Surprisal:

언어 이해 예측을 위한 도구

황동진, 2023.02.09, 2023 전산언어학 겨울학교

목차

1. Information-theoretical Complexity Metrics

Complexity Metric이란?

- “Generally speaking, a complexity metric is something that **quantifies how difficult it is to perceive** a linguistic expression.” (Hale, 2016)
  - Surprisal
  - Entropy Reduction

1. Information-theoretical Complexity Metrics

점증적(Incremental)

- 점증적(Incremental)
  - 각 단어의 인식이 얼마나 어려운지를 **실시간으로** 예측

1. Information-theoretical Complexity Metrics

점증적(Incremental)

선생님이 ➡️ 학교에서

선생님이 학교에서 ➡️ 학생들을

선생님이 학교에서 학생들을 ➡️ 가르친다

선생님이 학교에서 학생들을 가르친다

2. Surprisal이란?

Surprisal의 일반적인 정의

For a generic random variable Y, the surprisal of an outcome Y=y

$$\log_2\left(\frac{1}{P(y)}\right)$$

let y be the ratio of prefix probabilities

$$= \log_2\left(\frac{1}{\frac{\sum_{\text{after}}}{\sum_{\text{before}}}}\right)$$

$$= \log_2\left(\frac{\sum\text{before}}{\sum\text{after}}\right)$$

$$= -\log_2\left(\frac{\sum\text{after}}{\sum\text{before}}\right)$$

2. Surprisal이란?

확률 문법(Probabilistic Grammar)

- 언어 사용자들이 그들이 듣는 단어 연쇄의 구조를 설명하는 문법을 알고 있다고 가정
  - “지금까지 들은(말한) 단어들에 비추어 보아, 앞으로 어떤 구조의 단어 연쇄가 가능할까?”

    ➡️ 어떠한 단어의 연쇄에 이어, 어떤 단어를 만났을 때 ‘얼마나 놀라는지’를 수치화

2. Surprisal이란?

확률 문법(Probabilistic Grammar)

선생님이 학교에서 학생들을 ➡️ 가르친다

VS.

선생님이 학교에서 학생들을 ➡️ 준다

2. Surprisal이란?

언어 실험에서의 활용

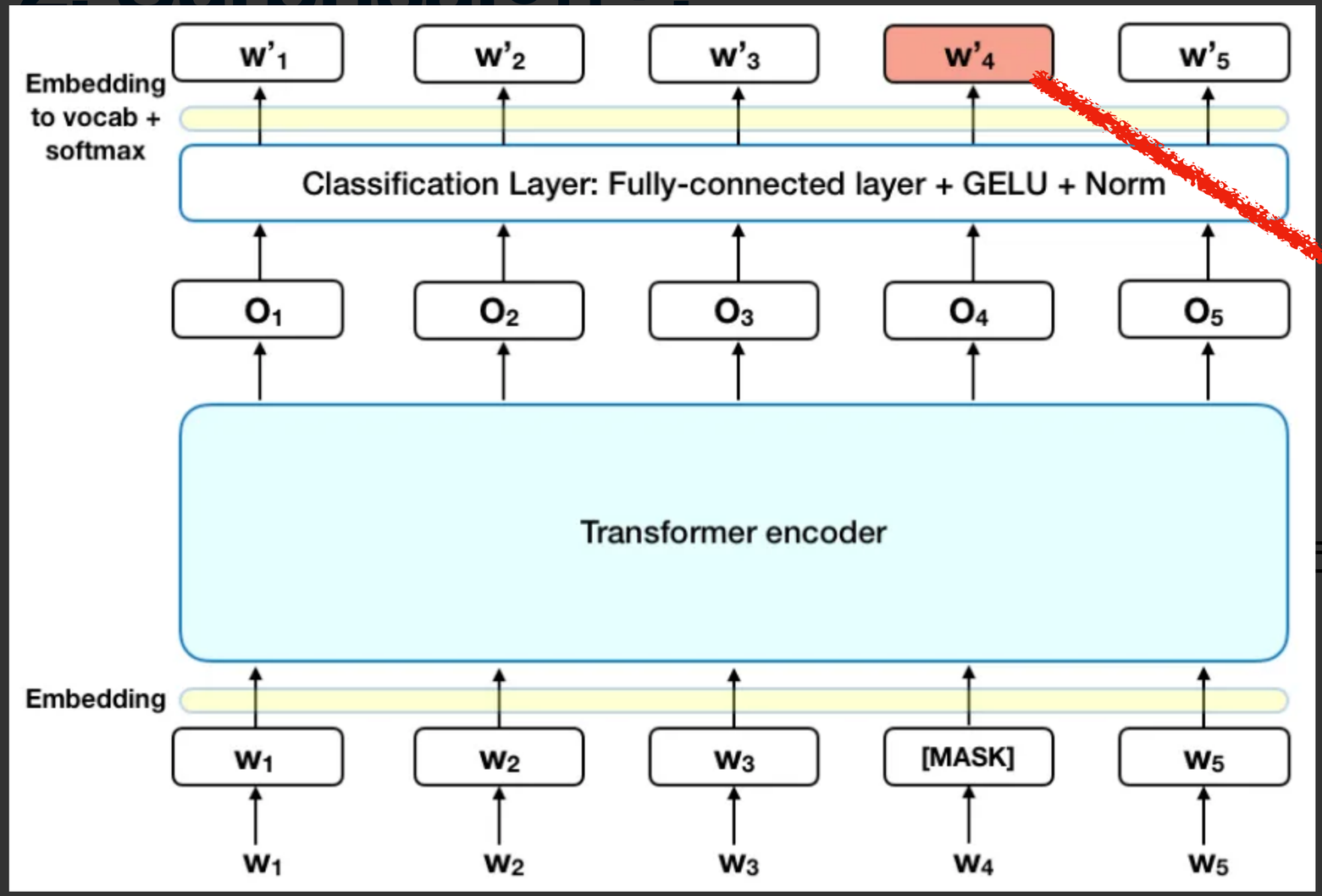
- Eye-tracking이나 수용성 판단 실험의 결과와 유의성을 가짐 (Boston et al. 2008; Demberg and Keller 2008)
- Reading time의 결과와 높은 유사성을 보임(Levy et al. 2012)
- 읽기에 있어 N400의 경향 또한 예측이 가능함을 보임(Frank et al. 2013)
- 수용성 판단 자체와도 선형적인 관계를 보임(Meister et al. 2021)

2. Surprisal이란?

딥러닝 언어 모델에서의 Surprisal

- ‘BERT for Masked LM’에서 surprisal을 확인하고자 하는 항목의 softmax 값을 추출한 후, 밑이 2인 음의 로그를 취함

2. Surprisal이란?



(Horev 2018)

$$Surprisal = log_2(\frac{1}{P(w_4')})$$

는 항목의 softmax 값을 추출

2. Surprisal이란?

딥러닝 언어 모델에서의 Surprisal

- 특정 단어가 출현할 확률이 높을수록 surprisal 값은 낮아지며, surprisal 값이 낮아질수록 모형 입장에서 해당 단어에 대한 수용성이 높아진다고 해석할 수 있음
- 인간의 입장에서 수용성이 높은 표현에 대해 모델이 출력한 surprisal 값이 낮게 나타나면, 이는 모델이 해당 구문을 적절히 학습했음을 암시
- 최근에는 surprisal에 근거한 BERT의 확률적 추론이 인간의 수용성 판단과 높은 상관관계를 가진다는 보고가 이루어짐

3. 언어 이해와 언어 모델

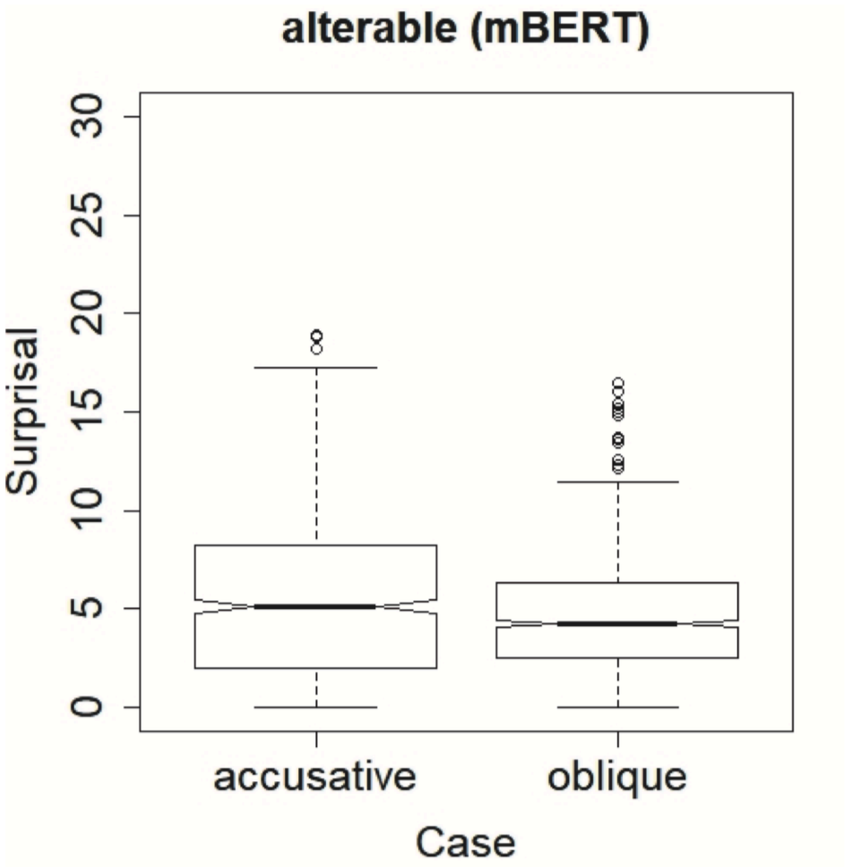
형태: 격표지 교체

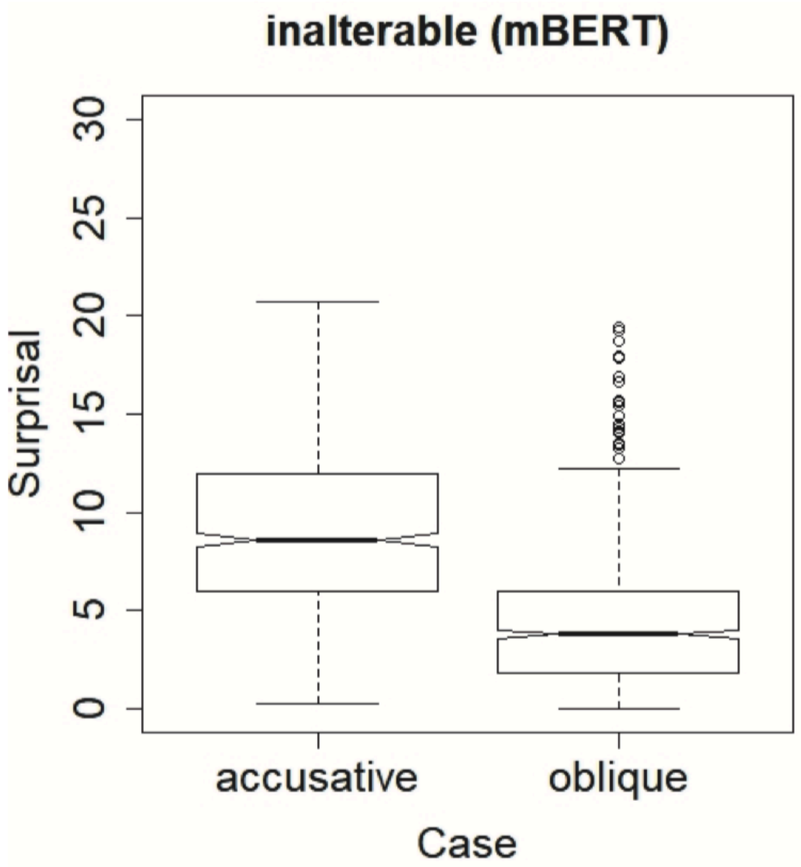
송상헌, 노강산, 박권식, 신운섭 and 황동식. (2022). 적대적 사례에 기반한 언어 모형의 한국어 격 교체 이해 능력 평가. 언어학, 30(1), 45-72.

- 교체 가능(alterable)
  - 철수가 학교{에/를} 갔다.
- 교체 불가능(inalterable)
  - 액자가 왼쪽{으로/*을} 기울었다.

3. 언어 이해와 언어 모델

형태: 격표지 교체





3. 언어 이해와 언어 모델

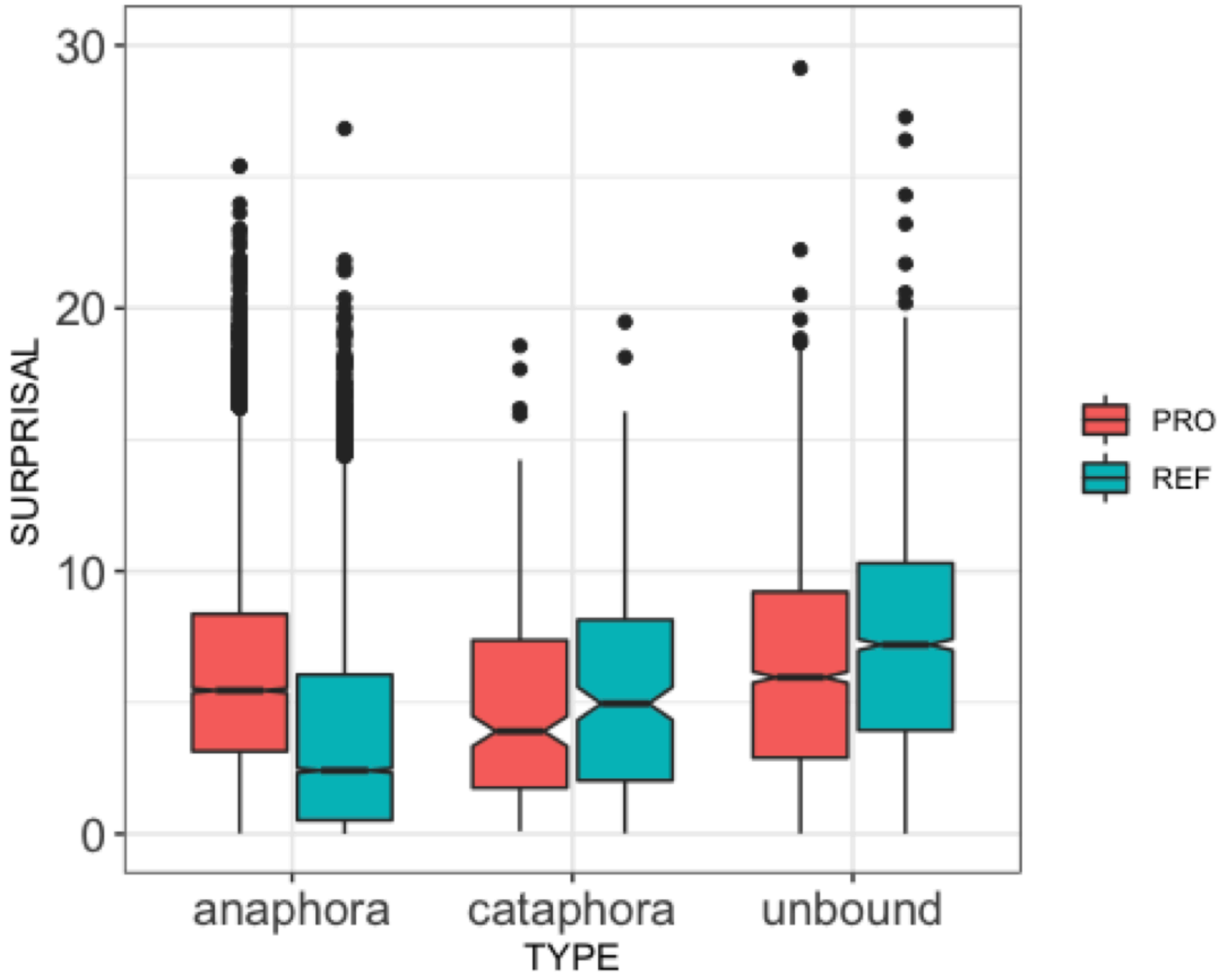
통사: 영어 재귀사

송상헌, 이규민, & 김경민. (2021). 딥러닝 언어모형을 활용한 영어 비결속 재귀사 검증. 언어과학, *28*(3), 51-78.

- 전방 조응사(anaphora): 선행하는 요소를 참조하는 표현.
- 후방 조응사(cataphora): 후행하는 요소를 참조하는 표현.
- 외적 조응사(exophora): 직접적으로 나타낸 대상이 아닌 상황적 문맥에서 추론 가능한 요소를 참조. 화자 혹은 청자 등을 포함한 인칭 화시(person deixis)와 관련을 지님.
- 인식 조응사(logophora): 특정 대명사 집합과 명시적으로 표현되지 않은 참조. 통상 관점의 변화로 서술상의 주체가 달라질 때 출현함.

3. 언어 이해와 언어 모델

형태: 격표지 교체



3. 언어 이해와 언어 모델

의미: 부정극어

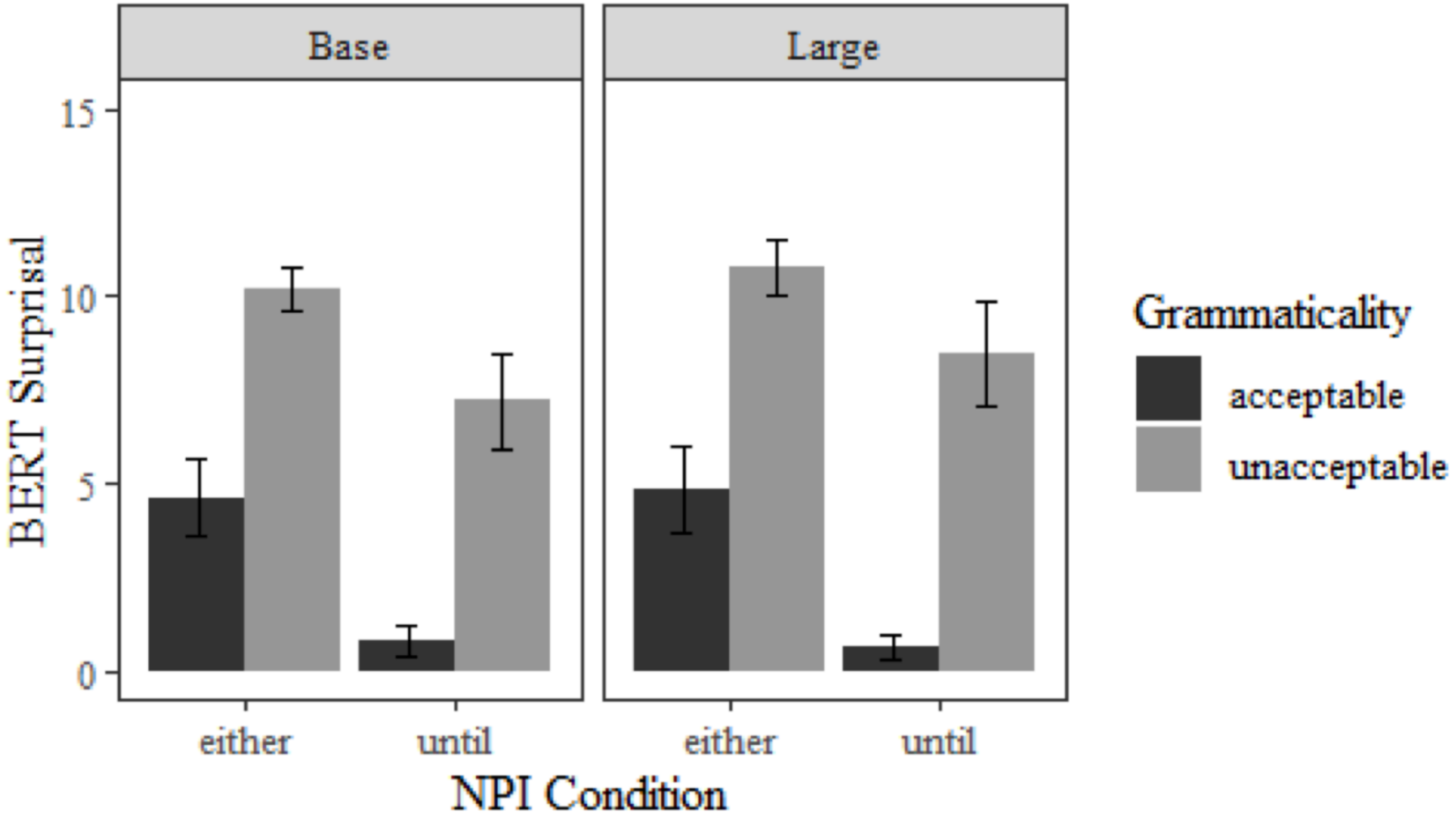
Shin, Unsub. (2022). Investigating neural network's sensitivity to negative polarity items, Master Thesis, Korea University.

Table 3.1. Examples of test sentences with the additive *either* and punctual *until*.

Condition	Sentence examples
Negative Polarity Acceptable	And the salesperson will **not** get the money *either*
Negative Polarity Unacceptable	*And the salesperson will **always** get the money *either*
Negative Polarity Acceptable	The forensic anthropologist **won't** finish the job *until* midday tomorrow
Negative Polarity Unacceptable	??The forensic anthropologist **will certainly** finish the job *until* midday tomorrow

3. 언어 이해와 언어 모델

형태: 격표지 교체



4. 실습