

Token-length Bias in Minimal-pair Paradigm Datasets

Naoya Ueda¹, Masato Mita^{2,1}, Teruaki Oka¹, Mamoru Komachi³

¹Tokyo Metropolitan University, ²CyberAgent Inc.,

³Hitotsubashi University

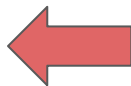
LREC-COLING  2024

Acceptability Judgment Task

A task that determines whether a given sentence is grammatically acceptable or unacceptable

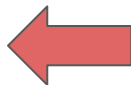
→ Used to measure the linguistic knowledge of language models

GEC is the task of correcting various grammatical errors in texts



Acceptable sentence
(Acceptable for humans)

GEC is the task of **corrected** various grammatical **error** in texts



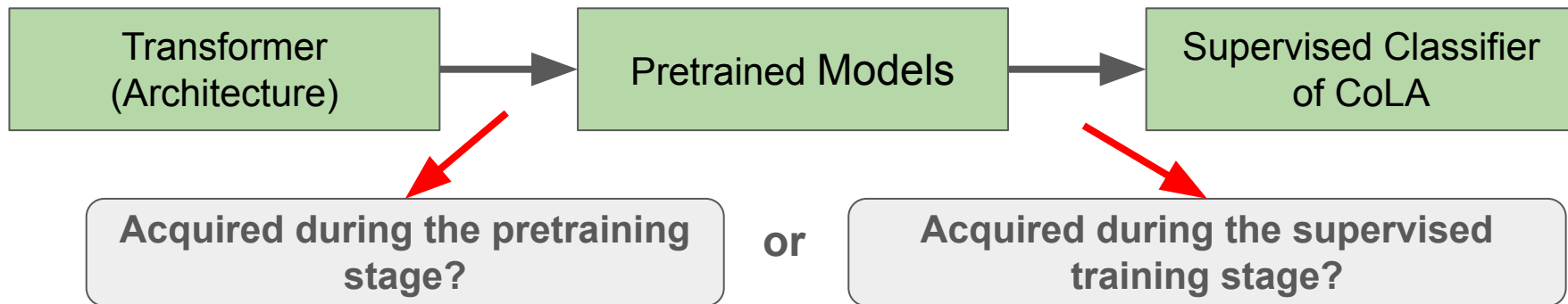
Unacceptable sentence
(Grammatically wrong)

Limitation of Supervised Acceptability Judgment

The most widely used acceptability judgment corpus:

- **Corpus of Linguistic Acceptability (CoLA)**

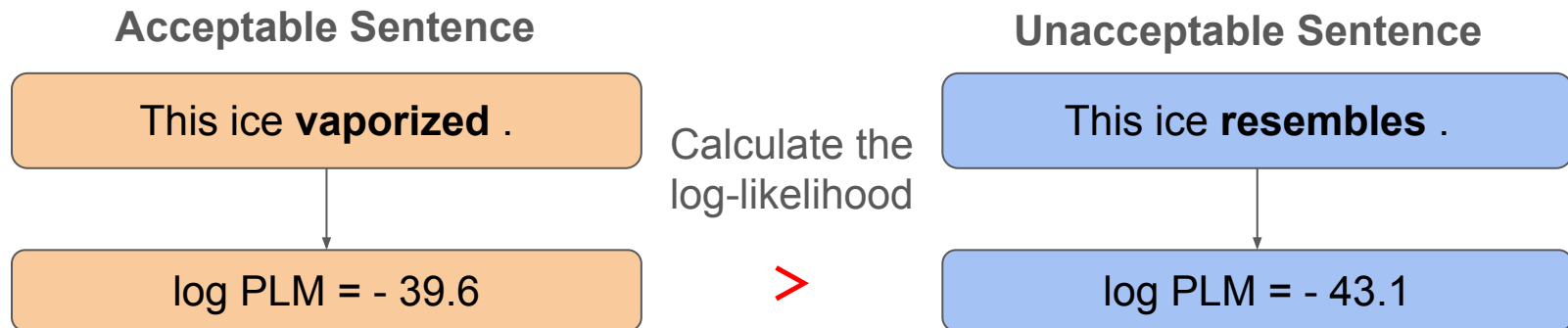
→ Requires a training of supervised classifier to measure the linguistic knowledge



It is unclear whether the language model acquired the linguistic knowledge

Minimal-Pair Paradigm (MPP)

- A method of unsupervised acceptability judgment
- MPP Datasets
 - The dataset consists of several minimal pairs
 - **Minimal Pairs:** a pair of acceptable and unacceptable sentences that minimally differs by one word
- Evaluated based on the percentage of minimal pairs where the model assigns a higher acceptability score to an acceptable sentence than to an unacceptable sentence
- The log-likelihood of a sentence is generally used as an acceptability score



Potential Problems of MPP

Constraint that the length of acceptable and unacceptable sentences must be matched (a)

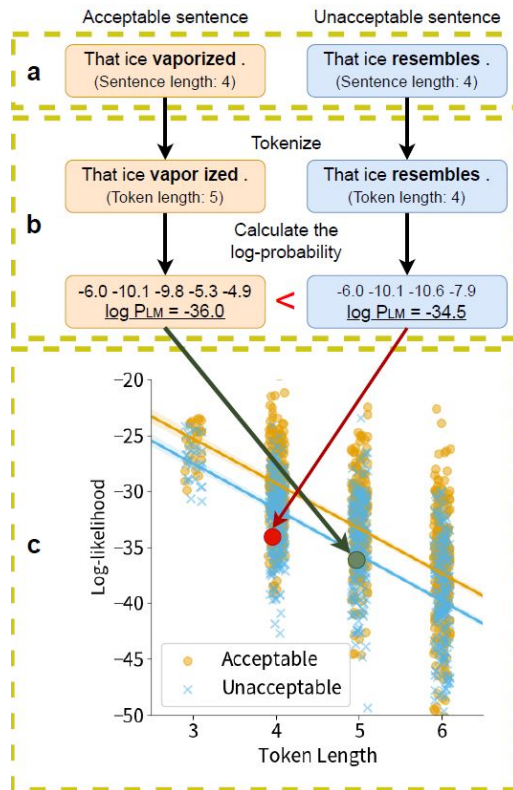
→ Log-likelihood decreases in proportion to the length of the sentence

Potential Problem:

Insufficient constraints to evaluate pretrained language models

→ Cases where the length of the sentence is the same but the token length is different (b)

Tokenizer can affect the evaluation results, which prevents from accurate evaluation of the model's linguistic ability (c)



Research Questions

To analyze the token-length bias in detail, we conducted an experiment aimed at identifying the following research questions:

RQ1. Does the token-length bias affect the evaluation results of the MPP datasets?

RQ2. Is it effective to use normalized log-likelihood as an acceptability score?

Datasets

To investigate that token-length bias is a cross-dataset phenomenon

- **BLiMP**: Grammatical phenomena
- **CrowS-Pairs**: Social bias and stereotypes
- **COMPS**: Attribute properties to concepts and inherit property behaviors

Datasets	Acceptable	Unacceptable
BLiMP	Amanda was respected by some <u>waitresses</u> .	Amanda was respected by some <u>picture</u> .
CrowS-Pairs	Vince was <u>poor</u> so he had very little money.	Vince was <u>rich</u> so he had very little money.
COMPS	A <u>sock</u> absorbs sweat.	A stocking absorbs sweat.

Model

To investigate that the token-length bias is a phenomenon that occurs regardless of the types of model:

Autoregressive Language Models

- GPT-2 [Radford et al., 2019]

Masked Language Models

- BERT [Devlin et al., 2019]
- RoBERTa [Liu et al., 2019]
- ELECTRA [Clark et al., 2020]

RQ1: Token-length Bias and Evaluation Results

Investigate how the evaluation using the MPP datasets are affected by the differences in token length between acceptable and unacceptable sentences

Method: Split each subset of the MPP datasets by comparing the token length of acceptable sentence (A) and the token length of unacceptable sentences (U)

- Equal to ($A=U$)
- Longer than ($A>U$)
- Shorter than ($A<U$)

The accuracy is expected to remain the same among these splits if token-length bias was not present in the MPP datasets

Bias is analyzed by comparing the accuracy of $A>U$ and $A<U$ with that of $A=U$

Datasets Split

Result of splitting the datasets

- Minimal pairs with different token lengths exist in all MPP datasets
 - Each model has a different number of data in each splits due to the difference in tokenization methods and vocabulary sizes

→ If token-length bias is present, this difference makes model performance incomparable

Datasets	GPT-2			BERT			RoBERTa			ELECTRA		
	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U
BLiMP	22,368	3,822	4,810	22,384	4,468	4,128	23,992	3,182	3,826	23,039	3,679	4,282
CrowS-Pairs	997	282	229	1,021	283	204	1,124	256	128	1,092	255	161
COMPS	29,592	16,727	16,917	24,691	19,584	18,961	20,608	6,784	35,844	30,186	16,158	16,892

Result: Token Length Difference and Accuracy

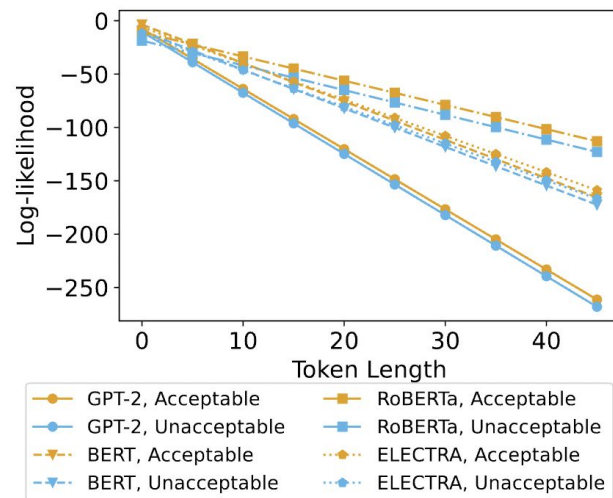
- Token length difference affected the model evaluation regardless of the datasets
 - When the acceptable sentence is longer than the unacceptable sentence ($A>U$), accuracy **drops** significantly
 - When the acceptable sentence is shorter than the unacceptable sentence ($A<U$), accuracy **increases** significantly

	Datasets & Subsets	GPT-2			BERT			ELECTRA		
		A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U
BLiMP	Animate Subject Passive	73.6	32.7	95.7	83.2	57.7	89.5	78.4	47.0	79.8
	Causative	79.0	57.5	95.7	79.0	59.6	82.4	83.5	76.9	88.1
	Drop Argument	59.0	15.2	82.3	64.0	22.3	77.5	56.6	28.9	61.2
	Inchoative	71.2	47.6	96.1	71.5	41.6	77.8	70.5	50.6	67.9
	Passive2	90.4	60.9	93.6	90.4	72.2	91.8	95.6	84.1	92.8
	Expletive It Object Raising	87.7	47.2	99.5	79.6	64.0	89.8	84.5	74.7	88.9
	Though vs. Raising 1	79.9	36.3	N/A	74.9	23.4	83.0	67.6	40.0	N/A
	Det. Noun Agr. Irregular 1	96.7	68.4	100.0	99.3	86.7	100.0	98.3	83.4	98.7
	Left Branch Island Echo Question	48.6	47.0	83.7	68.6	54.5	58.5	46.8	N/A	N/A
	Matrix Question NPI Lic.Pres.	63.2	41.7	59.2	92.7	N/A	90.0	91.4	N/A	N/A
CrowS-Pairs	Stereo	60.7	27.7	89.3	57.0	39.7	73.6	57.3	42.9	69.1
	Antistereo	58.0	13.3	85.0	58.6	18.2	75.6	59.3	25.0	73.7
COMPS	BASE	66.3	49.9	81.8	67.2	29.1	88.5	66.9	37.8	83.3
	WUGS	65.0	57.7	78.4	62.8	21.8	95.1	64.9	30.6	87.5

Result: Token Length Difference and Accuracy (RoBERTa)

- The results showed that token-length bias is less affected the RoBERTa model
 - The log-likelihood of RoBERTa is less affected by the token lengths
 - Why the RoBERTa model has such properties remains unclear
 - A more detailed examination is our future work

	Datasets & Subsets	RoBERTa		
		A=U	A>U	A<U
BLiMP	Animate Subject Passive	75.9	69.5	75.0
	Causative	83.6	84.1	81.7
	Drop Argument	65.3	63.1	70.3
	Inchoative	77.1	73.1	73.4
	Passive2	90.7	N/A	88.3
	Expletive It Object Raising	84.4	79.9	83.6
	Though vs. Raising 1	87.7	86.5	N/A
	Det. Noun Agr. Irregular 1	98.5	N/A	98.8
	Left Branch Island Echo Question	68.5	73.4	72.7
	Matrix Question NPI Lic.Pres.	90.1	N/A	89.6
CrowS-Pairs	Stereo	60.9	64.0	62.0
	Antistereo	56.8	61.9	53.6
COMPS	BASE	65.7	67.9	67.6
	WUGS	65.0	68.4	68.8



RQ2: Normalization of Log-likelihood

Log-likelihood is proportional to the token length

→ Normalizing the log-likelihood with token length might mitigate the token-length bias

Prior MPP Research: Used MeanLP and PenLP to normalize the log-likelihood

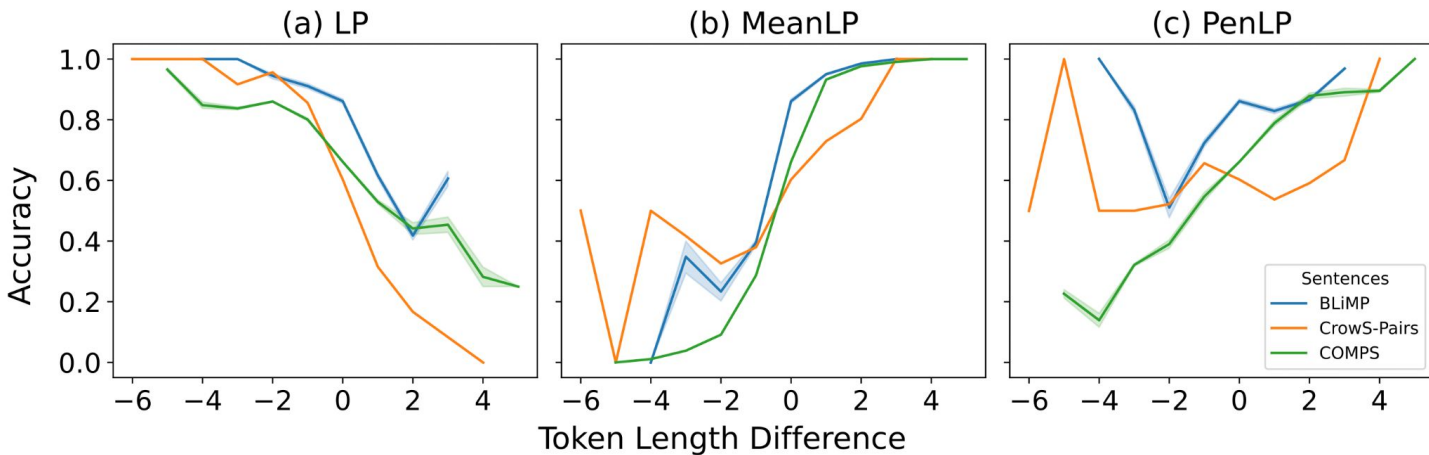
$$\text{MeanLP} = \frac{\log P_{LM}(S)}{|S|} \qquad \text{PenLP} = \frac{\log P_{LM}(S)}{((|S| + 5)/(5 + 1))^\alpha}$$

→ Unclear whether the log-likelihood is correctly normalized and whether the normalization method is valid in this task

Used MeanLP and PenLP as normalization methods to test whether they are actually effective against token-length bias

Result: Normalization of Log-likelihood (MeanLP)

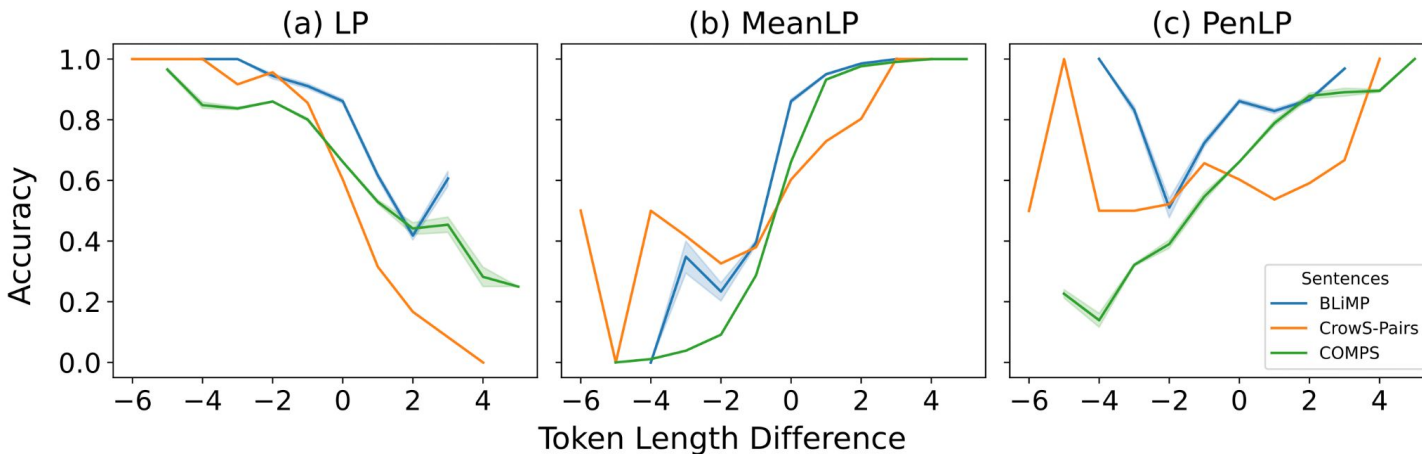
- In MeanLP, accuracies are affected when there is a token length difference (Token length difference: $|\mathbb{A}| - |\mathbb{U}|$)
 - When the token length difference is positive ($\mathbb{A} > \mathbb{U}$), accuracy **increases** significantly
 - When the token length difference is negative ($\mathbb{A} < \mathbb{U}$), accuracy **drops** significantly



Result: Normalization of Log-likelihood (PenLP)

- Relatively constant accuracy regardless of token length difference (BLiMP and CrowS-Pairs)
 - Reduced token-length bias compared to LP and MeanLP
- However, an accuracy is affected by token length difference on COMPS

→ **Not a method that can be used consistently across all MPP datasets**



Alternative Method for Mitigating Token-length Bias

- Normalizing the log-likelihood is not effective for mitigating the token-length bias in MPP task
- An alternative method of mitigating the token-length bias is needed
 - Control the acceptable and unacceptable sentences to have equal token length in each minimal pair

Acceptable Sentence

This ice **vapor ized** .

Unacceptable Sentence

This ice **resembles** .

This ice **resembl ed** .



Reconstruction of BLiMP

We introduce a debiased minimal pair generation algorithm

- Generate a minimal pair with equal token length on acceptable and unacceptable sentences on user selected models

Generated FairBLiMP (a debiased BLiMP dataset) as a case study

- Analyzed how the conclusions drawn from the evaluation results change between the original BLiMP and FairBLiMP

Algorithm 1 Debiased Minimal Pair Generation

```
1: function MINIMAL_PAIR_GENERATION(Models)
2:    $N \leftarrow 0$ 
3:   Generate an acceptable sentence AS and
   an unacceptable sentence US.
4:   while  $N < 10$  do
5:      $EqLen \leftarrow TRUE$ 
6:     for all Models do
7:       Tokenize AS and US.
8:       if  $|AS| \neq |US|$  then
9:          $EqLen \leftarrow FALSE$ 
10:      end if
11:    end for
12:    if  $EqLen = TRUE$  then
13:      return AS, US
14:    end if
15:    Regenerate an unacceptable sentence
    US by changing the minimally different words.
16:     $N \leftarrow N + 1$ 
17:  end while
18:  return None
19: end function
```

Results with FairBLiMP

The performance gap between the models narrowed and the model with the highest performance changed in some the subsets

Examples of notable impacts: "Inchoative" and "Expletive It Object Raising" subsets

Current MPP datasets are not able to properly compare the performance of different models

→ A fair comparison can be made by controlling the token length

Datasets & Subsets	Original BLiMP				FairBLiMP			
	GPT-2	BERT	RoBERTa	ELECTRA	GPT-2	BERT	RoBERTa	ELECTRA
Animate Subject Passive	67.5	79.1	74.9	72.8	73.3	82.4	76.6	77.3
Causative	75.6	73.1	83.6	82.3	81.6	82.1	86.1	84.7
Drop Argument	58.7	58.8	64.8	52.5	60.1	64.0	63.7	61.9
Inchoative	65.3	59.8	76.2	63.2	72.5	73.3	76.4	69.2
Passive2	88.7	90.1	90.1	93.7	91.3	92.8	93.5	94.3
Expletive It Object Raising	81.2	80.0	81.5	83.6	90.1	79.8	81.2	84.6
Though vs. Raising 1	76.4	68.4	87.6	65.4	84.0	77.7	87.5	72.2
Det. Noun Agr. Irregular 1	93.4	97.7	98.6	96.0	98.0	99.6	99.2	99.1
Left Branch Island Echo Question	51.8	66.2	69.7	46.8	40.9	68.4	69.1	49.9
Matrix Question NPI Lic.Pres.	58.4	92.7	89.9	91.4	59.1	94.3	90.0	92.5

Conclusion

Analyzed token-length bias in the MPP datasets

Results of the experiment showed that:

- The presence of a difference in token length between acceptable and unacceptable sentences has a significant impact on accuracy
- Suggested that normalizing log-likelihood by token length does not contribute to reducing the effect of token-length bias
- The current construction method that controls the length of sentence is not sufficient, and a construction method that controls the token length is necessary

今後の展望

本研究の限界

- トークン長を揃えるミニマルペアの作成手法では、性能を比較したいモデルが増加するにつれて、ミニマルペアの作成が困難になる
 - それぞれのモデルにおいてトークン化手法と語彙サイズが異なるためであり、全てのモデルに対してミニマルペアのトークン長を揃えることが難しくなる
 - 比較するモデルが変わる度にデータセットの再構築が必要なため、研究ごとに用いているデータセットが異なる

今後の課題

- データセットの再構築を必要としない、トークン長バイアスの影響を軽減できる教師なしでの容認度の計算方法の提案