Improving Automatic Evaluation of Acceptability Based on Language Models with a Coarse Sentence Representation

Vijay Daultani
Tokyo Institute of Technology
Meguro City, Tokyo
vijay.daultani@nlp.c.titech.ac.jp

Naoaki Okazaki
Tokyo Institute of Technology
Meguro City, Tokyo
okazaki@c.titech.ac.jp

Abstract

Motivated by recent findings on the probabilistic modeling of the acceptability judgments, several metrics have been proposed for its automatic evaluation. Frequently used metrics such as syntactic log odds ratio (SLOR) and its variants are based on utilizing probability from language model (LM) as a proxy for automatic acceptability evaluation of the generated text from an LM. Since one cannot use probability directly as a measure of acceptability, these metrics take steps to remove the confounding effects of noise from sentence length and lexical frequency to enable the usage of probability for acceptability evaluation. In this work, we argue that even though the effects are reduced, they still exist. We propose a data transformation strategy, Replace Named Entity (RNE), to get a coarse representation of a sentence to mitigate the remaining problems from lexical frequency. In RNE, we identify all proper nouns (i.e., NEs) in a sentence and classify them into one of eighteen types. Later RNE replaces all occurrences of NEs in a sentence with their identified type. We later trained three LMs (2, 3, 4-grams) and assessed their performance of five acceptability measures on four test datasets. We found that LMs trained on datasets preprocessed by RNE yield a significantly higher correlation (upto 52% on some datasets) with human acceptability judgment.

1 Introduction

Language Models (LMs) are often used to generate natural language text for NLP tasks — Machine Translation, Summarization, Question Answering, and many others. Moreover, intrinsic evaluation of the LMs often includes at least two characteristics (Mutton et al., 2007). First, how well the generated text represents the source data, whether it be the text in another language for machine translation, text to represent a summary of a document, or text to represent answers for a question, etc. Second, how well it conforms to regular human language use, a property we will refer to as acceptability of the sentence. Acceptability evaluation of a sentence is an essential task for automatically evaluating the quality of the text (to help filter unacceptable sentences) generated by the LMs.

Before moving forward, it is also essential to understand the difference between the usage of related words, i.e., fluency, readability, and grammaticality. Both fluency and readability are alternate words for acceptability, but the exact definition of these terms varies across the literature (Lau et al., 2017; Mutton et al., 2007; Kann et al., 2018; Storch, 2009; Pitler and Nenkova, 2008; Vadlapudi and Katragadda, 2010). However, we would like to differentiate acceptability from grammaticality. When a human evaluates the acceptability of a text, grammaticality is one of the possible factors, among others like semantic plausibility, processing difficulties, etc., that can also influence the acceptability of a given text. Though both 'acceptability' and 'grammaticality' have been used interchangeably, a sentence can be grammatical yet unacceptable and vice versa. A famous example is Chomsky's phrase, "Colorless green ideas sleep furiously." (Chomsky, 1957). Vice versa, acceptable sentences can be ungrammatical, e.g., in an informal context such as poems.

Sentence Type	Sentence
Original	Apple is set to hold its first event on Tuesday.
NER Result	[$_{\text{ORG}}$ Apple] is set to hold its [$_{\text{ORDINAL}}$ first] event of [$_{\text{DATE}}$ the year] on [$_{\text{DATE}}$ Tuesday].
Transformed	ORG is set to hold its ORDINAL event of DATE on DATE.

Table 1: Example sentence for motivation

Whether humans represent text acceptability evaluation as a binary classification (Warstadt et al., 2020) of acceptable vs. unacceptable class of sentence or as a probabilistic property (Lau et al., 2017) has been a subject of lengthy debate among cognitive scientists and linguists (Chomsky, 1957; Manning, 2002; Sprouse, 2007). Both of the above views have their strengths and weaknesses. On the one hand, binary classification models do not have the flexibility to distinguish text between varying degrees of acceptability. On the other hand, the acceptability of a sentence is not the same as the likelihood of its occurrence as determined by the probabilistic model, which depends on sentence length and lexical frequency. However, Lau et al. (2017) demonstrated it is possible to augment the probabilistic model to predict the acceptability of a sentence if one can normalize probability values from the LM to eradicate the confounding effects of noise introduced by length and lexical frequency, e.g., SLOR (Lau et al., 2017) and WP-SLOR (Kann et al., 2018).

In this article, we lean towards the view that acceptability is a probabilistic property. We depict that probability-based acceptability metrics SLOR and other variants, though they reduce the confounding effect of lexical frequency, do not resolve the problem entirely in Section 2. We later provide evidence that one reason for this problem is a granular-level representation of the sentence since LM has to predict the probabilities for all words in the sentence, including words for which LM has low confidence (i.e., rare words or out of vocabulary words). This work is motivated by how humans visualize the sentence (coarse-level representation) for acceptability evaluation. Our goal is to find how to generate such a sentence representation to enable LMs better correlate with human acceptability judgment for the existing metrics.

Table 1 presents a concrete example of our intuition and data transformation strategy. Original refers to the unprocessed sentence (granular-level representation). We consider replacing proper nouns, i.e., NEs, in a sentence to generate a coarse-level representation. We propose a two-step approach i.e., Replace Named Entity (RNE) (Section 3) to construct such a coarse-level representation. First, we employ Named Entity Recognition (NER), a task to identify the spans of text that constitute proper nouns in a sentence and classify each identified span into one of the NE types (i.e., subscript in the prefix) as shown in NER Result. Second, we replace each occurrence of a NE with its classified type within a sentence to generate the coarse-level representation (i.e., Transformed). We argue that the coarse-level representation closely resembles how humans will judge the Original sentence's acceptability. Therefore, it should be used as an input to train and test LMs. Our contributions are summarized as follows:

- We provide evidence that a popular probability-based acceptability evaluation metric SLOR has limitations since it is based on the lexical frequency of words in the training corpus.
- We demonstrate that the original sentence is not the best representation for training LM and propose RNE, a data transformation strategy to transform the sentences to a coarse-level representation.
- We present empirical evidence from our experiments that SLOR and other probability-based acceptability metrics correlate better with human judgment when LMs are trained on data processed with RNE.

To the best of our knowledge, ours is the first successful attempt to use NEs to find a coarse-level representation of a sentence that better resembles how humans evaluate acceptability to improve the correlation between existing automatic acceptability metrics and human acceptability judgment. In this paper, our target language is English. However, we believe the ideas and methods apply to other languages. Additionally, our proposed data transformation strategy is independent of both LM and the metric used to measure the acceptability.

2 The Problem

2.1 Problem Definition

Formally, a sentence S comprises of n words w_1, w_2, w_3, .. w_n. Each word w_i occurs with lexical frequency (count) f_i in the training corpus. The goal here is to find the acceptability $y \in \mathbf{R}_{\geq 0}$ of the sentence S.

2.2 Background

One might suggest treating the likelihood (probability) of a sentence S as its measure of acceptability, with 1 indicating completely acceptable and 0 means unacceptable sentence. Although, the idea seems enticing but will be an incorrect usage of the values in a probability distribution. The probability of a sentence, S, from an LM is the probability that a randomly selected sentence will be S and not a measure of its acceptability. Based on this observation, Lau et al. (2017) proposed several sentence and word level metrics to augment the probabilistic LM with an acceptability measure.

Among all proposed metrics, syntactic log odds ratio (SLOR) in Equation 1 has shown a good correlation with human acceptability judgment. SLOR is a function that normalizes the sentence probability and believed to eliminate the confounding factors of sentence length by dividing with the sentence length, i.e., $|S|$ and lexical frequency by subtracting the unigram probability of words comprising S. In Equation 1, $p_m(S)$ refers to the sentence probability of S, i.e., a product of probabilities assigned to each n-gram by the LM. $p_u(S)$ is the unigram probability for sentence S, i.e., a product of the unigram probabilities of the words comprised in the sentence.

$$\text{SLOR}(S) = \frac{\log p_m(S) - \log p_u(S)}{|S|} \qquad (1)$$

Against the expectation, we have observed that issues related to lexical frequency still persist in the formulation of SLOR. Essentially words lexical frequency in the training corpus can severely impact both sentence probability $p_m(S)$ and unigram probability $p_u(S)$, and therefore impairing the usage of SLOR for acceptability prediction.

To understand the impact of lexical frequency on SLOR, let's refer to three sentences, i.e., s1, s2, s3 in Table 2 with words France, Tuvalu, and Kiribati, referring to the names of three nations respectively. For our convenience, we will override the notation of $p_u(w)$ to refer to the unigram probability of the word w. To explain the issue, we have made two assumptions; first, let's assume that France occurs often and Tuvalu is a rare word in the training corpus. Second, let's assume the word 'Kiribati' never appears in the training corpus and is an out-of-vocabulary (OOV) word for the LM.

Index	Sentence
s1	He is a citizen of France.
s2	He is a citizen of Tuvalu.
s3	He is a citizen of Kiribati.

Table 2: Three sentences of equal length and equally acceptable

2.3 Unigram Probability

Based on our assumption about the frequencies of the word 'France' and 'Tuvalu', it will be safe to expect unigram probability $p_u(\text{France})$ to be higher than $p_u(\text{Tuvalu})$. In practice, to avoid the problem of 0 unigram probabilities with OOV words ('Kiribati'), it is common to replace them with UNK tokens in both training and test corpus and add UNK to the vocabulary. This will assign a tiny non zero unigram probability and therefore $p_u(\text{Kiribati}) \approx 0$. This tiny unigram probability for 'Kiribati' at first appears to do no harm, but voids the sole purpose of using unigram probabilities to counteract the higher sentence probability $p_m(s1)$ and $p_m(s2)$ in SLOR as we will show in the next section.

2.4 Sentence Probability

Now let us consider the sentence probability p_m from a 3-gram LM. Sentence probability is product of individual n-gram probabilities as described in Equation 2. Notice that all three sentences, s1, s2, and s3, in Table2 have a common prefix phrase '*He is a citizen of*' and differ only on the last word i.e., 'France', 'Tuvalu' and 'Kiribati'. The 3-gram LM will assign equal probabilities to all 3-grams ($p(a \mid He,\ is)$, $p(citizen \mid is,\ a)$, $p(of \mid a,\ citizen)$) within common prefix. Furthermore, based on our first assumption about words 'France' (i.e., high frequency) and 'Tuvalu' (i.e., rare) in

the training corpus we should expect 3-gram probability $p(France \mid citizen,\ of)$ to be higher than $p(Tuvalu \mid citizen,\ of)$.

$$p_m(S) = p_m(w_1^n) = \prod_{t=1}^{n} p(w_t \mid w_{t-2}, w_{t-1}) \quad (2)$$

2.5 Incompetence of SLOR

In the ideal world, $p_m(s1)$ should be equal to $p_m(s2)$ since both are equally acceptable sentences. However, due to the above two observed outcomes, first, equal 3-gram probabilities for the common prefix on both s1 and s2; second, a higher 3-gram probability for the word 'France' will result in $p_m(s1)$ higher than $p_m(s2)$. This observation motivated Lau et al. (2017) to propose SLOR, where they counteracted this behavior by subtracting the unigram probabilities from sentence probabilities to get similar acceptability scores for equally acceptable sentences.

However, subtracting unigram probabilities does not solve problems for all different cases. Why? Recall in section 2.3, we discovered that for an OOV word 'Kiribati' unigram probability is tiny, i.e. $p_u(\text{Kiribati}) \approx 0$. This tiny unigram probability for s3 will result in significantly different SLOR score for s3, therefore evaluating it as more acceptable sentence compared to s1 and s2. Hence $SLOR(s1) \approx SLOR(s2) \not\approx SLOR(s3)$ which is undesirable because the word choice ('France', 'Tuvalu' or 'Kiribati') should not lead to a different measure of acceptability.

3 Proposed Method

This observation that the lexical frequency of a word should not lead to a different measure of acceptability led us to think about how humans judge the acceptability of a sentence. We assert that human judgment of acceptability is only slightly influenced by word choice and is highly influenced by sentence structure.

Let us take an example from Table 3 to measure the acceptability of sentences s1 and s2. s1 and s2 are two original sentences with similar lengths and are equally acceptable. s3 and s4 represent sentences s1 and s2 with all identified NE spans and classified type as a subscript in prefix. s5 is the final transformed sentence (coarse-level representation) after replacing all identified NEs with the classified type for s1 and s2.

Index	Sentence
s1	Apple is set to hold its first event of the year on Tuesday.
s2	NEC is set to hold its second event of 2022 on Wednesday.
s3	[$_{\text{ORG}}$ Apple] is set to hold its [$_{\text{ORDINAL}}$ first] event of [$_{\text{DATE}}$ the year] on [$_{\text{DATE}}$ Tuesday].
s4	[$_{\text{ORG}}$ NEC] is set to hold its [$_{\text{ORDINAL}}$ second] event of [$_{\text{DATE}}$ 2022] on [$_{\text{DATE}}$ Wednesday].
s5	ORG is set to hold its ORDINAL event of DATE on DATE.

Table 3: Example sentences to explain the motivation and our proposed preprocessing data transformation strategy.

In Table 3 notice that sentences s1 and s2 are very similar in structure with few variations in the word choice, i.e. 'Apple' vs 'NEC', 'first' vs 'second', 'the year' vs '2022' and 'Tuesday' vs 'Wednesday'. Nonetheless, when it comes to a human judgment of acceptability for s1 and s2, one would rate both the sentences equally irrespective of different word type choices in a sentence. We argue that neither word choice nor lexical frequency should influence sentence acceptability. Therefore it does not matter if the word in the sentence is 'Apple' or 'NEC'; instead, the critical information is the fact that both the words refer to a single NE type, i.e., ORGANIZATION (ORG). Broadly we can think of any other ORG NE such as 'Microsoft', 'United Nations' etc, and it should not affect the measure of acceptability for the sentence. Similarly, the lexical frequency of phrases 'first' over 'second', 'the year' over '2022', and 'Tuesday' over 'Wednesday' is less critical than phrases referring to NE type ORDINAL, DATE, and DATE, respectively.

In a nutshell, to humans, the sentence's broad structure and transitions between POS (Lapata and Barzilay, 2005) are more critical than the lexical frequency of the words to determine the acceptability. Based on this motivation, if we were to replace phrases with their corresponding NE type, we can transform original (granular-level representation) sentences s1 and s2 to a standard (coarse-level representation) sentence s5. This transformation should help LMs overcome the issue of word choice and their lexical frequencies to influence sentences' measure of acceptability. If coarse-level rep-

resentation is helpful, why not replace the complete sentence with the corresponding POS instead of only replacing proper nouns? The reason is that replacing proper nouns with their NE Type generates an advantageous representation. On the one hand, it abstracts away details that are not critical for determining acceptability; on the other hand, it retains original words and sentence structure that highly influence acceptability i.e., rest of the POS classes (verb, adjective, adverb, preposition, conjunction, and interjection). Now we propose our two-step (Step I and Step II) solution Replace Named Entities (RNE) for data transformation.

3.1 Step I: Named Entity Identification

First, we segmented a sentence into words using spacy's (Honnibal and Montani, 2017) NLP English word segmenter. After completing the segmentation process, we scan the segmented sentence sequentially to find the consecutive words that constitute a NE. We used spacy's statistical entity recognition system with a default trained pipeline to assign one out of eighteen types (e.g., companies, locations, organizations, and products.) to an identified NE.

3.2 Step II: Replacing words with Named Entity Types

After identifying both NEs and their respective types over segmented input sentences, we then start the replacement process. In this step, we replace one or more consecutive words in a sentence previously identified as a NE in step I with its corresponding identified NE type both for training and test corpus.

3.3 Complete Pipeline

After transforming all the sentences in the training and test corpus with RNE, we train n-gram LM over the transformed training corpus. Such a sentence transformation (both independent of LM and the acceptability metric) will provide a coarse-level representation of a sentence to help LM focus on the transitions of POS without worrying about the words chosen for NEs. Furthermore, we believe this will enable a LM to generalize better into new domains with different vocabulary. Moreover, this abstract representation of a sentence will help all probability-based metrics, including SLOR as shown in section 5.

Description	Size	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 4: Details of the test corpus. Description and Size represents name of dataset and total number of sentences in the dataset. Followed by average number of words, NEs, and UNK tokens per sentence respectively

4 Experiment Setup

4.1 Dataset

We adopt the BNC corpus (BNC Consortium, 2007) that comprises 6.07M sentences for training LM. Moreover, to show the effectiveness of our proposed data transformation strategy, i.e., RNE, we evaluated the trained LMs on sentences that exhibited varying degrees of acceptability. Based on previous work of (Lau et al., 2017), we evaluated LMs on four English language datasets (BNC, ENWIKI, ADGER, and ADGER-FILTERED) within the Statistical Model of Grammaticality (SMOG) (The Center for Linguistic Theory and Studies in Probability, 2015) test corpus. Table 4 shares the detailed statistics on test datasets. Each sentence in the test dataset is associated with a human judgement of acceptability for further details on collection of the human ratings refer to Appendix A.1.

4.2 Baselines

We first preprocessed the training corpus following Standard Preprocess (SP) protocol as described in (Lau et al., 2017). SP comprises of three steps, first is to segment the sentences, second, filter out sentences with fewer then threshold (seven) words, third, replace rare words (i.e. with frequency less than threshold of four) with an unknown (UNK) token.

4.3 Language Models

We trained three n-gram i.e., 2-gram, 3-gram, and 4-gram LMs on BNC corpus though preprocessed differently for baseline (only SP with a vocabulary of 104,950) and our proposed work (i.e., SP + RNE with a vocabulary of 100,688). Each LM (for both SP and SP + RNE) was trained with Kneser-Key (Kneser and Ney, 1995) smoothing method.

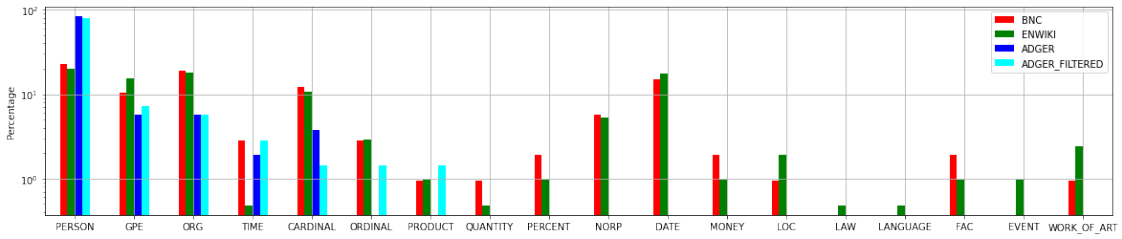

Figure 1: Percentage (Y axis in log scaled) of eighteen NE types (X axis) per sentence across four test datasets BNC, ENWIKI, ADGER and ADGER_FILTERED. Graph is sorted by the percentage of NE type on ADGER_FILTERED Dataset.

4.4 Metrics

To compare our results with previous work of Lau et al. (2017) we used pearson correlation between human judgement of acceptability and different probability scores (LogProb, Mean LP, Norm LP (DIV), Norm LP (SUB), SLOR) predicted to evaluate the performance of LMs. Due to the space limitation, we have only included the formula for SLOR in Section 2 for the formulation of rest of the metrics; refer to Appendix A.2.

Pearson Correlation We evaluated the performance of the LMs capability to predict the acceptability (X) by calculating it's pearson correlation with human judgement of acceptability (Y). In Equation 3 cov is the covariance. σ_X and σ_Y is the standard deviation of X and Y respectively.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (3)$$

5 Results and Discussion

We now discuss the experimental results, findings and their implications on acceptability evaluation.

Performance Comparison: Table 5, 6, 7, and 8 shows the performance on BNC, ENWIKI, ADGER, and ADGER_FILTERED test datasets respectively. As the test datasets were already processed via SP, we only preprocessed test datasets with RNE to evaluate LMs trained via SP + RNE.

We observed that LMs trained via SP + RNE have a higher or equal correlation with human judgments on all the measures for BNC, ENWIKI, and ADGER test datasets compared to LMs trained only via SP. The only exception is 2-gram LM for the BNC test

Measure	2-Gram		3-Gram		4-Gram	
	SP	SP+RNE	SP	SP + RNE	SP	SP + RNE
LogProb	0.33	0.35	0.40	0.50	0.42	0.65
Mean LP	0.46	0.36	0.52	0.55	0.55	0.67
Norm LP (Div)	**0.53**	0.43	0.57	**0.62**	0.60	**0.73**
Norm LP (Sub)	0.23	0.13	0.29	0.30	0.33	0.44
SLOR	**0.53**	0.44	0.55	0.61	0.57	0.69

Table 5: Pearson's r of acceptability measure and mean sentence rating for BNC. For BNC all the metrics are multiplied by factor of 10.

Measure	2-Gram		3-Gram		4-Gram	
	SP	SP+RNE	SP	SP + RNE	SP	SP + RNE
LogProb	0.22	**0.28**	0.24	0.32	0.24	0.33
Mean LP	0.14	0.22	0.19	0.28	0.20	0.30
Norm LP (Div)	0.19	0.27	0.24	**0.33**	0.25	**0.35**
Norm LP (Sub)	0.01	0.01	0.07	0.07	0.08	0.10
SLOR	0.20	0.27	0.24	**0.33**	0.24	0.34

Table 6: Pearson's r of acceptability measure and mean sentence rating for ENWIKI

Measure	2-Gram		3-Gram		4-Gram	
	SP	SP+RNE	SP	SP + RNE	SP	SP + RNE
LogProb	0.06	0.07	0.07	0.08	0.08	0.08
Mean LP	0.07	0.07	0.09	0.09	0.09	0.10
Norm LP (Div)	0.11	0.11	0.13	0.13	0.13	**0.14**
Norm LP (Sub)	0.09	0.10	0.12	0.12	0.12	0.12
SLOR	**0.12**	**0.12**	**0.14**	**0.14**	**0.14**	**0.14**

Table 7: Pearson's r of acceptability measure and mean sentence rating for ADGER

Measure	2-Gram		3-Gram		4-Gram	
	SP	SP+RNE	SP	SP + RNE	SP	SP + RNE
LogProb	0.30	0.31	0.32	0.33	0.33	0.35
Mean LP	0.23	0.22	0.25	0.26	0.26	0.28
Norm LP (Div)	0.32	0.30	0.35	0.34	**0.36**	**0.36**
Norm LP (Sub)	0.16	0.10	0.20	0.15	0.23	0.18
SLOR	**0.34**	0.30	**0.36**	0.33	**0.36**	0.34

Table 8: Pearson's r of acceptability measure and mean sentence rating for ADGER FILTERED

dataset. Furthermore, we got mixed improvement results on the ADGER_FILTERED dataset.

Quantitatively we observed an improvement in the range of 3% (LogProb) to 52% (LogProb) for 2-gram and 4-gram LM, respectively. For ENWIKI, we observed an improvement in the range of 2% (Norm LP Sub) to 50% (Mean LP) for the 2-gram LM. For ADGER, we observed an improvement in the range of 1% (SLOR) to 13% (Mean LP) for 3-gram and 2-gram LM, respectively. For ADGER_FILTERED, we observed an improvement of 2% (Norm LP Div) to 11% (Log Prob) for 4-gram LM.

RNE's impact on probability metrics other than SLOR: We verified our hypothesis that neither word choice nor lexical frequency for NEs is critical in determining the acceptability of the sentence as we saw consistent improvement in correlation for all probability related measures in addition to SLOR.

Impact of NE count on correlation: We investigated the impact of NE count on the correlation improvement. All four test datasets exhibited different sentence characteristics. On the one hand, BNC and ENWIKI comprised 1.07 and 2.07 NEs per sentence; on the other hand, ADGER and ADGER_FILTERED only comprised 0.53 and 0.68 NEs per sentence. In other words, only half of the sentences in the test corpus have one NE. This observation indicates that the higher the number of NEs in the sentence bigger the improvement in correlation with human acceptability judgment. E.g., ENWIKI enjoyed the maximum number of NEs (2.07) per sentence, resulting in the maximum gain (50% for Mean LP) in correlation. The above result is aligned with our hypothesis since the higher the number of NEs in a sentence, the more abstract the sentence representation, resulting in less dependency on word choice and their lexical frequencies for acceptability evaluation.

Impact of NE Type on Performance: Fig. 1 shows the distribution of different NE types across four test datasets. BNC and ENWIKI displayed different sentence characteristics from ADGER and ADGER_FILTERED. Y-axis (log scaled) is the percentage of NE types over all NE count from the dataset for eighteen NE types (X-axis) across four test datasets. We observed that NE type PERSON was the most prominent, i.e., $\approx$ 80% for ADGER and ADGER_FILTERED vs. $\approx$ 20% for BNC and ENWIKI. Furthermore, not a single sentence in ADGER and ADGER_FILTERED possessed the following eleven NE types starting from QUANTITY, to EVENT, WORK_OF_ART as shown in Fig. 1 leading to different performance across four test datasets.

6 Qualitative Analysis

To give an intuition for our proposed methodology, we present one example, sentence s.157, from the ENWIKI test dataset in Table 9. With the full range of values, we apply a Z-score transformation to each of the values in Y (acceptability score) by subtracting the mean of Y from each of the values and dividing them by the standard deviation of Y. We applied the Z-score transformation on human acceptability ratings for the original sentence. Furthermore, for sentences preprocessed via SP and SP + RNE, we applied the Z-score to the SLOR scores predicted from 4-gram LM.

In the first sentence, there exist three NEs 'Myrtle Beach', 'Coast RTA', and 'Pee Dee Regional Transportation Authority'. SP + RNE replaces NE 'Myrtle Beach' with GPE, and 'Coast RTA', 'Pee Dee Regional Transportation Authority' with ORG. Consequently generating a coarse representation of the sentence, allowing LM to focus on the POS transition rather than being swamped by the long-phrase corresponding to NEs. Z-score of 1.025 from LM trained via SP + RNE is comparable to 1.033 for human ratings (with + sign signifying acceptable sentence), unlike the Z-score of -1.070 from LM trained on SP (with - sign signifying unacceptable sentence), which further supports our above claim.

Preprocessing	Sentence	Z-score
- (Original)	Myrtle Beach is served by the Coast RTA and the Pee Dee Regional Transportation Authority .	1.033
SP	myrtle beach is served by the coast UNK and the pee dee regional transportation authority .	-1.070
SP + RNE (Our's)	GPE is served by ORG and ORG .	1.025

Table 9: Sentence s.157 from ENWIKI dataset with unprocessed, standard and NE replacement preprocessing methods and it's corresponding Z-score. UP, SP and RNE corresponds to UnProcessed, Standard Preprossed, and our's proposed NE Replaced data preprocessing methodologies. UNK (i.e., Unknown) corresponds to word 'RTA' as OOV word.

7 Related Work

A series of recent successes of LMs on several NLP tasks have raised the critical question of automatic acceptability tests. Although, there have been several studies to access the acceptability automatically. However, there has not been enough effort to evaluate the impact of sentence representation's on different levels (i.e., granular vs. coarse) on acceptability.

Wan et al. (2005) was the first work to evaluate sentence acceptability independent of the source content. The authors suggested using grammatical judgments of a parser to assess the sentence acceptability. The motivation behind the idea was that if the parser is trained on the appropriate corpus, the poor performance of the parser on one sentence relative to the other sentence will suggest the presence of ungrammaticality and unacceptability. (Mutton et al., 2007) later extended the idea by training the machine learners on top of several parser outputs and showing its correlation with the human judgment of acceptability test. Unlike this line of work, we do not rely on the grammatical assessments from the parser but instead rely on the probabilities assigned by the LM for acceptability test.

Lau et al. (2017) proposed the first work to hold a probabilistic view on linguistic knowledge. They proposed and experimented on a comprehensive list of different probability-based metrics at the sentence level and word level. Taking this work forward (Kann et al., 2018) further introduced WPSLOR, a WordPiece-based version of SLOR, to reduce the LM size. Complementary to this work of exploring different probability-based metrics, we focused on varying levels of sentence representation (granular vs. coarse). Furthermore, we demonstrate that our data transformation strategy can lead to an additional gain in PCC measure between the metrics proposed by (Lau et al., 2017) and human acceptability judgments.

Motivated by centering theory (Grosz et al., 1995), Lapata and Barzilay (2005) argue that patterns of local entity transitions specify how the focus of discourse changes from sentence to sentence. To expose the entity transition patterns of readable texts, they represented a text by an entity grid and showed coherent texts exhibits certain regularities reflected in the topology of grid columns (i.e., discourse entities). This work inspires our idea to use NEs transition, but with few differences. First, our job is to evaluate the acceptability at the intra-sentence level, whereas their goal was to assess the coherence at the inter-sentence level. Second, they created a new representation of the input text in the form of the grid and trained ML learners on top of the grid to evaluate the coherence. However, we replaced the entities with their respective class in the sentence, used this coarse abstract representation, and relied on the LM to learn the acceptability measure from sentence structure.

Brown et al. (1992) presented a statistical algorithm for assigning words to classes (clusters) based on their frequency of co-occurrence with other words. Their method extracted classes with syntactic or semantic-based groupings of words and later proposed class-based n-gram LMs. Though our work is a specific instance of their approach, it answers two crucial questions for sentence transformation required for acceptability evaluation. First, which type of words to replace? Second, what should a word be replaced with? As we explained in section 2 only replacing proper nouns with NE Types generates required coarse-level representation

without affecting the acceptability evaluation of the sentence. E.g., in the sentence "He sees the world from his *eyes*" randomly replacing the word 'eyes' with the other word 'mouth' from the same class or logical class name 'body parts' will drastically impact the acceptability measure of the sentence.

Pitler and Nenkova (2008) combined lexical, syntactic, and discourse features to produce a model to predict the human reader's judgments of text acceptability. They presented that discourse relations are strongly associated with the perceived quality of a text. Similarly, Vadlapudi and Katragadda (2010) proposed surface features like n-gram probabilities, 3-gram based class n-grams, hybrid model using both n-gram and class model on the POS-tag sequences and POS-chunk-tag sequences. They showed that their proposed models, especially the hybrid approach on the POS-chunk-tag sequence, can highly correlate with the human judgment of acceptability. Unlike this line of work, we kept the focus on NEs and proposed a data transformation methodology independent of both the LM and the metric used to measure the correlation with human acceptability judgment.

8 Conclusion

Several probability metrics have been proposed to conduct the reference less acceptability evaluation of a sentence automatically. SLOR in particular, has gained popularity given it removes the impact from the confounding factors of noise like sentence length and lexical frequency. We assert that the issues related to word choice and its lexical frequency persist for SLOR. We proposed a data preprocessing strategy motivated by humans who evaluate the sentence based on sentence structure and transitions between POS at a coarser sentence representation. RNE, our proposed method, identifies all NEs in a sentence and replaces it with the classified type of NE. Based on the results of the experiments, we found a correlation between NEs count in a sentence and improvement in LMs acceptability score. We observed an improvement (up to 52%) or equal performance for three (i.e., BNC, ENWIKI, and ADGER) out of four English datasets. In this work, we only focused on one class of POS, i.e., NE. In the future, we would like to find an optimal dynamic representation of a sentence based on its content to help LM predict acceptability scores better.

References

David Adger. 2003. Core syntax: A minimalist approach.

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.

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Comput. Linguistics*, 18:467–479.

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

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

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *CoNLL*.

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

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41 5:1202–1241.

David Manning. 2002. Probabilistic syntax.

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*, pages 344–351, Prague, Czech Republic, June. Association for Computational Linguistics.

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*, pages 186–195, Honolulu, Hawaii, October. Association for Computational Linguistics.

Jon Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*.

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*, 18:103–118.

The Center for Linguistic Theory and Studies in Probability. 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.

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*, pages 7–12, Los Angeles, CA, June. Association for Computational Linguistics.

Stephen Wan, R. Dale, and Mark Dras. 2005. Searching for grammaticality: Propagating dependencies in the viterbi algorithm. In *ENLG*.

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

A Metrics

A.1 Test Dataset Details

Humans annotated SMOG on three modes of presentation. (a) binary (MOP2), where annotators choose between two options: unacceptable and acceptable (b) four-category (MOP4), where annotators choose between four options: highly unacceptable, somewhat unacceptable, somewhat acceptable, and highly acceptable. (c) a sliding scale (MOP100) with two extremes, highly unacceptable and highly acceptable.

BNC test corpus comprised of 500 random sentences (in English) from the BNC training corpus with a length of 8-25 words. These 500 sentences were machine-translated (using Google Translate) to four target languages, i.e., Norwegian, Spanish, Chinese, and Japanese, and then back to English. This led to 2500 sentences, i.e. 500 *en* original and 500 each from back translation of original *en* to four target languages i.e, *es, no, zh, ja* and then back to en. BNC test corpus comprised 5250 annotated sentences, 2500 on MOP2, 2500 on MOP4, and 250 (10% randomly selected from 2500) on MOP100. To keep the training and test corpus distinct, we removed 500 English test sentences from the BNC training corpus.

Furthermore, SMOG comprised of linguistic sentences adopted from (Adger, 2003)'s syntax textbook. Lau et al. (2017) selected 100 random sentences from (Adger, 2003) where half of them were good (grammatical on author's judgment), and half of them starred (ungrammatical on author's judgment). To focus on syntactic violations, authors created another dataset, ADGER-FILTERED, after filtering out all sentences from (Adger, 2003) that were semantically or pragmatically anomalous. So that the left sentences only consisted of sentences that are either syntactic well-formed or syntactic violations.

A.2 Sentence Level Metrics Formulations

Log Probability In Equation 4 LogProb relates to the log of the sentence probability assigned by the LM.

$$\text{LogProb} = \log p_m(S) \tag{4}$$

Mean Log Probability In Equation 5 Mean LP relates to mean (i.e. average) log of the sentence probability. Which is calculated by dividing the log of the sentence probability with the length of the sentence.

$$\text{Mean LP } = \frac{\log p_m(S)}{|S|} \tag{5}$$

Normalized Log Probability Division In Equation 6 Norm LP (Div) relates to the normalized log probability which is calculated by diving the log of sentence probability with log of the sentence unigram probability.

$$\text{Norm LP (Div) } = -\frac{\log p_m(S)}{\log p_u(S)} \tag{6}$$

Normalized Log Probability Subtraction In Equation 7 Norm LP (Sub) relates to the normalized log probability which is calculated by subtracting the log of the sentence probability with log of sentence unigram probability. Which is also same as log of the division of the sentence probability with the sentence unigram probability.

$$\begin{aligned}\text{Norm LP (Sub)} &= \log p_m(S) - \log p_u(S) \\ &= \log \frac{p_m(S)}{p_u(S)}\end{aligned} \tag{7}$$