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
Type(English)	Doctoral Thesis

Doctoral Dissertation

EXAMINING IMPACT OF EVALUATION DATASET CHARACTERISTICS ON ACCEPTABILITY JUDGMENTS

Vijay Daultani

November 2024

Artificial Intelligence Course
Department of Computer Science
School of Computing
Institute of Science Tokyo

A Doctoral Dissertation
submitted to the School of Computing
Institute of Science Tokyo
in partial fulfillment of the requirements for the degree of
Doctor of PHILOSOPHY

Vijay Daultani

Thesis Committee:

Professor Naoaki Okazaki	(Supervisor)
Professor Takenobu Tokunaga	(Co-supervisor)
Professor Tsuyoshi Murata	(Co-supervisor)
Associate Professor Asako Kanezaki	(Co-supervisor)
Associate Professor Nakamasa Inoue	(Co-supervisor)

Abstract

Acceptability evaluation is a key aspect of assessing language models, focusing on how effectively a text conveys its intended meaning and resonates with native speakers. A dataset's characteristics can greatly influence model performance. While factors like lexical frequency and sentence length have been widely studied in tasks such as machine translation and text summarization, their role in acceptability evaluation has received limited attention. This study aims to fill that gap by investigating how these factors affect language model's ability to assess acceptability.

Our findings on lexical frequency indicate that out-of-vocabulary words undermines the reliability of several probability-based acceptability metrics, revealing limitations in using original sentences for training language models. To address this, we propose replacing proper nouns in a sentence with named entity categories to create more generalized sentence representations. This approach improves the alignment between model evaluations and human judgments of acceptability.

Additionally, our analysis of lexical frequency suggested that sentence length, another dataset characteristic, also plays a significant role in a model's capabilities of acceptability evaluation. This prompted a deeper investigation into its impact. Our analysis shows that commonly used datasets for acceptability evaluation do not accurately reflect the sentence length distribution of human-written language. These datasets often contain shorter sentences, likely due to their design, which focuses on testing specific linguistic phenomena in controlled settings. This bias inflates model performance, offering an inaccurate portrayal of how models handle more natural, human-like text. To address this issue, we introduce seven new datasets with more realistic sentence length distributions. These datasets provide a more naturalistic foundation for evaluating the acceptability of language models.

Our proposed methods, which involve improved preprocessing techniques and the development of better datasets, enhance the accuracy and reliability of acceptability evaluations. By tackling these crucial factors, our work advances the field, ensuring that language models generate text that is both meaningful and natural for the intended audience.

Acknowledgements

As I write this thesis, I reflect on the journey that began four years ago when I enrolled in the PhD program at the Institute of Science Tokyo. My research journey, however, began much earlier, following my graduation from IIT Delhi in 2014 and my subsequent work as a researcher at NEC in Japan. Over the years, I advanced my career, leading research teams at Rakuten and Amazon, all while nurturing a desire for higher education. For six years, I debated between pursuing a PhD or an MBA, ultimately choosing the former—a decision I am deeply grateful for. This path has taught me invaluable lessons about research, perseverance, and personal growth. As I move on to the next chapter of my life, I will fondly remember the stimulating academic environment I've been part of.

I owe my deepest gratitude to my advisor, Naoaki Okazaki, whose unwavering support and understanding, both academically and personally, have been indispensable. His mentorship taught me many things, but most importantly, the art of storytelling in research. While experiments and results are crucial, the ability to present them as a compelling narrative is a lesson I will carry forward. I am also sincerely thankful to the members of my thesis committee for their invaluable feedback and guidance throughout my research.

This journey would not have been the same without the support of Héctor Javier Vázquez Martínez. A single email sparked numerous insightful discussions and collaboration, and through it all, I gained a cherished friend. His unwavering support during the challenging moments of my PhD helped me think critically and explore new research directions.

I am incredibly fortunate to have been surrounded by supportive friends and family. I am especially grateful to Suryakant Soni, Vilas Sahu, Komal Bansal, Niranjan Viladkar, Harmeet Singh, Sameer Pandit, Rahul Nishant, and many others whose names I cannot list individually but who have been pillars of encouragement throughout this journey.

Finally, my deepest gratitude goes to my family. To my parents, Dilip and Vandana Daultani, my sisters, Neelam and Sonam, and my brother, Dinesh—your unwavering love and support have been the foundation of my success. A special thanks to my little champ, Aahan, for being my constant source of inspiration.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Language is the primary medium through which humans communicate, expressing complex thoughts, emotions, and ideas. In the field of natural language processing (NLP), replicating this human ability within machines is of paramount importance. Specifically, NLP systems must generate text that is not only grammatically accurate but also natural and contextually meaningful. This task, known as acceptability evaluation, is crucial for determining how comprehensible and natural a piece of text is to human readers.

The importance of acceptability evaluation lies in the fact that the quality of generated text extends beyond mere grammatical correctness. In machine translation, for instance, a grammatically correct translation can still feel awkward if the phrasing does not reflect how native speakers naturally communicate. This lack of naturalness can make the output harder for users to understand or trust, undermining the system's effectiveness. Therefore, it is essential that generated text not only be coherent but also closely reflect natural language patterns to enhance user experience.

In the context of machine learning and NLP, the characteristics of datasets significantly influence model performance. While considerable research has been conducted on how factors such as lexical frequency and sentence length impact tasks like machine translation and text summarization, their role in acceptability evaluation remains relatively underexplored. Lexical frequency, the occurrence rate of specific words, and sentence length, the statistical distribution of sentence lengths within a dataset,may directly affect how natural and comprehensible users perceive generated text. This research seeks to address this gap

by exploring how these factors influence the acceptability evaluation process, aiming to deepen the understanding of language models and improve their performance. Enhanced acceptability evaluation will ultimately enable NLP systems to produce text that is not only accurate but also more natural, fostering smoother interaction between machines and users.

This chapter will first provide a comprehensive overview of the research topic, setting the stage for a detailed discussion of the background and context of the study highlighting why investigating dataset characteristics like lexical frequency and sentence length is essential. Following this, the research problem, aims, and objectives will be articulated in detail, emphasizing the study's significance and its potential practical applications in improving NLP systems. The chapter will conclude by providing a clear roadmap for the structure of the thesis, outlining how each subsequent chapter will contribute to addressing the research questions and achieving the stated objectives.

1.1 Background

Acceptability evaluation is a critical aspect of natural language processing (NLP) that determines how natural, coherent, and comprehensible a text is to human readers. This task is integral to the development of language models that are capable of producing text indistinguishable from human writing, with applications ranging from machine translation to automated content generation and conversational AI systems. To understand the current state of acceptability evaluation, it is important to trace its evolution and examine the key developments that have shaped this field.

The origins of acceptability evaluation in NLP can be traced back to the 1950s and 1960s, a period marked by the rise of rule-based methods in computational linguistics. During this era, researchers focused on developing systems that could generate and evaluate text based on predefined grammatical rules. One of the pivotal figures in early research was Noam Chomsky, whose introduction of generative grammar in 1957 revolutionized the way linguists and computational researchers approached language (Chomsky, 1957). Chomsky's distinction between grammaticality—the structural correctness of a sentence—and acceptability—the naturalness and comprehensibility of a sentence to native speakers—laid the groundwork for future research in acceptability evaluation.

As computational power and data availability increased in the 1970s and 1980s, the field of computational linguistics began to mature. Early efforts in this period still primarily focused on rule-based systems, which were effective at identifying grammatical errors but struggled with the more subjective and context-dependent aspects of language, such as naturalness and acceptability. The limitations of these early systems highlighted the need for more sophisticated approaches that could capture the nuances of human language more effectively.

The 1990s and 2000s marked a significant shift in the field, driven by the advent of large corpora and the rise of statistical methods in NLP. Researchers began leveraging vast amounts of text data to develop models that could predict the likelihood of a sentence based on its frequency and structure within these datasets. Techniques such as n-grams, Hidden Markov Models (HMMs), and Maximum Entropy models became prevalent, enabling more accurate and context-sensitive evaluations of text. During this time, metrics like perplexity were introduced to assess the fluency of language models, providing an indirect measure of text acceptability.

However, the fundamental transformation in acceptability evaluation came in the 2010s with the advent of neural networks and deep learning models. Innovations such as Word2Vec, GloVe, and Transformers (e.g., BERT, GPT) allowed for the creation of sophisticated language representations that captured both the syntactic and semantic properties of text. These models could generate text that was not only grammatically correct but also contextually appropriate and natural-sounding. As a result, acceptability evaluation became a more nuanced and integral part of NLP, essential for refining and advancing these models.

In the current landscape, acceptability evaluation is an critical task due to the diverse applications of language models and the growing demand for high-quality text generation across different domains. Modern language models are deployed in a wide range of settings, from chatbots and virtual assistants to automated content creation and translation services. This broad applicability has driven the need for more robust and reliable methods of evaluating text acceptability, as the expectations for naturalness and coherence have risen significantly.

1.2 Research Problem

The evaluation of machine-generated text has long been a central focus in NLP. Early methods, rooted in rule-based systems and generative grammar, have gradually evolved into more sophisticated approaches, including statistical models and neural networks. Recent breakthroughs, particularly with transformer-based models like BERT and GPT, have not only improved text generation but also enhanced the capacity of language models to assess text acceptability.

Despite these advancements, a notable gap remains in the current research. While many studies have examined how dataset characteristics impact performance in tasks such as machine translation and text summarization, less attention has been given to how these same factors influence language models' ability to evaluate text acceptability. Specifically, critical factors like lexical frequency (Jean et al., 2015; Koehn and Knowles, 2017) and sentence length (Provilkov and Malinin, 2021; Xuewen et al., 2021; Lu et al., 2022; Liangm et al., 2022), though extensively studied in other NLP contexts, have yet to be fully explored in the domain of acceptability assessment. This lack of focus highlights the need for further investigation into the influence of these factors on how language models evaluate the quality of generated text.

Addressing this gap is crucial because the quality of acceptability evaluations directly impacts the reliability of language models in practical applications. Without a clear understanding of how dataset characteristics affect evaluation outcomes, there is a risk of producing biased or inconsistent results, undermining the trustworthiness of NLP systems. Filling this research gap will help develop more robust and generalizable evaluation methods, ultimately improving the performance and applicability of language models across a wide range of tasks.

1.3 Research Aims

This research seeks to address a gap by investigating how specific characteristics of evaluation datasets—namely, lexical frequency and sentence length—affect the acceptability judgments made by language models. The first objective is to examine how lexical

frequency, particularly the occurrence of out-of-vocabulary (OOV) words, influences the model's assessment of sentence acceptability. The second goal is to evaluate the impact of sentence length distribution within the dataset on the model's acceptability judgments. Ultimately, the broader goal is to understand the extent of these influences and develop strategies to mitigate their effects in the evaluation process.

1.4 Significance

Evaluating linguistic acceptability plays a crucial role in enhancing the accuracy, reliability, and human-like performance of language processing systems. This acceptability evaluation process offers a number of benefits, some of which are outlined below.

Firstly, improving NLP systems through acceptability evaluation allows for the improvement of models used in different NLP tasks that require text generation. By ensuring that the sentences generated or analyzed by these models are not only grammatically accurate but also natural and contextually appropriate for native speakers, we can create language systems that are more aligned with human communication. This alignment is crucial for the practical deployment of NLP technologies in real-world scenarios, where the nuances of natural language play a significant role in the effectiveness of the system.

Secondly, the process of acceptability evaluation provides deep insights into human language understanding, which is invaluable for both the development of linguistic theory and the improvement of computational models. By examining how sentences are judged as acceptable or unacceptable, researchers can gain a better understanding of the underlying principles that govern language use. These insights, in turn, can inform the design of more sophisticated NLP models that better mimic human cognitive processes, leading to improved performance across a range of language tasks.

Lastly, the impact of acceptability evaluation extends to enhancing the user experience in applications like chatbots, virtual assistants, and other interactive language technologies. When these systems are able to generate responses that are not only accurate but also sound natural and human-like, the quality of interaction improves significantly. This leads to higher user satisfaction, as the interactions feel more intuitive and less mechanical, thereby bridging the gap between human and machine communication.

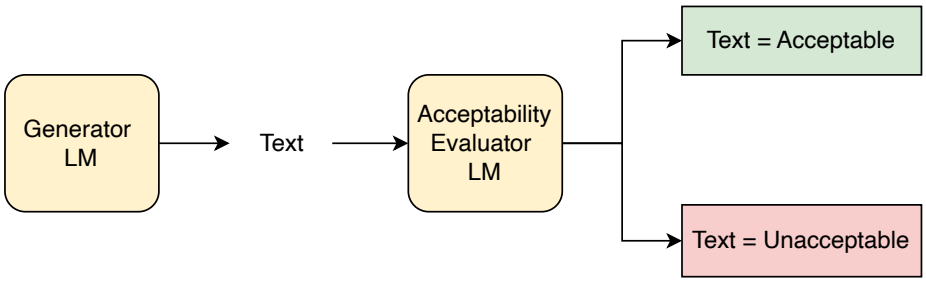


Figure 1.1: Two practical applications of acceptability evaluation

Overall, the evaluation of linguistic acceptability is a critical component of NLP that drives improvements in system accuracy, enriches linguistic understanding, and enhances user experiences. By focusing on how well language models align with human language norms, we can develop more sophisticated and user-friendly language technologies that better serve the needs of their users.

1.5 Practical Applications

In light of the growing use of generative language models, **Fig. 1.1** illustrates a typical setup for evaluating the acceptability of generated text in practical applications. As depicted, a generator language model produces natural language text. The objective of this model is to generate text that is both accurate and fluent, ensuring clarity and ease of understanding for readers. This generated text is then assessed by an evaluator language model, which aims to determine the acceptability of the input. Broadly, there are two key applications of acceptability evaluation:

1. **Evaluating the Text Generation Ability of Generator LM**: This involves examining the quality and overall acceptability of the text generated by the Generator LM.

2. **Evaluating the Judgment Accuracy of Evaluator LM**: This assesses how well the Evaluator LM can differentiate between acceptable and unacceptable text.

Although both tasks are crucial, this thesis concentrates on the second—assessing the Evaluator LM's ability to classify text as either acceptable or unacceptable. Enhancing the

evaluator's accuracy in making this distinction is essential, as it will indirectly improve the quality of text generated by language models, pushing development toward models that produce more acceptable output.

1.6 Limitations

This study provides an in-depth examination of how two key dataset features—lexical frequency and sentence length—influence the performance of language models in assessing sentence acceptability. While we briefly acknowledge the limitations here, a more detailed discussion will be presented in the concluding chapter (§5.7) of this thesis.

First, the study focuses solely on two characteristics: lexical frequency and sentence length. Other important factors, such as syntactic complexity and semantic content, were not explored, potentially limiting the scope of the findings. Second, the analysis is confined to English-language datasets. Given the structural and lexical diversity across languages, applying similar methods to non-English datasets may yield different results. Third, while we propose a preprocessing method to address the issue of out-of-vocabulary (OOV) words by transforming the data into a coarser level of representation, this method is based on intuition about human acceptability judgments. We did not conduct a detailed investigation into the impact of varying levels of granularity in these text transformations. Fourth, our evaluation relies on traditional probability-based metrics for acceptability judgments, without considering more recently introduced metrics, such as the BARTScore. Future research into these newer methods could provide a more nuanced understanding of sentence acceptability. Lastly, to address the effect of sentence length distribution on model performance, we performed dataset transformations that resulted in a reduction in dataset size. This reduction may have limited the variability present in the original dataset, possibly affecting the generalizability of our results.

1.7 Thesis Outline

The thesis is organized as follows: The opening chapter sets the stage by introducing the research topic, emphasizing its importance, and outlining the research questions addressed.

The second chapter focuses on the concept of acceptability evaluation, discussing its various dimensions. It contrasts different interpretations of acceptability, such as whether it should be viewed as a gradient or binary property, and reviews the two primary paradigms used in its assessment.

In the third chapter, the influence of lexical frequency on acceptability evaluation is examined, with particular attention to challenges posed by OOV words. While previous research acknowledges the role of lexical frequency in these evaluations and has proposed various metrics to address this, we demonstrate that the issue remains unresolved, as OOV words continue to impact probability-based metrics. To address this, we introduce a novel method called 'Replaced Named Entity' and provide experimental evidence of its effectiveness. The analysis also identifies sentence length as a critical factor, which is explored in greater depth in the following chapter.

Chapter four shifts focus to the role of sentence length in acceptability evaluation. We observe significant differences between the sentence length distributions in human-written corpora and standard acceptability datasets, which tend to be skewed towards shorter sentences. Through rigorous experimentation, we show that this skew leads to inflated performance estimates for language models in acceptability tasks. To address this issue, we introduce seven new datasets (six derived and one novel) that better reflect natural language usage, aiming to improve the accuracy of acceptability evaluations. We also propose strategies for refining evaluation datasets to mitigate the bias introduced by sentence length distributions in commonly used datasets.

The final chapter concludes the thesis by restating the research objectives, summarizing the key findings, and discussing their broader implications. It reflects on the practical applications of the findings, how they contribute to the field, and suggests potential directions for future research.

Chapter 2

Background

To begin, I will provide an overview of text evaluation in NLP and clarify the role of acceptability evaluation within the broader NLP pipeline. I will illustrate the significance of acceptability evaluation through specific examples, demonstrating its practical relevance. I will also give a few examples of how acceptability evaluation is often included when evaluating the text generation capabilities of a new language model.

Next, I will discuss the different terms related to 'acceptability' frequently appearing in the literature, such as fluency, readability, and grammaticality. I will explain how these terms are used interchangeably and highlight their subtle distinctions. Additionally, I will differentiate between acceptability and grammaticality, drawing from foundational studies, including Chomsky's work, to provide a clearer understanding of these concepts.

I will then examine the traditional view, as proposed by Chomsky, which treats acceptability as a gradient property and contrasts it with the contemporary perspective that considers acceptability as a binary property. Finally, I will review the two primary paradigms of acceptability evaluation: the single sentence paradigm and the minimal pair paradigm. While this study focuses on the judgment of the single-sentence approach, understanding both paradigms is essential for a thorough comprehension of acceptability evaluation.

2.1 Overview of Text Quality Evaluation in NLP

To understand the significance of acceptability evaluation, it is essential to recognize its role within the broader framework of text quality assessment in NLP pipelines. Language models are frequently used for tasks such as machine translation, text summarization, and chatbot interactions, each requiring text generation. For example, machine translation provides a text in the target language corresponding to an input source text, text summarization condenses the original text, and chatbots produce responses to user queries.

Generally, evaluating the performance of a language model involves two stages: intrinsic and extrinsic evaluation. Intrinsic evaluation examines the quality of the generated text itself, focusing on aspects such as acceptability, coherence, and perplexity. Acceptability ensures that the generated text is natural and coherent. On the other hand, extrinsic evaluation assesses how well the text performs in practical applications. For instance, in machine translation, BLEU (Papineni et al., 2002) scores compare generated translations to human translations, and ROUGE (Lin, 2004) scores are used in text summarization to measure the overlap between generated summaries and human-written ones.

2.1.1 Motivation

Consider the example shown in **Fig. 2.1** to grasp the importance of acceptability evaluation in machine translation. This figure demonstrates a machine translation system where the input is a sentence in the Japanese language, and the system's task is to translate this Japanese sentence into English. Now, let's suppose we have two machine translation outputs. The first translation is the sentence: "He was reading a book at the library yesterday". The second translation is the sentence: "He was at the library reading a book yesterday".

While both translations are grammatically correct, there is a notable difference in their acceptability. The first translation, "He was reading a book at the library yesterday" adheres well to English grammatical rules and is an acceptable sentence that makes it easy to understand. On the other hand, the second translation, "He was at the library reading a book yesterday" while also grammatically accurate, feels awkward and less natural. Its structure—[Subject (He), Auxiliary Verb (was), Prepositional Phrase (at the library), Gerund Phrase (reading a book), Adverb of Time (yesterday)]—is technically correct but results in

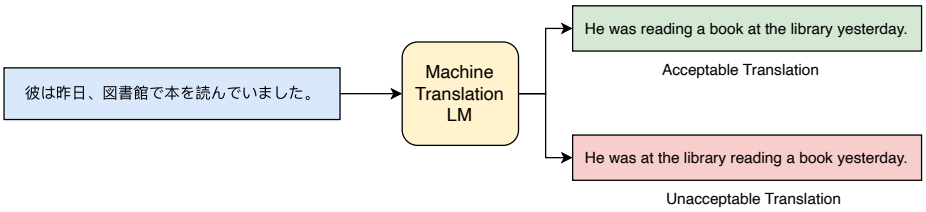


Figure 2.1: Sample machine translation for motivation of acceptability evaluation

a sentence that lacks acceptability.

Even though both translations are grammatically sound, the acceptability and naturalness of a translation are essential for clear and effective communication. This underscores the importance of acceptability evaluation in text evaluation systems.

2.1.2 In Practise

In NLP, it is standard practice to complement quantitative automatic evaluations on extensive benchmarks with qualitative manual evaluations when assessing a language model's performance on a dataset. A key aspect of this qualitative evaluation is often the assessment of acceptability. For instance, in evaluating various language models on the Newsroom dataset (Grusky et al., 2018), the qualitative assessment included syntactic aspects like acceptability and coherence, as well as semantic aspects like informativeness and relevance. Similarly, BARTScore (Yuan et al., 2021), a metric designed to evaluate machine-generated text, emphasizes the importance of mirroring the text generation process during evaluation. To highlight BARTScore's relevance, the authors demonstrate its Spearman correlation with various metrics, including fluency, on multiple human judgment datasets. There are several other examples where qualitative manual evaluation, particularly of acceptability, is employed to demonstrate model improvements.

From this, it is clear that acceptability evaluation is essential in NLP research, with many researchers incorporating it as part of language model evaluation. Given the relevance of this task, developing automated systems for acceptability evaluation is crucial. Automatic acceptability evaluation in this area could reduce the time and cost associated with manual assessment and facilitate the development of systems capable of generating

more acceptable text.

2.2 Terms Similar to Acceptability

In NLP, terms like fluency and readability are often synonymously used with acceptability. While these terms highlight slightly distinct aspects of language, they all share a common goal: ensuring that a text is clear, understandable, and appropriate for its audience. These concepts often overlap despite their nuances, as a well-written text must be fluent, readable, and acceptable. Therefore, they are frequently considered equivalent when evaluating a text's overall quality and effectiveness in communicating to its audience. In this thesis, we build on previous research (Mutton et al., 2007; Pitler and Nenkova, 2008; Storch, 2009; Vadlapudi and Katragadda, 2010; Lau et al., 2016; Kann et al., 2018), treating acceptability, fluency, and readability as synonymous for evaluation purposes.

2.2.1 Acceptability vs. Grammaticality

Theoretical linguists and researchers have long debated the distinction between 'acceptability' and 'grammaticality'. In this thesis, we adopt the view of Chomsky (1965), who emphasizes the importance of distinguishing between these two concepts. While grammaticality refers to a sentence's adherence to the formal rules of syntax, acceptability encompasses a broader range of factors, including grammaticality, semantic plausibility, and ease of processing.

In practice, a sentence can be grammatically correct but still unacceptable to native speakers due to issues such as semantic incongruence or cognitive complexity. For instance, Chomsky's famous example, "Colorless green ideas sleep furiously" (Chomsky, 1957) is grammatically sound but semantically nonsensical, rendering it unacceptable. Conversely, a sentence might be deemed acceptable in informal contexts, such as poetry, even if it violates grammatical norms. This distinction between grammaticality and acceptability has been further explored by scholars, including Radford (1988) and Haegeman (1994), who further emphasize that grammaticality pertains to the conformity of a sentence to linguistic rules, whereas acceptability relates to how native speakers perceive a sentence. In essence,

'grammaticality' focuses on rule-based correctness, while 'acceptability' assesses a sentence's overall fit within natural language use, taking into account various factors (e.g., semantic plausibility, and processing ease) beyond mere grammar. Below are relevant excerpts from these works that illustrate this distinction.

> **Chomsky, N. (1965).** ***Aspects of the Theory of Syntax***
> "The notion 'acceptable' is not to be confused with 'grammatical'. Acceptability is a concept that belongs to the study of performance, whereas grammaticalness belongs to the study of competence... Grammaticalness is only one of many factors that interact to determine acceptability." (p. 10)

> **Radford, A. (1988).** ***Transformational Grammar: A First Course***
> "It is important to distinguish between grammaticality and acceptability. A sentence may be grammatically correct according to the rules of a language, yet be unacceptable to speakers due to various performance factors." (p. 16)

> **Haegeman, L. (1994).** ***Introduction to Government and Binding Theory***
> "Grammaticality refers to the conformity of a sentence to the rules of a language, whereas acceptability pertains to how a sentence is perceived by native speakers. A sentence can be grammatical but not necessarily acceptable." (p. 12)

Although grammaticality and acceptability are distinct concepts, it is important to recognize that grammaticality is one of several factors that contribute to a text's overall acceptability.

2.3 Acceptability Evaluation

Acceptability evaluation is a linguistic assessment method used to determine how native speakers perceive the naturalness or appropriateness of a sentence or utterance within their language. Unlike grammaticality, which focuses strictly on adherence to formal syntactic rules, acceptability evaluation considers a broader range of factors including semantic

coherence, pragmatics, cognitive processing ease, and contextual appropriateness. It involves eliciting judgments from native speakers about whether a sentence 'sounds right' or is likely to be used in actual language situations, even if it violates certain grammatical norms. This evaluation can be influenced by factors such as frequency of usage, discourse context, and the speaker's tolerance for non-standard forms.

2.3.1 Granularity for Acceptability Evaluation

Acceptability in NLP can be assessed across various levels, from individual phrases to entire documents. However, due to practical challenges—particularly the complexity of compiling large-scale datasets for longer text segments—evaluations are often limited to the sentence level. As a result, most existing datasets and research in this area focus primarily on sentence-level acceptability. In line with these constraints, this thesis emphasizes the evaluation of acceptability at the sentence level.

Linguistic acceptability at the sentence level does not necessarily imply that each constituent phrase within the sentence is independently acceptable. A sentence can achieve overall linguistic coherence even when certain phrases within it may seem less acceptable on their own or require contextual cues to be fully understood. The acceptability of a sentence thus often arises from how its elements interact holistically, influenced by factors such as syntax, semantics, and pragmatics. In this study, however, we focus exclusively on sentence-level acceptability without assessing the independent acceptability of constituent phrases.

2.3.2 Gradient vs Binary Property

Cognitive scientists and linguists have widely debated the concept of text acceptability. Several scholars (Chomsky, 1965; Sprouse, 2007; Lau et al., 2016) propose that acceptability is a gradient property, existing on a continuum rather than as a binary distinction. Lau et al. (2016) supports this view with empirical evidence, showing that humans typically evaluate text acceptability on a scale. In contrast, other researchers, including (Warstadt and Bowman, 2020; Warstadt et al., 2020), advocate for a binary approach, asserting a clear division between acceptable and unacceptable sentences.

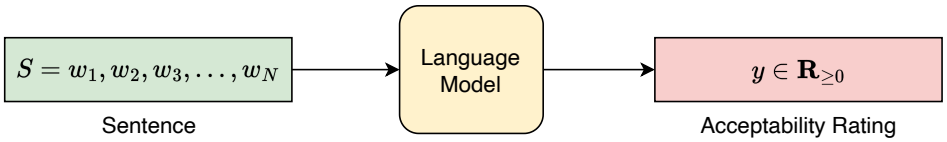


Figure 2.2: Evaluation of acceptability as a gradient property

Both perspectives offer valuable insights, each with distinct strengths and limitations. On one hand, treating acceptability as a gradient allows language models to leverage text likelihood—the probability assigned to tokens within a text—to estimate varying degrees of acceptability. However, it is crucial to distinguish between acceptability and likelihood, as the latter reflects factors like sentence length and lexical frequency rather than the inherent quality of the text. On the other hand, while binary classification models provide a simpler approach, they often fail to capture the subtleties of acceptability that can vary within a text in real world.

Creating datasets with gradient-based annotations presents several challenges. First, assembling a collection of sentences that span a range of acceptability is both time-intensive and costly. Unlike binary classification, where sentences are simply marked as acceptable or not, gradient annotation requires evaluators to assess sentences along a continuous scale, demanding significantly more effort. Additionally, the inherent subjectivity in these annotations can lead to inconsistencies, as different annotators may assign widely varying scores to the same sentence. These factors contribute to the complexity and expense of developing such datasets. As a result, recent research has increasingly adopted binary classification approaches for acceptability evaluation, as seen in datasets like CoLA (Warstadt et al., 2019) and BLiMP (Warstadt et al., 2020), which employ binary classification methods.

We align with Chomsky's 1965 view that acceptability is a gradient property. However, for our second study on how sentence length distributions affect acceptability evaluation, it was crucial to use larger datasets capable of capturing a broad range of sentence lengths. Many existing datasets for gradient acceptability are relatively small, often containing fewer than 5,000 sentences. In contrast, larger datasets exist for binary acceptability evaluation, such as CoLA (10,657 sentences) and BLiMP (67,000 sentences). As previously

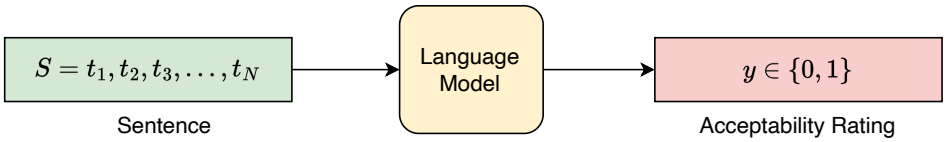


Figure 2.3: Evaluation of acceptability as a binary property

mentioned, developing large-scale datasets with gradient annotations is resource-intensive. As a result, we had to rely on these larger binary datasets to gain meaningful insights.

This limitation led us to examine acceptability from both gradient and binary perspectives. In the first part of our study, which focused on how lexical frequency influences language models' acceptability judgments, we treated acceptability as a gradient property. In the second part, where we explored the influence of sentence length on acceptability judgments, we adopted a binary approach. This dual perspective allowed us to utilize larger, more established datasets like CoLA and BLiMP. Below, we will define the task of acceptability evaluation within both gradient and binary frameworks.

- **Gradient Property**: **Figure 2.2** illustrates the approach of evaluating acceptability as a gradient property. In this case, the problem is framed as a regression task. The input to the language model is a sentence S, composed of N words $\{w_1, w_2, \ldots, w_N\}$. The model outputs a continuous value y, representing the degree of acceptability assigned to the sentence S. A higher y value corresponds to higher levels of acceptability.

- **Binary Property**: **Figure 2.3** presents the problem of evaluating acceptability as a binary property. Here, the task is formulated as a classification problem. The input, as in the previous case, is a sentence S consisting of N words $\{w_1, w_2, \ldots, w_N\}$. The model assigns a binary label y, where 0 denotes an unacceptable sentence and 1 denotes an acceptable one.

2.4 Why Lexical Frequency and Sentence Length?

An evaluation dataset in NLP can have several different characteristics like lexical frequency, sentence length distribution, vocabulary diversity, syntactic complexity, and semantic diversity etc. While each of the several dataset characteristics can potentially affect how language models perform in acceptability assessments, this study specifically investigates the influence of two key characteristics—lexical frequency and sentence length—on model performance in these evaluations. The rationale behind selecting these characteristics is worth examining.

Lexical frequency, especially the occurrence of out-of-vocabulary (OOV) words, has been extensively examined in various NLP tasks, including machine translation and text summarization. Given its well-documented importance, we initially investigated the role of lexical frequency in shaping the performance of language models during acceptability evaluation. Our findings revealed preliminary evidence that, in addition to lexical frequency, sentence length might significantly influence model outcomes in this context. This prompted a deeper exploration into how sentence length distribution impacts the acceptability evaluation of language models.

While previous research has explored the effects of lexical frequency (Jean et al., 2015; Koehn and Knowles, 2017) and sentence length (Provilkov and Malinin, 2021; Xuewen et al., 2021; Lu et al., 2022; Liangm et al., 2022) on various NLP tasks, our study offers a unique perspective by focusing specifically on their influence on acceptability evaluation. By examining both factors, this research aims to provide new insights into how these characteristics affect language model performance in this specific task. Below, we outline our motivations and objectives for studying each of these characteristics in greater detail.

2.4.1 First Characteristic: Lexical Frequency

Lexical frequency refers to the frequency of word occurrences within a dataset and is closely related to the long-tail distribution problem, where a small number of words occur frequently while many others are rare. Language models are often trained with a fixed, limited vocabulary to optimize computational efficiency and memory usage. As a result,

infrequent words, including rare ones, may be excluded from the model's vocabulary, leading to OOV issues during inference. Previous studies, such as Jean et al. (2015), have shown that the performance of language models in tasks like machine translation declines with an increase in OOV words. Koehn and Knowles (2017) further found that while neural machine translation systems using subword-level techniques, such as byte-pair encoding, outperform statistical machine translation models on rare words, they still struggle with highly inflected words, such as verbs. These insights into the relationship between lexical frequency and model performance motivated us to explore its impact on acceptability evaluation.

2.4.2 Second Characteristic: Sentence Length

Our first study on lexical frequency revealed that the number of named entities in a sentence, which typically increases with sentence length, correlates with improvements in human acceptability judgments when using the proposed preprocessing method. Furthermore, disparities in input sequence lengths between training and testing phases are a known issue in NLP evaluation datasets. For instance, Provilkov and Malinin (2021) identified a bias in machine translation datasets toward shorter input sequences, exacerbating the beam-search problem where increasing the beam size beyond a certain point degrades translation quality. In contrast, Bando et al. (2012) found that the TREC Novelty track dataset for text summarization favors longer sentences, as they are often deemed more relevant than shorter ones. These observed discrepancies, along with our own findings, motivated us to investigate the impact of sentence length distribution on the performance of acceptability evaluations.

2.5 Related Work on Impact of Additional Dataset Characteristics

This thesis focuses on examining the impact of two key dataset characteristics—lexical frequency and sentence length. However, recent studies have explored a broader range of dataset characteristics and their influence on language models' acceptability judgments.

For instance, Kann et al. (2019) assessed the capacity of embeddings to distinguish between acceptable and unacceptable verb-frame combinations in English. Using both word-level and sentence-level embeddings, they found that while models could reliably classify some verbal alternations, others proved more challenging, suggesting that although embeddings encode detailed lexical and structural information, certain frames are more accessible to models. Their work indicates that while some verb-argument structures are reliably represented in model embeddings, consistency across all types remains limited.

In a related line of inquiry, Zhang et al. (2024) demonstrated that the linguistic and syntactic diversity within training datasets significantly affects models performance in multilingual acceptability evaluations. Their research revealed that language models acceptability judgments vary across language families, suggesting that models perform better when trained on datasets with linguistic structures similar to the target language. This finding emphasizes the role of syntactic alignment in improving cross-linguistic model accuracy for acceptability tasks. These studies collectively underscore the importance of dataset alignment with linguistic structures in both monolingual and multilingual contexts to optimize model judgments on acceptability tasks.

2.6 Acceptability Evaluation Paradigms

The evaluation of sentence acceptability in linguistic research typically follows two main paradigms: the Single Sentence Paradigm (SSP) and the Minimal Pair Paradigm (MPP). Below, we outline the key characteristics of each. In SSP, the model processes a single sentence, S, composed of N words $\{w_1, w_2, \ldots, w_N\}$, and produces an acceptability rating y for that sentence. This provides a direct assessment of how well the sentence adheres to the language norms. By contrast, MPP involves comparing two sentences: S, which is acceptable, and S', which is unacceptable. The model assigns preference scores to both, with the goal of giving the acceptable sentence a higher score. This paradigm offers a more nuanced assessment, as it tests the model's ability to distinguish between varying levels of sentence acceptability.

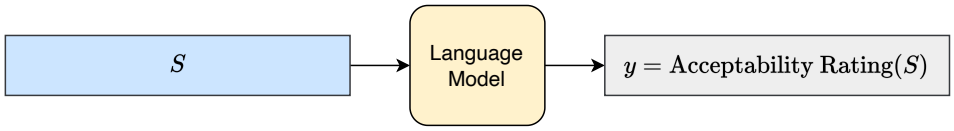


Figure 2.4: Single Sentence Paradigm for acceptability evaluation

2.6.1 Single Sentence Paradigm

Figure 2.4 illustrates the Single Sentence Paradigm (SSP), in which a language model assesses the acceptability of a single input sentence S. The model generates an acceptability rating y, which can be either a continuous value $y \in \mathbf{R}_{\geq 0}$ for gradient-based acceptability evaluation or a binary value $y \in \{0, 1\}$ for a categorical judgment of acceptability.

Evaluating Acceptability as a Gradient Property

Researchers often rely on measures like likelihood or log probability when evaluating sentence acceptability as a gradient property. This is typically achieved by using probabilistic models to estimate the likelihood $P(S)$ of a sentence S.

$$P(S) = P(w_1, w_2, \ldots, w_N) \tag{2.1}$$

This likelihood is calculated by decomposing the joint probability of the words into a series of conditional probabilities:

$$P(S) = \prod_{i=1}^{N} P\left(w_i \mid w_1, w_2, \ldots, w_{i-1}\right) \tag{2.2}$$

Likelihood values can become extremely small, particularly when working with long sequences of probabilities. Multiplying these small values can lead to numbers that approach zero, increasing the risk of underflow during computations. To avoid this issue, it is standard practice to convert likelihoods into log probabilities. By taking the logarithm of the likelihood, the small values are transformed into more manageable negative numbers, thus improving numerical stability. The log probability of a sentence S can be expressed as:

$$\log P(S) = \sum_{i=1}^{N} \log P\left(w_i \mid w_1, w_2, \ldots, w_{i-1}\right) \tag{2.3}$$

Finally, the log probability can be used to represent a sentence's acceptability rating:

$$\text{Acceptability Rating}(S) = \log P(S) \tag{2.4}$$

While log probability is a commonly used measure for sentence acceptability, as seen in Eq. 2.4, it is not without limitations. As discussed in Chapter 3, this approach can be affected by factors such as lexical frequency and sentence length. To overcome these issues, recent research has introduced alternative metrics that integrate unigram probabilities and sentence length, offering a more nuanced and accurate assessment of sentence acceptability.

Evaluating Acceptability as a Binary Property

The evaluation is typically framed as a binary classification problem in binary acceptability rating tasks for SSP. A common approach involves fine-tuning a language model, followed by applying a linear classifier. This classifier maps the probability distribution over the model's output vocabulary into a 2-dimensional vector corresponding to two classes: acceptable and unacceptable. Here, p_1 represents the probability of a sentence being classified as acceptable, and p_0 represents the probability of it being classified as unacceptable. A sentence is considered acceptable if p_1 is greater than p_0, as outlined in Eq. 2.5.

$$\text{Acceptability Rating}(S) = \begin{cases} 1, & \text{if } p_1 > p_0 \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

2.6.2 Minimal Pair Paradigm

Figure 2.5 depicts the Minimal Pair Paradigm (MPP), which involves presenting two sentences: S (acceptable) and S' (unacceptable). The language model evaluates these sentences and assigns preference scores, represented as $\mathcal{P}$. The objective is for the model to give a higher preference score to the acceptable sentence compared to the unacceptable one, specifically, $\mathcal{P}(S) > \mathcal{P}(S')$. Various methods exist for calculating this preference score,

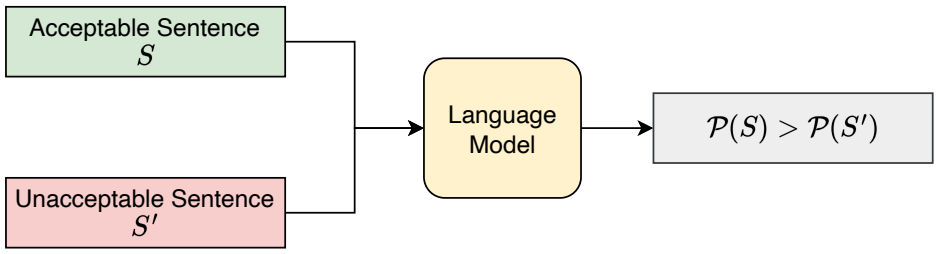


Figure 2.5: Minimal Pair Paradigm for acceptability evaluation

with one prevalent approach being a variation of the sentence log probability, as shown in Eq. 2.6.

$$\mathcal{P}(S) = \exp\left(-\frac{1}{N}\sum_{n=1}^{N}\log p(w^n)\right) \tag{2.6}$$

2.6.3 Single Sentence Paradigm vs. Minimal Pair Paradigm

The Single Sentence Paradigm and Minimal Pair Paradigm offer distinct advantages and serve different purposes in acceptability evaluation. Consequently, both types of datasets are commonly found in practice. For instance, the Adger dataset (Lau et al., 2016) utilizes SSP with acceptability as a gradient property, while CoLA (Warstadt and Bowman, 2020) uses SSP with binary acceptability, and BLiMP (Warstadt et al., 2020) follows the MPP approach.

In SSP, a language model evaluates each sentence independently without considering other sentences in the dataset. This straightforward approach allows for an absolute measure of sentence acceptability based on isolated judgments. However, normalization is often needed for meaningful interpretation when acceptability is treated as a gradient property in SSP. By contrast, MPP involves comparing pairs of minimally different sentences, enabling a more nuanced analysis of linguistic contrasts. MPP can better capture subtle differences in acceptability by focusing on how two sentences vary in grammaticality or naturalness.

Given the differences between these paradigms, SSP is more suitable for situations

where the goal is to assess the acceptability of individual sentences without comparison to unacceptable counterparts. This makes SSP particularly useful in real-world applications, where sentences are typically assessed in isolation and paired sentences are rarely available. Consequently, this thesis focuses exclusively on SSP to explore how lexical frequency and sentence length impact language models' acceptability evaluations.

Chapter 3

Impact of Lexical Frequency

In this study, we investigate how two key factors—lexical frequency and sentence length—affect the performance of language models in evaluating sentence acceptability. This chapter focuses on lexical frequency and its influence on acceptability judgments when approached as a gradient property. We begin by defining the problem and presenting sample sentences demonstrating varying acceptability levels. Following this, we review the relevant literature, highlighting why relying solely on sentence likelihood is inadequate for assessing acceptability, as these metrics are influenced by both lexical frequency and sentence length.

Over time, researchers have introduced probability-based metrics like the syntactic log odds ratio (SLOR) (Pauls and Klein, 2012; Lau et al., 2016) to minimize these confounding factors. However, we argue that while SLOR reduces these effects, it does not completely eliminate them. A key issue lies in the granular-level representation used by language models, where probabilities for even out-of-vocabulary words contribute to the overall sentence log probability. I will present scenarios to demonstrate that despite advancements, the confounding effects of lexical frequency and sentence length remain problematic. Drawing on how humans evaluate acceptability—by considering sentences more holistically—we propose that input sentences should be adopted for a coarser-level representation to better align with human judgments.

To address this, I introduce the 'Repalce Named Entity' method, a data transformation technique aimed at preprocessing datasets before training language models for acceptability evaluation. The chapter concludes with a discussion of the experimental setup, results, and

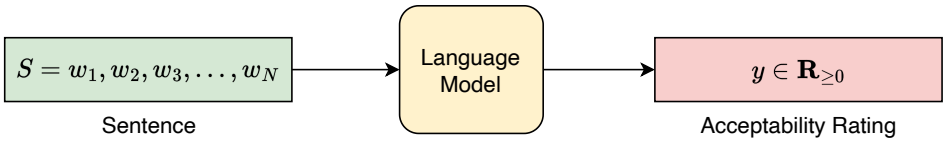


Figure 3.1: Problem statement for evaluating gradient acceptability rating

qualitative analysis, followed by a summary of the key findings.

3.1 Problem Statement

Figure 3.1 illustrates the problem setting when evaluating acceptability as a gradient property. Specifically, a sentence S consists of N words $\{w_1, w_2, \ldots, w_N\}$, with each word w_i having a lexical frequency $freq_i$ in the training corpus. Our objective is to determine the acceptability rating $y \in \mathbf{R}_{\geq 0}$ for the sentence S. A higher value of y indicates greater acceptability of the language model assigned to the sentence.

To illustrate the motivation behind acceptability evaluations, **Table 3.1** presents two sentences with contrasting acceptability ratings. The table is divided into three columns: 'ID', 'Sentence', and 'Acceptability Rating'. The 'ID' column uniquely identifies each sentence, the 'Sentence' column displays the text, and the 'Acceptability Rating' column reflects human judgments of the sentence's acceptability. The first sentence, S_1, "My first real friend was probably a boy called Adam", is rated highly, with an acceptability score of 4. In contrast, the second sentence, S_2, "Wonder why you - But mibbe I should thank ma", receives a much lower score of 1.

The significant difference in ratings is mainly due to multiple issues in S_2. First, it contains grammatical errors: the phrase "Wonder why you" is incomplete, lacking both a subject and a predicate, making the intended meaning unclear. The word 'mibbe' is an informal, nonstandard spelling of 'maybe', which may be confusing in more formal writing. Additionally, 'ma' is an ambiguous colloquialism, often short for 'my' or 'mother', but its meaning is unclear without further context. Overall, the sentence feels disjointed and fragmented, with an abrupt transition between "Wonder why you" and "But mibbe I should

ID	Sentence	Acceptability Rating
S_1	My first real friend was probably a boy called Adam.	4
S_2	Wonder why you - But mibbe I should thank ma.	1

Table 3.1: Sample sentences for gradient acceptability evaluation

thank ma", disrupting the logical flow and making it difficult to interpret the intended message.

3.2 Background

In evaluating the acceptability of a sentence S, one might consider using its likelihood $P(S)$ as a measure of acceptability as defined in Eq. 3.1. A highest likelihood value of 1 would imply the sentence is entirely acceptable, while a lowest likelihood value of 0 would suggest it is a completely unacceptable sentence. Intermediate likelihood values could be interpreted as corresponding to varying levels of acceptability. However, this interpretation of likelihood is fundamentally flawed. The likelihood assigned to a sentence by a language model represents the probability that S appears in the training corpus, not its inherent acceptability.

$$P(S) = \prod_{i=1}^{N} P\left(w_i \mid w_1, w_2, \ldots, w_{i-1}\right) \tag{3.1}$$

The likelihood of a sentence is the product of the probabilities assigned to each word in that sentence. This measure is influenced by factors such as word frequency and sentence length. Sentences containing less common words will have a lower probability, leading to a lower likelihood overall. Likewise, longer sentences accumulate more multiplications of word probabilities, which further reduces the likelihood. An effective acceptability evaluation metric should mitigate the confounding impacts of lexical frequency and sentence length. In other words, such a metric must rate sentences of equal acceptability similarly, regardless of the frequency of the words they contain. Moreover, the metric should base its evaluation solely on the sentence's acceptability rather than being skewed by sentence length.

Researchers have proposed several metrics to evaluate sentence acceptability, both at the sentence and word levels (Pauls and Klein, 2012; Lau et al., 2016; Kann et al., 2018). Before discussing these metrics in detail, it is crucial to establish some basic concepts. For simplicity, we will define the likelihood of a sentence as its sentence probability, $p_m(S)$:

$$p_m(S) = P(S) \tag{3.2}$$

Additionally, the unigram probability $p_u(S)$ of the sentence is the product of the unigram probabilities of each word:

$$p_u(S) = \prod_{t=1}^{n} p_u\left(w_t\right) \tag{3.3}$$

Finally, the length of the sentence S is denoted by $|S|$

3.2.1 Gradient Acceptability Metrics

This section outlines several probability-based metrics proposed by Pauls and Klein (2012); Lau et al. (2016) for evaluating the gradient acceptability of sentences.

1. **Log Probability** (LP) of a sentence is calculated as the logarithm of likelihood or the probability assigned to the sentence by the language model, as shown in Eq. 3.4.

$$\text{LP}\ (S) = \log p_m(S) \tag{3.4}$$

2. **Mean Log Probability** (Mean LP) is determined by dividing the Log Probability of the sentence by its length $|S|$, as given in Eq. 3.5.

$$\text{Mean LP}\ (S) = \frac{\log p_m(S)}{|S|} \tag{3.5}$$

3. **Normalized Log Probability Division** (Norm LP Div) addresses the impact of lexical frequency. This metric divides the Log Probability of the sentence by the Log

Probability of its unigram probability, as shown in Eq. 3.6.

$$\text{Norm LP Div } (S) = -\frac{\log p_m(S)}{\log p_u(S)} \tag{3.6}$$

4. **Normalized Log Probability Subtraction** (Norm LP Sub) also addresses the impact of the lexical frequency and is calculated by subtracting the unigram probability's Log Probability from the sentence's Log Probability. This is equivalent to taking the log of the ratio between the sentence and the unigram probability, as shown in Eq. 3.7.

$$\text{Norm LP Sub } (S) = \log p_m(S) - \log p_u(S) = \log \frac{p_m(S)}{p_u(S)} \tag{3.7}$$

5. **Syntactic Log Odds Ratio** (SLOR), formulated in Eq. 3.8, shows strong correlation with human acceptability judgments. SLOR addresses both lexical frequency and sentence length by subtracting the unigram probability and dividing by sentence length.

$$\text{SLOR } (S) = \frac{\log p_m(S) - \log p_u(S)}{|S|} \tag{3.8}$$

3.3 Limitations of Probability-Based Metrics

The metrics proposed above aim to evaluate acceptability as a gradient property, addressing the challenge of adapting sentence probability (p_m) for this task while reducing the confounding effects of lexical frequency and sentence length through the inclusion of unigram probability (p_u) and sentence length ($|S|$). While this approach helps mitigate these factors somewhat, it does not entirely eliminate the issue.

To illustrate, consider three sentences, S_3, S_4, and S_5, presented in **Table 3.2**: (1) "He is a citizen of France" (2) "He is a citizen of Tuvalu" and (3) "He is a citizen of Kiribati" Each sentence shares the prefix "He is a citizen of" followed by the names of different countries: 'France', 'Tuvalu', and 'Kiribati'. Let $freq(w)$ represent the frequency of a word w.

Assume, for simplicity, that in the training corpus, 'France' is a high-frequency word, 'Tuvalu' is less common, and 'Kiribati' is an OOV word. The probability of the shared prefix "He is a citizen of" will be the same for all three sentences. Now let's see how the

ID	Sentence
S_3	He is a citizen of France
S_4	He is a citizen of Tuvalu
S_5	He is a citizen of Kiribati

Table 3.2: Three sentences of equal length and acceptability

difference in sentence and unigram probabilities due to the words 'France', 'Tuvalu' and 'Kiribati' will influence a metric like SLOR.

Since $freq(\text{France}) > freq(\text{Tuvalu})$, for a bi-gram language model this results in $p(\text{France} \mid \text{citizen}) > p(\text{Tuvalu} \mid \text{citizen})$ and $p_u(\text{France}) > p_u(\text{Tuvalu})$, ultimately leading to $\text{SLOR}(S_3) \approx \text{SLOR}(S_4)$. Conversely, since $freq(\text{Kiribati}) = 0$ (as it is an OOV word), $p_u(\text{Kiribati}) \approx 0$. This occurs because OOV words are typically replaced by a UNK token in training and test corpora, with a small, non-zero unigram probability assigned to such words. While this approach is meant to prevent zero probabilities for OOV words, it still is not completely effective since $\text{SLOR}(S_3) \not\approx \text{SLOR}(S_5)$.

This discrepancy is problematic because ideally, $\text{SLOR}(S_3) = \text{SLOR}(S_4) = \text{SLOR}(S_5)$, given that all three sentences share the same structure and differ only in the country names, which should not affect their overall acceptability. However, due to the differences in word frequency and the presence of an OOV word, we observe $\text{SLOR}(S_3) \approx \text{SLOR}(S_4) \not\approx \text{SLOR}(S_5)$.

Despite introducing unigram probability (p_u) to mitigate the confounding effects of lexical frequency and sentence length $|S|$ to account for sentence length, probability-based metrics like SLOR still do not fully resolve these issues. This challenge is not unique to SLOR but is inherent to all probability-based metrics. While unigram probabilities help address the influence of lexical frequency to some extent, they are insufficient, mainly when dealing with OOV words, which can significantly distort acceptability ratings.

3.4 Proposed Method

The previous section highlights that simply subtracting the unigram probability (p_u) from the sentence probability helps reduce the confounding effects of lexical frequency but does

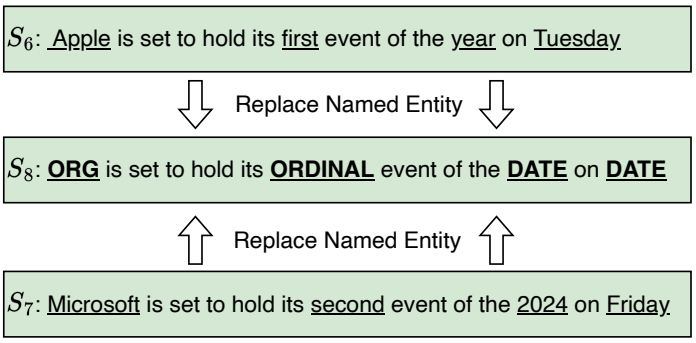


Figure 3.2: Motivation for the proposed method of Replaced Named Entity

not fully resolve the issue. This raises a crucial question: Can we further minimize the influence of lexical frequency? The answer is yes.

To explore this idea, consider two sentences S_6 and S_7 as shown in **Fig. 3.2**. Sentence S_6 reads: "Apple is set to hold its first event of the year on Tuesday", and sentence S_7 reads: "Microsoft is set to hold its second event of 2024 on Friday". Both sentences have similar structures and the same length, with the primary difference being the named entities. In S_6, the named entities are 'Apple', 'first', 'year', and 'Tuesday', corresponding to the named entity types ORG, ORDINAL, DATE, and DATE. Similarly, S_7 contains 'Microsoft', 'second', '2024', and 'Friday', matching the same entity types as those in S_6.

As human readers, we tend to rate these sentences similarly in terms of acceptability because their structural composition is the same despite the change in named entities. Our judgment focuses on the syntactic structure rather than the specific entities. Whether 'Apple' is replaced by 'Microsoft' or 'Google' has no significant impact on evaluating the sentence. In other words human judgments of sentence acceptability primarily focus on the overall syntactic structure and the transitions between parts of speech (POS), rather than the frequency or familiarity of individual words (Lapata and Barzilay, 2005). Building on this observation, we propose a method 'Repalce Named Entity' (RNE) where named entities in a sentence (e.g., sentences S_6 and S_7) are replaced with their respective entity types, creating more generalized, abstract representations (e.g., sentence S_8). This approach helps language models evaluate sentence acceptability with reduced bias from specific word choices or their frequencies. By standardizing sentences based on entity types, we ensure a more

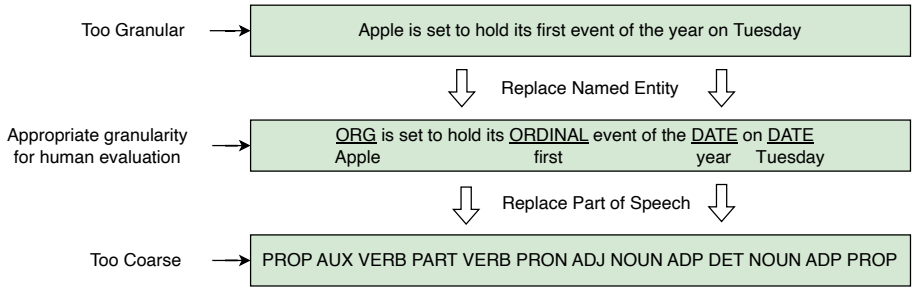


Figure 3.3: Different levels of granularity for sentence representation

consistent representation of similar syntactic structures, mitigating the impact of lexical frequency.

A natural question arises: why not replace every word in a sentence with its corresponding POS tag instead of only substituting named entities (e.g., proper nouns) with their types? We contend that replacing only named entities while preserving the rest of the sentence strikes a balance, maintaining the sentence's structure while abstracting its specific details. For example, consider the sentence: "Apple is set to hold its first event of the year on Tuesday" (denoted as S_6 and illustrated in **Fig 3.3**). Using our proposed method, RNE, this transforms into "ORG is set to hold its ORDINAL event of the DATE on DATE." This level of abstraction offers a suitable balance for human evaluation and is likely optimal for language models.

In contrast, replacing every word with its POS tag would yield: "PROP AUX VERB PART VERB PRON ADJ NOUN ADP DET NOUN ADP PROP." While this is a more generalized form, it is difficult for humans to interpret and does not support effective evaluation of sentence acceptability. Similarly, such a broad abstraction may obscure essential details for a language model, making assessing the sentence's acceptability challenging. In summary, we contend that substituting only named entities offers a balanced representation, maintaining sufficient detail while avoiding the extremes of overly granular or overly coarse representations.

Dataset	Sentences	Avg. Words	Avg. NE's	Avg UNK's
BNC	5250	17.81	1.07	0.96
ENWIKI	2500	17.21	2.07	0.23
ADGER	300	7.30	0.53	0.04
ADGER-FILTERED	133	8.02	0.68	0.00

Table 3.3: Details of Testing Data for gradient acceptability evaluation

3.5 Experiments

3.5.1 Training Data

We trained our language models using the British National Corpus (BNC) (BNC Consortium, 2007), following a method similar to that of Lau et al. (2016). The BNC contains approximately 6 million sentences and 100 million words, offering a broad and representative snapshot of British English across various contexts. Developed by the University of Oxford, it features a diverse range of text types, including spoken dialogues, newspapers, magazines, fiction, and academic writing.

3.5.2 Testing Data

In line with the methodology outlined by Lau et al. (2016), we evaluated the trained language models (LMs) using sentences with varying degrees of acceptability. Our analysis utilized four English language datasets: BNC, EnWiki, Adger, and Adger-Filtered. These datasets are components of the Statistical Model of Grammaticality (SMOG) test corpus (Lau et al., 2015). Detailed statistics for each dataset are presented in **Table 3.3**. Human-annotated acceptability judgments accompany each sentence within these datasets.

3.5.3 Language Models

We trained three variations of n-gram language models—2-gram, 3-gram, and 4-gram—using the BNC training corpus. We chose these models because our approach builds on the work of Lau et al. (2016), who demonstrated that n-gram models are highly competitive and among the most effective for evaluating linguistic acceptability. Additionally, note that our

process involves applying dataset transformations before both training and testing, which mirrors the token masking technique commonly employed during the pre-training of modern large language models.

3.5.4 Methods

We compared the baseline model (BL for simplicity) and a model trained on a dataset preprocessed with our proposed RNE method (Ours). Below, we outline the specifics of each approach.

1. **Baseline**: For our baseline (BL) model, we preprocessed the training corpus according to the protocol outlined by Lau et al. (2016). This process involved three main steps. First, we segmented the text into sentences. Next, we excluded sentences containing fewer than seven words. Finally, we replaced any rare words—those occurring less than four times in the corpus—with an unknown (UNK) token. This preprocessing resulted in a vocabulary of 104,950 words. To address the zero unigram probabilities for UNK tokens, we applied Kneser-Ney smoothing (Kneser and Ney, 1995).

2. **Replace Named Entity**: We followed the same initial preprocessing steps as the Baseline model in the Replace Named Entity approach. This included segmenting the text into sentences, filtering out sentences with fewer than seven words, and substituting rare words with a UNK token. Additionally, we enhanced this method by substituting all named entities with their respective types. We utilized spaCy's entity recognition system (Honnibal et al., 2020) to categorize named entities into one of 18 predefined types CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK OF ART, and MISC. A comprehensive description of each entity type can be found in the Appendix in **Table A.1**. Following the baseline model, Kneser-Ney (Kneser and Ney, 1995) smoothing was applied to manage zero unigram probabilities for UNK tokens.

3.5.5 Metrics

We assessed the performance of the trained language models on test data processed using both the Baseline method and our proposed Replace Named Entity approach. To evaluate the effectiveness of these methods, we employed all five probability-based metrics (§3.2.1). We used the language model's probability distribution over words for every sentence to compute acceptability ratings for all five metrics.

We used the Pearson correlation coefficient (PCC) to gauge the significance of each metric and track improvements offered by our method. This statistical measure quantifies the relationship between human acceptability judgments and the acceptability scores predicted by each metric. By calculating PCC, we can evaluate how closely the model's predictions align with human evaluations. As outlined in Eq. 3.9, PCC measures the correlation between human judgments of acceptability (x) and the model's predictions (y) for a given sentence (S). Here, σ_x and σ_y denote the standard deviations of human judgments and model predictions, while Cov refers to the covariance between these two sets of values, as shown in Eq. 3.10. In this equation, $\bar{x}$ and $\bar{y}$ represent the mean values of x and y, respectively.

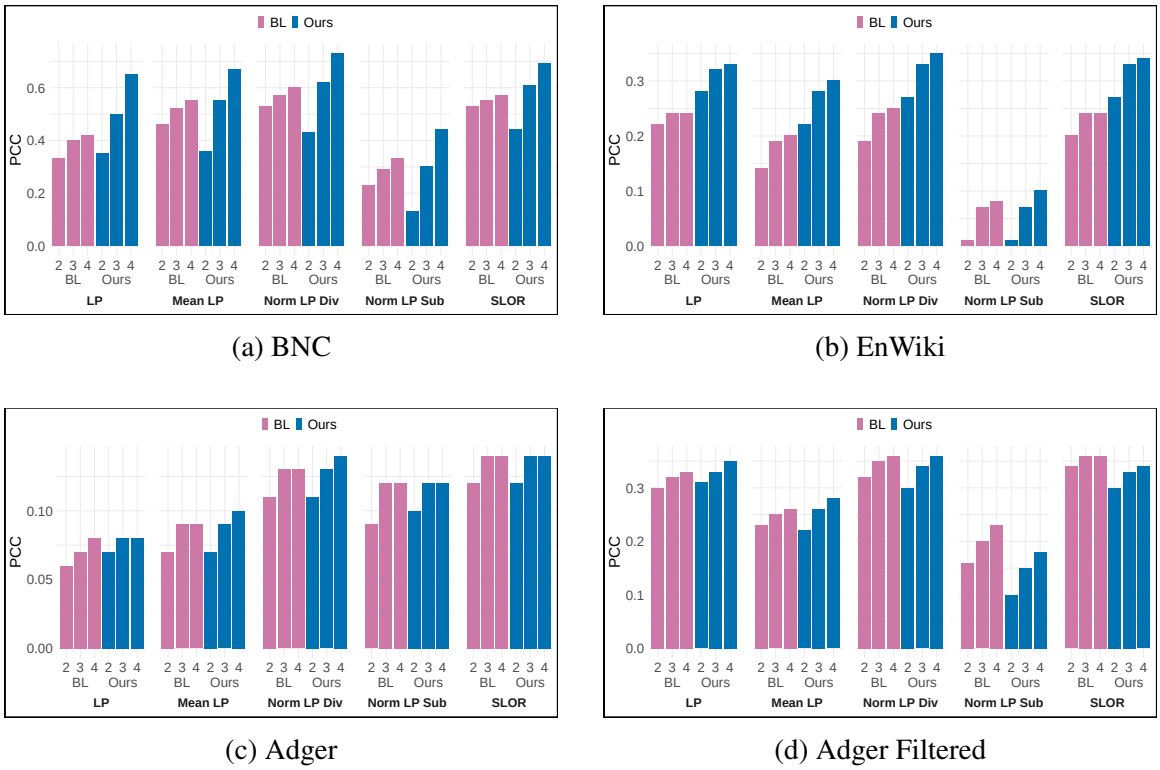
$$\rho_{x,y} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} \tag{3.9}$$

$$\mathrm{Cov}(x,y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})\,(y_i - \bar{y}) \tag{3.10}$$

3.6 Main Results

Figure 3.4 presents the performance of the model across five probability-based evaluation metrics (§3.2.1) for four different test datasets: BNC (a), EnWiki (b), Adger (c), and Adger Filtered (d). The x-axis lists the evaluation metrics, while the y-axis depicts the PCC, which measures the degree of alignment between the model's predictions and human acceptability ratings. A higher PCC suggests a stronger positive correlation, indicating better alignment with human judgments.

In our analysis, we compare two approaches for each dataset: the Baseline (BL) and



(a) BNC

(b) EnWiki

(c) Adger

(d) Adger Filtered

Figure 3.4: Comparison of Baseline and Ours (RNE) across four test datasets

our proposed method, referred to as Ours (RNE method discussed in Section §3.4). The results are presented for three types of language models: 2-gram, 3-gram, and 4-gram. In the accompanying bar graphs, pink bars indicate the performance of the Baseline method, while blue bars represent the outcomes of the RNE method.

The results consistently show that language models trained using the RNE approach either match or surpass the Baseline method in correlating with human judgments across almost all metrics. The most significant improvements are observed for the metrics Norm LP Div and SLOR. The RNE method achieves higher PCC scores, demonstrating more substantial alignment with ground truth acceptability ratings than the Baseline method. While some dataset-specific differences exist, the overall trend confirms the superior performance of the RNE method across all datasets. This suggests that existing metrics, which adjust for factors like lexical frequency and sentence length, do not fully resolve these issues and

underscore the effectiveness of RNE in enhancing the predictive accuracy of probability-based acceptability metrics.

3.7 Analysis

3.7.1 Impact Across Individual Datasets

In the BNC dataset, our model consistently outperforms the BL across most metrics across all configurations—2, 3, and 4-gram language models. The most notable improvements is seen in Norm LP Div and SLOR, with Norm LP Div showing a particularly strong correlation with human acceptability ratings, achieving a peak PCC above 0.7. Similarly, on the EnWiki dataset, our approach demonstrates superior performance, especially in Norm LP Div and SLOR, where the PCC values are significantly higher than those of the Baseline. However, the performance gap is smaller compared to the BNC results.

For the Adger dataset, the overall PCC values are lower than those of BNC and EnWiki, with both models showing diminished performance. Despite this, our model maintains an edge over the Baseline, particularly in the Norm LP Div metric. In the Adger Filtered dataset, our model surpasses the Baseline in all metrics except Norm LP Sub and SLOR, with the most significant improvements observed in LP and Mean LP.

3.7.2 Impact of Named Entity Type

Figure 3.5 illustrates the distribution of named entity types across four test datasets: BNC, EnWiki, Adger, and Adger Filtered. As previously discussed, the x-axis represents 18 different named entity types (§3.5.4). At the same time, the y-axis, presented on a logarithmic scale, shows the percentage of each named entity type within each dataset. This scaling highlights the wide variation in named entity occurrences across the datasets. For instance, the PERSON category is overwhelmingly dominant in the Adger and Adger Filtered datasets, accounting for around 80% of all named entities, compared to only about 20% in BNC and EnWiki. In contrast, categories such as QUANTITY, EVENT, and WORK OF ART are entirely absent in Adger and Adger Filtered, underscoring the diverse nature of these datasets.

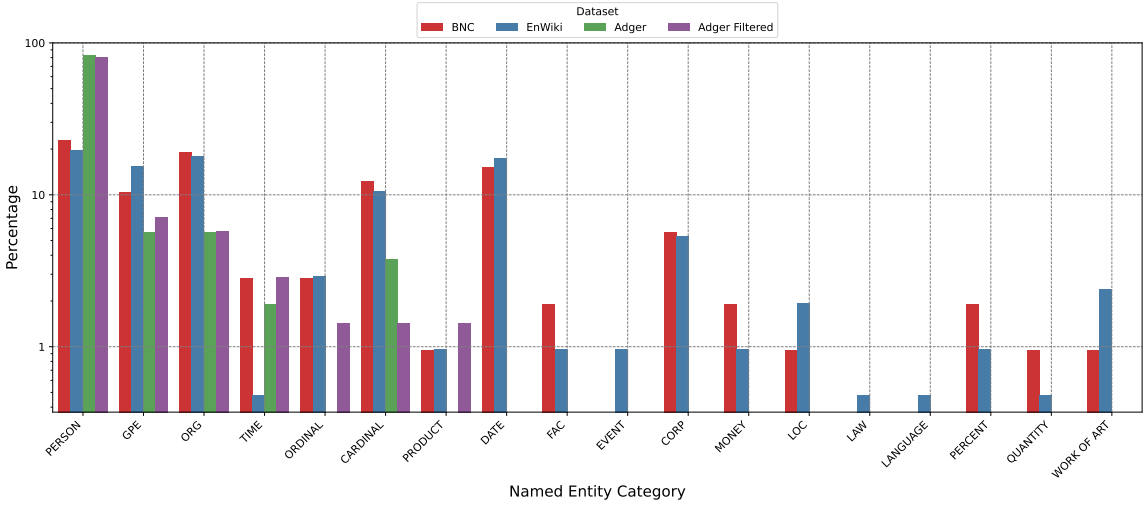


Figure 3.5: Percentage of 18 named entity types per sentences across four test datasets

These disparities in named entity distribution reflect the differing purposes behind the creation of each dataset. BNC and EnWiki consist of naturally occurring human text. In contrast, Adger and Adger Filtered are drawn from linguistic journals and syntax textbooks, prioritizing specific linguistic examples and often using minimalistic language. This distinction likely explains the grouping of BNC and EnWiki together and Adger and Adger Filtered as separate clusters. The differences in named entity types across the datasets provide an early indication of the performance variation observed, with more significant improvements seen in BNC and EnWiki. At the same time, Adger and Adger Filtered show less pronounced gains.

3.7.3 Impact of Named Entity Count

In the previous section, we examined the distribution of 18 named entity types across four datasets. Here, we focus on the differences in named entity count and their potential effect on the language model's performance in acceptability evaluation. As shown in Table 3.3, the average number of named entities per sentence varies notably across the test datasets. For example, BNC and EnWiki contain 1.07 and 2.07 named entities per sentence, respectively, while Adger and Adger Filtered feature only 0.53 and 0.68 named entities per

sentence. This variation partly stems from differences in sentence length, with BNC and EnWiki averaging around 17 words per sentence, compared to the shorter 7-word sentences in Adger and Adger Filtered.

The observed disparity reveals that sentences with a higher concentration of named entities are more aligned with human acceptability judgments. This pattern suggests that more named entities in a sentence provide a more abstract linguistic structure, minimizing the influence of specific word choices and lexical frequency on how acceptable the sentence appears. This conclusion is consistent with our earlier findings (§3.7.1), where we noted more significant performance improvements for the BNC and EnWiki datasets compared to Adger and Adger Filtered datasets.

3.8 Qualitative Analysis

3.8.1 Expectation from SLOR

To demonstrate the effectiveness of our proposed method, we use the SLOR metric, as defined in Eq. 3.8, as an example. Let's first consider the expected behavior of SLOR for an acceptable sentence and how this desired behavior changes for an unacceptable one. These characteristics are visually summarized in **Fig. 3.6**, where green upward arrows denote a higher value is desirable, and red downward arrows indicate a preference for lower values.

Since probability-based metrics measure acceptability rating y, higher values of any of these metrics suggest greater model-assigned acceptability to a sentence. In essence, for an acceptable sentence, we aim for a high SLOR, whereas a low SLOR is expected for an unacceptable sentence. A high SLOR occurs when the numerator is large, and the denominator is relatively small, and conversely, a low SLOR results from a small numerator and large denominator. Using qualitative examples, we further demonstrate that when a sentence undergoes preprocessing using RNE, these behaviors become even more pronounced.

3.8.2 Example Sentences

To evaluate the effectiveness of our proposed methodology, we conducted a detailed analysis of a sentence from the EnWiki test dataset: S_{1679}. Sentence S_{1679} reads, "A native

(a) Acceptable (b) Unacceptable

Figure 3.6: Expectation from SLOR for acceptable and unacceptable sentences

of Rayville in Richland Parish in northeastern Louisiana, McConnell had seven children". **Table 3.4** provide a comparative analysis of the acceptability rating between the Baseline and Ours.

The table contains five columns: 'Preprocessing', 'Evaluator', 'Sentence', 'Absolute Acceptability Rating', and 'Normalized Acceptability Rating'. These represent the preprocessing method used, the evaluator assessing the sentence, the sentence under evaluation, its absolute acceptability score, and the normalized acceptability score calculated using z-scores for relative comparison. 'Normalized Acceptability Rating' is computed by applying the z-score, which adjusts the absolute ratings relative to the dataset's mean and standard deviation. The table contains three rows, corresponding to: no preprocessing + human evaluator, baseline preprocessing + 4-gram language model, and RNE preprocessing + 4-gram LM. In cases where words in the sentence are OOV, they are represented by the special UNK token. Notice in case of RNE, named entities are replaced with their corresponding entity types.

The probability distributions assigned by the 4-gram language model for each word in the sentence are illustrated in **Fig. 3.7**. The table in the figure include columns: 'Word Index', 'Word', 'Quadgram Probability', 'Unigram Probability', and the difference between the 'Quadgram' and 'Unigram' log probabilities, comparing the Baseline and proposed RNE preprocessing methods. The Quadgram probability and Unigram probability are calculated based as we previously described in Eq. 3.2 and Eq. 3.3 respectively. For clarity, named entities are highlighted—red for the Baseline method and green for the RNE method—facilitating a visual comparison between the two approaches.

For sentence S_{1679}: "A native of Rayville in Richland Parish in northeastern Louisiana,

Preprocessing	Evaluator	Sentence S	Absolute Acceptability Rating y	Normalized Acceptability Rating z-score(y)
-	Human	A native of Rayville in Richland Parish in northeastern Louisiana, McConnell had seven children.	3.76	1.05
Baseline	4-Gram LM	a native of UNK in UNK parish in northeastern louisiana , mcconnell had seven children .	0.61	-0.17
RNE	4-Gram LM	a native of GPE in GPE in northeastern GPE , GPE had CARDINAL children .	1.06	0.96

Table 3.4: Comparison of acceptability ratings for Baseline and RNE for sample a sentence from EnWiki

McConnell had seven children" when preproocessing with Baseline lead to two UNK tokens corresponding to words 'Rayville' and 'Richland' and corresponding transformed sentence is shown in row 2 of Table 3.4. On the other hand the same sentence consisted of five named entities: 'Rayville', 'Richland', 'Louisiana', 'McConnell', and 'seven' which were replaced with corresponding named entities types: 'GPE', 'GPE', 'GPE', 'GPE', and 'CARDINAL', the transformed sentence is shown in row 3 of Table 3.4.

Expected Outcome

The sentence S_{1679} is completely acceptable, and we expect a language model to assign it a high acceptability score, reflected in a higher SLOR value. As illustrated in **Fig 3.6a**, this can be achieved by maximizing the difference between quadgram and unigram probabilities (numerator) while minimizing the sentence length (denominator). Next, we examine how the two preprocessing methods, Baseline and RNE, affect both the SLOR score and the differences across individual tokens.

Observed Outcome

Notice in Fig 3.7 for words replaced by corresponding named entity types, the RNE method systematically increases the difference between the sentence probability and unigram probability $\log(p_m) - \log(p_u)$, which forms the numerator in the SLOR formulation. While the sum of these differences is 12.39 in the Baseline method, RNE yields a higher sum of 20.22, as shown in the last row of the table. The results in Table 3.4 demonstrate that the normalized acceptability rating given by humans for the sentence is 1.05. When the

	Baseline				Replace Named Entity			
Word Index	**Word**	**Quadgram probability** $\log(p_m)$	**Unigram probability** $\log(p_u)$	**Difference** $\log(p_m) - \log(p_u)$	**Word**	**Quadgram probability** $\log(p_m)$	**Unigram probability** $\log(p_u)$	**Difference** $\log(p_m) - \log(p_u)$
w_1	a	-1.68	-2.61	0.93	a	-1.72	-2.63	0.91
w_2	native	-3.64	-4.37	0.74	native	-3.61	-4.36	0.75
w_3	of	-0.21	-2.33	2.12	of	-0.09	-2.37	2.28
w_4	UNK	-1.45	-2.79	1.34	GPE	-0.48	-2.94	2.46
w_5	in	-1.59	-2.19	0.60	in	-2.15	-2.26	0.11
w_6	UNK	-1.90	-2.79	0.89	GPE	-0.64	-2.94	2.30
w_7	parish	-3.49	-4.29	0.80				
w_8	in	-2.04	-2.19	0.15	in	-150	-2.26	0.76
w_9	northeastern	-6.37	-5.77	-0.60	northeastern	-7.77	-5.68	-2.09
w_{10}	louisiana	-5.70	-5.35	-0.35	GPE	-0.12	-2.94	2.82
w_{11}	,	-0.84	-1.91	1.07	,	-0.52	-2.04	1.52
w_{12}	mcconnell	-7.47	-5.73	-1.74	GPE	-0.92	-2.94	2.02
w_{13}	had	-2.84	-2.57	-0.27	had	-2.72	-2.72	0.00
w_{14}	seven	-3.65	-3.93	0.28	CARDINAL	-1.42	-2.72	1.30
w_{15}	children	-0.93	-3.52	2.59	children	-2.31	-3.57	1.26
w_{16}	.	-0.88	-1.98	1.10	.	-0.86	-2.09	1.23
w_{17}	<\s>	-0.01	-2.76	2.75	</s>	-0.01	-2.60	2.59
Total				**12.39**				**20.22**

Figure 3.7: Detailed probability analysis for Baseline vs Ours (RNE) on sample sentence from EnWiki

language model processes the sentence using the proposed RNE method, it produces a normalized rating of 0.96, which is much closer to the human rating compared to the -0.17 rating generated by the LM when the sentence is preprocessed using the Baseline method.

In summary (§3.8.1), applying RNE to an acceptable sentence increases the difference between the sentence and unigram probabilities, thereby raising the overall SLOR acceptability rating.

3.9 Related Work

Brown et al. (1992) proposed a statistical algorithm for grouping words based on their co-occurrence frequencies, leading to the development of class-based n-gram language models (LMs). While our approach can be viewed as a specialized case of class-based n-gram LMs, it diverges in two key aspects. First, our method addresses the question of which phrase in the input sentence should be replaced. Second, we determine what the selected phrase should be replaced with, providing a distinct contribution compared to the traditional class-based n-gram LMs.

In their work inspired by centering theory, Grosz et al. (1995); Lapata and Barzilay (2005) explored how local entity transitions affect discourse coherence by representing texts through a two-dimensional entity grid. This approach uses entity transitions as features from the input text to identify patterns in coherent discourse at the intra-sentence level. Our approach diverges from theirs in two significant ways. First, we transform the input sentence by replacing named entities with their corresponding categories, which are then fed into a language model for further analysis. Second, we focus on evaluating the acceptability of a single sentence rather than multi-sentence coherence.

Wan et al. (2005) introduced the idea of using grammatical judgments from parsers to evaluate sentence acceptability, suggesting that a parser trained on a well-constructed corpus would identify ungrammatical sentences as unacceptable. Mutton et al. (2007) later expanded on this by demonstrating that machine learning models trained on parser outputs align with human judgments of acceptability. Our approach departs from theirs by employing a different input for assessing acceptability. While previous studies relied on text parser outputs, we use sentence likelihood to determine acceptability.

Pitler and Nenkova (2008); Vadlapudi and Katragadda (2010) used surface-level features such as lexical, syntactic, and discourse elements to model human judgments of text acceptability. Their findings showed that these features helped improve alignment with human assessments. In contrast, our approach does not involve feature extraction. Instead, we focus on transforming the text directly by replacing named entities.

To address the confounding effects of lexical frequency and sentence length, which limit the effectiveness of sentence likelihood as an acceptability metric, several probability-based metrics, such as SLOR and WP-SLOR, have been proposed over the years (Pauls and Klein, 2012; Lau et al., 2016; Kann et al., 2018). While these metrics reduce the influence of lexical frequency and sentence length, they do not fully eliminate their impact. Building on these previous approaches, our research introduces a novel data transformation method that further enhances the correlation between these probability-based metrics and human judgments of sentence acceptability.

Recent advances in transformer-based model training have introduced objectives like Masked Language Modeling (MLM) (Devlin et al., 2019) and Entity-Level Masking (ELM) (Sun et al., 2020a). While both MLM and ELM share similarities with our proposed method, RNE, key distinctions set these methods apart. First, while MLM and ELM are training objectives, RNE is instead a data preprocessing step, not designed to directly optimize model parameters. Second, RNE incorporates a more informed masking approach compared to MLM and ELM.

MLM has shown effectiveness in enhancing language comprehension but has a notable limitation: it masks words independently, often leading to partial masking of multi-word entities. To overcome this, Sun et al. (2020a) proposed ELM, which masks entire named entities, including multi-word expressions, improving model performance on tasks like entity recognition and relation extraction.

Building on MLM and ELM, our method, RNE, further enhances entity context by not only masking entities but also replacing them with corresponding entity types (e.g., PERSON, GPE, DATE, ORG) rather than using a generic [MASK] token. This approach offers entity-specific context that supports more nuanced understanding of entity relationships in text. For instance, given the sentence "Albert Einstein moved to Germany in 1914 to work at the Prussian Academy of Sciences," MLM might output "MASK Einstein moved

to Germany in 1914 to work at the MASK Academy of Sciences," while ELM could yield "MASK moved to MASK in MASK to work at the MASK." In contrast, RNE would output "PERSON moved to GPE in DATE to work at ORG," showing how RNE preserves entity type information, which can improve context-sensitive comprehension.

Over the years, several approaches have been developed to assess acceptability as a gradient property. However, to our knowledge, this study is the first to explore the use of coarse sentence representations by replacing named entities in the input, with the aim of enhancing the correlation between probability-based metrics and human judgments of acceptability. While our experiments are conducted on English language datasets, we believe that the methodology we propose has wider applicability and is not tied to specific language models or acceptability evaluation metrics.

3.10 Conclusion

Over the years, various probability-based metrics have been developed to assess acceptability ratings as a gradient property without relying on reference texts. These metrics modify the likelihood of a sentence based on the probabilities assigned to individual tokens, aiming to reduce the confounding effects of factors like lexical frequency and sentence length. While these adjustments improve the evaluation of acceptability, they do not fully eliminate the influence of lexical frequency, particularly for OOV words, as we demonstrate through examples.

To address these limitations, we draw inspiration from how humans perform acceptability evaluations in practice and demonstrate that the original sentence may not always provide the most optimal representation for language model training. As a solution, we propose the RNE strategy, which generates a more effective, coarse-grained representation by replacing named entities with their corresponding entity types. This transformation enhances the robustness of acceptability evaluations by addressing issues present in the original sentence representations used during training.

Our experimental results provide strong empirical support for the effectiveness of the RNE strategy. Language models trained on RNE-processed data showed improved performance across multiple probability-based metrics and datasets in acceptability evaluation

tasks. In a qualitative analysis, we establish expectations for probability-based metrics like SLOR in the context of acceptability evaluation and demonstrate how RNE helps these metrics better meet those expectations.

In addition, we observed early indications that another dataset characteristic, specifically the length of input sentences, may also influence the language model's ability to evaluate acceptability. This insight forms the foundation for the next phase of our study, which will be discussed in the following chapter.

Chapter 4

Impact of Sequence Length

In the previous chapter, we explored the effect of lexical frequency on the language model's ability to evaluate sentence acceptability. One key takeaway from those findings was that sentence length within a dataset might also influence acceptability judgments. This observation prompted further investigation into the role of sentence length distribution and its impact on acceptability evaluation, which is the focus of this chapter.

As highlighted earlier (§2.3.2), the scientific community has an ongoing debate about whether acceptability should be viewed as a binary or gradient property. Most recent evaluation datasets have leaned towards treating acceptability as binary, and these datasets tend to be larger compared to those adopting a gradient perspective. In light of this difference, the current chapter examines how sentence length affects acceptability judgments when framed as a binary property.

This chapter begins by formally defining the problem statement and presenting examples of sentences with binary acceptability judgments. Following this, we review relevant literature on the influence of sentence length on various NLP tasks. Ideally, evaluation datasets should mirror the data's characteristics, which models are likely to encounter in real-world applications. For sentence acceptability assessments, datasets should closely resemble naturally occurring, human-authored text. However, our analysis reveals that commonly-used datasets for this task often diverge significantly from authentic human text, especially in their sentence length distribution.

To address this bias in sentence length found in commonly-used datasets, we introduce

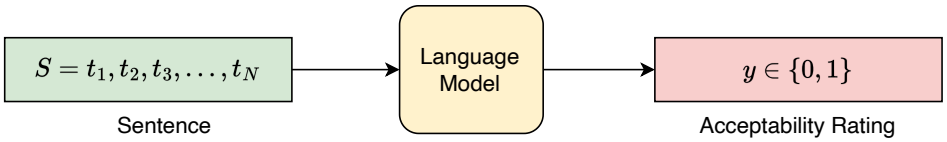


Figure 4.1: Problem statement for evaluating binary acceptability rating

seven new datasets—six adapted from existing sources and one novel. We then describe our experimental setup, which explores how sentence length distribution impacts the performance of language models in acceptability evaluations. We assess whether language models show biases or preferences for particular sentence types (shorter or longer) within these datasets through a series of tests. Our findings reveal that shorter sentences are overrepresented in most existing datasets, leading to inflated performance estimates for language models on acceptability tasks. This overemphasis on shorter sentences distorts the models' true performance when confronted with a more diverse range of sentence lengths. Finally, we recommend improving evaluation datasets to assess language models' abilities on human-like text more accurately.

4.1 Problem Statement

Figure 4.1 illustrates the problem of acceptability evaluation framed as a binary classification task. Given an input sentence S, composed of N tokens $\{t_1, t_2, \ldots, t_N\}$, the goal of the language model is to assign an acceptability label $y \in \{0, 1\}$ to the sentence. Here, $y = 0$ denotes an unacceptable sentence, while $y = 1$ indicates an acceptable one.

Table 4.1 provides examples of both an acceptable and an unacceptable sentence. The first example, S_1, "With what did the baby eat the food?", is grammatically correct and acceptable as it adheres to standard English grammar and structure. The phrase "with what" appropriately inquires about the instrument or tool used for an action, and the sentence maintains a clear subject-verb-object structure.

In contrast, the second sentence, S_2, "The authorities blamed Greenpeace with the bombing", is considered unacceptable due to the incorrect use of the preposition 'with'. In

ID	Sentence	Acceptability
S_1	With what did the baby eat the food?	1
S_2	The authorities blamed Greenpeace with the bombing.	0

Table 4.1: Sample sentences for binary acceptability evaluation

standard English, 'blame' typically pairs with the preposition 'for' (e.g., "blamed Greenpeace for the bombing"), as 'for' indicates the reason or cause of the blame. The Incorrect use of 'with' disrupts this clarity, making the sentence confusing and unnatural.

4.2 Motivation

Moving forward in this study, we will refer to "**human-written corpora**" to collectively refer to corpora like Books (Yukun et al., 2015), Wikipedia (Foundation, 2023a), and CNN-DM (Hermann et al., 2015), which consist of naturally occurring text that reflects real-world language use, with diverse sentence lengths and complexity. And we will collectively call the widely used datasets in the field of NLP for acceptability evaluation task, i.e., CoLA (Warstadt and Bowman, 2020) and BLiMP (Warstadt et al., 2020) as "**commonly-used datasets**". Before further investigating the datasets mentioned above, we will define each.

4.2.1 Human-Written Corpora

Naturally written text, or human-written corpora, refers to text produced organically by humans in real-world contexts without artificial constraints or prompts. Human-written corpora are collections of texts authored by humans, not generated by machines or algorithms. This type of text typically reflects the full range of linguistic diversity seen in everyday communication, such as variations in sentence length, complexity, and style. These corpora are valuable for analyzing and understanding human language due to their natural and varied language use. They exhibit complex syntactic structures and diverse grammatical patterns, showcasing the richness of human communication. Our study incorporates corpora from a broad spectrum of text types, including books, wikipedia articles, and news articles, to ensure a comprehensive representation of written language.

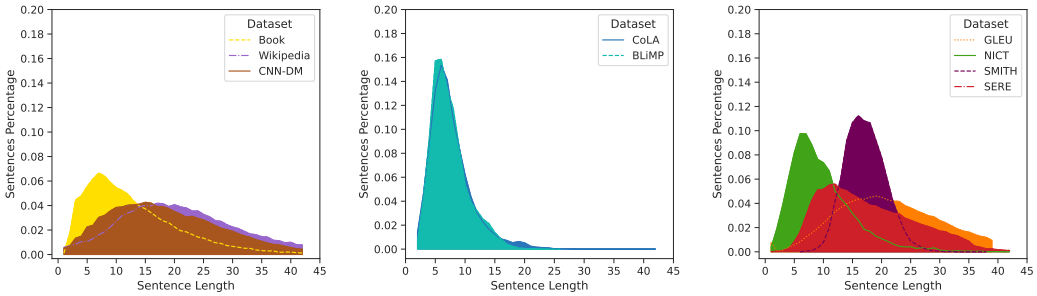
- **Book**: BookCorpus (Yukun et al., 2015) consists of over 11,000 books from diverse genres and topics. These books are sourced from the public domain and cover a wide range of literary styles, including fiction, non-fiction, and various sub-genres. The dataset is designed to provide a rich source of text for natural language processing tasks, featuring long passages and diverse sentence structures that reflect different writing styles and narrative techniques. The content is aimed at supporting research and development in language modeling and other text-based applications, offering a comprehensive resource for studying language use in extended and varied contexts.

- **CNN-DM**: CNN-DailyMail (Hermann et al., 2015) dataset is a collection of news articles and their corresponding summaries, sourced from CNN and the Daily Mail. It contains over 300,000 news articles paired with human-written summaries covering a wide range of topics and events. The dataset is designed for text summarization tasks, including extractive and abstractive methods. Each entry includes the full text of the article and its summary, making it a valuable resource for developing and evaluating models that generate concise and coherent summaries of news content.

- **Wikipedia**: Wikipedia (Foundation, 2023a) dataset comprises text extracted from Wikipedia articles, offering a comprehensive snapshot of a wide range of topics. While the dataset includes articles in various languages, for our study, we focused exclusively on English and excluded content in other languages. The texts are formatted to include article content while stripping out non-essential elements like headers and footnotes. It provides extensive coverage of comprehensive knowledge, making it valuable for various natural language processing tasks, including text classification, information retrieval, and language modeling.

4.2.2 Commonly-Used Datasets

Commonly-used datasets refer to the most commonly used in literature when assessing the acceptability evaluation capabilities of language models. Our study focuses on two such prominent datasets: CoLA and BLiMP. Below, we provide detailed information on each dataset.



(a) Human-Written Corpora (b) Commonly-Used Datasets (c) Proposed Datasets

Figure 4.2: Sentence length distribution for (a) human-written corpora, (b) commonly-used datasets, and (c) proposed datasets

- **CoLA**: Corpus of Linguistic Acceptability (Warstadt et al., 2019) is a set of 10,657 English sentences labeled as acceptable or unacceptable from published linguistics literature. A collection of sentences from the linguistics literature with expert acceptability labels. CoLA aims to represent a wide variety of phenomena of interest in theoretical linguistics.

- **BLiMP**: The Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020) offers a rigorous evaluation dataset for assessing the proficiency of language models in comprehending diverse grammatical phenomena of English. The dataset is meticulously crafted using expert-designed grammar rules and undergoes automatic template-based generation. The assigned acceptability labels show high accuracy, with an average human agreement rate of 96.4%

Figures 4.2a and **4.2b** present the sentence length distributions for human-written corpora and commonly-used datasets. In both subfigures, the x-axis represents sentence length, while the y-axis shows the percentage of sentences of a given length within the dataset. A clear difference emerges when comparing the two: the sentence lengths in human-written corpora are more evenly distributed. At the same time, those in the commonly-used datasets are heavily skewed towards shorter sentences. This highlights a notable contrast in the distribution patterns between the two types of datasets.

4.3 Sentence Length Bias in Commonly-Used Datasets

Evaluating text acceptability plays a crucial role in assessing the quality of machine-generated text (§1.5). For such evaluations to be meaningful, the generated text must closely resemble human-written language, reflecting natural variations in sentence length and linguistic patterns. This helps gauge a language model's ability to differentiate between acceptable and unacceptable text.

As highlighted earlier, there is a noticeable discrepancy in sentence length distribution between human-written corpora and commonly-used datasets, as illustrated in Fig. 4.2. The clear contrast in these distributions raises the question: Why does this disparity exist? An explanation lies in the differing objectives of these corpora and datasets.

Human-written corpora are general-purpose, naturally occurring corpus where sentence length varies based on the content being conveyed. Longer sentences in human-written corpora often reflect the need to provide detailed information, combine multiple ideas, or include clarifying clauses, which are common in human-authored text aiming to explain complex topics. The variety of sentence lengths in these corpora mirrors real-world language use, where short and long sentences are naturally mixed depending on the context and communicative intent.

In contrast, commonly-used datasets like CoLA and BLiMP primarily contain shorter sentences because they are designed to test specific linguistic phenomena in a controlled manner. Short sentences allow for more precise and more focused evaluation of grammatical structures, minimizing complexity and ensuring that the acceptability judgment hinges on particular syntactic or semantic features. These datasets often isolate grammatical contrasts, making it easier for models to identify errors or distinctions without being influenced by extraneous factors that longer sentences might introduce.

Consequently, the sentence length distribution in commonly-used datasets is often skewed toward shorter sentences, with an average length of around 7 words. In contrast, human-written corpora exhibit a broader range, with average sentence lengths ranging from 13 to 21 words. This difference reflects the distinct purposes and design of these corpora and datasets.

4.4 Proposed Datasets

In the previous section, we highlighted the significant differences in sentence length distribution between human-written corpora and commonly-used datasets and explored potential reasons behind these variations. Given that our primary goal is to assess the language model's ability to distinguish between acceptable and unacceptable text and considering the bias towards shorter sentences in commonly-used datasets, a logical question arises. If these datasets are not representative of human-written corpora, and we aim to evaluate the acceptability of text similar to human writing, why not simply use the human-written corpora for this task?

The issue with this approach is that human-written corpora only contain acceptable sentences. There is no access to a corresponding set of unacceptable sentences essential for assessing a model's ability to evaluate acceptability. To examine the model's capabilities in this regard, we need both acceptable and unacceptable examples, making human-written corpora unsuitable for this specific evaluation task.

To address this limitation, we propose the creation of datasets that contain both acceptable and unacceptable sentences while maintaining the naturalness and sentence length distribution typical of human-written text. In this study, we introduce seven new datasets (six derived and one novel) to provide a more reliable framework for evaluating the ability of language models to assess sentence acceptability. These datasets, referred to as the '**proposed datasets**', offer a balanced representation of acceptable and unacceptable sentences, overcoming the shortcomings of traditional human-written corpora and commonly-used datasets. Unlike commonly-used datasets, these are specifically tailored for acceptability evaluation, yet they retain the sentence length characteristics of natural text. Figure 4.2 shows the sentence length distribution in the proposed datasets, comparing them with both human-written corpora and commonly-used datasets. The figure reveals that the sentence length distribution of the proposed datasets more closely resembles that of human-written corpora than the commonly-used datasets.

Table 4.2 presents a systematic comparison of human-written corpora, commonly-used datasets, and the proposed datasets. The table is structured with the following columns: 'Acceptable Source', 'Unacceptable Source', 'Acc. vs Una. Approx. Ratio', 'Sentence

Length Mean (SD)', 'Train', 'Dev', 'Test', 'Total', and 'Avg. KL Similarity Rank'.

The columns 'Acceptable Source' and 'Unacceptable Source' refer to the origin of acceptable and unacceptable sentences. The 'Unacceptable Source' column with the postfix (Syn) indicates synthetically generated, non-human sentences. NA denotes information that is unavailable in the relevant columns. The 'Acc. vs Una. Approx. Ratio' column shows the ratio between the percentage of acceptable and unacceptable sentences in each dataset. The 'Sentence Length Mean (SD)' column presents the average sentence length and its standard deviation for the sentences in the dataset. The 'Train', 'Dev', and 'Test' columns show the number of sentences in each dataset split (70%, 20%, 10%), followed by the total number of sentences across all splits. Finally, the 'Avg. KL Similarity Rank' column reflects the average Kullback–Leibler (KL) (Kullback and Leibler, 1951) divergence of the dataset compared to human-written corpora, i.e., Book, Wikipedia, and CNN-DM. For reference, we have provided sample sentences from both commonly-used datasets and our proposed datasets in the appendix (see **Table B.1**).

Below, we provide a concise overview of the collection and transformation processes to prepare each of the seven proposed datasets for the acceptability evaluation task.

- **NICT**: The National Institute of Information and Communications Technology Japanese Learner English Corpus (Izumi et al., 2004) is a learner corpus containing transcripts of 1,281 audio-recorded English oral proficiency interviews and standard speaking tests, totaling 1.2 million words and 300 hours. It includes annotations with 47 error tags for grammatical and lexical errors. We parsed these transcripts and formed minimal pairs of acceptable (error phrase replaced with correct phrase) and unacceptable sentences (original sentence with error phrase), resulting in a dataset of 15,164 sentences.

- **SMITH**: The Set of Modified Incomplete TecHnical paper sentences (Ito et al., 2019) comprises naturally written English sentences from published papers, considered grammatical and acceptable. To create unacceptable counterparts, the sentences were translated using Neural Machine Translation and then back-translated to English by native Japanese speakers. Nearly 94.8% of the draft sentences were evaluated as less fluent than their original versions, indicating a high quality of the annotation labels.

Dataset	Acceptable Source	Unacceptable Source	Acc. vs Una. Approx. Ratio	Sentence Length Mean (SD)	Train	Dev	Test	Total	Avg. KL Similarity Rank
Book	Text Books	NA	NA	13 (8)	NA	NA	NA	NA	NA
CNN-DM	CNN-DM Articles	NA	NA	18 (9)	NA	NA	NA	NA	NA
Wikipedia	Wikipedia Articles	NA	NA	21 (9)	NA	NA	NA	NA	NA
CoLA	Linguistics Books	Linguistics Books	70:30	8 (4)	8551	527	515	9593	8.4
BLiMP	Template (Syn)	Template (Syn)	50:50	7 (3)	70,000	20,000	10,000	100,000	9.5
NICT	Audio Transcription	Corrected Errors	50:50	10 (5)	10,614	3034	1516	15,164	7.0
SMITH	ACL Publications	Back Translation	60:40	17 (3)	12,599	3601	1799	17,999	9.8
GRAM	ACL Publications	Encoder-Decoder (Syn)	50:50	19 (7)	70,000	20,000	10,000	100,000	5.2
STYL	ACL Publications	Encoder-Decoder (Syn)	50:50	21 (8)	70,000	20,000	10,000	100,000	5.8
ENTA	ACL Publications	Encoder-Decoder (Syn)	50:50	16 (7)	70,000	20,000	10,000	100,000	6.2
SERE	ACL Publications	GRAM + STYL + ENTA	25:75	18 (8)	70,000	20,000	10,000	100,000	4.5
GLEU	Wikipedia Articles	Concatenate Sequences (Syn)	50:50	21 (8)	70,000	20,000	10,000	100,000	6.4

Table 4.2: Details of human-written corpora, commonly-used datasets, and proposed datasets

- **GRAM**: This dataset includes acceptable sentences from SMITH, with unacceptable sentences generated by an encoder-decoder model trained to introduce synthetic GRAMmatical errors (Ito et al., 2019). The model was trained using a dataset from GEC (clean to erroneous), including Lang-8, AESW (Alikaniotis and Raheja, 2019), and JFLEG datasets (Napoles et al., 2017).

- **STYL**: This dataset contains acceptable sentences from SMITH, with unacceptable sentences generated by an encoder-decoder model trained to create STYListically unnatural sentences in the academic domain via paraphrasing (STYLe) (Ito et al., 2019), using the ParaNMT-50M (Wieting and Gimpel, 2018) dataset.

- **ENTA**: This dataset comprises acceptable sentences from SMITH, with unacceptable sentences generated by an encoder-decoder model trained to simulate missing words for ENTAilment (Ito et al., 2019). The model used entailed sentence pairs extracted from the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets.

- **SERE**: To enhance the representation of diverse phenomena that render sentences unacceptable, we created the SEntence REvision dataset by integrating unacceptable sentences from GRAM, STYL, and ENTA.

- **GLEU**: Mutton et al. (Mutton et al., 2007) proposed a method to synthetically generate unacceptable sentences by concatenating sequences from different acceptable sentences. Using Wikipedia Featured articles (Foundation, 2023b), we extracted and cleaned acceptable sentences. We then applied the method from Mutton et al. (2007) to generate unacceptable sentences by concatenating sequences, selecting the sequence length as:

$$\text{Sequence Length} = \frac{\text{Desired Length of Unacceptable Sentence}}{2}$$

- **NewsRoom**: Grusky et al. (2018) provide human evaluations for 60 articles, 7 summarization systems, and 3 ratings per article. Each annotation includes a single rating (1-5) across four dimensions: coherence, fluency, informativeness, and relevance.

We used only the fluency rating, labeling sentences with ratings above three as acceptable and the rest as unacceptable. NewsRoom comprised 178 sentences and was used solely for evaluation, not fine-tuning.

4.5 Quantifying Distance Between Distributions

In the previous section, we discussed the issue of sentence length bias in commonly-used datasets (§4.3), primarily through a visual comparison of sentence length distributions. Figure 4.2 demonstrates that the proposed datasets align more closely with human-written corpora than the commonly used ones. However, it is crucial to quantify this resemblance to draw meaningful conclusions.

To evaluate the similarity between the sentence length distributions of the commonly used and proposed datasets compared to human-written corpora, we use KL divergence, as described in Eq. 4.1. In this equation, p represents the true or the target distribution, and q represents the approximated distributions, respectively. The goal is to calculate the distance between the distribution of q and p, while keeping p as the reference distribution. The term x_i refers to the i^{th} sentence length out of a total of N. In practice, a small smoothing factor is often applied to prevent issues related to missing data points, which can result in infinite values. Accordingly, we use a smoothing factor of $\epsilon = 10^{-5}$.

$$D_{KL}(p \,||\, q) = \sum_{i=1}^{N} p\left(x_i\right) \log\left(\frac{p\left(x_i\right)}{q\left(x_i\right)}\right) \tag{4.1}$$

Following the formulation in Eq. 4.1, we calculated the KL divergence between all dataset pairs, as presented in **Fig. 4.3**. The datasets are organized sequentially, starting with human-written corpora, followed by the proposed datasets, and then commonly-used datasets. The total number of datasets analyzed is 12, comprising 3 human-written corpora, 2 commonly-used datasets, and seven proposed datasets. This results in a 12x12 matrix (144 pairwise comparisons) in Figure 4.3. The KL divergence between a dataset p and all other datasets q is used to create a similarity rank, where lower KL divergence values indicate greater similarity, represented by a lower rank. Each cell in the matrix corresponds to the similarity rank between two datasets: the row represents the true distribution (p) and

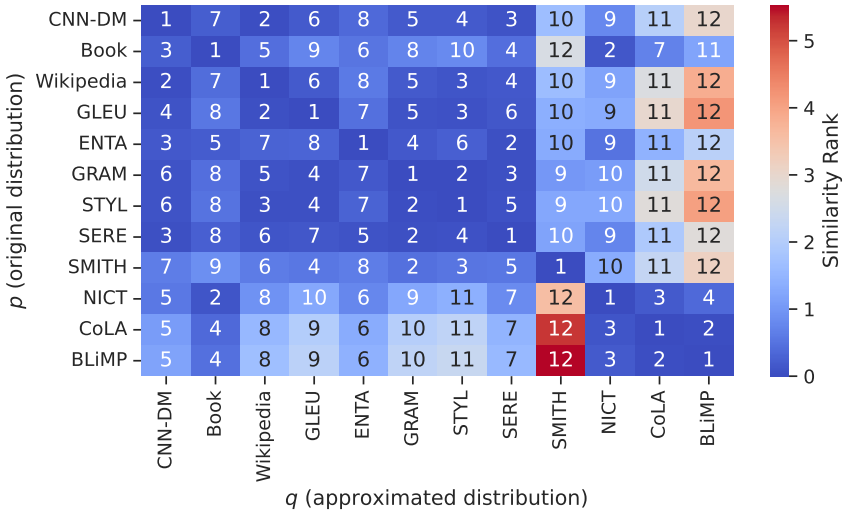


Figure 4.3: Similarity Rank based on KL distance between dataset pairs

the column represents the approximated distribution (q). As expected, the diagonal values are 1, indicating that each dataset is most similar to itself. The row values range from 1 to 12, representing the relative similarity rankings of q from the perspective of p.

A clear pattern emerges in Figure 4.3, consistent with our earlier observation of a bias towards shorter sentence lengths in commonly-used datasets. These datasets display higher similarity ranks compared to the lower ranks observed for the proposed datasets. Given the 2 commonly-used datasets and seven proposed datasets, and to focus the analysis, it is crucial to select one representative from each category. The KL divergence calculated in this section aids in making this informed selection.

To make these choices, we computed the average similarity rank of each dataset, both from the commonly-used and proposed categories, relative to the human-written corpora (Books, CNN-DM, and Wikipedia). The average similarity rank is presented in the last column of Table 4.2. Based on these values, we selected the SERE dataset as the representative for the proposed datasets, as it has the lowest average KL similarity rank of 4.5 (averaging ranks of 3, 4, and 4 across the human-written corpora). Similarly, CoLA was chosen as the representative for the commonly-used datasets, with an average similarity rank of 8.4 (averaging ranks of 7, 11, and 11).

In subsequent experiments, we will use SERE to generalize findings for the proposed datasets and CoLA for the commonly-used datasets. This selection is reasonable, as SERE's sentence length distribution is closer to that of the human-written corpora, and it captures a broad range of linguistic features, including sentence acceptability, fluency, and stylistic variations that impact sentence quality.

4.6 Experiments

4.6.1 Experiments Overview

The main objective of this research is to investigate how the distribution of sentence lengths within a dataset influences the performance of language models on the task of acceptability evaluation. To address this, we conducted six experiments, each designed to examine different aspect of language model behavior under varying conditions. The key research questions that shape these experiments are outlined below.

Due to hardware limitations, we had to limit the number of models to evaluate acorss all experiments. Although high-performing language models have been developed for commonly-used acceptability evaluation datasets, and one can know the best performing model on commonly-used datasets based on public leaderboard rankings. However since we introduced new datasets for acceptability evaluation, there is no benchmark available to determine which language model performs best across both existing and newly proposed datasets. This raises an important question: Is there a single model that consistently performs well on both widely-used datasets and the newly introduced ones (§4.7.1)?

Before exploring the impact of specific dataset characteristics, it is essential to first investigate potential interactions between dataset pairs. This raises the question: Are there shared characteristics between datasets that enable a language model to perform better than random chance on the test split of an unseen dataset (§4.7.2)? If such shared characteristics do exist, a logical follow-up is to assess whether sentence length distribution is one of these factors and to what degree it influences the language model's performance on acceptability evaluation tasks (§4.7.3).

Following the analysis of sentence length distribution within individual datasets, the

next question is: Does the language model exhibit any bias or preference towards sentences of certain lengths, such as shorter or longer ones, within the same dataset (Section 4.7.4)? To address this, we adjusted datasets to match the sentence length distribution of another dataset, a transformation that inevitably reduced their size. This leads to a crucial consideration: How does the number of training examples affect the language model's performance and learning capability (§4.7.5)?

Finally, we seek to determine whether the observed performance loss due to downsampling is a result of the increased complexity of grammatical structures in longer sentences or simply a byproduct of training on fewer, shorter sentences (§4.7.6).

4.6.2 Language Models

We evaluated a series of language models, including ERNIE (Sun et al., 2020b), BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and TDA-BERT (Cherniavskii et al., 2022), selected for their balance of accuracy and efficiency. At the time of our experiments, ERNIE 2.0 was the highest-performing model on the CoLA benchmark, as reported by the official GLUE leaderboard (Wang et al., 2023). For implementation, we used the standard versions of these models available through HuggingFace's library (Wolf et al., 2020), with PyTorch (Paszke et al., 2019) as the computational backend.

4.6.3 Hardware

We fine-tuned all language models on a cluster of four NVIDIA GeForce A6000 GPUs, each equipped with 48GB of GDDR6 memory. The optimization process utilized the AdamW optimizer, set with a learning rate of 2e-5 and a weight decay of 0.01. Fine-tuning was conducted over ten epochs, with early stopping employed to terminate training if performance declined across two consecutive evaluation steps.

4.6.4 Metrics

The Matthews Correlation Coefficient (MCC) (Matthews, 1975) and Accuracy are widely used metrics for assessing the performance of language models on acceptability evaluation tasks. These metrics are described in Eq. 4.2 and Eq. 4.3 respectively. In these equations, TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively. In this work, we primarily rely on MCC (Matthews, 1975) to facilitate comparison with previous studies (Warstadt et al., 2019; Warstadt and Bowman, 2020). MCC, a correlation coefficient ranging from -1 to +1, measures the relationship between observed and predicted binary classifications and is particularly useful when there is a class imbalance.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.2}$$

It is important to note that MCC is not suitable for evaluating performance on a single class (e.g., only acceptable sentences). In such cases, we use Accuracy as the evaluation metric.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

4.7 Results and Analysis

4.7.1 Preliminary Ranking of Language Models

To address the computational resource limitations, our initial experiment aimed to establish a relative ranking of the language models under consideration using a set of acceptability datasets. The objective was to identify a single model that consistently outperformed the others. For this experiment, we fine-tuned and evaluated the base (non-degenerative) versions of the language models three times on a train split of each dataset.

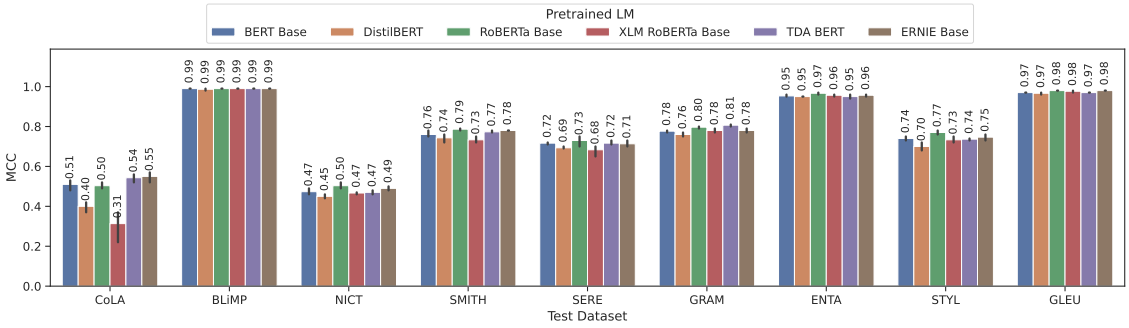


Figure 4.4: Performance of different pre-trained language models when fine-tuned and evaluated across train/test splits of a respective dataset

Results

Figure 4.4 presents the ranking results for the language models across individual datasets. The x-axis represents the test datasets, the y-axis shows the MCC scores, and the hue color indicates the different pre-trained language models. Among the models evaluated, RoBERTa, TDA-BERT, and ERNIE consistently achieved superior performance across all datasets.

In line with the official GLUE benchmark leaderboard, ERNIE demonstrated the highest performance on the CoLA dataset. We were able to replicate ERNIE 2.0's MCC score of 0.55 on CoLA. However, we did not test ERNIE 3.0 because its English version was not available on HuggingFace at the time of our experiments. We chose ERNIE for further experiments for two primary reasons. First, to address potential biases toward commonly-used datasets. Second, ERNIE outperformed other models on three of the nine acceptability datasets and matched RoBERTa's performance on the remaining six.

Analysis

Despite its skewed distribution towards shorter sentence lengths, CoLA appears systematically more challenging than the other datasets for our current choice of language models. On the other hand, all language models quickly solved BLiMP, ENTA, and GLEU, achieving almost perfect MCC scores. We suspect that the synthetic methods used to generate

these datasets, as discussed in detail in Section 4.7.3, may provide supplementary signals that contribute to the high performance of the models.

4.7.2 Controlling for Interactions Across Datasets

In this experiment, we explore whether certain shared characteristics between pairs of datasets can lead to better-than-chance performance of a language model on an unseen test set. To investigate this, we generate all possible pairwise combinations of training and testing splits from our dataset collection and evaluate ERNIE's performance for each pairing.

Following this, we conduct a binary classification analysis—acceptable vs. unacceptable. For each dataset, we fine-tune ERNIE on the training split and then evaluate its performance on acceptable and unacceptable sentences from the test split separately. This analysis helps determine whether language models display any inherent bias towards one of the two classes when fine-tuned on specific datasets.

Results

The results of the pairwise fine-tuning and testing experiments are illustrated in **Fig. 4.5**. In this figure, the x-axis represents the test datasets, while the y-axis shows the MCC. The hue color indicates the dataset used for fine-tuning. ERNIE performed best when fine-tuned and tested on the same dataset, highlighting the unique nature of each dataset. For instance, when fine-tuned on CoLA, ERNIE achieved an MCC of 0.55, outperforming models fine-tuned on other datasets. This result is expected, as the train and test data share similar characteristics.

Interestingly, despite the variation in datasets, ERNIE fine-tuned on one acceptability evaluation dataset can still capture generalizable features, resulting in performance better than random (MCC > 0) when tested on different datasets. For instance, fine-tuning on CoLA led to an MCC of 0.40 when tested on NICT, showing that the model retains some transferable knowledge.

Figure 4.6 presents the results on acceptable vs. unacceptable sentences of a test split from ERNIE when fine-tuned over a train split of the same dataset. It is evident that CoLA,

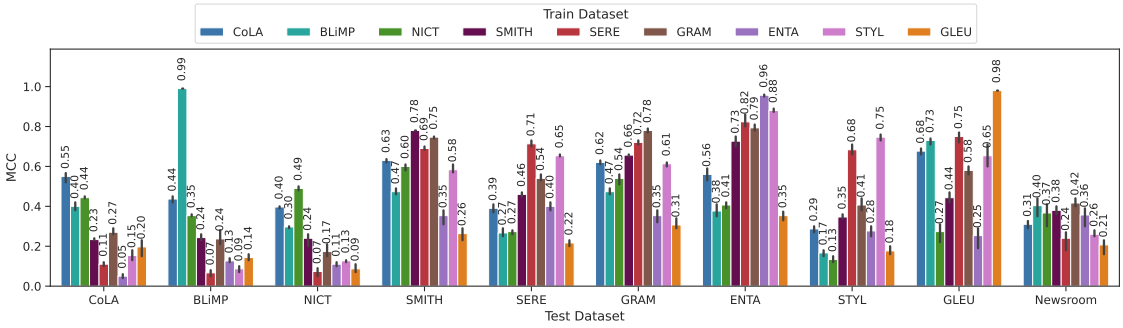


Figure 4.5: Performance of ERNIE when fine-tuned on train split of a dataset (hue in legend) and evaluated on test split of a dataset (x-axis)

NICT, GRAM, and STYL datasets all exhibit better performance on acceptable sentences compared to unacceptable ones. Notably, even in datasets like NICT, GRAM, and STYL, which have an equal distribution of acceptable and unacceptable sentences (50:50, see Table 4.2), the model consistently identifies acceptable sentences more accurately. This discrepancy is most pronounced in CoLA, where the performance gap between acceptable and unacceptable sentence classification is as high as 40 points.

In summary, while the model performs best when fine-tuned and tested on the same dataset, it also demonstrates a surprising level of generalization across different datasets. Moreover, across various datasets, the model consistently favors acceptable sentences over unacceptable ones, particularly in the CoLA dataset.

Discussion

In this control experiment, we see signs that sentence length distribution may influence ERNIE's performance. Specifically, ERNIE performs better when fine-tuned on a train set with a sentence length distribution similar to the test set. For example, CoLA, BLiMP, and NICT share similar sentence length distributions, as illustrated in Figures 4.2 and 4.3. As a result, when fine-tuned on CoLA, ERNIE demonstrates better performance on the test splits of CoLA, BLiMP, and NICT than when fine-tuned on other datasets. We observe a similar pattern with SERE, GRAM, STYL, and ENTA, which also have comparable sentence length distributions.

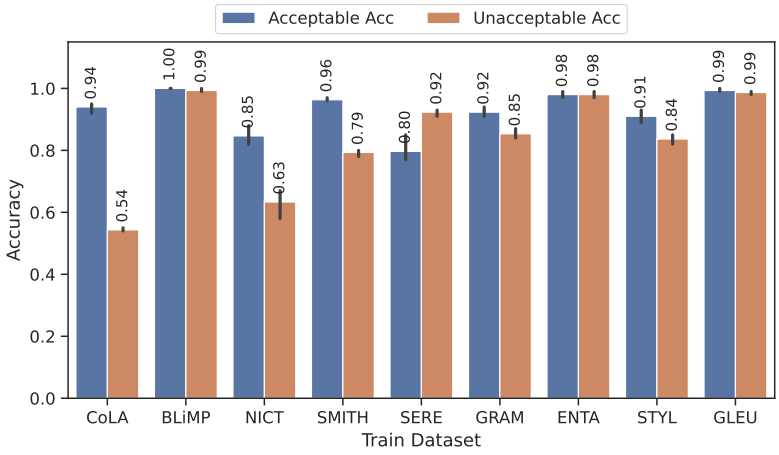


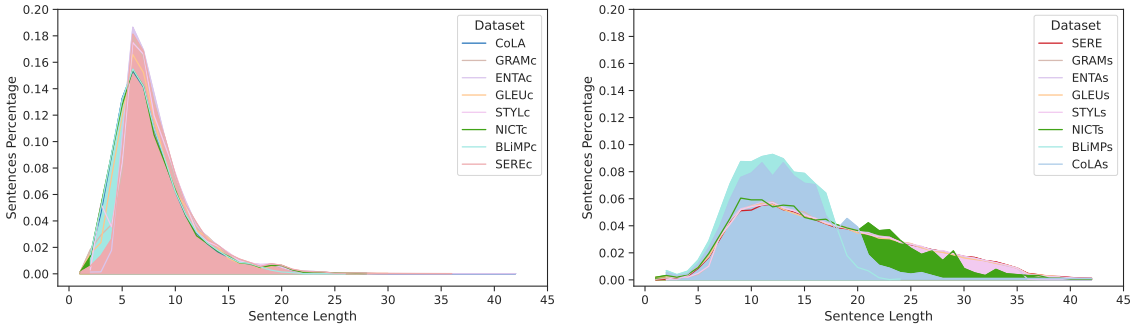
Figure 4.6: Performance across acceptable vs unacceptable sentences of a test split

It is worth noting that, as previously mentioned (§4.4), NewsRoom is a relatively small dataset (178 sentences) and was therefore used only for evaluation, not fine-tuning. Despite this, models fine-tuned on other datasets performed moderately well on NewsRoom, with GRAM and BLiMP fine-tuned ERNIE models achieving the highest performance, with an average MCC score of 0.40.

4.7.3 Does Sentence Length Distributions Introduce Bias

In a previous experiment, we observed that datasets often share common characteristics that can lead to above-chance performance when a model is fine-tuned on one dataset and tested on another. This raised an important question: can sentence length be one of these characteristics, potentially introducing bias in a language model's performance on acceptability evaluation tasks?

To investigate this, we applied three different transformations to the datasets. Before the transformation, the train, dev, and test sets were combined into a single dataset, which was then adjusted to align with specific target sentence length distributions before being split back into 70/20/10 train/dev/test sets. The first transformation adjusted the dataset to match the sentence length distribution of the **C**oLA dataset, with these versions labeled using the suffix '-c' (e.g., NICTc). The second transformation aligned the dataset with the sentence



(a) CoLA (b) SERE

Figure 4.7: Sentence length distribution of datasets when transformed to emulate (a) CoLA and (b) SERE

length distribution of the **S**ERE dataset, designated by the suffix '-s' (e.g., NICTs). Finally, we randomly sampled from both CoLA (representing commonly-used datasets) and SERE (representing proposed datasets), retaining their original sentence length distributions. These **D**ownsampled datasets, equal in sentence count to CoLAs and SEREc, are labeled with the suffix '-d' (e.g., CoLAd).

Both CoLAs and CoLAd contain 1,321 sentences from the original CoLA dataset. The difference between them lies in their sentence length distributions: CoLAs mimics the distribution of SERE, while CoLAd retains the original CoLA distribution. A similar distinction applies between the SEREc and SEREd datasets. To create these transformations, we used a distribution-based sampling algorithm. **Figures 4.7a** and **4.7b** illustrate the sentence length distributions in the train splits of the datasets transformed to resemble CoLA (-c) and SERE (-s), and can be compared to the original distributions in Figure 4.2. The sizes of the transformed datasets are detailed in **Table 4.3**. We assessed the performance of both the base and large versions of ERNIE on the modified datasets.

Results

Figure 4.8 illustrates ERNIE's performance (both base and large versions) on datasets adjusted to match the sentence length distributions of CoLA (-c) and SERE (-s). These results are compared against ERNIE's performance on the original datasets from section 4.7.2. A significant performance drop was observed for ERNIE Base on the SEREc dataset

Dataset	Train	Dev	Test	Total
BLiMPc	29,239	8355	4177	41,771
NICTc	6812	1947	973	9732
SEREc	6297	1801	899	8997
GRAMc	2360	675	337	3372
ENTAc	8164	2333	1166	11,663
STYLc	2053	588	293	2934
GLEUc	3474	994	496	4964
CoLAs	924	265	132	1321
BLiMPs	4342	1241	620	6203
NICTs	1538	441	219	2198
GRAMs	42,215	12,063	6030	60,308
ENTAs	48,909	13,974	6987	69,870
STYLs	35,871	10,250	5124	51,245
GLEUs	26,699	7629	3814	38,142
CoLAd	924	265	132	1321
SEREd	6297	1801	899	8997

Table 4.3: Size of train, dev, and test split for the transformed datasets

(SERE transformed to CoLA's distribution), where its MCC score decreased by 18 points, from 71 to 53. Similar declines were noted across other transformed datasets, with the sharpest drop on STYLc. ERNIE Base's MCC score fell by 18 points on GRAMc (from 78 to 60), 38 points on STYLc (from 75 to 37), 11 points on ENTAc (from 96 to 85), and 3 points on GLEUc (from 98 to 95). No substantial changes were observed for BLiMPc or NICTc.

When datasets were adjusted to SERE's distribution, ERNIE Base's performance also declined, most notably by 48 points on CoLAs (from 55 to 7). Less severe declines were recorded on other datasets: 23 points on NICTs (from 49 to 26), 5 points on GRAMs (from 78 to 73), 6 points on STYLs (from 75 to 69), 2 points on GLEUs (from 98 to 96), and 20 points on BLiMPs (from 99 to 79). ERNIE Base's performance on ENTAs remained stable. Interestingly, ERNIE Base showed an 11-point improvement on CoLAd compared to CoLAs (from 7 to 18), and a 10-point increase on SEREd compared to SEREc (from 53 to 63).

Discussion

This experiment evaluates ERNIE's performance before and after modifying a dataset to match a specific sentence length distribution. The baseline performance was obtained using

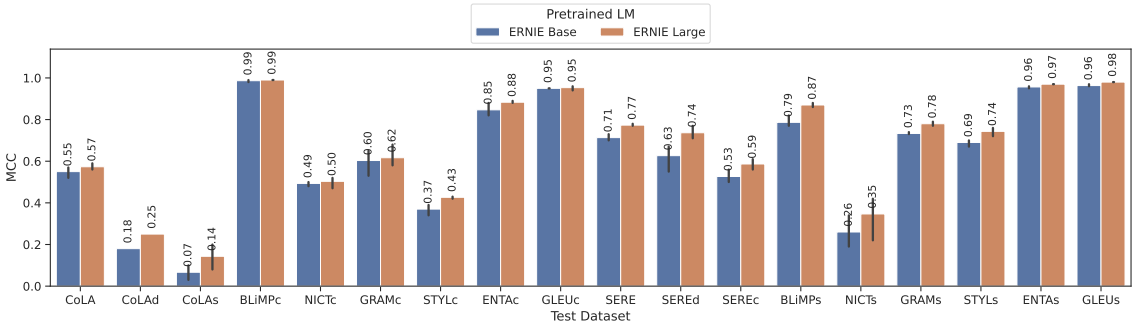


Figure 4.8: Performance of Base and Large versions of ERNIE across different train and test datasets

the original training and test sets (§4.7.1). In the later phase, the same datasets are transformed to match the sentence length distributions of either CoLA or SERE, and ERNIE is re-evaluated.

Importantly, while the domain of the datasets remains constant (i.e., no new sentences are introduced), the availability of sentences during training and evaluation is altered. Despite this, ERNIE's performance shifts asymmetrically depending on the specific distribution applied. For most datasets, transforming to align with SERE's distribution does not produce notable changes in performance, as their baseline sentence length distributions are already well-aligned with SERE's.

However, the transformation to match CoLA's distribution reveals a sharp contrast. CoLA's dataset, characterized by a high density of short sentences, substantially boosts model performance, particularly on sentences around 5 words in length, which dominates CoLA's distribution. Once this advantage is removed, significant performance drops are observed, as depicted in Fig. 4.8.

4.7.4 Performance as a Function of Sentence Length

In previous experiments, we observed that the distribution of sentence lengths within a dataset influences the performance of the ERNIE model. However, to better understand this relationship, we need to examine how ERNIE's performance varies across different sentence lengths. To this end, we assess the model's performance for each dataset while

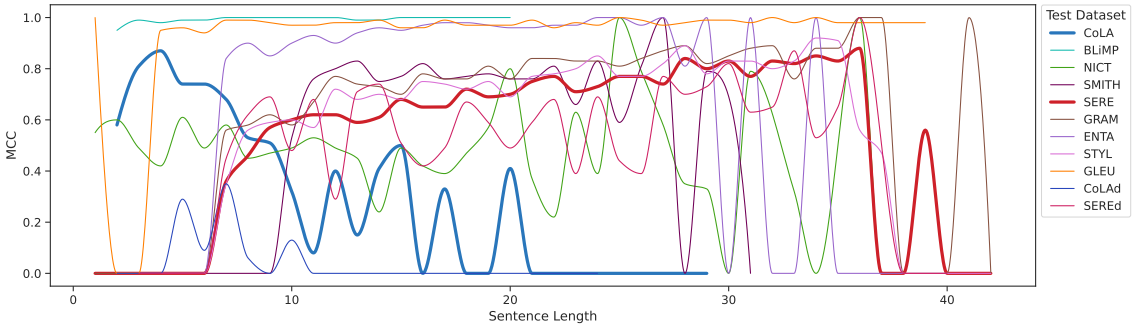


Figure 4.9: ERNIE's Performance (MCC) vs Sentence Length with PCHIP interpolation

keeping sentence length constant. To address the discontinuities in sentence length distribution observed in datasets like CoLA, we apply smoothed interpolation using Piecewise Cubic Hermite Interpolating Polynomials (PCHIP). This approach helps create clearer, more interpretable performance curves.

Results

The smoothed performance curves are shown in **Fig. 4.9**, with bold lines emphasizing ERNIE's performance on the CoLA and SERE datasets across varying sentence lengths. Consistent with prior findings (Warstadt and Bowman, 2020), we observed a negative correlation between MCC scores and sentence length for CoLA, with a significant decline in performance within the typical range of human-written sentences (13-21 words, see Table 4.2). In contrast, ERNIE maintained consistently high performance across all sentence lengths on the BLiMP dataset. Performance on the GLUE and ENTA datasets was strong only for sentences exceeding 5 and 7 words, respectively. For the NICT dataset, performance remained stable regardless of sentence length. Similar patterns emerged in the SERE, GRAM, and STYL datasets, where ERNIE's performance improved as sentence length increased.

Discussion

Our analysis revealed significant differences in ERNIE's performance when fine-tuned on SERE compared to CoLA, particularly for sentences in the 13-21 words range, which is typical of human-written corpora (see Table 4.2). These differences indicate that the sentence length distribution in the fine-tuning data significantly impacts the model's performance. Furthermore, the pronounced performance fluctuations—both improvements and declines—across various datasets are attributed to the uneven distribution of sentence lengths, creating a long-tail effect that causes sudden changes in the moving MCC average. To mitigate overfitting during fine-tuning, we employed early stopping criteria, as detailed in Section 4.6.3. Nevertheless, the near-perfect MCC scores on BLiMP, ENTA, and GLEU suggest that models like ERNIE may be highly effective at leveraging the limited diversity present in these datasets, raising questions about the model's robustness when applied to more complex or diverse linguistic structures.

4.7.5 Effect of the Number of Examples

When controlling for sentence length distribution, we observe notable shifts in ERNIE's performance across the CoLA and SERE datasets. These variations can be attributed to two main factors: differences in sentence length distributions within the fine-tuning data or fluctuations in data volume caused by downsampling. In earlier experiments, we examined how sentence length impacts ERNIE's performance on individual datasets, but the role of data volume was not fully addressed. This experiment seeks to evaluate the influence of fine-tuning data size on ERNIE's performance in the acceptability task.

To investigate this, we created alternative fine-tuning datasets of different sizes—100, 300, 924 (CoLAs's full dataset), 1000, 3000, 5000, 8551 (CoLA's full dataset), 10,663 (NICT's full dataset), 12,599 (SMITH's full dataset), 15,000, 20,000, 30,000, and 70,000—by randomly downsampling the original SERE and CoLA datasets. ERNIE's performance was then evaluated using these reduced datasets.

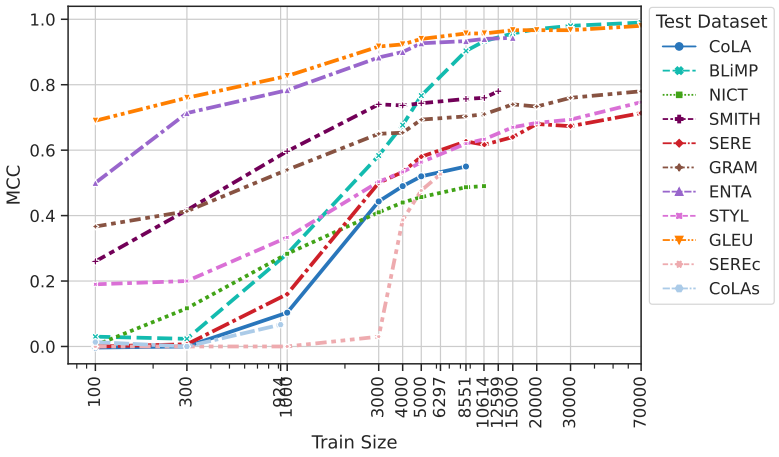


Figure 4.10: ERNIE's Performance (MCC) vs Train split size for fine-tuning

Results

Figure 4.10 presents the results, with the x-axis showing the size of the fine-tuning dataset and the y-axis depicting the MCC. The color hue distinguishes different datasets. The point at which ERNIE's performance, as measured by MCC, stabilizes during fine-tuning varies significantly across datasets. For ENTA and GLUE, high performance is reached with as few as 1,000 samples, while for GRAM, this occurs around 10,000 samples. Notably, when trained on smaller datasets, ERNIE initially performs better on CoLA and NICT than on SERE. However, around 3,000 samples, an inflection point is observed where ERNIE fine-tuned on SERE starts to outperform its performance on the other datasets.

It is worth noting that CoLAd and SEREd are subsets of CoLA and SERE, respectively, and share similar sentence length distributions. Therefore, the performance trends seen in CoLA and SERE also apply to CoLAd and SEREd, particularly for smaller training sets. In general, ERNIE's performance across most datasets plateaus with around 20,000 samples, peaking at 30,000, with no significant gains beyond this point.

Discussion

ERNIE's performance on the CoLA dataset, across varying training sizes from 0 to 5,000 samples (Figure 4.10), highlights that its poor results on CoLAs (a downsampled version of CoLA matching SERE's distribution (§4.7.4) may be partly attributed to the smaller dataset size (924-265-132 for train-dev-test; see Table 4.3). However, this does not fully explain the substantial performance drop of 48 points from CoLA to CoLAs (from 0.55 to 0.07 MCC), which is much greater than the corresponding drop of 18 points from SERE to SEREc (from 0.71 to 0.53 MCC).

If dataset size alone accounted for the performance decline, we would expect a similar trend between SERE and SEREc. Yet, both datasets, when downsampled to 3,000 sentences, yielded the same MCC score of 0.42. This suggests that factors beyond dataset size, such as sentence length distribution, play a significant role in the performance disparity. Interestingly, despite SEREc containing 6,297 samples, ERNIE's MCC should theoretically have dropped to around 0.4—consistent with the downsampled SERE results—but, as previously discussed (§4.7.3), ERNIE performed better on SEREc, achieving an MCC of 0.53.

These findings point to an additional hypothesis regarding CoLA's sentence length distribution: longer sentences in CoLA may capture more complex grammatical structures that shorter sentences (typically 5–8 words) do not. Consequently, the longer sentences retained during downsampling could present greater challenges than the shorter sentences that were excluded. This hypothesis will be further tested in the next experiment.

4.7.6 Effect Across Grammatical Features

In this final experiment, we examine how sentence length variations within a dataset affect the performance of specific grammatical features using the grammatically annotated development set of CoLA (Warstadt and Bowman, 2020). The objective is to determine whether ERNIE's lower performance on CoLA—after downsampling it to resemble SERE—can be attributed to the greater grammatical complexity of longer sentences or to ERNIE's training on shorter sentences, which may have led to a bias against longer ones.

To address this question, we fine-tuned ERNIE separately on CoLA and SERE and

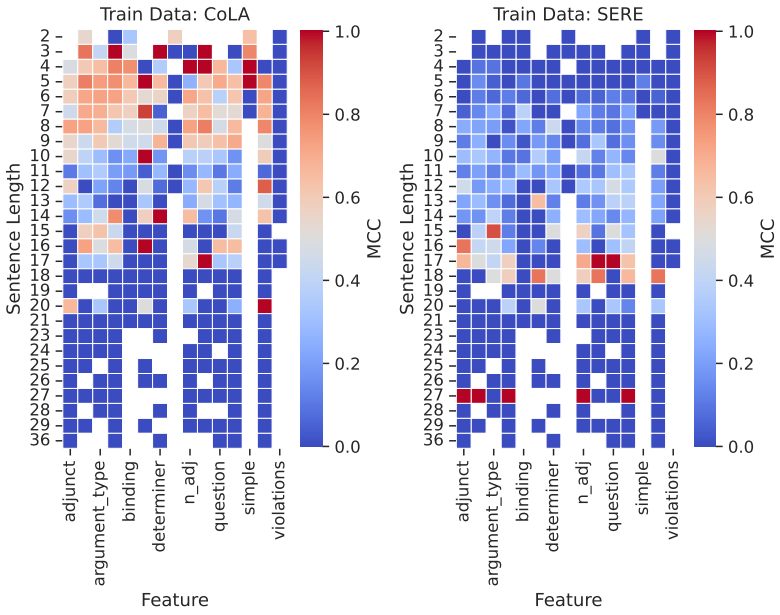


Figure 4.11: ERNIE's Performance (MCC) across sentence length vs major features on CoLA's development set

then evaluated its performance using the CoLA development set. This set is annotated for both major and minor grammatical features, with each sentence labeled to indicate the presence or absence of specific grammatical constructions. The annotations cover 15 major features, representing broad grammatical categories, and 63 minor features representing more detailed grammatical phenomena.

Results

Figure 4.11 presents a heatmap illustrating the model's performance across different combinations of major grammatical features (x-axis) and sentence lengths (y-axis). The color intensity represents the MCC score, where warmer colors (red) indicate higher MCC values, and cooler colors (blue) reflect lower values. Non-white squares indicate the presence of sentences for the corresponding feature and sentence length combination, while white

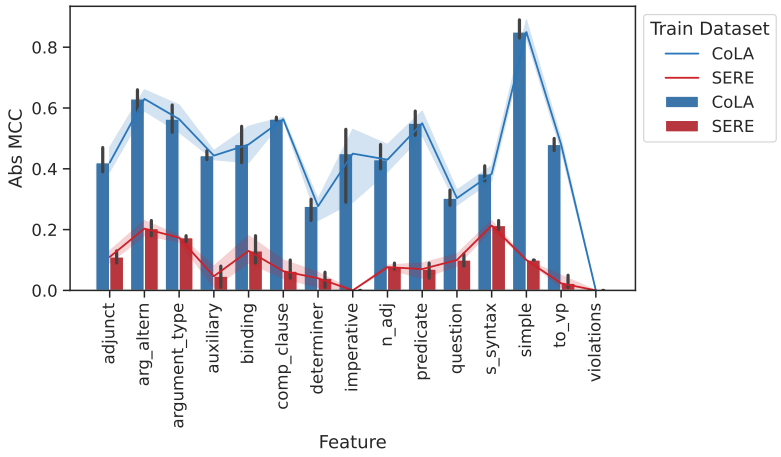


Figure 4.12: ERNIE's performance across major features on CoLA's development set

squares indicate no available data for that pairing. The left heatmap shows ERNIE fine-tuned on CoLA, while the right shows its performance when fine-tuned on SERE.

Additionally, **Fig. 4.12** presents the average absolute MCC (y-axis) for each major grammatical feature (x-axis), with error bars and shaded regions representing the variation in MCC scores across different sentences. A comparison of ERNIE fine-tuned on CoLA versus SERE highlights performance differences across grammatical features. For instance, when fine-tuned on CoLA, ERNIE exhibits lower MCC scores for the binding feature compared to complement clauses, while SERE shows the opposite trend. Similar patterns are observed with other grammatical features, such as determiners and imperatives. These trends remain consistent across both major and minor grammatical features, as well as varying sentence lengths.

Discussion

When fine-tuned on the CoLA dataset, ERNIE shows significantly better performance on shorter sentences across all major linguistic features. This is likely due to the scarcity of longer sentences in the CoLA development set, a known issue often referred to as the long tail problem. For these longer sentences, ERNIE consistently performs poorly, with an MCC score of 0 across all major features. In contrast, when fine-tuned on SERE, ERNIE

demonstrates improved performance, though on fewer linguistic phenomena. However, its performance remains consistent around the average sentence length (13-21 words) typical of human-written corpora.

These results raise important questions about how much of the information in the training dataset is being incorporated during the fine-tuning process, a recognized challenge in the language model probing approach. This makes it difficult to fully determine whether the behavior of the language models is driven by sentence length distribution, domain variation, or an imbalanced representation of sentences with different grammatical features. Despite these confounding factors, the experiment highlights that SERE, which lacks the rich linguistic diversity of CoLA, biases the model to perform better on a different range of sentence lengths. Along with prior findings, this serves as strong evidence that the sentence length distribution in training data significantly influences how language models handle future sentences. Consequently, sentence length distribution should be considered alongside domain relevance when fine-tuning language models.

4.8 Related Work

The influence of sentence length distribution on language model performance across various NLP tasks has been extensively studied. For instance, Bando et al. (2012) found that text summarization systems on the TREC Novelty track dataset tend to favor longer sentences, often considering them more relevant than shorter ones. In machine translation, Provilkov and Malinin (2021); Xuewen et al. (2021); Lu et al. (2022); Liangm et al. (2022) demonstrated that length bias in datasets can significantly degrade translation quality, especially with large beam sizes. Notably, neural machine translation (NMT) datasets are often strongly biased towards shorter sentences.

Similarly, datasets used for acceptability evaluations have faced criticism for their limitations. Vázquez Martínez (2021); Vázquez Martínez et al. (2023) highlighted issues with commonly-used datasets like CoLA and BLiMP, including limited coverage of linguistic phenomena and the generation of semantically implausible sentences.

However, no specific research has investigated the role of sentence length bias in acceptability evaluation datasets. Our earlier work (Daultani and Okazaki, 2022), which explored the influence of lexical frequency on acceptability judgments, revealed that sentence length, particularly when proportional to the number of named entities, affected evaluation outcomes. This led us to explore how variations in sentence length distribution affect language model performance in acceptability evaluations, addressing a previously unexamined aspect of this field.

Sinha et al. (2023) explored how context influences language models' acceptability judgments, defining context as a prefix leading into the test sentence. Their study specifically investigated how the content and length of this prefix affect the model's ability to accurately assess acceptability. They found that models performed better with longer, syntactically similar prefixes to the test sentence. In contrast, our study focuses solely on the length of the test sentence itself, without generating or examining any prefix or assessing the impact of prefix length on the language model's acceptability judgments.

In response to the need for multilingual acceptability evaluation, researchers have expanded upon the original CoLA dataset—designed exclusively for English—by developing similar resources for other languages. For example, Jentoft and Samuel (2023) introduced NoCoLA, a Norwegian version of CoLA, while Hu et al. (2023) created a version for Chinese. Similarly, Someya et al. (2024) adapted CoLA for Japanese, and Bel et al. (2024) introduced EsCoLA for Spanish. These language-specific versions retain CoLA's focus on acceptability judgments, facilitating cross-linguistic comparisons in NLP research.

In addition, the BLiMP dataset, originally developed for English linguistic evaluation, has also inspired multilingual adaptations. Xiang et al. (2021) released a Chinese version called CLiMP, while Someya and Oseki (2023) developed a Japanese adaptation, JBLiMP. Most recently, Taktasheva et al. (2024) introduced RuBLiMP, a version designed to evaluate linguistic acceptability in Russian. These efforts collectively expand the accessibility and applicability of acceptability evaluation benchmarks across diverse linguistic contexts.

Despite the availability of versions of CoLA and BLiMP datasets across multiple languages, our study focuses exclusively on English. The primary objective of this study was to investigate how sentence length influences language model acceptability judgments. To

the best of our knowledge, none of the existing versions of CoLA and BLiMP were specifically designed to reflect the sentence length distributions typically found in human-written corpora. By tailoring our proposed dataset to mirror these natural length distributions, we aim to provide a more accurate and practical assessment of model performance. This approach aligns our experimental conditions with real-world text characteristics, enhancing the relevance of our findings for acceptability evaluations in English-language contexts.

4.9 Conclusion

In the previous chapter, we explored how lexical frequency influences the language model's ability to evaluate acceptability. This prompted a deeper investigation into the effect of sentence length distribution on the model's performance on the task of acceptability evlauation. We begin by highlighting prior research showing that sentence length impacts language model performance across various NLP tasks, such as machine translation and text summarization. Building on this, we introduce two key concepts: human-written corpora, which refer to naturally produced text, and commonly-used datasets for acceptability evaluation, frequently employed in the literature to assess language model performance.

For meaningful acceptability evaluation, we argue that datasets should closely reflect the characteristics of human-written language, including natural variations in sentence length and linguistic patterns. However, upon analyzing the sentence length distribution in commonly-used datasets compared to human-written corpora, we found that these datasets are often skewed towards shorter sentences. This skewness is likely due to their design, which often focuses on testing specific linguistic phenomena in a controlled way. To overcome this limitation, we introduced seven new datasets (six derived and one novel) that are specifically designed for acceptability evaluation but maintain the sentence length properties of natural text.

We measured the differences in sentence length distributions between human-written corpora and both the commonly-used and proposed datasets using KL divergence. Our analysis revealed that the proposed datasets align much more closely with human-written corpora than the commonly-used datasets. We then conducted a series of experiments, reaching the conclusion that sentence length distribution significantly influences language

model performance on acceptability tasks, consistent with findings from other NLP tasks. For optimal use of language models in acceptability evaluation, it is essential that evaluation benchmarks accurately reflect the domain and expected sentence length distribution in real-world applications.

Based on these findings, we propose two strategies: (1) Expanding the CoLA dataset by incorporating longer, linguistically rich sentences, thus offering a more representative challenge while preserving the dataset's original purpose. This would enable models to tackle more complex linguistic features. (2) Curating multiple datasets from diverse domains, including written text, spoken language, and verse, to distribute evaluation tasks across varied contexts, similar to the structure of the GLUE Leaderboard. We believe that combining these approaches offers the most promising avenue for future research, and we are currently exploring these next steps.

Chapter 5

Conclusion

This concluding chapter begins by summarizing the key research aims and questions addressed in the study. Following this, I will reflect on the gaps identified in existing literature, highlighting how this research has aimed to fill those gaps. I will then present the key findings and discuss the contributions of this work, emphasizing its practical significance. A brief overview of the study's limitations and recommendations for future research will follow. Finally, the chapter will conclude with a summary of the key points covered throughout the preceding chapters.

5.1 Research Problem

Recent advances in transformer-based large language models have greatly enhanced text generation capabilities. Consequently, evaluating the quality of the generated text has become increasingly important. Performance on evaluation datasets helps assess these models' current effectiveness on various tasks. It informs future improvements to narrow the gap between human and machine-generated language.

Evaluation datasets possess several characteristics—lexical frequency, sentence length, semantic content, and syntactic complexity—that can significantly affect the model's performance. Previous research has highlighted the impact of lexical frequency and sentence length on tasks like machine translation and text summarization. However, no studies have specifically examined how these factors influence a language model's ability to perform

acceptability evaluations, leaving a gap in the literature.

5.2 Research Aims

This study investigates how two key characteristics of evaluation datasets—lexical frequency and sentence length—affect the language model's performance on acceptability evaluation tasks. The first aim is to explore how lexical frequency, particularly the presence of out-of-vocabulary (OOV) words, influences the model's evaluation of sentence acceptability. The second aim is to assess how sentence length distribution in the evaluation dataset impacts the model's acceptability judgments. The broader goal is to understand the extent of these influences and develop strategies to mitigate their effects in the evaluation process.

5.3 Research Gaps

While prior studies have examined the influence of lexical frequency and sentence length on various NLP tasks, such as machine translation and summarization, there has been no focused investigation into how these characteristics affect acceptability evaluation. Research has shown that rare words can introduce OOV issues, which in turn degrade language model performance in certain tasks. Similarly, sentence length has been found to affect model performance across multiple domains. However, its impact on acceptability evaluation remains underexplored.

Additionally, limited research has addressed how the characteristics of evaluation datasets compare to natural human language patterns, highlighting the need for further investigation into the alignment between these datasets and real-world language use. Understanding this alignment is crucial for improving language models' performance on tasks requiring human-like judgment.

5.4 Key Findigs

Our research yielded two key insights. First, we discovered that lexical frequency, particularly for OOV words, adversely affects language model performance in acceptability evaluation. When OOV words are replaced with UNK tokens, models assign probability mass to these tokens without fully capturing the context they represent. This undermines the reliability of probability-based metrics in evaluating linguistic acceptability. Although languages like English have a vast number of words, practical limitations on vocabulary size prevent models from covering all words. Notably, we noticed that a significant portion of these OOV words are proper nouns. To address this issue, we developed a preprocessing method called 'Replace Named Entity' (RNE), which substitutes each named entity in the text with its corresponding entity type. Incorporating this preprocessing step greatly enhances the alignment between the model's judgment of acceptability and human judgment across various probability-based metrics and the majority of datasets.

Additionally, our analysis of lexical frequency suggested that sentence length is another crucial factor influencing language model performance in acceptability evaluation. This observation led us to conduct a more detailed investigation into how sentence length affects model performance in this task.

Second, we identified a notable discrepancy in sentence length distribution between commonly-used datasets for acceptability evaluation and naturally occurring, human-written corpora. Specifically, these standard evaluation datasets tend to skew toward shorter sentences. We hypothesize this arises from the differing goals of these datasets. Human-authored corpora, being general-purpose, exhibit a natural variability in sentence length, as they often include detailed information, complex ideas, or clarifying clauses. In contrast, shorter sentences in evaluation datasets are often used deliberately to highlight grammatical contrasts and simplify the assessment of grammatical structures. This bias toward shorter sentences inflates model performance on these datasets, providing an inaccurate picture of how models perform on human-written texts. To mitigate this bias, we proposed seven new datasets that closely follow the sentence length distribution found in human-written corpora.

By aligning evaluation datasets with real-world sentence distributions and addressing

the issue of OOV words, our work provides a more accurate framework for assessing language model performance in acceptability evaluation.

5.5 Contributions

1. **Critique of Existing Metrics**: This study highlights key limitations in standard probability-based acceptability evaluation metrics, such as Normalized Log Probability Division and Syntactic Log Odds Ratio (§3.3). These metrics are often influenced by frequency of words in the training corpus, which can compromise their effectiveness, particularly for OOV words. This dependency on word frequency can skew results, undermining their reliability in assessing sentence acceptability.

2. **Improving Reliability of Existing Metrics**: To address the issue of frequency dependence for OOV words, we propose a novel data preprocessing technique called Replace Named Entity (§3.4). This method transforms sentences into more generalized, abstract representations, reducing the impact of lexical frequency. Our results show that using RNE improves the alignment of probability-based metrics with human judgment, offering a more robust and reliable evaluation of sentence acceptability.

3. **Insight into Sentence Length Bias**: Our analysis reveals that sentence length, another characteristic of evaluation datasets, also influences acceptability assessments. Further investigation shows that many commonly-used datasets are biased towards shorter sentences, which do not accurately represent the distribution found in naturally occurring, human-written text (§4.3). This skew towards shorter sentences can distort the evaluation of language models.

4. **Development of New Datasets**: To address the bias towards shorter sentences in existing evaluation datasets, we created seven new datasets designed to better reflect the sentence length and complexity of human-written corpora (§4.4). These datasets provide a more realistic benchmark for evaluating language models. We argue that for language models to perform optimally in acceptability evaluation, benchmarks

must reflect natural sentence distributions and complexity found in real-world text. Moreover, we highlight that commonly-used datasets like CoLA and BLiMP, which often favor shorter sentences, can misrepresent language model performance.

5.6 Application in Practise

There are two main practical applications of acceptability evaluation systems, as discussed earlier (§1.5). The first is to assess the generation capabilities of a language model by evaluating the quality of its output. The second is to evaluate the effectiveness of an evaluator language model in distinguishing between acceptable and unacceptable sentences. In this study, where we examine the influence of both lexical frequency and sentence length on a language model's ability to judge acceptability, the focus aligns with the second application—evaluating the effectiveness of the evaluator model in differentiating between acceptable and unacceptable text.

Improved dataset preprocessing techniques, particularly in handling the lexical frequency of OOV words, along with the development of new datasets more closely aligned with human-written corpora, play a crucial role in refining acceptability evaluation systems. These improvements make the evaluation systems more reflective of natural language patterns. Enhancing the accuracy of acceptability evaluation has broad implications, including improving the fluency, coherence, and appropriateness of machine-generated content. This, in turn, fosters greater user trust and expands the potential applications of language models across various sectors, ultimately making interactions with machine more natural and reliable.

5.7 Limitations

While this study provides an in-depth analysis of the impact of two dataset characteristics—lexical frequency and sentence length—on the performance of language models in acceptability evaluation, few limitations remain, as outlined below.

1. **Limited Scope of Dataset Characteristics**: This study focuses on two specific

dataset characteristics: lexical frequency and sentence length. While these features are crucial, other potentially influential factors, such as syntactic complexity and semantic content, were not examined. The omission of these elements may impact the generalizability of the results, as interactions between the studied characteristics and unexamined factors could affect the robustness of the findings.

2. **Language-Specific Constraints**: The research was conducted exclusively on English datasets. Since languages vary significantly in their structural and lexical properties, different observations may emerge when studying the impact of these characteristics and proposed methodologies to non-English datasets, which may restrict the cross-linguistic applicability of the findings.

3. **Granularity of Lexical Frequency Representation**: In exploring lexical frequency, we noted that OOV words negatively impact probability-based metrics used for acceptability evaluation. We proposed a coarser representation by substituting named entities with their categories to mitigate this. Although this approach addresses the issues related to proper nouns, it leaves unexplored the potential benefits of alternative levels of granularity in sentence representation. A more thorough investigation into different levels of granularities could provide a deeper understanding of their impact on acceptability evaluation.

4. **Metric Selection for Acceptability Rating**: Our evaluation of acceptability used traditional probability-based metrics. Recently developed semantic similarity based metrics, such as the BERTScore and BARTScore, could offer a more nuanced assessment of acceptability. Incorporating these newer metrics might provide a richer and more detailed evaluation of sentence acceptability, enhancing the precision of the results.

5. **Impact of Sentence Length Transformations**: In addressing sentence length, we transformed the datasets, which led to a reduction in size due to adjustments in sentence length distributions. This transformation may affect the generalizability of the findings, as the reduced dataset might not fully represent the variability present in larger or differently distributed datasets.

5.8 Recommendations for Future Research

1. **Expanding Sentence Representation**: Our solution to mitigate the issue from the lexical frequency of OOV words highlights the need for more comprehensive investigation of different sentence representations. Future research should explore dynamic granularity of the sentence representation or explore more content-rich sentence representations that include semantic content, syntactic complexity, and contextual factors. Incorporating these elements would allow language models to better capture the nuances of sentence structure, leading to more accurate predictions and a deeper understanding of how different aspects of language contribute to acceptability judgments.

2. **Augmenting Existing Datasets**: To address the limitations related to sentence length distribution, future research should focus on expanding existing commonly-used datasets like CoLA by incorporating longer sentences drawn from a variety of linguistic sources. This would help create a more balanced dataset that better mirrors the natural variation in sentence lengths found in human-written texts. By doing so, the evaluation of language models would become more robust and reflective of real-world language use, ultimately leading to more accurate assessments of model performance across different sentence structures.

3. **Diversifying Evaluation Metrics**: While this study primarily used the Pearson Correlation Coefficient (PCC) and Matthews Correlation Coefficient (MCC) for evaluation, future research should consider integrating a broader range of metrics, including recent advancements like BARTScore or other semantic similarity based metrics. Expanding the evaluation framework to include these additional metrics, along with qualitative assessments, would provide a more nuanced understanding of model performance. This approach would help capture the complex, multifaceted nature of language acceptability, enabling researchers to assess models in a more holistic and comprehensive manner.

5.9 Summary

This research investigates the influence of two key dataset characteristics—lexical frequency and sentence length—on the acceptability evaluation of language models. It addresses a crucial but often neglected aspect of NLP.

Chapter 1 - Introduction provides a historical overview and the basics of acceptability evaluation. The chapter introduces the research problem and outlines its aims and significance, followed by a discussion of the practical applications of acceptability evaluation.

Chapter 2 - Background offers a broader view of text quality evaluation in NLP, with a focus on the acceptability evaluation task. It differentiates between acceptability and related concepts like grammaticality, and explains the motivation for sentence-level acceptability evaluation. The chapter also contrasts traditional views of acceptability as a gradient with modern perspectives that treat it as a binary property. The motivation for exploring lexical frequency and sentence length is discussed, along with an introduction to two paradigms for acceptability evaluation.

Chapter 3 - Impact of Lexical Frequency examines the impact of lexical frequency on acceptability evaluation. It begins by defining acceptability as a gradient property and reviewing related research. The chapter introduces various metrics used to assess acceptability as a gradient, highlighting the limitations of probability-based metrics. The Replace Named Entity method is proposed for dataset preprocessing to address these issues. Experimental settings are described, followed by an analysis of the results. The findings show that lexical frequency, particularly for OOV words, significantly affects model performance. However, the proposed preprocessing method mitigates this impact.

Chapter 4 - Impact of Sentence Length explores the effect of sentence length on acceptability evaluation. The problem is framed in binary terms, followed by a review of relevant literature and the motivation for studying this characteristic. The chapter contrasts sentence length distributions in commonly-used datasets with those in human-written corpora, highlighting a bias toward shorter sentences. To address this, seven new datasets are introduced. These datasets are evaluated against human-written corpora using KL distance, and the experimental setup is discussed. The analysis reveals that sentence length significantly affects model performance, especially in the range of 13-21 tokens, which is typical

of human-written text. Models trained on commonly-used datasets perform poorly in this range.

Chapter 5 - Conclusion revisits the research problem, objectives, and gaps. It summarizes the findings on lexical frequency and sentence length, their impact on acceptability evaluation, and the contributions of this research. The chapter also outlines the limitations of the study and offers recommendations for future work.

In summary, this research demonstrates that both lexical frequency and sentence length affect the performance of language models in acceptability evaluation. However, these effects can be mitigated using the proposed methodologies. Overall, this study not only identifies the limitations of current evaluation practices but also provides practical solutions to improve the accuracy and reliability of acceptability assessments. By addressing biases introduced by lexical frequency and sentence length, it contributes to the development of more refined and human-like language models.

Appendix A

Named Entity Types

The spaCy named entity recognition (NER) system is capable of identifying 18 distinct types of entities (Honnibal et al., 2020). A detailed summary of these entity types is provided in Table A.1, which includes three key columns: 'Index', 'Named Entity Category', and 'Description'. The Index column assigns a unique identifier to each entity type, the Named Entity Category specifies the nature of the entity, and the Description offers a brief explanation of each category. In Chapter 3, we explored how lexical frequency influences acceptability judgments by replacing proper nouns in the text with their corresponding named entity categories, as listed in the second column of the table.

Index	Named Entity Category	Description
1	PERSON	People, including fictional.
2	NORP	Nationalities or religious or political groups.
3	FAC	Buildings, airports, highways, bridges, etc.
4	ORG	Companies, agencies, institutions, etc.
5	GPE	Countries, cities, states.
6	LOC	Non-GPE locations, mountain ranges, bodies of water.
7	PRODUCT	Objects, vehicles, foods, etc. (Not services.)
8	EVENT	Named hurricanes, battles, wars, sports events, etc.
9	WORK OF ART	Titles of books, songs, etc.
10	LAW	Named documents made into laws.
11	LANGUAGE	Any named language.
12	DATE	Absolute or relative dates or periods.
13	TIME	Times smaller than a day.
14	PERCENT	Percentage
15	MONEY	Monetary values, including unit.
16	QUANTITY	Measurements, as of weight or distance.
17	ORDINAL	“first”, “second”, etc.
18	CARDINAL	Numerals that do not fall under another type.

Table A.1: Description of named entity types supported by spaCy

Appendix B

Sample Sentences from Acceptability Datasets

Table B.1 presents sample sentences from both commonly-used and proposed datasets analyzed in Chapter 4. The table contains four columns: 'Dataset', 'Sentence ID', 'Acceptability', and 'Sentence'. The CoLA and BLiMP datasets are from commonly-used datasets, while the other seven—NICT, SMITH, GRAM, STYL, ENTA, SERE, and GLEU—are newly proposed. The 'Sentence ID' column indicates the source of each sentence, while 'Acceptability' represents the judgment label (1 = acceptable, 0 = unacceptable). The 'Sentence' column contains the actual sentence text. Each combination of Sentence Identifier and Acceptability forms a unique pair.

Dataset	Sentence ID	Acceptability	Sentence
CoLA	ks08	1	With what did the baby eat the food?
	ks08	0	The authorities blamed Greenpeace with the bombing.
BLiMP	d_n_a_i_1_194	1	A company doesn't upset this woman.
	l_b_i_e_q_144	0	What was Suzanne watching actresses?
NICT	E_file_00580_000000111	1	I think because of that audio guidance, the station clerks are not so kind.
	E_file_00839_000000058	0	And, as soon as they get there, they started to make a tent by themselves.
SMITH	test_000006055	1	Each Wikipedia article is an entity in a general knowledge base (KB) .
	test_000006911	0	Section 4 shows an evaluation of this approach as a baseline , and shows it does not move well .
GRAM	gegn_001001978	1	In a plagiarism detection system , every incoming document is compared with all registered non - plagiarized documents .
	gegn_002887874	0	Due to the stochastic training and uncertainty in the hyperparameter value pattern robust across the ensemble are very likely to reflect useful regularities than individual models .
STYL	srn_001184806	1	Our basic model architecture is similar to that of the ByteNet encoder , except that the inputs to our model are tokens and not bytes .
	srn_001118566	0	Represents the class of which can be pimped together .
ENTA	esgn_001707027	1	These relations are modelled in the CDM , which is a unified and simplified version of the logical data model represented in the database .
	esgn_004758769	0	The score of capitalization may be set to the morpulent names after scholaries are speculated .
SERE	esgn_003907072	1	One is sentence alignment error , and the other is English parse error .
	srn_000037827	0	A two - layer LSET had been used for FTTT while the LSNUs was used on another room .
GLEU	1489_0072	1	The character of a nation as a people of great deeds is one, it appears to me, that should not be lost sight of.
	000032047	0	His brother Henrik at the end of theseason after losing in the championship game to Louisville.

Table B.1: Sample acceptable and unacceptable sentences from commonly-used and proposed datasets

Bibliography

Dimitris Alikaniotis and Vipul Raheja. 2019. The unreasonable effectiveness of transformer language models in grammatical error correction. In Helen Yannakoudakis, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, and Torsten Zesch, editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Florence, Italy, pages 127–133.

Lorena Leal Bando, Falk Scholer, and Andrew Turpin. 2012. Sentence length bias in TREC novelty track judgements. In Andrew Trotman, Sally Jo Cunningham, and Laurianne Sitbon, editors, *The Seventeenth Australasian Document Computing Symposium, ADCS '12, Dunedin, New Zealand, December 5-6, 2012*. ACM, pages 55–61.

Nuria Bel, Marta Punsola, and Valle Ruíz-Fernández. 2024. EsCoLA: Spanish corpus of linguistic acceptability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, pages 6268–6277.

BNC Consortium. 2007. The british national corpus, version 3 (bnc xml edition). distributed by oxford university computing services on behlaf of the bnc consortium. `http://www.natcorp.ox.ac.uk/`, Last accessed on 2022-03-21.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Lluís Màrquez,

Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 632–642.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics* 18(4):467–480.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pages 88–107.

Noam Chomsky. 1957. *Syntactic Structures*. De Gruyter Mouton.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 8440–8451.

Vijay Daultani and Naoaki Okazaki. 2022. Improving automatic evaluation of acceptability based on language models with a coarse sentence representation. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics, Manila, Philippines, pages 109–118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pages 4171–4186.

Wikimedia Foundation. 2023a. Wikimedia downloads.

Wikipedia Foundation. 2023b. Wikipedia featured articles.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203–225.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 708–719.

Liliane Haegeman. 1994. *Introduction to Government and Binding Theory*. Blackwell Publishers Ltd.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., volume 28.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python .

Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Patterson, Jiahui Huang, Peng Zhang, Chien-Jer Charles Lin, and Rui Wang. 2023. Revisiting acceptability judgements.

Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the rough: Generating fluent sentences

from early-stage drafts for academic writing assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, pages 40–53.

Emi Izumi, Kiyotaka Utimoto, and Hitoshi Isahara. 2004. The nict jle corpus: Exploiting the language learners' speech database for research and education. *Special Issues of International Journal of the Computer* 1:31–48.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1–10.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. University of Tartu Library, Tórshavn, Faroe Islands, pages 610–617.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In Anna Korhonen and Ivan Titov, editors, *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Brussels, Belgium, pages 313–323.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In Gaja Jarosz, Max Nelson, Brendan O'Connor, and Joe Pater, editors, *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*. pages 287–297.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*. volume 1, pages 181–184 vol.1.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, pages 28–39.

Solomon Kullback and Richard Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. *Proceedings of the 19th International Joint Conference on Artificial Intelligence* pages 1085–1090.

Jey Han Lau, Clark Alexander, and Lappin Shalom. 2016. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41(5):1202–1241.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Statistical model of grammaticality. distributed by the center for linguistic theory and studies in probability. `https://gu-clasp.github.io/projects/smog/experiments/`, Last accessed on 2022-03-21.

Bowen Liangm, Pidong Wang, and Yuan Cao. 2022. The implicit length bias of label smoothing on beam search decoding .

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692.

Zhiyun Lu, Yanwei Pan, Thibault Doutre, Parisa Haghani, Liangliang Cao, Rohit Prabhavalkar, Chao Zhang, and Trevor Strohman. 2022. Input length matters: Improving rnn-t and mwer training for long-form telephony speech recognition.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405(2):442–451.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 344–351.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 229–234.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pages 8024–8035.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Jeju Island, Korea, pages 959–968.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 186–195.

Ivan Provilkov and Andrey Malinin. 2021. Multi-sentence resampling: A simple approach to alleviate dataset length bias and beam-search degradation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 8612–8621.

Andrew Radford. 1988. *Transformational Grammar*. Cambridge University Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2023. Language model acceptability judgements are not always robust to context. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 6043–6063.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, Dubrovnik, Croatia, pages 1581–1594.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2024. JCoLA: Japanese corpus of linguistic acceptability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro

Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, pages 9477–9488.

Jon Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1:118–129.

Neomy Storch. 2009. The impact of studying in a second language (l2) medium university on the development of l2 writing. *Journal of Second Language Writing* 18(2):103–118.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020a. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(05):8968–8975.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020b. ERNIE 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(05):8968–8975.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs.

Ravikiran Vadlapudi and Rahul Katragadda. 2010. On automated evaluation of readability of summaries: Capturing grammaticality, focus, structure and coherence. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*. Association for Computational Linguistics, Los Angeles, CA, pages 7–12.

Héctor Javier Vázquez Martínez. 2021. The acceptability delta criterion: Testing knowledge of language using the gradience of sentence acceptability. In Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic, pages 479–495.

Héctor Javier Vázquez Martínez, Annika Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni, editors, *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*. Association for Computational Linguistics, Singapore, pages 48–64.

Stephen Wan, Robert Dale, and Mark Dras. 2005. Searching for grammaticality: Propagating dependencies in the Viterbi algorithm. In Graham Wilcock, Kristiina Jokinen, Chris Mellish, and Ehud Reiter, editors, *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*. Association for Computational Linguistics, Aberdeen, Scotland.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2023. Glue leaderboard. `https://gluebenchmark.com/leaderboard`, Last accessed on 2023-09-04.

Alex Warstadt and Samuel R. Bowman. 2020. Linguistic analysis of pretrained sentence encoders with acceptability judgments.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7:625–641.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 451–462.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, pages 38–45.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, pages 2784–2790.

Shi Xuewen, Huang Heyan, Jian Ping, and Tang Yi-Kun. 2021. Reducing length bias in scoring neural machine translation via a causal inference method. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Huhhot, China, pages 874–885.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., volume 34, pages 27263–27277.

Zhu Yukun, Kiros Ryan, Zemel Rich, Salakhutdinov Ruslan, Urtasun Raquel, Torralba

Antonio, and Fidler Sanja. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual evaluation of linguistic acceptability. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, pages 2658–2674.