나, 한국어와 같이 빈번한 형태적 굴절을 허용하고, 격 표지로 논항을 나타낼 수 있는 언어는 동일한 모형으로 학습하기 어렵다. 따라서, 대량의 원시 텍스트에서 신경망이 스스로 언어 자질을 학습하는 엔드 투 엔드(end-to-end) 방식에서도 기존의 기계학습과 동일한 문제가 발생하게 된다. 결과적으로 특정 매개 언어에 과적합된 모형은 유형론적 특성이 다른 매개 언어에 학습한 규칙을 일반화하기 어렵다는 정리를 다시 확인하게 된다(Bender, 2009). 이는 신경망 언어 모형이 표층적인 어휘적 패턴이 아닌 형태통사적 굴절과 같은 언어 심층적인 이해를 하였는지 평가하는 절차가 필요함을 보여준다. 그러므로 언어 심층적인 구조를 연구하는 언어학자가 신경망 언어 모형의 평가 체계를 연구하여 개발하려는 노력이 요청된다.

5.3. 출현 빈도와 규칙의 학습

마지막으로 어휘의 통계적 빈도와 신경망 언어 모형의 통사 규칙 학습의 상관관계를 언급하고자 한다. 신경망 언어 모형의 학습에 비판적인 논거는 대개 어휘의 통계적 빈도가 모형의 성능에 크게 영향을 미친다는 점에 근거한다. 예를 들어, 흔하게 쓰이는 단수형 동사 ‘walks’의 선행사가 복수형 명사 ‘they’가 아니라는 규칙은 추론할 수 있지만, 말뭉치에서 학습할 수 없는 ‘wuz’ 또는 ‘wug’가 단수형인지 복수형인지 추론할 수 없다는 것이다. 그러나 Yu et al. (2020)에 따르면, 신경망 언어 모형은 새로운 어휘 ‘wug’와 ‘wuz’의 용례 4~5건을 추가로 학습하는 것만으로도 ‘wug’와 ‘wuz’의 형태통사적 규칙을 정확하게 추론하였다. 이를 소량 학습(few-shot learning)이라고 한다. 특히, 선행 명사가 1만 분의 1 내외로 희소하게 출현하더라도 모형의 성능과는 아무런 상관관계가 없었다. 이는 신경망 언어 모형이 낮은 수준의 통계적 학습자가 아닌, 고의적 잡음에 취약하지만 형태통사적 규칙을 스스로 추론할 수 있는 학습자임을 뒷받침하는 실험이다(Wei et al., 2021).

이상과 같은 정리는 본 실험의 결과와 관련하여 자못 시사하는 바가 크다. 신경망 언어 모형을 구성할 때, 학습 데이터가 되는 한국어 코퍼스 안에서는 목적격으로 격 교체가 이루어진 변이형보다 부사격으로 논항이 나타난 정규형의 출현 빈도가 더 높을 것이다. 그러하

다면 surprisal 등의 결과치에 그 빈도의 차이가 드러나야 맞지만, 그 차이는 통계적으로 유의미하지 않은 수준이었다. 이는 딥러닝 학습이 단순히 빈도 계수에만 의존하는 것이 아니라 더 체계적인 규칙을 획득하는 장치라는 점을 의미한다. 신경망 언어 모형에서 빈도와 규칙의 관계에 관해서는 추가적인 연구가 필요하다.

6. 결론

본고는 인공지능 언어 모형의 평가를 위해 언어학적 관점에서 적대적 사례를 생성하여 검증을 시도하였다. 일반적으로 적대적 사례를 비교적 쉬운 수준에서 생성할 때, 특정 형태를 잡음에 가까운 변이형으로 '교체'하는 방식을 사용한다. 이를 적용하여, 언어의 여러 층위와 현상 가운데 형태통사론 단계에서 한국어의 격 교체 현상을 다루었다.

실험의 결과는 두 가지 측면으로 분석되었다. 첫째, 다국어 언어 모형인 mBERT와 한국어 모형인 KR-BERT 모두 한국어 격 교체를 잘 파악하고 있었다. 이는 언어 모형이 형태통사적 변이형을 정확히 포착하지 못할 것이라는 당초의 예상을 벗어나는 결과이다. 한국어에서 적대적 사례를 생성하려면 격조사 등을 교체하는 단순한 방식으로는 소기의 성과를 거두기 어렵다는 것을 의미한다. 이후 한국어에 대한 언어 모형 평가를 위해 적대적 사례를 생성하려면 더 심층적인 수준에서 언어 제약이 정교하게 다루어져야 한다는 결론이다. 둘째, 다국어 모형보다는 한국어 특징적인 모형이 격 교체에 대한 포착을 더 수월하게 하였다. 이는 매개 언어의 형태통사적 특징은 그 언어를 전문으로 하는 모형에서 더 확실하게 인식된다는 일반적 예상과 부합하는 결과이다. 이는 역으로 말하여 언어 모형에 대한 제대로 된 평가를 위해서라도 한국어 현상과 제약에 대한 전문 지식이 중요함을 방증한다. 즉, 한국어로 구성된 적대적 사례를 생성하기 위해서는 한국어 자체에 대한 이해가 밑바탕이 되어야 한다.

다음 단계의 연구로 통사의미 또는 의미화용의 층위에서 적대적 사례를 생성하여 실험을 수행하고자 한다. 즉, 표층에서의 유사점과 차이점을 다루는 것이 아니라 그 이면에서 작동하는 한국어 표현의 특성을 반영하여 적대적 사례군을 생성할 것이다. 이를 통해 자연어 추론 과제의 일환으로 언어 모형이 이른바 '행간'에서 발생하는 미묘한 의미 차이를 적절히 인식할 수 있는가에 주목할 것이다.

딥러닝 기반의 자연어 이해 및 생성 과제 전반에서 언어 모형 평가는 언어학자들이 크게 이바지할 수 있는 대상이다. 위 실험의 결과는, 적대적 사례 생성에서 언어학자들이 중심적 역할을 할 수 있고 또 해야만 한다는 사실을 보여준다. 매개 언어에 대한 폭넓은 이해 없이 딥러닝 신경망을 '속일 수 있는' 공격형 평가 데이터를 구성하는 것은 어렵기 때문이다. 전술한 바와 같이, 한국어는 영어와 같은 타 언어와 비교하면 적대적 사례를 생성하기

가 비교적 어려운 언어에 속한다고 판단한다. 한국어가 가지는 본연의 특성과 표현의 다양성을 충분히 반영한 적대적 사례 생성 연구가 이어지길 기대한다.

참고문헌

김미령. (2004). 격교체 양상에 따른 동사 분류에 대한 연구. *한국어학*, *25*, 161-190.

박권식, 김성태, 송상헌. (2021). 최소대립 문장쌍을 활용한 한국어 사전학습모델의 통사 연구 활용 가능성 검증. *언어와 정보*, *25*(3), 1-21.

송창선. (2019). 격조사 교체 현상을 통해 본 국어의 격 기능. *국어교육연구*, *71*, 21-38.

우형식. (1996). *국어의 타동사 구문 연구*. 서울: 도서출판 박이정.

이규민, 김성태, 김현수, 박권식, 신운섭, 왕규현, 박명관, 송상헌. (2021). DeepKLM - 통사 실험을 위한 전산 언어모델 라이브러리. *언어사실과 관점*, *52*, 265-306.

이종근. (2006). 한국어 동사와 대격에 관한 연구. *언어학*, *14*(1), 223-242.

이홍식. (2004). 조사 '을'의 의미에 대하여. *한국어 의미학*, *15*, 303-327.

한국전자통신연구원. (2019). KorBERT(Korean Bidirectional Encoder Representations from Transformers). https://aiopen.etri.re.kr/service_dataset.php.

홍재성, 이성헌 (2007). 세종 전자사전 : 전산어휘부로서의 특성과 의의. *한국정보과학회 언어공학 연구회 학술발표 논문집*, 323-331.

Bender, E. M. (2009). *Linguistically naïve != language independent: Why NLP needs linguistic typology*. Paper presented at the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?, 26-32.

Da Costa, J. K., & Chaves, R. P. (2020). *Assessing the ability of Transformer- based Neural Models to represent structurally unbounded dependencies*. Paper presented at the Society for Computation in Linguistics, *3*(1), 189-198.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Paper presented at the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186.

Ebrahimi, J., Lowd, D., & Dou, D. (2018). *On adversarial examples for character- level neural machine translation*. Paper presented at the 27th International Conference on Computational Linguistics, 653-663.

Fukuda, S. (2020). The syntax of variable behavior verbs: Experimental evidence from

the accusative-oblique alternations in Japanese. *Journal of Linguistics*, *56*(2), 269-314.

Garg, S., & Ramakrishnan, G. (2020). *BAE: BERT-based adversarial examples for text classification*. Paper presented at the 2020 Conference on Empirical Methods in Natural Language Processing, 6174-6181.

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

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. Paper presented at the ICLR 2015.

Hale, J. (2001). *A probabilistic Earley parser as a psycholinguistic model*. Paper presented at the Second meeting of the north American chapter of the association for computational linguistics.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). *A systematic assessment of syntactic generalization in neural language models*. Paper presented at the Association for Computational Linguistics, 1725-1744.

Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020). *Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition.* Paper presented at the 58th Annual Meeting of the Association for Computational Linguistics, 8690-8705.

Jiang, N., & de Marneffe, M. C. (2021). He Thinks He Knows Better than the Doctors: BERT for Event Factuality Fails on Pragmatics. *Transactions of the Association for Computational Linguistics*, *9*, 1081-1097.

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). *Is BERT really robust? a strong baseline for natural language attack on text classification and entailment.* Paper presented at the AAAI conference on artificial intelligence, *34*(5), 8018-8025.

Lee, S., Jang, H., Baik, Y., Park, S., & Shin, H. (2020). KR-BERT: A small-scale korean-specific language model. *arXiv preprint* arXiv:2008.03979.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.

Marvin, R., & Linzen T. (2018). *Targeted syntactic evaluation of language models*. Paper presented at the 2018 Conference on Empirical Methods in Natural Language Processing, 1192-1202.

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the uniform information density hypothesis. *arXiv preprint* arXiv:2109.11635.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). *Adversarial NLI: A new benchmark for natural language understanding*. Paper presented at the 58th Annual Meeting of the Association for Computational Linguistics, 4885-4901.

Park, K., Park, M.-K., & Song, S. (2021). Deep learning can contrast the minimal pairs of syntactic data. *Linguistic Research*, *38*(2), 395-424.

Park, S.-H., & Yi, E. (2021). Perception-production asymmetry for Korean double accusative ditransitives. *Linguistic Research, 38*(1), 27-52.

Pires, T., Schlinger, E., & Garrette, D. (2019). *How multilingual is multilingual BERT?*. Paper presented at the 57th Annual Meeting of the Association for Computational Linguistics, 4996-5001.

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., & Kiela, D. (2021). *Masked language modeling and the distributional hypothesis: Order word matters pre-training for little*. Paper presented at the 2021 Conference on Empirical Methods in Natural Language Processing, 2888-2913.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks*. Paper presented at International Conference on Learning Representations (ICLR).

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415-433.

Wei, J., Garrette, D., Linzen, T., & Pavlick, E. (2021). F*requency Effects on Syntactic Rule Learning in Transformers.* Paper presented at the 2021 Conference on Empirical Methods in Natural Language Processing, 932-948.

Wilcox, E., Levy, R., & Futrell, R. (2019). Hierarchical representation in neural language models: Suppression and recovery of expectations. *arXiv preprint* arXiv:1906.04068.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint* arXiv:1609.08144.

Yanaka, H., & Mineshima, K. (2021). *Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference*. Paper presented at the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 337-349.

Yu, C., Sie, R., Tedeschi, N., & Bergen, L. (2020). *Word frequency does not predict*

grammatical knowledge in language models. Paper presented at the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4040-4054.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). *SWAG: A large-scale adversarial dataset for grounded commonsense inference.* Paper presented at the 2018 Conference on Empirical Methods in Natural Language Processing, 93-104.

부록

https://bit.ly/3D0bf3P

송상헌
02841 서울시 성북구 안암로 145
고려대학교 문과대학 언어학과 부교수
전화: (02)3290-2177
이메일: sanghoun@korea.ac.kr

노강산
02841 서울시 성북구 안암로 145
고려대학교 문과대학 언어학과 석사과정
전화: (02)3290-2170
이메일: kasan1998@korea.ac.kr

박권식
02841 서울시 성북구 안암로 145
고려대학교 문과대학 언어학과 박사과정
전화: (02)3290-1648
이메일: oneiric66@korea.ac.kr

신운섭
02841 서울시 성북구 안암로 145
고려대학교 문과대학 언어학과 박사과정
전화: (02)3290-1648
이메일: prab35@korea.ac.kr

황동진
02841 서울시 성북구 안암로 145
고려대학교 문과대학 언어학과 석사과정
전화: (02)3290-2170
이메일: no02041@gmail.com

Received on February 10, 2022
Revised version received on March 16, 2022
Accepted on March 31, 2022