T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Examining Impact of Evaluation Dataset Characteristics on Acceptability Judgments
著者(和文)	ヴィジャイ ドルタニ
Author(English)	Vijay Daultani
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第33号, 授与年月日:2024年12月31日, 学位の種別:課程博士, 審査員:岡崎 直観,徳永 健伸,村田 剛志,金﨑 朝子,井上 中順
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第33号, Conferred date:2024/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

Powered by T2R2 (Science Tokyo Research Repository)

（博士課程）

Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 知能情報	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（学術 ）
学生氏名： Student's Name	Vijay Daultani		審査員主査： Chief Examiner	岡崎直観	

要旨（英文 800 語程度）

Thesis Summary（approx.800 English Words ）

Acceptability evaluation is a fundamental aspect of assessing the linguistic capabilities of language models. It centers on how effectively a text communicates its intended meaning and resonates with native speakers. However, the characteristics of datasets used in these evaluations can significantly influence the performance of language models. While factors such as lexical frequency and sentence length have been extensively studied in areas like machine translation and text summarization, their specific impact on acceptability evaluation remains underexplored. This study aims to bridge this gap by examining how lexical frequency and sentence length affect a language model's ability to assess the acceptability of a sentence.

Sentence likelihood is a measure of how likely a sentence appears in a language model's training data, yet this does not represent its acceptability. External factors such as lexical frequency and sentence length influence a sentence's likelihood, resulting in situations where equally acceptable sentences may receive different likelihood scores. Over time, researchers have developed various probability-based metrics to deal with this issue, including the Syntactic Log Odds Ratio, aimed at leveraging sentence likelihood while mitigating the confounding effects of lexical frequency and sentence length.

Despite the development of these metrics, our research shows that lexical frequency continues to present challenges, particularly when handling out-of-vocabulary (OOV) words. OOV words, or words not present in the model's training data, can negatively impact the reliability of acceptability metrics. While existing methods have made progress in addressing this issue, they have not fully resolved it, leaving room for further improvement. A key limitation identified in our research is the reliance on original sentence representations for training language models. Specifically, when OOV words are substituted with unknown (UNK) tokens, the models tend to allocate probability mass to these tokens without adequately capturing the contextual significance of the original words.

A significant source of OOV words in many sentences comes from proper nouns. To address this challenge, we introduce a novel data transformation technique called Replace Named Entity (RNE). This method preprocesses both training and test datasets by replacing proper nouns with their corresponding named entity categories (e.g., replacing "1st January" with "DATE"). The rationale behind this approach is that the syntactic structure of a sentence often carries more weight in acceptability evaluations than the specific proper nouns it contains. Our experiments show that RNE improves the alignment between model predictions and human judgments, enhancing the accuracy of probability-based acceptability metrics across the board.

In addition to lexical frequency, our analysis underscores the importance of sentence length as another critical characteristic influencing a model's acceptability evaluation capabilities. This discovery prompted us to conduct a thorough examination of how sentence length impacts model performance. Our findings indicate that standard commonly-used datasets for acceptability evaluation often fail to represent the natural distribution of sentence lengths found in everyday human writing. Many of these datasets are skewed toward shorter sentences, likely due to their design, which aims to test specific linguistic phenomena in controlled settings. However, this bias toward shorter sentences can inflate the performance of language models, making it appear as though the models are more effective than they truly are when dealing with longer, more complex sentences.

To address this issue, we propose seven datasets (six derived and one novel) that better reflect realistic sentence length distributions. These datasets were carefully curated to mirror the sentences people use in everyday communication, thereby providing a more authentic basis for evaluating the acceptability of text generated by language models. By testing language models on these new datasets, we were able to offer a clearer and more accurate picture of how sentence length impacts model performance.

In conclusion, this study demonstrates that both lexical frequency and sentence length significantly influence the ability of language models to evaluate the acceptability of sentences. However, these effects can be mitigated by introducing improved methodologies, such as the RNE data transformation technique, and developing more representative datasets. Our research not only identifies the limitations of current evaluation practices but also offers practical solutions for improving the accuracy and reliability of acceptability assessments.

The techniques proposed in this study enhance the language model's ability to generate text that is not only syntactically coherent and semantically meaningful but also aligned with human expectations. By addressing key factors like OOV words and sentence length distributions, our work contributes to the ongoing advancement of NLP. These improvements will likely foster greater user trust in language models and expand their potential applications across various fields, especially content creation. As language models continue to evolve, ensuring that they produce text that aligns with human linguistic standards will be crucial for making human-computer interactions more seamless and natural. Ultimately, the methodologies proposed in this research pave the way for more robust, accurate, and reliable acceptability evaluations, contributing to the creation of language models that better meet the needs of their intended users.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).