

Surprisal:

언어 이해 예측을 위한 도구

강의 목표

- Surprisal의 특징 이해
- Surprisal을 이용한 언어 실험 진행 방법 이해

목차

1. Information-theoretical Complexity Metrics
2. Surprisal이란?
3. 언어 연구와 언어 모델
4. 실습: 실험 설계 및 진행

1. Information-theoretical Complexity Metrics

1. Information-theoretical Complexity Metrics

Complexity Metric이란?

- **“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(in time) 예측

1. Information-theoretical Complexity Metrics

점증적(Incremental)

선생님이 ➡ 학교에서

선생님이 학교에서 ➡ 학생들을

선생님이 학교에서 학생들을 ➡ 가르친다

선생님이 학교에서 학생들을 가르친다

2. Surprisal이란?

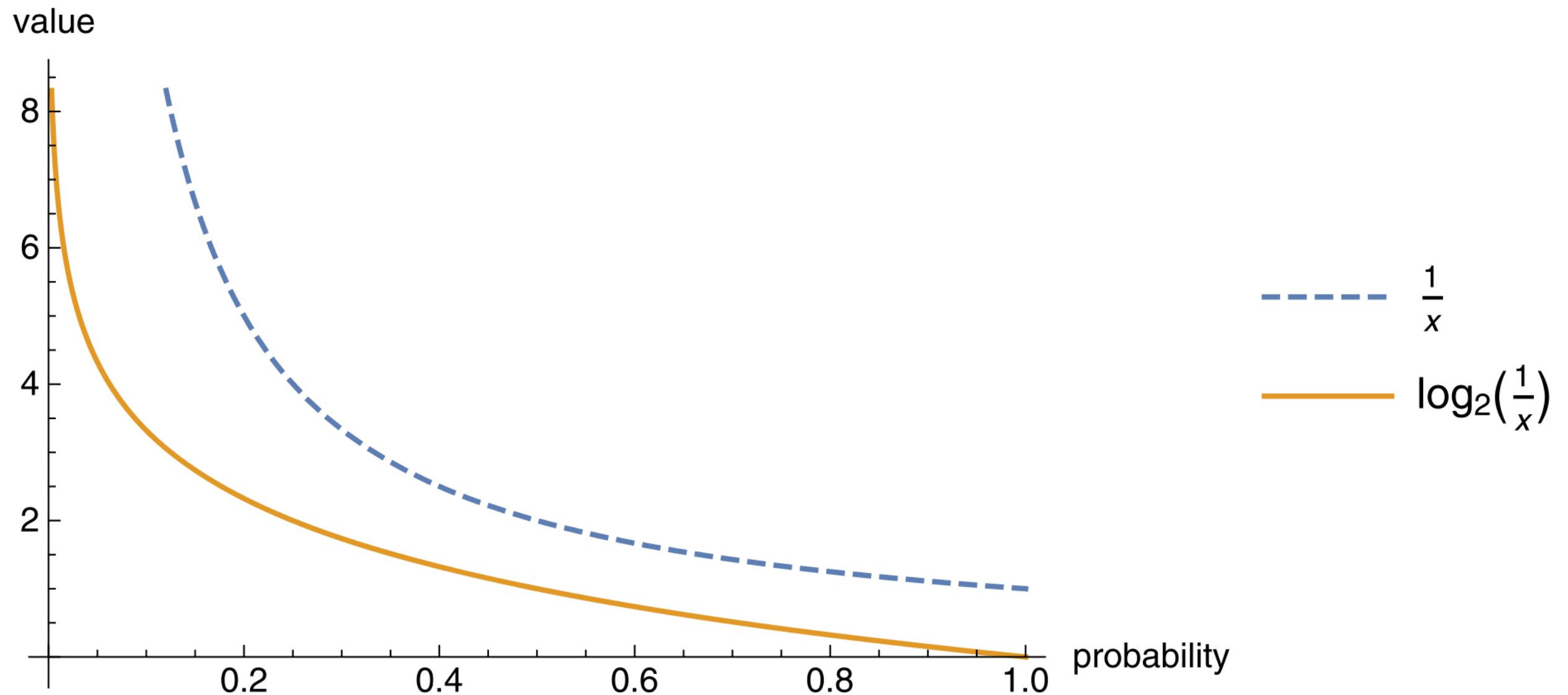
2. Surprisal이란?

Surprisal의 일반적인 정의

$$\begin{aligned} C_{surprisal}(w_i | w_{1:i-1}) &= \\ -\log P(w_i | w_{1:i-1}) &= \\ \log \frac{1}{P(w_i | w_1, \dots, w_{i-1})} \end{aligned}$$

2. Surprisal이란?

Surprisal의 일반적인 정의



(Hale 2016)

2. Surprisal이란?

확률 문법(Probabilistic Grammar)

“지금까지 들은(말한) 단어들에 비추어 보아,
앞으로 어떤 구조의 단어 연쇄가 가능할까?”

➡ 특정한 단어의 연쇄에 이어, 어떤 단어가 등장
했을 때 ‘얼마나 놀라운지’를 수치화

2. Surprisal이란?

확률 문법(Probabilistic Grammar)

선생님이 학교에서 학생들을 ➡ 가르친다

VS.

선생님이 학교에서 학생들을 ➡ 준다

2. Surprisal이란?

언어 실험에서의 활용

- **Eye-tracking** 실험의 결과와 유의성 (Boston et al. 2008; Demberg and Keller 2008)
- **Reading time**과 높은 유사성 (Levy et al. 2012)
- 읽기에 있어 **N400**의 경향 예측 (Frank et al. 2013)
- 수용성 판단과 선형적인 관계 (Meister et al. 2021)

2. Surprisal이란?

딥러닝 언어 모델에서의 Surprisal

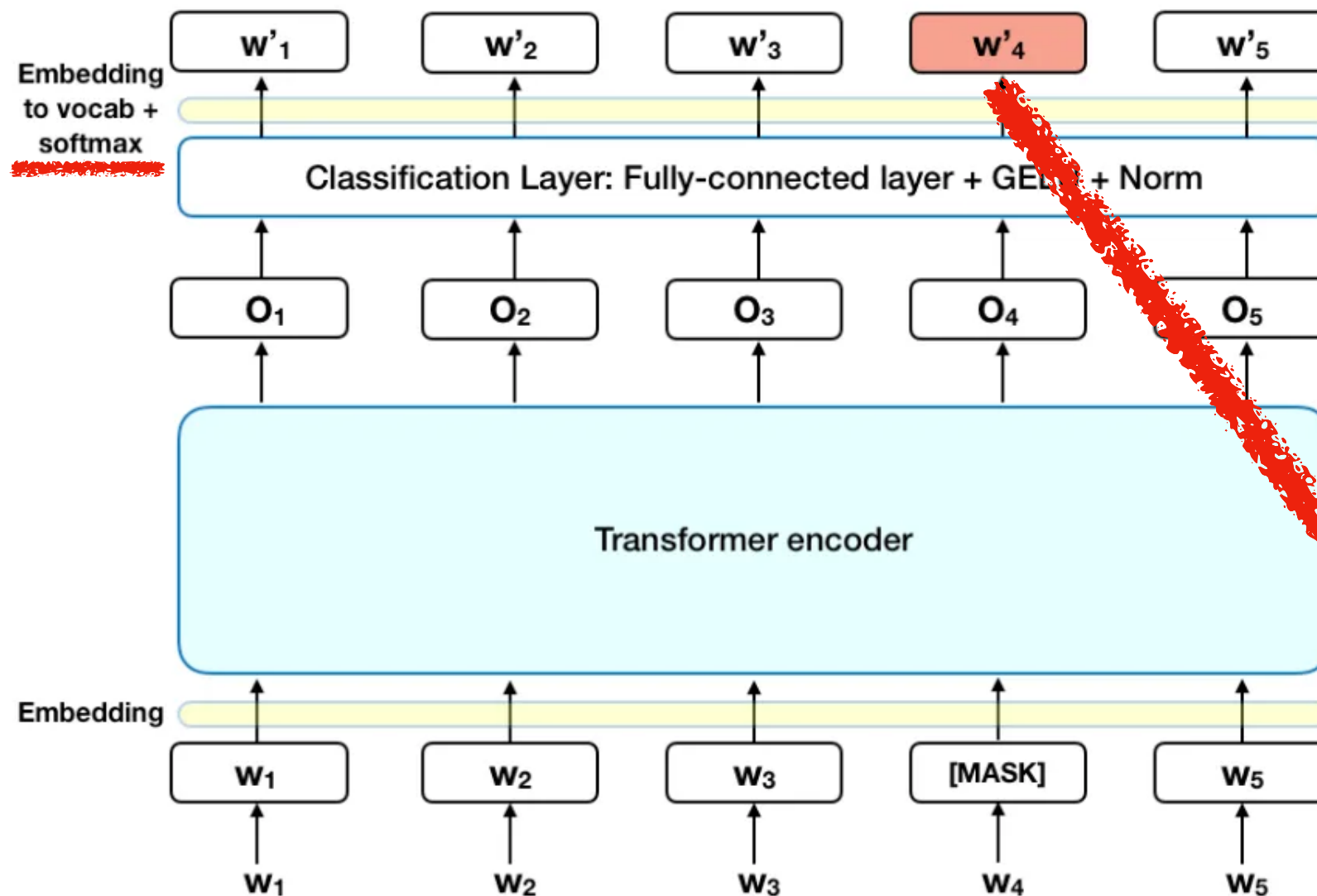
A. 언어 모델에서

B. Surprisal을 확인하고자 하는 항목의 **softmax** 값을 추출

C. 밑이 2인 음의 로그를 취함

2. Surprisal이란?

딥러닝 언어 모델에서의 Surprisal








(이규민 2021)

$$Surprisal = \log_2\left(\frac{1}{w'_4}\right)$$

2. Surprisal이란?

Surprisal 값의 해석

- 단어가 출현 확률 , Surprisal 
- Surprisal , 단어 수용성 
- 수용성이 높은 표현 Surprisal 
- 해당 구문을 적절히 학습

2. Surprisal이란?

Surprisal 값의 해석

- 단, 특정한 구문의 적형성(well-formedness)을 보장하는 기준은 존재하지 않음

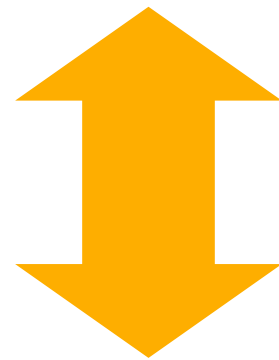
IDX	SEN	ITEM	SURPRISAL
1	철수가 영희[MASK] 좋아한다.	을	9.349677
2	철수가 영희[MASK] 좋아한다.	를	0.01812660

3. 언어 연구와 언어 모델

3. 언어 연구와 언어 모델

호혜성

언어 모델의 언어 능력 → 언어 이론



언어 이론 → 언어 모델의 언어 능력

3. 언어 연구와 언어 모델

언어 모델의 언어 능력 ➡ 언어 이론

Wilcox, E. G., Futrell, R., & Levy, R. (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1-88.

- Surprisal을 활용하여 딥러닝 언어 모델들이 filler-gap dependency, 섬 제약(island constraints) 등을 학습했는지 확인
- 확인 결과, “weakly biased models”(i.e., 딥러닝 언어 모델)들이 이러한 현상들을 잘 학습한 것을 확인

➡ “Our results provide **empirical evidence against the Argument from the Poverty of the Stimulus** for this particular structure.”

3. 언어 연구와 언어 모델

언어 이론 ➡ 언어 모델의 언어 능력(형태: 격표지 교체)

송상헌, 노강산, 박권식, 신운섭, 황동진. (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.

➡ 딥러닝 언어 모델이 이러한 한국어의 통사적 특성을 이미 잘 이해하고 있기 때문에, 인공지능 언어능력 평가를 위한 적대적 접근이 필요

3. 언어 연구와 언어 모델

언어 이론 → 언어 모델의 언어 능력(형태: 격표지 교체)

송상헌. (2022). 딥러닝 언어모델과 Surprisal을 활용한 언어분석. 인공지능인문학연구, 12(0), 9-39.

4. 실습: 실험 설계 및 진행

4. 실습: 실험 설계 및 진행

가설

H_0 = 언어 모델(KR-BERT)은 한국어의 격교체 현상을 제대로 표상하지 못할 것이다.

H_1 = 언어 모델(KR-BERT)은 한국어의 격교체 현상을 제대로 표상할 것이다.

4. 실습: 실험 설계 및 진행

요인설계

2 × 2

요인1: 동사 논항의 격표지 교체 가능 여부

수준1: 교체 가능(alterable)

수준2: 교체 불가능(inalterable)

요인2: 격표지의 종류

수준1: 목적격(e.g., -을/-를)

수준2: 사격(e.g., -에게, -으로, ...)

4. 실습: 실험 설계 및 진행

예문 구성 시 주의점

- 최소대립쌍
- 토큰나이저(Tokenizer)
- 인간 대상 실험과의 차이점

4. 실습: 실험 설계 및 진행

예문 구성 시 주의점: 최소대립쌍

- 컴퓨터의 입장에서 최소대립쌍이 맞는지 확인
 - 철수는 영화를 {좋아한다/좋아하지 않는다}.
 - 철수는 {영화를/수진을} 좋아한다.
 - 철수는 영화{를/을} 좋아한다.

4. 실습: 실험 설계 및 진행

예문 구성 시 주의점: 토큰나이저

- **Tokenization**이 제대로 이루어졌는지 확인

- 철수가 영희[MASK] 좋아한다.

➡ ['[CLS]', '철수', '##가', '영', '##희', '[MASK]', '좋아', '##한다', '.', '[SEP]']

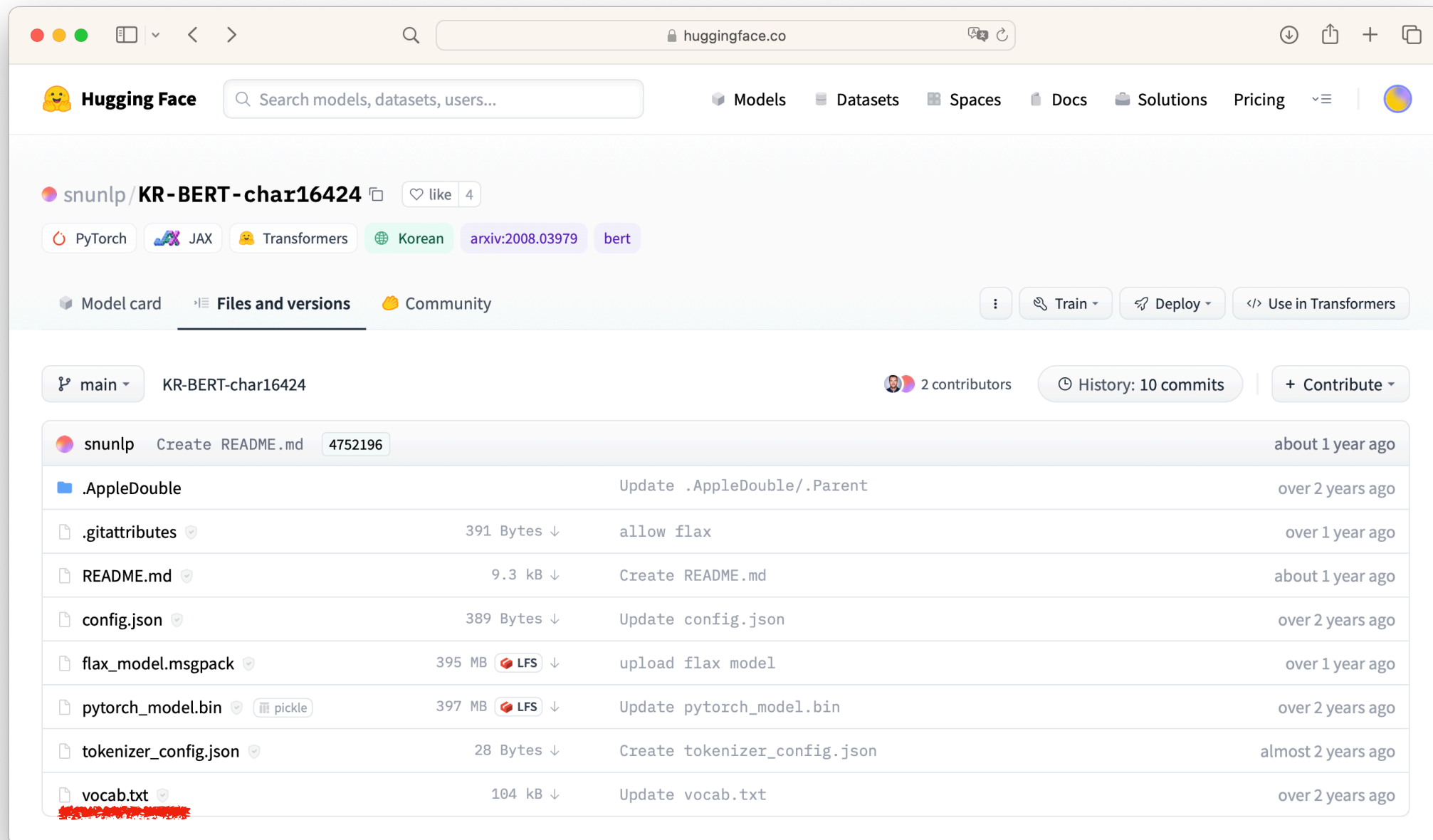
- 철수가 영희[MASK] 혐모한다.

➡ ['[CLS]', '철수', '##가', '영', '##희', '##를', '[MASK]', '혐', '##모', '##한다', '.', '[SEP]']

4. 실습: 실험 설계 및 진행

예문 구성 시 주의점: 토큰나이저

- 사용하는 모델의 어휘 목록 확인
 - <https://huggingface.co/snunlp/KR-BERT-char16424/tree/main>



4. 실습: 실험 설계 및 진행

예문 구성 시 주의점: 토큰나이저

- 많은 수의 예문 사용

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

4. 실습: 실험 설계 및 진행

예문 구성 시 주의점: 인간 대상 실험과의 차이점

- 필러를 사용하지 않음
- 문장의 제시 순서가 상관 없음

4. 실습: 실험 설계 및 진행

예문 구성

	목적격(accusative)	사격(oblique)
교체 가능	철수가 학교를 갔다.	철수가 학교에 갔다.
교체 불가능	*액자가 왼쪽을 기울었다.	액자가 왼쪽으로 기울었다.

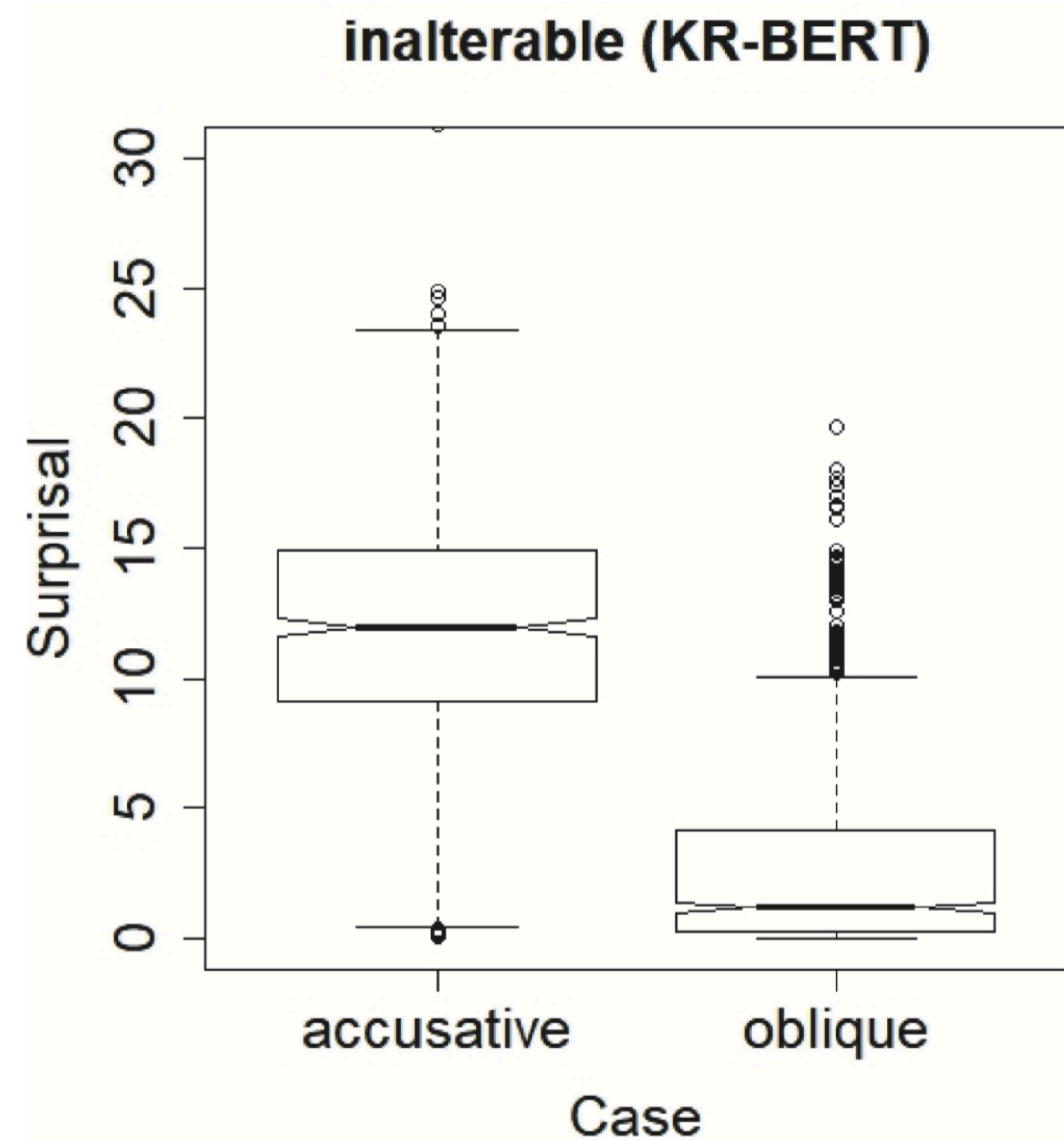
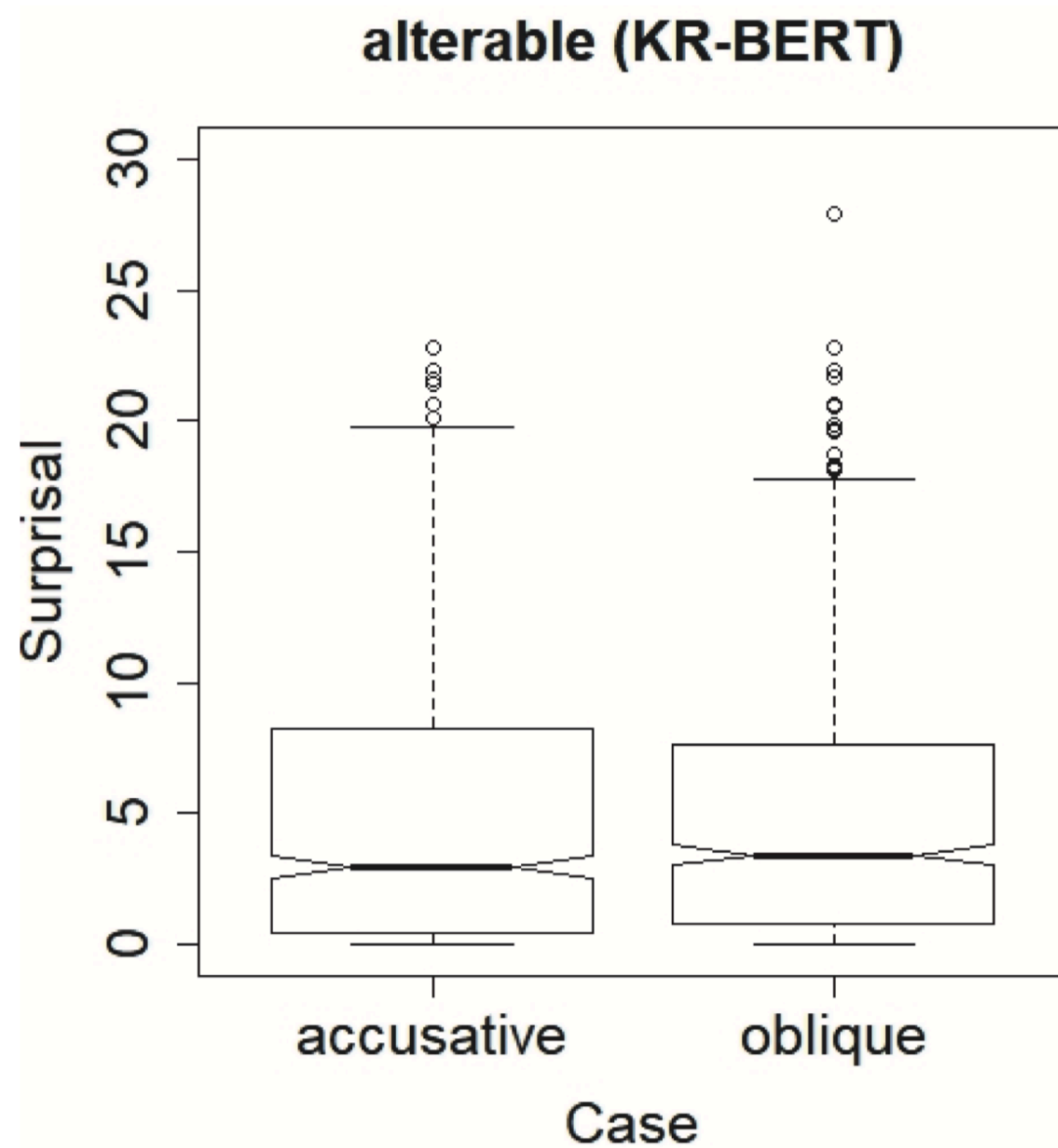
4. 실습: 실험 설계 및 진행

실험 및 분석

Colab + Excel을 활용하여 실험 및 분석 진행

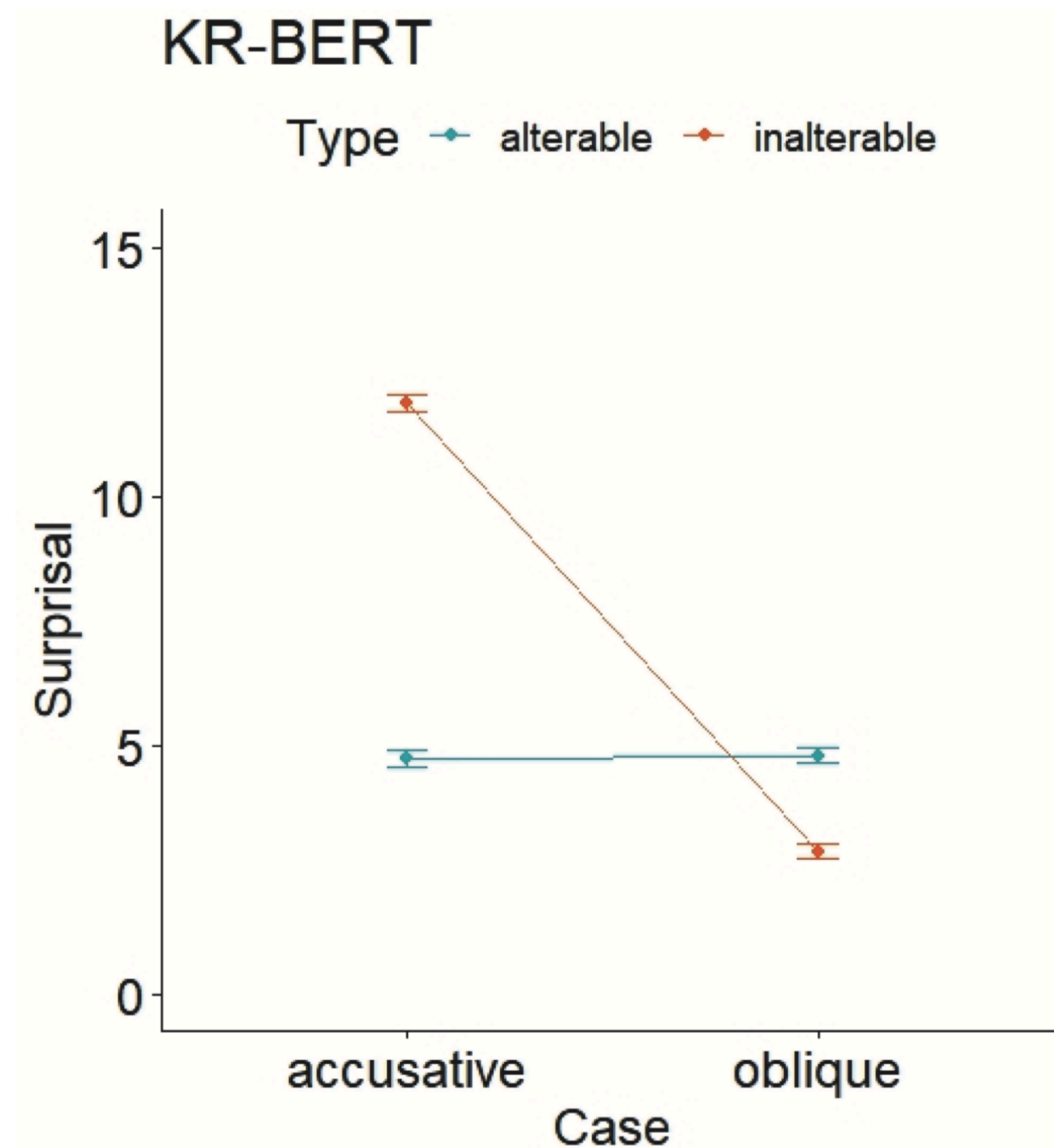
4. 실습: 실험 설계 및 진행

결과



4. 실습: 실험 설계 및 진행

결과



참고문헌 및 참고자료

- 이규민, 김성태, 김현수, 박권식, 신운섭, 왕규현, 박명관 and 송상헌. (2021). DeepKLM - 통사 실험을 위한 전산 언어모델 라이브러리 -. 언어사실과 관점, 52, 265-306.
- 송상헌, 노강산, 박권식, 신운섭, 황동진. (2022). 적대적 사례에 기반한 언어 모형의 한국어 격 교체 이해 능력 평가. 언어학, 30(1), 45-72.
- 송상헌. (2022). 딥러닝 언어모델과 Surprisal을 활용한 언어분석. 인공지능인문학연구, 12(0), 9-39.
- Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading.
- Horev, R. (2018, November 17). Bert explained: State of the art language model for NLP. Medium. Retrieved February 5, 2023, from <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>
- Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9), 397-412.
- Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition*, 122(1), 12-36.
- Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the uniform information density hypothesis. *arXiv preprint arXiv:2109.11635*.
- Wilcox, E. G., Futrell, R., & Levy, R. (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1-88.

코드 및 파이썬 모듈: <https://github.com/gyulukeyi/DeepKLM>