How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context

Jey Han Lau[1,7] **Carlos Armendariz**[2] **Shalom Lappin**[2,3,4]
Matthew Purver[2,5] **Chang Shu**[6,7]

[1]The University of Melbourne [2]Queen Mary University of London
[3]University of Gothenburg [4]King's College London
[5]Jožef Stefan Institute [6]University of Nottingham Ningbo China [7]DeepBrain
jeyhan.lau@gmail.com, c.santosarmendariz@qmul.ac.uk
shalom.lappin@gu.se, m.purver@qmul.ac.uk, scxcs1@nottingham.edu.cn

Abstract

We study the influence of context on sentence acceptability. First we compare the acceptability ratings of sentences judged in isolation, with a relevant context, and with an irrelevant context. Our results show that context induces a cognitive load for humans, which compresses the distribution of ratings. Moreover, in relevant contexts we observe a discourse coherence effect that uniformly raises acceptability. Next, we test unidirectional and bidirectional language models in their ability to predict acceptability ratings. The bidirectional models show very promising results, with the best model achieving a new state-of-the-art for unsupervised acceptability prediction. The two sets of experiments provide insights into the cognitive aspects of sentence processing and central issues in the computational modeling of text and discourse.

1 Introduction

Sentence *acceptability* is the extent to which a sentence appears natural to native speakers of a language. Linguists have often used this property to motivate grammatical theories. Computational language processing has traditionally been more concerned with *likelihood*—the probability of a sentence being produced or encountered. The question of whether and how these properties are related is a fundamental one. Lau et al. (2017b) experiment with unsupervised language models to predict acceptability, and they obtained an encouraging correlation with human ratings. This raises foundational questions about the nature of linguistic knowledge: If probabilistic models can acquire knowledge of sentence acceptability from raw texts, we have prima facie support for an alternative view of language acquisition that does not rely on a categorical grammaticality component.

It is generally assumed that our perception of sentence acceptability is influenced by *context*. Sentences that may appear odd in isolation can become natural in some environments, and sentences that seem perfectly well formed in some contexts are odd in others. On the computational side, much recent progress in language modeling has been achieved through the ability to incorporate more document context, using broader and deeper models (e.g., Devlin et al., 2019; Yang et al., 2019). While most language modeling is restricted to individual sentences, models can benefit from using additional context (Khandelwal et al., 2018). However, despite the importance of context, few psycholinguistic or computational studies systematically investigate how context affects acceptability, or the ability of language models to predict human acceptability judgments.

Two recent studies that explore the impact of document context on acceptability judgments both identify a *compression* effect (Bernardy et al., 2018; Bizzoni and Lappin, 2019). Sentences perceived to be low in acceptability when judged without context receive a boost in acceptability when judged within context. Conversely, those with high out-of-context acceptability see a reduction in acceptability when context is presented. It is unclear what causes this compression effect. Is it a result of cognitive load, imposed by additional

Transactions of the Association for Computational Linguistics, vol. 8, pp. 296–310, 2020. https://doi.org/10.1162/tacl_a_00315
Action Editor: George Foster. Submission batch: 10/2019; Revision batch: 1/2020; Published 6/2020.

processing demands, or is it the consequence of an attempt to identify a discourse relation between context and sentence?

We address these questions in this paper. To understand the influence of context on human perceptions, we ran three crowdsourced experiments to collect acceptability ratings from human annotators. We develop a methodology to ensure comparable ratings for each *target sentence* in isolation (without any context), in a relevant three-sentence context, and in the context of sentences randomly sampled from another document. Our results replicate the compression effect, and careful analyses reveal that both cognitive load and discourse coherence are involved.

To understand the relationship between sentence acceptability and probability, we conduct experiments with unsupervised language models to predict acceptability. We explore traditional unidirectional (left-to-right) recurrent neural network models, and modern bidirectional transformer models (e.g., BERT). We found that bidirectional models consistently outperform unidirectional models by a wide margin, calling into question the suitability of left-to-right bias for sentence processing. Our best bidirectional model achieves simulated human performance on the prediction task, establishing a new state-of-the-art.

2 Acceptability in Context

2.1 Data Collection

To understand how humans interpret acceptability, we require a set of sentences with varying degrees of well-formedness. Following previous studies (Lau et al., 2017b; Bernardy et al., 2018), we use round-trip machine translation to introduce a wide range of infelicities into naturally occurring sentences.

We sample 50 English (target) sentences and their contexts (three preceding sentences) from the English Wikipedia.[1] We use Moses to translate the target sentences into four languages (Czech, Spanish, German, and French) and then back to English.[2] This produces 250 sentences in total (5 languages including English) for our *test* set. Note that we only do round-trip translation for the target sentences; the contexts are not modified.

We use Amazon Mechanical Turk (AMT) to collect acceptability ratings for the target sentences.[3] We run three experiments where we expose users to different types of context. For the experiments, we split the test set into 25 HITs of 10 sentences. Each HIT contains 2 original English sentences and 8 round-trip translated sentences, which are different from each other and not derived from either of the originals. Users are asked to rate the sentences for naturalness on a 4-point ordinal scale: bad (1.0), not very good (2.0), mostly good (3.0), and good (4.0). We recruit 20 annotators for each HIT.

In the first experiment we present only the target sentences, without any context. In the second experiment, we first show the context paragraph (three preceding sentences of the target sentence), and ask users to select the most appropriate description of its topic from a list of four candidate topics. Each candidate topic is represented by three words produced by a topic model.[4] Note that the context paragraph consists of original English sentences which did not undergo translation. Once the users have selected the topic, they move to the next screen where they rate the target sentence for naturalness.[5] The third experiment has the same format as the second, except that the three sentences presented prior to rating are randomly sampled from another Wikipedia article.[6] We require annotators to perform a topic identification task prior to rating the target sentence to ensure that they read the context before making acceptability judgments.

For each sentence, we aggregate the ratings from multiple annotators by taking the mean. Henceforth we refer to the mean ratings collected from the first (no context), second (real context), and third (random context) experiments as $\text{H}^{\varnothing}$,

[1] We preprocess the raw dump with WikiExtractor (https://github.com/attardi/wikiextractor), and collect paragraphs that have ≥ 4 sentences with each sentence having ≥ 5 words. Sentences and words are tokenized with spaCy (https://spacy.io/) to check for these constraints.

[2] We use the pre-trained Moses models from http://www.statmt.org/moses/RELEASE-4.0/models/ for translation.

[3] https://www.mturk.com/.

[4] We train a topic model with 50 topics on 15 K Wikipedia documents with Mallet (McCallum, 2002) and infer topics for the context paragraphs based on the trained model.

[5] Note that we do not ask the users to judge the naturalness of the sentence *in context*; the instructions they see for the naturalness rating task is the same as the first experiment.

[6] Sampled sentences are sequential, running sentences.

H^+, and H^-, respectively. We rolled out the experiments on AMT over several weeks and prevented users from doing more than one experiment. Therefore a disjoint group of annotators performed each experiment.

To control for quality, we check that users are rating the English sentences ≥ 3.0 consistently. For the second and third experiments, we also check that users are selecting the topics appropriately. In each HIT one context paragraph has one real topic (from the topic model), and three fake topics with randomly sampled words as the candidate topics. Users who fail to identify the real topic above a confidence level are filtered out. Across the three experiments, over three quarters of workers passed our filtering conditions.

To calibrate for the differences in rating scale between users, we follow the postprocessing procedure of Hill et al. (2015), where we calculate the average rating for each user and the overall average (by taking the mean of all average ratings), and decrease (increase) the ratings of a user by 1.0 if their average rating is greater (smaller) than the overall average by 1.0.[7] To reduce the impact of outliers, for each sentence we also remove ratings that are more than 2 standard deviations away from the mean.[8]

2.2 Results and Discussion

We present scatter plots to compare the mean ratings for the three different contexts ($\text{H}^\varnothing$, H^+, and H^-) in Figure 1. The black line represents the diagonal, and the red line represents the regression line. In general, the mean ratings correlate strongly with each other. Pearson's r for H^+ vs. $\text{H}^\varnothing = 0.940$, H^- vs. $\text{H}^\varnothing = 0.911$, and H^- vs. $\text{H}^+ = 0.891$.

The regression (red) and diagonal (black) lines in H^+ vs. $\text{H}^\varnothing$ (Figure 1a) show a compression effect. Bad sentences appear a little more natural, and perfectly good sentences become slightly less natural, when context is introduced.[9] This is the same compression effect observed by Bernardy et al. (2018). It is also present in the graph for H^- vs. $\text{H}^\varnothing$ (Figure 1b).

Two explanations of the compression effect seem plausible to us. The first is a *discourse coherence* hypothesis that takes this effect to be caused by a general tendency to find infelicitous sentences more natural in context. This hypothesis, however, does not explain why perfectly natural sentences appear less acceptable in context. The second hypothesis is a variant of a *cognitive load* account. In this view, interpreting context imposes a significant burden on a subject's processing resources, and this reduces their focus on the sentence presented for acceptability judgments. At the extreme ends of the rating scale, as they require all subjects to be consistent in order to achieve the minimum/maximum mean rating, the increased cognitive load increases the likelihood of a subject making a mistake. This increases/lowers the mean rating, and creates a compression effect.

The discourse coherence hypothesis would imply that the compression effect should appear with real contexts, but not with random ones, as there is little connection between the target sentence and a random context. By contrast, the cognitive load account predicts that the effect should be present in both types of context, as it depends only on the processing burden imposed by interpreting the context. We see compression in both types of contexts, which suggests that the cognitive load hypothesis is the more likely account.

However, these two hypotheses are not mutually exclusive. It is, in principle, possible that both effects—discourse coherence and cognitive load—are exhibited when context is introduced.

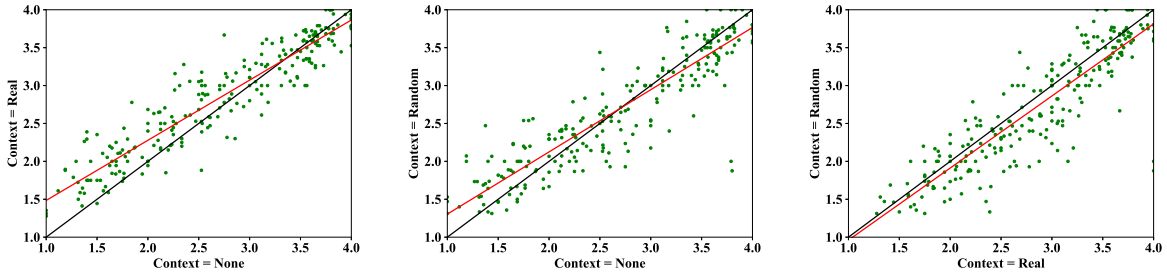
To better understand the impact of discourse coherence, consider Figure 1c, where we compare H^- vs. H^+. Here the regression line is parallel to and below the diagonal, implying that there is a consistent decrease in acceptability ratings from H^+ to H^-. As both ratings are collected with some form of context, the cognitive load confound is removed. What remains is a discourse coherence effect. Sentences presented in relevant contexts undergo a consistent increase in acceptability rating.

To analyze the significance of this effect, we use the non-parametric Wilcoxon signed-rank test (one-tailed) to compare the difference between H^+ and H^-. This gives a p-value of 1.9×10^{-8},

[7] No worker has an average rating that is greater or smaller than the overall average by 2.0.

[8] This postprocessing procedure discarded a total of 504 annotations/ratings (approximately 3.9%) over 3 experiments. The final average number of annotations for a sentence in the first, second, and third experiments is 16.4, 17.8, and 15.3, respectively.

[9] On average, good sentences (ratings ≥ 3.5) observe a rating reduction of 0.08 and bad sentences (ratings ≤ 1.5) an increase of 0.45.



(a) H^{+} vs. $H^{\varnothing}$ (b) H^{-} vs. $H^{\varnothing}$ (c) H^{-} vs. H^{+}

Figure 1: Scatter plots comparing human acceptability ratings.

indicating that the discourse coherence effect is significant.

Returning to Figures 1a and 1b, we can see that (1) the offset of the regression line, and (2) the intersection point of the diagonal and the regression line, is higher in Figure 1a than in Figure 1b. This suggests that there is an increase of ratings, and so, in addition to the cognitive load effect, a discourse coherence effect is also at work in the real context setting.

We performed hypothesis tests to compare the regression lines in Figures 1a and 1b to see if their offsets (constants) and slopes (coefficients) are statistically different.[10] The p-value for the offset is 1.7×10^{-2}, confirming our qualitative observation that there is a significant discourse coherence effect. The p-value for the slope, however, is 3.6×10^{-1}, suggesting that cognitive load compresses the ratings in a consistent way for both H^{+} and H^{-}, relative to $H^{\varnothing}$.

To conclude, our experiments reveal that context induces a cognitive load for human processing, and this has the effect of compressing the acceptability distribution. It moderates the extremes by making very unnatural sentences appear more acceptable, and perfectly natural sentences slightly less acceptable. If the context is relevant to the target sentence, then we also have a discourse coherence effect, where sentences are perceived to be generally more acceptable.

[10]We follow the procedure detailed in `https://statisticsbyjim.com/regression/comparing-regression-lines/` where we collate the data points in Figures 1a and 1b and treat the in-context ratings (H^{+} and H^{-}) as the dependent variable, the out-of-context ratings ($H^{\varnothing}$) as the first independent variable, and the type of the context (real or random) as the second independent variable, to perform regression analyses. The significance of the offset and slope can be measured by interpreting the p-values of the second independent variable, and the interaction between the first and second independent variables, respectively.

3 Modeling Acceptability

In this section, we explore computational models to predict human acceptability ratings. We are interested in models that do not rely on explicit supervision (i.e., we do not want to use the acceptability ratings as labels in the training data). Our motivation here is to understand the extent to which sentence probability, estimated by an unsupervised model, can provide the basis for predicting sentence acceptability.

To this end, we train language models (Section 3.1) using unsupervised objectives (e.g., next word prediction), and use these models to infer the probabilities of our test sentences. To accommodate sentence length and lexical frequency we experiment with several simple normalization methods, converting probabilities to *acceptability measures* (Section 3.2). The acceptability measures are the final output of our models; they are what we use to compare to human acceptability ratings.

3.1 Language Models

Our first model is an LSTM language model (LSTM: Hochreiter and Schmidhuber, 1997; Mikolov et al., 2010). Recurrent neural network models (RNNs) have been shown to be competitive in this task (Lau et al., 2015; Bernardy et al., 2018), and they serve as our baseline.

Our second model is a joint topic and language model (TDLM: Lau et al., 2017a). TDLM combines topic model with language model in a single model, drawing on the idea that the topical context of a sentence can help word prediction in the language model. The topic model is fashioned as an auto-encoder, where the input is the document's word sequence and it is processed by convolutional layers to produce a topic vector to predict the input words. The language model

functions like a standard LSTM model, but it incorporates the topic vector (generated by its document context) into the current hidden state to predict the next word.

We train LSTM and TDLM on 100K uncased English Wikipedia articles containing approximately 40M tokens with a vocabulary of 66K words.[11]

Next we explore transformer-based models, as they have become the benchmark for many NLP tasks in recent years (Vaswani et al., 2017; Devlin et al., 2019; Yang et al., 2019). The transformer models that we use are trained on a much larger corpus, and they are four to five times larger with respect to their model parameters.

Our first transformer is GPT2 (Radford et al., 2019). Given a target word, the input is a sequence of previously seen words, which are then mapped to embeddings (along with their positions) and fed to multiple layers of ''transformer blocks'' before the target word is predicted. Much of its power resides in these transformer blocks: Each provides a multi-headed self-attention unit over all input words, allowing it to capture multiple dependencies between words, while avoiding the need for recurrence. With no need to process a sentence in sequence, the model parallelizes more efficiently, and scales in a way that RNNs cannot.

GPT2 is trained on WebText, which consists of over 8 million web documents, and uses Byte Pair Encoding (BPE: Sennrich et al., 2016) for tokenization (casing preserved). BPE produces sub-word units, a middle ground between word and character, and it provides better coverage for unseen words. We use the released medium-sized model (''Medium'') for our experiments.[12]

Our second transformer is BERT (Devlin et al., 2019). Unlike GPT2, BERT is not a typical language model, in the sense that it has access to both left and right context words when predicting the target word.[13] Hence, it encodes context in a bidirectional manner.

To train BERT, Devlin et al. (2019) propose a masked language model objective, where a random proportion of input words are masked and the model is tasked to predict them based on non-masked words. In addition to this objective, BERT is trained with a next sentence prediction objective, where the input is a pair of sentences, and the model's goal is to predict whether the latter sentence follows the former. This objective is added to provide pre-training for downstream tasks that involve understanding the relationship between a pair of sentences (e.g., machine comprehension and textual entailment).

The bidirectionality of BERT is the core feature that produces its state-of-the-art performance on a number of tasks. The flipside of this encoding style, however, is that BERT lacks the ability to generate left-to-right and compute sentence probability. We discuss how we use BERT to produce a probability estimate for sentences in the next section (Section 3.2).

In our experiments, we use the largest pre-trained model (''BERT-Large''),[14] which has a similar number of parameters (340M) to GPT2. It is trained on Wikipedia and BookCorpus (Zhu et al., 2015), where the latter is a collection of fiction books. Like GPT2, BERT also uses sub-word tokenization (WordPiece). We experiment with two variants of BERT: one trained on cased data (BERT_{CS}), and another on uncased data (BERT_{UCS}). As our test sentences are uncased, a comparison between these two models allows us to gauge the impact of casing in the training data.

Our last transformer model is XLNET (Yang et al., 2019). XLNET is unique in that it applies a novel permutation language model objective, allowing it to capture bidirectional context while preserving key aspects of unidirectional language models (e.g., left-to-right generation).

The permutation language model objective works by first generating a possible permutation (also called ''factorization order'') of a sequence. When predicting a target word in the sequence, the context words that the model has access to are determined by the factorization order. To illustrate this, imagine we have the sequence $\mathbf{x} = [x_1, x_2, x_3, x_4]$. One possible factorization order is: $x_3 \rightarrow x_2 \rightarrow x_4 \rightarrow x_1$. Given this order, if predicting target word x_4, the model only has access to context words $\{x_3, x_2\}$; if the target word is x_2, it sees only $\{x_3\}$. In practice, the target word is set to be the last few words in the factorization

[11] We use Stanford CoreNLP (Manning et al., 2014) to tokenize words and sentences. Rare words are replaced by a special UNK symbol.

[12] https://github.com/openai/gpt-2.

[13] Note that *context* is burdened with two senses in the paper. It can mean the preceding sentences of a target sentence, or the neighbouring words of a target word. The intended sense should be apparent from the usage.

[14] https://github.com/google-research/bert.

Model	Configuration			Training Data			
	Architecture	Encoding	#Param.	Casing	Size	Tokenization	Corpora
LSTM	RNN	Unidir.	60M	Uncased	0.2GB	Word	Wikipedia
TDLM	RNN	Unidir.	80M	Uncased	0.2GB	Word	Wikipedia
GPT2	Transformer	Unidir.	340M	Cased	40GB	BPE	WebText
BERT_{CS}	Transformer	Bidir.	340M	Cased	13GB	WordPiece	Wikipedia, BookCorpus
BERT_{UCS}	Transformer	Bidir.	340M	Uncased	13GB	WordPiece	Wikipedia, BookCorpus
XLNET	Transformer	Hybrid	340M	Cased	126GB	Sentence-Piece	Wikipedia, BookCorpus, Giga5 ClueWeb, Common Crawl

Table 1: Language models and their configurations.

order (e.g., x_4 and x_1), and so the model always sees some context words for prediction.

As XLNET is trained to work with different factorization orders during training, it has experienced both full/bidirectional context and partial/unidirectional context, allowing it to adapt to tasks that have access to full context (e.g., most language understanding tasks), as well as those that do not (e.g., left-to-right generation).

Another innovation of XLNET is that it incorporates the segment recurrence mechanism of Dai et al. (2019). This mechanism is inspired by truncated backpropagation through time used for training RNNs, where the initial state of a sequence is initialized with the final state from the previous sequence. The segment recurrence mechanism works in a similar way, by caching the hidden states of the transformer blocks from the previous sequence, and allowing the current sequence to attend to them during training. This permits XLNET to model long-range dependencies beyond its maximum sequence length.

We use the largest pre-trained model (''XLNet-Large''),[15] which has a similar number of parameters to our BERT and GPT2 models (340M). XLNET is trained on a much larger corpus combining Wikipedia, BookCorpus, news and web articles. For tokenization, XLNET uses SentencePiece (Kudo and Richardson, 2018), another sub-word tokenization technique. Like GPT2, XLNET is trained on cased data.

Table 1 summarizes the language models. In general, the RNN models are orders of magnitude smaller than the transformers in both model parameters and training data, although they are trained on the same domain (Wikipedia), and use uncased data as the test sentences. The RNN models also operate on a word level, whereas the transformers use sub-word units.

[15] https://github.com/zihangdai/xlnet.

3.2 Probability and Acceptability Measure

Given a unidirectional language model, we can infer the probability of a sentence by multiplying the estimated probabilities of each token using previously seen (left) words as context (Bengio et al., 2003):

$$\overrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i|w_{<i}) \quad (1)$$

where s is the sentence, and w_i a token in s.

LSTM, TDLM, and GPT2 are unidirectional models, so they all compute sentence probability as described. XLNET's unique permutational language model objective allows it to compute probability in the same way, and to explicitly mark this we denote it as $\text{XLNET}_{\text{UNI}}$ when we infer sentence probability using only left context words.

BERT is trained with bidirectional context, and as such it is unable to compute left-to-right sentence probability.[16] We therefore compute sentence probability as follows:

$$\overleftrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i|w_{<i}, w_{>i}) \quad (2)$$

With this formulation, we allow BERT to have access to both left and right context words when predicting each target word, since this is consistent with the way in which it was trained. It is important to note, however, that sentence probability computed this way is not a *true probability value*: These probabilities do not sum to 1.0 over all sentences. Equation (1), in contrast, does guarantee true probabilities. Intuitively, the sentence probability computed with this bidirectional formulation is a measure

[16] Technically we can mask all right context words and predict the target words one at a time, but because the model is never trained in this way, we found that it performs poorly in preliminary experiments.

of the model's confidence in the likelihood of the sentence.

To compute the true probability, Wang and Cho (2019) show that we need to sum the pre-softmax weights for each token to score a sentence, and then divide the score by the total score of all sentences. As it is impractical to compute the total score of all sentences (an infinite set), the true sentence probabilities for these bidirectional models are intractable. We use our non-normalized confidence scores as stand-ins for these probabilities.

For XLNET, we also compute sentence probability this way, applying bidirectional context, and we denote it as XLNET_{BI}. Note that $\text{XLNET}_{\text{UNI}}$ and XLNET_{BI} are based on the same trained model. They differ only in how they estimate sentence probability at test time.

Sentence probability (estimated either using unidirectional or bidirectional context) is affected by its length (e.g., longer sentences have lower probabilities), and word frequency (e.g., *the cat is big* vs. *the yak is big*). To modulate for these factors we introduce simple normalization techniques. Table 2 presents five methods to map sentence probabilities to *acceptability measures*: *LP*, *MeanLP*, *PenLP*, *NormLP*, and *SLOR*.

LP is the unnormalized log probability. Both *MeanLP* and *PenLP* are normalized on sentence length, but *PenLP* scales length with an exponent (α) to dampen the impact of large values (Wu et al., 2016; Vaswani et al., 2017). We set $\alpha = 0.8$ in our experiments. *NormLP* normalizes using unigram sentence probability (i.e., $P_{\text{u}}(s) = \prod_{i=0}^{|s|} P(w_i)$), while *SLOR* utilizes both length and unigram probability (Pauls and Klein, 2012).

When computing sentence probability we have the option of including the context paragraph that the human annotators see (Section 2). We use the superscripts $\varnothing$, $+$, $-$ to denote a model using no context, real context, and random context, respectively (e.g., $\text{LSTM}^{\varnothing}$, LSTM^{+}, and LSTM^{-}). Note that these variants are created at test time, and are all based on the same trained model (e.g., LSTM).

For all models except TDLM, incorporating the context paragraph is trivial. We simply prepend it to the target sentence before computing the latter's probability. For TDLM^{+} or TDLM^{-}, the context paragraph is treated as the document context, from which a topic vector is inferred and fed to the language model for next-word prediction. For $\text{TDLM}^{\varnothing}$, we set the topic vector to zeros.

Acc. Measure	Equation
LP	$\log P(s)$
MeanLP	$\frac{\log P(s)}{\lvert s\rvert}$
PenLP	$\frac{\log P(s)}{((5+\lvert s\rvert)/(5+1))^{\alpha}}$
NormLP	$-\frac{\log P(s)}{\log P_{\text{u}}(s)}$
SLOR	$\frac{\log P(s) - \log P_{\text{u}}(s)}{\lvert s\rvert}$

Table 2: Acceptability measures for predicting the acceptability of a sentence; $P(s)$ is the sentence probability, computed using Equation (1) or Equation (2) depending on the model; $P_{\text{u}}(s)$ is the sentence probability estimated by a unigram language model; and α =0.8.

3.3 Implementation

For the transformer models (GPT2, BERT, and XLNET), we use the implementation of *pytorch-transformers*.[17]

XLNET requires a long dummy context prepended to the target sentence for it to compute the sentence probability properly.[18] Other researchers have found a similar problem when using XLNET for generation.[19] We think that this is likely due to XLNET's recurrence mechanism (Section 3.1), where it has access to context from the previous sequence during training.

For TDLM, we use the implementation provided by Lau et al. (2017a),[20] following their optimal hyper-parameter configuration without tuning.

We implement LSTM based on Tensorflow's Penn Treebank language model.[21] In terms of

[17] https://github.com/huggingface/pytorch-transformers. Specifically, we employ the following pre-trained models: gpt2-medium for GPT2, bert-large-cased for BERT_{CS}, bert-large-uncased for BERT_{UCS}, and xlnet-large-cased for $\text{XLNET}_{\text{UNI}}$/$\text{XLNET}_{\text{BI}}$.

[18] In the scenario where we include the context paragraph (e.g., $\text{XLNET}^{+}_{\text{UNI}}$), the dummy context is added before it.

[19] https://medium.com/@amanrusia/xlnet-speaks-comparison-to-gpt-2-ea1a4e9ba39e.

[20] https://github.com/jhlau/topically-driven-language-model.

[21] https://github.com/tensorflow/models/blob/master/tutorials/rnn/ptb/ptb_word_lm.py.

hyper-parameters, we follow the configuration of TDLM where applicable. TDLM uses Adam as the optimizer (Kingma and Ba, 2014), but for LSTM we use Adagrad (Duchi et al., 2011), as it produces better development perplexity.

For *NormLP* and *SLOR*, we need to compute $P_{\mathrm{u}}(s)$, the sentence probability based on a unigram language model. As the language models are trained on different corpora, we collect unigram counts based on their original training corpus. That is, for LSTM and TDLM, we use the 100K English Wikipedia corpus. For GPT2, we use an open source implementation that reproduces the original WebText data.[22] For BERT we use the full Wikipedia collection and crawl `smashwords.com` to reproduce BookCorpus.[23] Finally, for XLNET we use the combined set of Wikipedia, WebText, and BookCorpus.[24]

Source code for our experiments is publicly available at: `https://github.com/jhlau/acceptability-prediction-in-context`.

3.4 Results and Discussion

We use Pearson's r to assess how well the models' acceptability measures predict mean human acceptability ratings, following previous studies (Lau et al., 2017b; Bernardy et al., 2018). Recall that for each model (e.g., LSTM), there are three variants with which we infer the sentence probability at test time. These are distinguished by whether we include no context ($\text{LSTM}^{\varnothing}$), real context ($\text{LSTM}^{+}$), or random context ($\text{LSTM}^{-}$). There are also three types of human acceptability ratings (ground truth), where sentences are judged with no context, ($\text{H}^{\varnothing}$), real context (H^{+}), and random context (H^{-}). We present the full results in Table 3.

To get a sense of what the correlation figures indicate for these models, we compute two human performance estimates to serve as upper bounds on the accuracy of a model. The first upper bound (UB_1) is the one-vs-rest annotator correlation, where we select a random annotator's rating and compare it to the mean rating of the rest, using Pearson's r. We repeat this for a large number of trials (1,000) to get a robust estimate of the mean correlation. UB_1 can be interpreted as the average human performance working in isolation. The second upper bound (UB_2) is the half-vs.-half annotator correlation. For each sentence we randomly split the annotators into two groups, and compare the mean rating between groups, again using Pearson's r and repeating it (1,000 times) to get a robust estimate. UB_2 can be taken as the average human performance working collaboratively. Overall, the simulated human performance is fairly consistent over context types (Table 3), for example, $\text{UB}_1 = 0.75$, 0.73, and 0.75 for $\text{H}^{\varnothing}$, H^{+}, and H^{-}, respectively.

When we postprocess the user ratings, remember that we remove the outlier ratings (≥ 2 standard deviation) for each sentence (Section 2.1). Although this produces a cleaner set of annotations, this filtering step does (artificially) increase the human agreement or upper bound correlations. For completeness we also present upper bound variations where we do not remove the outlier ratings, and denote them as $\text{UB}_1^{\varnothing}$ and $\text{UB}_2^{\varnothing}$. In this setup, the one-vs.-rest correlations drop to 0.62–0.66 (Table 3). Note that all model performances are reported based on the outlier-filtered ratings, although there are almost no perceivable changes to the performances when they are evaluated on the outlier-preserved ground truth.

Looking at Table 3, the models' performances are fairly consistent over different types of ground truths ($\text{H}^{\varnothing}$, H^{+}, and H^{-}). This is perhaps not very surprising, as the correlations among the human ratings for these context types are very high (Section 2).

We now focus on the results with $\text{H}^{\varnothing}$ as ground truth (''Rtg'' $= \text{H}^{\varnothing}$). *SLOR* is generally the best acceptability measure for unidirectional models, with *NormLP* not far behind (the only exception is $\text{GPT2}^{\varnothing}$). The recurrent models (LSTM and TDLM) are very strong compared with the much larger transformer models (GPT2 and $\text{XLNET}_{\text{UNI}}$). In fact TDLM has the best performance when context is not considered ($\text{TDLM}^{\varnothing}$, $SLOR = 0.61$), suggesting that model architecture may be more important than number of parameters and amount of training data.

For bidirectional models, the unnormalized *LP* works very well. The clear winner here, however,

[22] `https://skylion007.github.io/OpenWebTextCorpus/`.

[23] We use the scripts in `https://github.com/soskek/bookcorpus` to reproduce BookCorpus.

[24] XLNET also uses Giga5 and ClueWeb as part of its training data, but we think that our combined collection is sufficiently large to be representative of the original training data.

Rtg	Encod.	Model	*LP*	*MeanLP*	*PenLP*	*NormLP*	*SLOR*
$H^{\varnothing}$	Unidir.	$\text{LSTM}^{\varnothing}$	0.29	0.42	0.42	0.52	**0.53**
		LSTM^{+}	0.30	0.49	0.45	0.61	**0.63**
		$\text{TDLM}^{\varnothing}$	0.30	0.49	0.45	0.60	**0.61**
		TDLM^{+}	0.30	0.50	0.45	0.59	**0.60**
		$\text{GPT2}^{\varnothing}$	0.33	0.34	**0.56**	0.38	0.38
		GPT2^{+}	0.38	0.59	0.58	**0.63**	0.60
		$\text{XLNET}^{\varnothing}_{\text{UNI}}$	0.31	0.42	0.51	0.51	**0.52**
		$\text{XLNET}^{+}_{\text{UNI}}$	0.36	0.56	0.55	0.61	**0.61**
	Bidir.	$\text{BERT}^{\varnothing}_{\text{CS}}$	0.51	0.54	**0.63**	0.55	0.53
		$\text{BERT}^{+}_{\text{CS}}$	0.53	0.63	**0.67**	0.64	0.60
		$\text{BERT}^{\varnothing}_{\text{UCS}}$	0.59	0.63	**0.70**	0.63	0.60
		$\text{BERT}^{+}_{\text{UCS}}$	0.60	0.68	**0.72**	0.67	0.63
		$\text{XLNET}^{\varnothing}_{\text{BI}}$	0.52	0.51	**0.66**	0.53	0.53
		$\text{XLNET}^{+}_{\text{BI}}$	0.57	0.65	**0.73**	0.66	0.65
	—	UB_1 / $\text{UB}_1^{\varnothing}$			0.75 / 0.66		
		UB_2 / $\text{UB}_2^{\varnothing}$			0.92 / 0.88		
H^{+}	Unidir.	$\text{LSTM}^{\varnothing}$	0.29	0.44	0.43	**0.52**	**0.52**
		LSTM^{+}	0.31	0.51	0.46	**0.62**	**0.62**
		$\text{TDLM}^{\varnothing}$	0.30	0.50	0.45	**0.59**	**0.59**
		TDLM^{+}	0.30	0.50	0.46	**0.58**	**0.58**
		$\text{GPT2}^{\varnothing}$	0.32	0.33	**0.56**	0.36	0.37
		GPT2^{+}	0.38	0.60	0.59	**0.63**	0.60
		$\text{XLNET}^{\varnothing}_{\text{UNI}}$	0.30	0.42	0.50	0.49	**0.51**
		$\text{XLNET}^{+}_{\text{UNI}}$	0.35	0.56	0.55	0.60	**0.61**
	Bidir.	$\text{BERT}^{\varnothing}_{\text{CS}}$	0.49	0.53	**0.62**	0.54	0.51
		$\text{BERT}^{+}_{\text{CS}}$	0.52	0.63	**0.66**	0.63	0.58
		$\text{BERT}^{\varnothing}_{\text{CS}}$	0.58	0.63	**0.70**	0.63	0.60
		$\text{BERT}^{+}_{\text{CS}}$	0.60	0.68	**0.73**	0.67	0.63
		$\text{XLNET}^{\varnothing}_{\text{BI}}$	0.51	0.50	**0.65**	0.52	0.53
		$\text{XLNET}^{+}_{\text{BI}}$	0.57	0.65	**0.74**	0.65	0.65
	—	UB_1 / $\text{UB}_1^{\varnothing}$			0.73 / 0.66		
		UB_1 / $\text{UB}_2^{\varnothing}$			0.92 / 0.89		
H^{-}	Unidir.	$\text{LSTM}^{\varnothing}$	0.28	0.44	0.43	**0.50**	**0.50**
		LSTM^{-}	0.27	0.41	0.40	**0.47**	**0.47**
		$\text{TDLM}^{\varnothing}$	0.29	0.52	0.46	**0.59**	0.58
		TDLM^{-}	0.28	0.49	0.44	**0.56**	0.55
		$\text{GPT2}^{\varnothing}$	0.32	0.34	**0.55**	0.35	0.35
		GPT2^{-}	0.30	0.42	**0.51**	0.44	0.41
		$\text{XLNET}^{\varnothing}_{\text{UNI}}$	0.30	0.44	**0.51**	0.49	0.49
		$\text{XLNET}^{-}_{\text{UNI}}$	0.29	0.40	**0.49**	0.46	0.46
	Bidir.	$\text{BERT}^{\varnothing}_{\text{CS}}$	0.48	0.53	**0.62**	0.53	0.49
		$\text{BERT}^{-}_{\text{CS}}$	0.49	0.52	**0.61**	0.51	0.47
		$\text{BERT}^{\varnothing}_{\text{UCS}}$	0.56	0.61	**0.68**	0.60	0.56
		$\text{BERT}^{-}_{\text{UCS}}$	0.56	0.58	**0.66**	0.57	0.53
		$\text{XLNET}^{\varnothing}_{\text{BI}}$	0.49	0.48	**0.62**	0.49	0.48
		$\text{XLNET}^{-}_{\text{BI}}$	0.50	0.51	**0.64**	0.51	0.50
	—	UB_1 / $\text{UB}_1^{\varnothing}$			0.75 / 0.68		
		UB_2 / $\text{UB}_2^{\varnothing}$			0.92 / 0.88		

Table 3: Modeling results. Boldface indicates optimal performance in each row.

is *PenLP*. It substantially and consistently outperforms all other acceptability measures. The strong performance of *PenLP* that we see here illuminates its popularity in machine translation for beam search decoding (Vaswani et al., 2017). With the exception of *PenLP*, the gain from normalization for the bidirectional models is small, but we don't think this can be attributed to the size of models or training corpora, as the large unidirectional models (GPT2 and $\text{XLNET}_{\text{UNI}}$) still benefit from normalization. The best model without considering context is $\text{BERT}^{\varnothing}_{\text{UCS}}$ with a correlation of 0.70 (*PenLP*), which is very close to the idealized single-annotator performance UB_1 (0.75) and surpasses the unfiltered performance $\text{UB}^{\varnothing}_1$ (0.66), creating a new state-of-the-art for unsupervised acceptability prediction (Lau et al., 2015, 2017b; Bernardy et al., 2018). There is still room to improve, however, relative to the collaborative UB_2 (0.92) or $\text{UB}^{\varnothing}_2$ (0.88) upper bounds.

We next look at the impact of incorporating context at test time for the models (e.g., $\text{LSTM}^{\varnothing}$ vs. LSTM^{+} or $\text{BERT}^{\varnothing}_{\text{UCS}}$ vs. $\text{BERT}^{+}_{\text{UCS}}$). To ease interpretability we will focus on *SLOR* for unidirectional models, and *PenLP* for bidirectional models. Generally, we see that incorporating context always improves correlation, for both cases where we use $\text{H}^{\varnothing}$ and H^{+} as ground truths, suggesting that context is beneficial when it comes to sentence modeling. The only exception is TDLM, where $\text{TDLM}^{\varnothing}$ and TDLM^{+} perform very similarly. Note, however, that context is only beneficial when it is relevant. Incorporating random contexts (e.g., $\text{LSTM}^{\varnothing}$ vs. LSTM^{-} or $\text{BERT}^{\varnothing}_{\text{UCS}}$ vs. $\text{BERT}^{-}_{\text{UCS}}$ with H^{-} as ground truth) reduces the performance for all models.[25]

Recall that our test sentences are uncased (an artefact of Moses, the machine translation system that we use). Whereas the recurrent models are all trained on uncased data, most of the transformer models are trained with cased data. BERT is the only transformer that is pre-trained on both cased (BERT_{CS}) and uncased data (BERT_{UCS}). To understand the impact of casing, we look at the performance of BERT_{CS} and BERT_{UCS} with $\text{H}^{\varnothing}$ as ground truth. We see an improvement of 5–7 points (depending on whether context is incorporated), which suggests that casing has a significant impact on performance. Given that $\text{XLNET}^{+}_{\text{BI}}$ already outperforms $\text{BERT}^{+}_{\text{UCS}}$ (0.73 vs. 0.72), even though $\text{XLNET}^{+}_{\text{BI}}$ is trained with cased data, we conjecture that an uncased XLNET is likely to outperform $\text{BERT}^{\varnothing}_{\text{UCS}}$ when context is not considered.

To summarize, our first important result is the exceptional performance of bidirectional models. It raises the question of whether left-to-right bias is an appropriate assumption for predicting sentence acceptability. One could argue that this result may be due to our experimental setup. Users are presented with the sentence in text, and they have the opportunity to read it multiple times, thereby creating an environment that may simulate bidirectional context. We could test this conjecture by changing the presentation of the sentence, displaying it one word at a time (with older words fading off), or playing an audio version (e.g., via a text-to-speech system). However, these changes will likely introduce other confounds (e.g., prosody), but we believe it is an interesting avenue for future work.

Our second result is more tentative. Our experiments seem to indicate that model architecture is more important than training or model size. We see that TDLM, which is trained on data orders of magnitude smaller and has model parameters four times smaller in size (Table 1), outperforms the large unidirectional transformer models. To establish this conclusion more firmly we will need to rule out the possibility that the relatively good performance of LSTM and TDLM is not due to a cleaner (e.g., lowercased) or more relevant (e.g., Wikipedia) training corpus. With that said, we contend that our findings motivate the construction of better language models, instead of increasing the number of parameters, or the amount of training data. It would be interesting to examine the effect of extending TDLM with a bidirectional objective.

Our final result is that our best model, BERT_{UCS}, attains a human-level performance and achieves a new state-of-the-art performance in the task of unsupervised acceptability prediction. Given this level of accuracy, we expect it would be suitable for tasks like assessing student essays and the quality of machine translations.

[25] There is one exception: $\text{XLNET}^{\varnothing}_{\text{BI}}$ (0.62) vs. $\text{XLNET}^{-}_{\text{BI}}$ (0.64). As we saw previously in Section 3.3, XLNET requires a long dummy context to work, and so this observation is perhaps unsurprising, because it appears that context—whether it is relevant or not—seems to always benefit XLNET.

4 Linguists' Examples

One may argue that our dataset is potentially biased, as round-trip machine translation may introduce particular types of infelicities or unusual features to the sentences (Graham et al., 2019). Lau et al. (2017b) addressed this by creating a dataset where they sample 50 grammatical and 50 ungrammatical sentences from Adger (2003)'s syntax textbook, and run a crowdsourced experiment to collect their user ratings. Lau et al. (2017b) found that their unsupervised language models (e.g., simple recurrent networks) predict the acceptability of these sentences with similar performances, providing evidence that their modeling results are robust.

We test our pre-trained models using this linguist-constructed dataset, and found similar observations: GPT2, BERT_{CS}, and XLNET_{BI} produce a *PenLP* correlation of 0.45, 0.53, and 0.58, respectively. These results indicate that these language models are able to predict the acceptability of these sentences reliably, consistent with our modeling results with round-trip translated sentences (Section 3.4). Although the correlations are generally lower, we want to highlight that these linguists' examples are artificially constructed to illustrate specific syntactic phenomena, and so this constitutes a particularly strong case of out-of-domain prediction. These texts are substantially different in nature from the natural text that the pre-trained language models are trained on (e.g., the linguists' examples are much shorter—less than 7 words on average—than the natural texts).

5 Related Work

Acceptability is closely related to the concept of grammaticality. The latter is a theoretical construction corresponding to syntactic well-formedness, and it is typically interpreted as a binary property (i.e., a sentence is either grammatical or ungrammatical). Acceptability, on the other hand, includes syntactic, semantic, pragmatic, and non-linguistic factors, such as sentence length. It is gradient, rather than binary, in nature (Denison, 2004; Sorace and Keller, 2005; Sprouse, 2007).

Linguists and other theorists of language have traditionally assumed that context affects our perception of both grammaticality (Bolinger, 1968) and acceptability (Bever, 1970), but surprisingly little work investigates this effect systematically, or on a large scale. Most formal linguists rely heavily on the analysis of sentences taken in isolation. However, many linguistic frameworks seek to incorporate aspects of context-dependence. Dynamic theories of semantics (Heim, 1982; Kamp and Reyle, 1993; Groenendijk and Stokhof, 1990) attempt to capture intersentential coreference, binding, and scope phenomena. Dynamic Syntax (Cann et al., 2007) uses incremental tree construction and semantic type projection to render parsing and interpretation discourse dependent. Theories of discourse structure characterize sentence coherence in context through rhetorical relations (Mann and Thompson, 1988; Asher and Lascarides, 2003), or by identifying open questions and common ground (Ginzburg, 2012). While these studies offer valuable insights into a variety of context related linguistic phenomena, much of it takes grammaticality and acceptability to be binary properties. Moreover, it is not formulated in a way that permits fine-grained psychological experiments, or wide coverage computational modeling.

Psycholinguistic work can provide more experimentally grounded approaches. Greenbaum (1976) found that combinations of particular syntactic constructions in context affect human judgments of acceptability, although the small scale of the experiments makes it difficult to draw general conclusions. More recent work investigates related effects, but it tends to focus on very restricted aspects of the phenomenon. For example, Zlogar and Davidson (2018) investigate the influence of context on the acceptability of gestures with speech, focussing on interaction with semantic content and presupposition. The *priming* literature shows that exposure to lexical and syntactic items leads to higher likelihood of their repetition in production (Reitter et al., 2011), and to quicker processing in parsing under certain circumstances (Giavazzi et al., 2018). Frameworks such as ACT-R (Anderson, 1996) explain these effects through the impact of cognitive activation on subsequent processing. Most of these studies suggest that coherent or natural contexts should increase acceptability ratings, given that the linguistic expressions used in processing become more activated. Warner and Glass (1987) show that such syntactic contexts can indeed affect grammaticality judgments in the expected way for

garden path sentences. Cowart (1994) uses comparison between positive and negative contexts, investigating the effect of contexts containing alternative more or less acceptable sentences. But he restricts the test cases to specific pronoun binding phenomena. None of the psycholinguistic work investigates acceptability judgments in real textual contexts, over large numbers of test cases and human subjects.

Some recent computational work explores the relation of acceptability judgments to sentence probabilities. Lau et al. (2015, 2017b) show that the output of unsupervised language models can correlate with human acceptability ratings. Warstadt et al. (2018) treat this as a semi-supervised problem, training a binary classifier on top of a pre-trained sentence encoder to predict acceptability ratings with greater accuracy. Bernardy et al. (2018) explore incorporating context into such models, eliciting human judgments of sentence acceptability when the sentences were presented both in isolation and within a document context. They find a compression effect in the distribution of the human acceptability ratings. Bizzoni and Lappin (2019) observe a similar effect in a paraphrase acceptability task.

One possible explanation for this compression effect is to take it as the expression of cognitive load. Psychological research on the cognitive load effect (Sweller, 1988; Ito et al., 2018; Causse et al., 2016; Park et al., 2013) indicates that performing a secondary task can degrade or distort subjects' performance on a primary task. This could cause judgments to regress towards the mean. However, the experiments of Bernardy et al. (2018) and Bizzoni and Lappin (2019) do not allow us to distinguish this possibility from a coherence or priming effect, as only coherent contexts were considered. Our experimental setup improves on this by introducing a topic identification task and incoherent (random) contexts in order to tease the effects apart.

6 Conclusions and Future Work

We found that processing context induces a cognitive load for humans, which creates a compression effect on the distribution of acceptability ratings. We also showed that if the context is relevant to the sentence, a discourse coherence effect uniformly boosts sentence acceptability. Our language model experiments indicate that bidirectional models achieve better results than unidirectional models. The best bidirectional model performs at a human level, defining a new state-of-the art for this task.

In future work we will explore alternative ways to present sentences for acceptability judgments. We plan to extend TDLM, incorporating a bidirectional objective, as it shows significant promise. It will also be interesting to see if our observations generalize to other languages, and to different sorts of contexts, both linguistic and non-linguistic.

Acknowledgments

We are grateful to three anonymous reviewers for helpful comments on earlier drafts of this paper. Some of the work described here was presented in talks in the seminar of the Centre for Linguistic Theory and Studies in Probability (CLASP), University of Gothenburg, December 2019, and in the Cambridge University Language Technology Seminar, February 2020. We thank the participants of both events for useful discussion.

Lappin's work on the project was supported by grant 2014-39 from the Swedish Research Council, which funds CLASP. Armendariz and Purver were partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

References

David Adger. 2003. *Core Syntax: A Minimalist Approach*, Oxford University Press, United Kingdom.

John R. Anderson. 1996. ACT: A simple theory of complex cognition. *American Psychologist*, 51:355–365.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*, Cambridge University Press.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural

probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. 2018. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 456–461. Melbourne, Australia.

Thomas G. Bever. 1970, The cognitive basis for linguistic structures, J. R. Hayes, editor, *Cognition and the Development of Language*, Wiley, New York, pages 279–362.

Yuri Bizzoni and Shalom Lappin. 2019. The effect of context on metaphor paraphrase aptness judgments. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 165–175. Gothenburg, Sweden.

Dwight Bolinger. 1968. Judgments of grammaticality. *Lingua*, 21:34–40.

Ronnie Cann, Ruth Kempson, and Matthew Purver. 2007. Context and well-formedness: the dynamics of ellipsis. *Research on Language and Computation*, 5(3):333–358.

Mickaël Causse, Vsevolod Peysakhovich, and Eve F. Fabre. 2016. High working memory load impairs language processing during a simulated piloting task: An ERP and pupillometry study. *Frontiers in Human Neuroscience*, 10:240.

Wayne Cowart. 1994. Anchoring and grammar effects in judgments of sentence acceptability. *Perceptual and Motor Skills*, 79(3):1171–1182.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860.

David Denison. 2004. *Fuzzy Grammar: A Reader*, Oxford University Press, United Kingdom.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186. Minneapolis, Minnesota.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Maria Giavazzi, Sara Sambin, Ruth de Diego-Balaguer, Lorna Le Stanc, Anne-Catherine Bachoud-Lévi, and Charlotte Jacquemot. 2018. Structural priming in sentence comprehension: A single prime is enough. *PLoS ONE*, 13(4):e0194959.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*, Oxford University Press.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Sidney Greenbaum. 1976. Contextual influence on acceptability judgements. *Linguistics*, 15(187):5–12.

Jeroen Groenendijk and Martin Stokhof. 1990. Dynamic Montague grammar. L. Kalman and L. Polos, editors, In *Proceedings of the 2nd Symposium on Logic and Language*, pages 3–48. Budapest.

Irene Heim. 1982. The Semantics of Definite and Indefinite Noun Phrases. Ph.D. thesis, University of Massachusetts at Amherst.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Aine Ito, Martin Corley, and Martin J. Pickering. 2018. A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition*, 21(2):251–264.

Hans Kamp and Uwe Reyle. 1993. *From Discourse To Logic*, Kluwer Academic Publishers.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294. Association for Computational Linguistics, Melbourne, Australia.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. Brussels, Belgium.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017a. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365. Vancouver, Canada.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the Joint conference of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, pages 1618–1628. Beijing, China.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017b. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41:1202–1241.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048. Makuhari, Japan.

Hyangsook Park, Jun-Su Kang, Sungmook Choi, and Minho Lee. 2013. Analysis of cognitive load for language processing based on brain activities. In *Neural Information Processing*, pages 561–568. Springer Berlin Heidelberg, Berlin, Heidelberg.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968. Jeju Island, Korea.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Daivd Reitter, Frank Keller, and Johanna D. Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive Science*, 35(4):587–637.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Berlin, Germany.

Antonella Sorace and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115:1497–1524.

Jon Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, 1123–134.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36. Association for Computational Linguistics, Minneapolis, Minnesota.

John Warner and Arnold L. Glass. 1987. Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language*, 26(6):714 – 738.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27. Washington, DC, USA.

Christina Zlogar and Kathryn Davidson. 2018. Effects of linguistic context on the acceptability of co-speech gestures. *Glossa*, 3(1):73.