Do We Need Neural Models to Explain Human Judgments of Acceptability?

Wang Jing (wangjingrs@bisu.edu.cn)
Beijing International Studies University, Beijing, China

M. A. Kelly (mak582@psu.edu)
David Reitter (reitter@google.com)
The Pennsylvania State University, University Park, PA, USA

Abstract

Native speakers can judge whether a sentence is an acceptable instance of their language. Acceptability provides a means of evaluating whether computational language models are processing language in a human-like manner. We test the ability of language models, simple language features, and word embeddings to predict native speakers' judgments of acceptability on English essays written by non-native speakers. We find that much sentence acceptability variance can be captured by a combination of misspellings, word order, and word similarity ($r = 0.494$). While predictive neural models fit acceptability judgments well ($r = 0.527$), we find that a 4-gram model is just as good ($r = 0.528$). Thanks to incorporating misspellings, our 4-gram model surpasses both the previous unsupervised state-of-the art ($r = 0.472$), and the average native speaker ($r = 0.46$), demonstrating that acceptability is well captured by n-gram statistics and simple language features.

Keywords: acceptability judgments; language models; neural networks; word embeddings; statistical models

Introduction

Proficient language users, when given a sentence in their language, are able to judge the *acceptability* of the sentence when asked whether the sentence is natural, well-formed, or grammatical. Acceptability is of interest to cognitive scientists because it provides a means for evaluating whether computational models of language are processing language in a manner similar to humans.

At the same time, models of acceptability have applications in natural language processing. For example, they can be used to evaluate the fluency of machine translation outputs, of answers produced by question answering systems, and automatically generated language snippets quite generally (Lau, Clark, & Lappin, 2015). In computer-assisted learning, acceptability models can help grade essays and provide feedback to native speakers as well as language learners. Acceptability may also be a more precise training signal for general-purpose language models (albeit a costly-to-obtain one).

Recently, great progress has been made in language modeling. But do the models process language in a manner similar to humans? And are the models good at discriminating between acceptable and unacceptable language, or are the models merely sensitive to the distinction between probable and improbable language? Is acceptability statistical in nature, or best understood through the rules of a formal grammar?

In this paper, we investigate what models and language features can capture the acceptability judgements of native English speakers on sentences of English. We train different word embeddings to investigate the role of semantic features in acceptability. We train n-gram models, simple recurrent neural network language models (RNN), and long-short term memory language models (LSTM), to obtain sentence-level probabilities. We explore how individual word frequency, n-gram frequency, spelling errors, the order of words in sentences, and the semantic coherence of a sentence, contribute to human judgments of sentence acceptability.

Acceptability exhibits gradience (Keller, 2000; Lau, Clark, & Lappin, 2014). Accordingly, we treat acceptability as a continuous variable. We evaluate model performance by measuring the correlation between each model's acceptability prediction and a gold standard based on human judgments.

We are concerned with models that learn in an *unsupervised* fashion. Supervised training requires copious labelled data and explicit examples of both acceptable and unacceptable sentences. Conversely, we use only the type of data that is most available to a human learner. Humans learn mainly through exposure to a language. While explicit training on what is and is not acceptable often occurs in school, this training is unnecessary for native language competence.

In what follows, we describe how we calculate an acceptability estimate from the models, the English-language training corpora for the models, and the test data set we use to assess the ability of the models to judge acceptability, before presenting the results of our experiments with word embeddings and language models.

Related Work

Prior work has focused on predicting sentence acceptability using syntactic parsing (Blache, Hemforth, & Rauzy, 2006; Wong & Dras, 2010; Ferraro, Post, & Van Durme, 2012; Clark, Giorgolo, & Lappin, 2013a), statistical language models (Clark et al., 2013a; Lau et al., 2015), and linguistic features (Heilman et al., 2014; Clark, Giorgolo, & Lappin, 2013b). Neural language models have shown evidence of acquiring deeper grammatical competence beyond mere n-gram statistics (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018), suggesting that the models are a good basis for modelling human acceptability judgements (Lau et al., 2015).

Vecchi, Marelli, Zamparelli, and Baroni (2017) use word embeddings to predict human judgements of semantic acceptability on adjective - noun pairs (e.g., *remarkable onion* vs. *legislative onion*). Similarly, Marelli and Baroni (2015)

Measures	Equation
Mis	$Score(\xi)^{(m+\alpha)}$
NormMul	$Mis \times P_u(\xi)$
NormSub	$Mis - P_u(\xi)$
SLOR	$\frac{NormSub}{\lvert\xi\rvert}$

Table 1: Measures for predicting the acceptability of a sentence. Notation: *Score* is an estimate produced by a model (i.e., semantic coherence or sentence log probability), and normalized to [0,1]; ξ is the sentence; $|\xi|$ is the sentence length; m is the number of misspelled words in a sentence; α is a fitting parameter; $Pu(\xi)$ is the unigram probability of the sentence. We train the unigram model on English Wikipedia.

compose morpheme embeddings to model the acceptability of novel word forms (e.g., *undiligent* vs. *unthird*). However, little work has been done on using distributional semantic models to predict sentence acceptability.

Methodology

Acceptability Measures

To estimate acceptability using word embeddings, we take the cosine similarity between word embeddings in a sentence and normalize to the range [0,1] to get the model's *Score*. For language models, we exponentiate a sentence's log probability to yield a probability, which we use as the *Score*. We then convert the *Score* into an acceptability measure using one of the methods in Table 1. We introduce one more variable: misspellings (*Mis*), which reflects the effect of misspellings.

Because Lau et al. (2015) found that sentence acceptability is invariant with respect to sentence length or word frequency, we normalize to produce a set of final scores based on the measures proposed by Lau et al.. The Syntactic Log Odds Ratio (SLOR) measure, proposed by Pauls and Klein (2012), normalizes both word frequency and sentence length. To estimate acceptability, SLOR takes the probability of a sentence, subtracts out the frequency of the individual words in the sentence, so that sentences are not penalized for using rare words, and divides by sentence length, so that sentences are not penalized for being long. Lau et al. (2015) found that *SLOR* was the best predictor of acceptability. But, as we find that *SLOR* does not perform well using our models, we include *NormMul*, and *NormSub*, variants of the measures proposed by Lau et al.. We normalize by word frequency in the *NormMul* (normalized multiply) and *NormSub* (normalized subtract) measures. We compute each measure's Pearson correlation coefficient with the sentence's gold standard to evaluate effectiveness in predicting acceptability.

Training Corpora

We train our models on two large corpora of "acceptable" English: the British National Corpus (**BNC**: 130K unique words, 100M tokens; British National Corpus Consortium, 2007) and on a corpus of novels (**NC**: 39K unique words, 145M tokens; Johns, Jones, & Mewhort, 2016). The corpora are tokenized using NLTK[1] and all words are lower case.

The GUG Dataset

To test the ability of the models to predict acceptability, we need a collection of sentences that exhibit varying degrees of acceptability with ratings from native speakers. We use the Grammatical versus Ungrammatical (GUG) data set built by Heilman et al. (2014). The GUG data set contains 3129 sentences randomly selected from essays written by non-native speakers of English as part of a test of English language proficiency (see Table 2). Heilman et al. (2014) crowd-sourced acceptability ratings for the sentences on a 1 (incomprehensible) to 4 (perfect) scale, obtaining five ratings for each sentence. Each sentence also has an "expert" rating from a linguist. Heilman et al. (2014) randomly split the data into training (50%), development (25%), and test (25%) sets.

In order to compare to Lau et al. (2015), we only use the GUG test set. We remove 23 sentences from GUG that have less than 5 words, lower-case all words, and extract 744 sentences for our test set. We take the average of the crowd-sourced ratings (across 5 workers) as the gold standard. To evaluate the models, we compute the correlation of the predicted ratings and the gold standard ratings. We correct misspelled words using the PyEnchant[2] spell-checker. We use PyEnchant's first suggestion as the corrected spelling. We count the number of misspelled words in every sentence, which serves as a feature for the *Mis* measure (see Table 1).

To illustrate the difficulty of predicting acceptability, we compute the Pearson's correlation coefficient between the ratings of each human rater and the mean acceptability rating. We find that the correlations for crowd workers range from 0.440 to 0.485. However, the correlation between the expert and the average of all non-expert ratings is high, $r = 0.753$. The high correlation could be an artifact of crowd worker selection. Crowd workers were screened using a qualifying test that assessed the agreement between their ratings and the expert ratings on a set of trial sentences, which could force a correlation. Conversely, the higher correlation for the expert and the average of the workers could reflect the expert's language expertise and the wisdom of the crowd.

In sum, to achieve non-expert performance at predicting acceptability, computational models need a correlation to the mean acceptability rating of at least $r = 0.440$, but to achieve expert performance may require a much higher correlation.

[1] https://www.nltk.org

[2] http://pythonhosted.org/pyenchant/

Table 2: Example sentences from Heilman et al. (2014)'s GUG data set with acceptability ratings.

sentence	expert	workers	mean
For not use car.	1	[3, 4, 3, 4, 3]	3.0
I would like to initiate, myself , whatever I do on my trip to get much out of my trip.	2	[4, 3, 2, 3, 3]	2.8
These kind of fish can't live so long in water that contain salt.	3	[3, 3, 4, 3, 4]	3.3
So if you want me to choose right now, I will choose ordinary milk instead of that special kind.	4	[3, 4, 3, 4, 4]	3.7

Experiments

Acceptability as Semantic Coherence

Measures of semantic coherence are used in models of document-level topic generation (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) and visual scene generation (Vertolli, Kelly, & Davies, 2018) to ensure, respectively, that the topics and scenes "make sense." We hypothesize that more acceptable sentences have higher semantic coherence. We propose a novel metric for the semantic coherence of an individual sentence. We quantify semantic coherence between a word and its context as the cosine similarity between the word's embedding and the sentence's embedding without the word. The semantic coherence of the sentence as a whole is then computed as either the minimum or average of each word's similarity to the rest of the sentence.

The word embeddings we train are *Word2Vec* (*skip-gram*, i.e., SK, and *continuous bag of words*, i.e., CBOW; Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), the *GloVe* model (Pennington, Socher, & Manning, 2014), and the *Hierarchical Holographic Model* (HHM; Kelly, Reitter, & West, 2017),

Word2Vec Using CBOW, Word2Vec predicts the current word from a window of surrounding context words, while in SK, Word2Vec uses the current word to predict the surrounding context (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). We use *Gensim Word2vec* to train CBOW and SK models with 300 dimensions and a window size of 5.[3]

GloVe *GloVe* takes aggregate word-word co-occurrence statistics from a corpus (Pennington et al., 2014) and performs a dimensional reduction on the co-occurrence matrix to produce a set of word embeddings. We build 300 dimensional GloVe embeddings on the BNC and NC corpora[4].

Hierarchical Holographic Model GloVe and Word2vec treat sentences as unordered sets of words. In English, word order conveys much of the meaning of the sentence, and is critical in constructing a grammatical sentence. To account for this, we include in our analysis the Hierarchichal Holographic Model (HHM; Kelly et al., 2017), a model sensitive to the order of words in a sentence. HHM generates multiple levels of representations, such that higher levels are sensitive to more abstract relationships between words, such as part-of-speech relationships (Kelly et al., 2017). We trained three levels of HHM representations with 1024 dimensions and a context window of 5 words to the left and right of each target word. Dimensionality, level, and window size are HHM's only hyper-parameters, and as such, the number of hyper-parameters is comparable to the other word embeddings.

Semantic Coherence The semantic coherence of a sentence is either:

$$Score = \min(\text{cosine}(word_i, context_i)) \quad (1)$$

$$Score = \text{avg}(\text{cosine}(word_i, context_i)) \quad (2)$$

where min is the minimum, avg is the average, w_i is the i-th word's representation in a sentence and $context_i$ is the context representation of the sentence without w_i. If a word in the GUG test set is not in the corpus (169 test set words not in NC, 0 words not in BNC), we use a random embedding instead. We have two methods for computing the context representation $context_i$ of word w_i. One method is to sum:

$$context_i = \frac{\sum(w_1, ..., w_{i-1}, w_{i+1}, ...w_n)}{n} \quad (3)$$

We can also get the context by building a holographic representation using HHM's environment vectors via a method described in Jones and Mewhort (2007) and Kelly et al. (2017). By using aperiodic convolution to combine environment vectors into bigrams, trigrams, tetragrams, etc., up to n-grams of sentence length, we can construct a representation of the sentence that accounts for the ordering of the words within it.

Results Table 3 shows that semantic coherence, by itself, correlates weakly with acceptability when the sentence context is computed as a sum (as in Eqn. 3). Correlations for the *Score* range from $r = -0.181$ (CBOW on BNC) to $r = 0.185$ (SK on NC) with an average correlation across word embeddings and corpora of only $r = 0.058$.

We can improve the correlation by incorporating misspellings and unigram probability. GloVe is the best performing model when using the combined measures. We experiment with different values of the fitting parameter α for the *Mis* measure and find the highest correlation for $\alpha = 0$ and using the minimum (rather than average) semantic coherence of the sentence (GloVe, $r = 0.449$ on NC and $r = 0.446$

[3]Gensim Word2vec from: https://radimrehurek.com/gensim/

[4]Our GloVe implementation comes from https://github.com/stanfordnlp/GloVe. Other parameter values of the GloVe vectors: $x_{max} = 100; window\_size = 10; iter = 100$

on BNC), which can be further improved using *NormMul* by multiplying by the unigram probability (GloVe, $r = 0.464$ on NC and $r = 460$ on BNC). However, an α of zero indicates heavy reliance on the misspellings to predict acceptability.

When the context vector is computed holographically, such that the ordering of the words in the sentence is preserved, we find that the semantic coherence score is a stronger predictor of acceptability then when using a sum to construct the context (see Table 4). *Score* ranges from $r = 0.140$ to $r = 0.314$ and an average of $r = 0.201$. By incorporating misspellings ($\alpha = 0.3$), the correlation for average coherence increases to $r = 0.471$ (NC) and 0.457 (BNC) for Level 1 of HHM and to 0.467 (NC) and 0.494 (BNC) for Level 2 of HHM.

In sum, while semantic coherence is not a strong predictor of acceptability, when combined with sensitivity to the order of the words in the sentence and the number of misspellings, it can act as an effective predictor of acceptability ($r = 0.494$).

Acceptability using Language Models

Language models can predict the probability of a sequence of words. Lau et al. (2015) compared acceptability judgments against predictions by language models including *n*-gram models, Hidden Markov Models, latent Dirichlet allocation, a Bayesian chunker, and a recurrent neural network language model (RNNLM). Lau et al. (2015) obtained acceptability scores using crowd-sourcing and a corpus of sentences using round-robin machine translation from English to a second language, and back. On this (unpublished) data, the neural language model (RNNLM) performed best.

However, a 4-gram model trained on the British National Corpus (BNC) beat the RNNLM when tested on the GUG test data ($r = 0.472$; Lau et al., 2015).

We take the 4-gram model's performance ($r = 0.472$) as the baseline in the following experiments. The primary difference between the translation dataset and the GUG dataset is that sentences in the translation dataset are produced by Google Translate automatically, while sentences in the GUG dataset are produced by a non-native speaker. We suspect that a language model may perform better on a dataset produced by machine translation than on a naturalistic dataset comprised of essays from L2 speakers, which may contain greater linguistic variability.

Training Using the BNC and NC corpora, we produce lexical 4- and 5-gram language models using Kneser-Ney smoothing (Stolcke, 2002), and a basic RNNLM [5] (Elman, 1998; Mikolov, 2012). We also train the RNNLM and an LSTMLM on NC using Tensorflow. The embedding layer is initialized either to a Gaussian distribution of values with 1024 dimensions or a pre-trained embedding with 3072 dimensions (we use a concatenation of HHM1, HHM2 and HHM3 here). We set the LSTM's hidden layer size to 1024, the projection layer size to 128, and the maximum sequence length to 25. In the GUG test set, we replace out-of-vocabulary words (i.e., words not in the corpora) and low-frequency (< 5) words with the `<unk>` token. Sentences with less than 8 words are removed.

Results Score_C in Table 6 equals the log probability of sentences in the spell-corrected test data and Score_O is the log probability of sentences in the original test data.

Setting the *Mis* score's fitting parameter $\alpha = 1.3$ (see Table 1) yielded best results for the lexical 4-gram model on BNC. The 4-gram model's correlation of $r = 0.528$ improves notably upon Lau et al.'s best correlation of 0.472. The RNNLM trained on NC performs similarly, at $r = 0.527$.

Table 5 shows the performance of the LSTM and RNN trained on NC with an embedding layer. We find that the best performing model is the LSTM with a pre-trained embedding layer that is not fixed and can be trained further.

Table 6 shows that the correlation of *Mis* exceeds the original log probability no matter whether the log probability is computed on corrected test data or the original test data. The *Mis* score's high correlation shows that using misspellings allows for a better translation of log probability to acceptability.

Unlike Lau et al. (2015), who found *SLOR* to be the best measure, we find that *NormSub* performs better than other measures across all models (Table 6). *NormSub* removes the influence of unigram probability from the acceptability score. Word frequency has little influence on acceptability ($r = 0.2$ between acceptability and unigram probability), but greatly affects the log probability computed by the language models.

To test the robustness of our methodology on the test data, we also test the models on the GUG's development set (747 sentences after preprocessing) and find that a 5-gram model trained on the BNC gets the best correlation at $r = 0.467$.

Discussion

Modeling acceptability provides a window into the human brain's language engine. We explore what aspects of language are important to account for what people consider acceptable, well-formed sentences. In particular, we examine the importance of misspellings, semantic coherence, and *n*-gram probability using simple features, word embeddings, *n*-gram models, and predictive neural language models.

We find that accounting for misspellings considerably improves the ability of unsupervised models to capture acceptability judgments. Prior work by Heilman et al. (2014) with supervised models found that the number of misspelled words was an important feature for their models. We develop a technique for incorporating misspellings into unsupervised models by correcting all spelling errors so that the model works with clean data, and then raising the model's acceptability estimate to the power of the number of misspelled words. Our *Mis* measure is not, itself, intended to be a cognitive model, but rather an indication that cognitive models of human judgements of acceptability need to be highly sensitive to misspellings.

[5] We use the Mikolov, Karafiát, Burget, Černockỳ, and Khudanpur (2010) implementation for the RNNLM: `http://www.fit.vutbr.cz/~imikolov/rnnlm/`. Meta-parameter values of the RNNLM included the following: number of classes = 550; bptt = 4; bptt-block = 100, hidden = 600.

measure	CBOW (NC)		SK (NC)		GloVe (NC)		CBOW (BNC)		SK (BNC)		GloVe (BNC)	
	min	avg	min	avg	min	avg	min	avg	min	avg	min	avg
Score	0.029	-0.176	0.185	-0.028	0.089	-0.071	0.035	-0.181	0.173	-0.029	0.064	-0.068
Mis	0.444	0.41	0.427	0.397	0.449	0.426	0.445	0.41	0.425	0.399	0.446	0.426
NormMul	0.458	0.44	0.445	0.436	**0.464**	0.445	0.459	0.440	0.444	0.437	**0.460**	0.445
NormSub	0.346	0.273	0.332	0.238	0.356	0.339	0.348	0.275	0.327	0.243	0.351	0.339
SLOR	0.288	0.239	0.298	0.242	0.323	0.276	0.290	0.240	0.301	0.248	0.318	0.276

Table 3: Pearson's r between semantic coherence and acceptability with $\alpha = 0$ for *Mis*. Semantic coherence computed using minimum and average similarity between word and context representations. Context representation obtained via Equation 3. Word embeddings were trained on NC (left) and BNC (right). Boldface indicates the best performing measure.

measure	HHM1 (NC)		HHM2 (NC)		HHM3 (NC)		HHM1 (BNC)		HHM2 (BNC)		HHM3 (BNC)	
	min	avg	min	avg	min	avg	min	avg	min	avg	min	avg
Score	0.153	0.198	0.19	0.27	0.191	0.232	0.097	0.129	0.217	0.314	0.14	0.278
Mis	0.454	0.471	0.459	0.467	0.447	0.462	0.435	0.457	0.475	**0.494**	0.425	0.482
NormMul	0.457	**0.476**	0.453	0.462	0.436	0.455	0.438	0.459	0.463	0.479	0.432	0.464
NormSub	0.346	0.36	0.365	0.38	0.363	0.378	0.308	0.318	0.381	0.404	0.274	0.409
SLOR	0.239	0.176	0.243	0.204	0.293	0.219	0.131	0.096	0.183	0.169	0.155	0.24

Table 4: Context representations obtained via the holographic method for HHMs. Measures computed with $\alpha = 0.3$. Boldface indicates the best-performing measure. Trained on NC (left) and BNC (right).

measure	LSTM fixed	LSTM train	RNN random	LSTM random
Score	0.42	0.408	0.307	0.344
Mis	0.493	0.49	0.418	0.445
NormMul	0.453	0.447	0.385	0.406
NormSub	0.509	**0.522**	0.453	0.489
SLOR	0.38	0.4	0.339	0.374

Table 5: Pearson's r between model predictions and mean acceptability for language models trained on the NC corpus with $\alpha = 2.1$. The models are an LSTM with fixed, pretrained HHM embeddings, an LSTM with trainable HHM embeddings, an RNN with Gaussian, randomly initialized trainable embeddings, and an LSTM with randomly initialized trainable embeddings. Boldface indicates the best performing models and measures.

measure	4-gram		5-gram		RNN	
	BNC	NC	BNC	NC	BNC	NC
Score_O	0.284	0.250	0.252	0.311	0.337	0.284
Score_C	0.315	0.302	0.315	0.302	0.295	0.314
Mis	0.465	0.460	0.464	0.459	0.448	0.461
NormMul	0.424	0.418	0.423	0.417	0.410	0.421
NormSub	**0.528**	0.518	0.527	0.518	0.510	**0.527**
SLOR	0.469	0.457	0.474	0.460	0.472	0.485

Table 6: Pearson's r between model predictions and mean acceptabiliy for language models trained on BNC and NC with $\alpha = 1.3$. Boldface indicates best models and measures.

We propose a novel metric of the semantic coherence of a sentence and we examine the contribution of semantic coherence with respect to acceptability. By itself, semantic coherence correlates poorly with acceptability, providing evidence that meaning constitutes only a small part of what makes a sentence acceptable or unacceptable, in keeping with Chomsky (1956)'s arguments for a distinction between syntactic and semantic well-formedness.

However, when semantic coherence is combined with misspellings and unigram probabilities (i.e., word frequencies), we can account for much of the variability in acceptability ($r = 0.46$). We can further improve the correlation by incorporating the order of the words in the sentence into our measure of semantic coherence ($r = 0.49$). These results suggest that acceptability is not wholly independent of semantics. The role of semantic coherence in linguistic acceptability warrants further investigation. The validity of metrics of semantic coherence need to be assessed against human judgements, but we leave such explorations for future work.

Language models provide a means of estimating the probability of a sentence, which in turn can be used to predict acceptability. We replicate prior work (Lau et al., 2015) in finding that the RNNLM is a good language model for accounting for acceptability ($r = 0.53$). Yet, a simple 4- or 5-gram model with statistical smoothing is just as good as an RNNLM. So, 4- or 5-gram frequencies may be what the neural network model primarily learns to rely upon to model human language with respect to acceptability judgments.

We replicate Lau et al. (2015)'s finding that the log probabilities produced by language models need to be normalized by subtracting the unigram probability. This normalization, *NormSub*, prevents the models from underestimating the acceptability of sentences with low frequency words. Lau et al. (2015) find that the Syntactic Log Odds Ratio (SLOR; Pauls & Klein, 2012), which further normalizes by dividing by sentence length, is the best method for converting language model log probabilities into predicted acceptability. However, we consistently find that *SLOR* performs less well at predicting acceptability than *NormSub* (see Tables 6 and 5).

The difference between our findings and Lau et al.'s (2015) may be due to the different datasets we use: Lau et al. primarily report results for a machine translation dataset, whereas we evaluate results on a dataset of sentences produced by second language speakers. The appropriateness of normalizing by sentence length thus may be dependent on the characteristics of the language being evaluated for acceptability.

In sum, the best and simplest model of acceptability judgment consists of only three statistical language features: 4- or 5-gram frequency (with statistical smoothing), misspellings, and word (or unigram) frequency. The correlation of non-expert human raters to mean acceptability ranges from 0.440 to 0.485. Thus, our models' correlation with the mean acceptability exceeds the rater reliability.

Yet even our best model is not as good as the expert rater (r=0.75). Do experts have more experience and thus, different assumptions of lexical or syntactic distributions, or do they just interpret the task differently, perhaps isolating grammaticality from the less well-defined acceptability? Do expert judgments represent less dialectal variety, or are they correlated with ease-of-processing? Future models will, hopefully, capture expert-level performance.

We have considered only unsupervised models of acceptability on the assumption that unsupervised data best reflects the learning environment of the average native speaker of the language. However, supervised approaches may be appropriate for modelling expert performance, as language experts have likely had the benefit of explicit training on what is and is not acceptable and "proper" language. Though it remains an open question whether capturing such expertise is even desirable given that "proper" language is typically associated with a specific region or class, rather than the population as a whole (McArthur, 1992, pp. 984-985).

Conclusion

While more sophisticated neural language models may have lower perplexity than simpler language models, we find that simple *n*-gram models perform just as well at predicting acceptability. By incorporating a count of misspellings, our 4-gram model ($r = 0.53$) surpasses the previous unsupervised state-of-the art (Lau et al, 2015, $r = 0.47$), reducing the gap to expert performance ($r = 0.75$) and surpassing the average non-expert native speaker ($r = 0.46$). A high correlation to acceptability can also be achieved without a predictive language model, by combining number of misspellings, semantic similarity, and word order information ($r = 0.49$).

Our results suggest that the layperson's ability to judge whether or not a given sentence is acceptable may be derived from sensitivity to simple, statistical language features, without necessarily requiring syntactic rules or even a complex machine learning model. Cognitively, these results suggest that humans may learn what is an acceptable sentence of a language in a simple, statistical, rather than rule-based, manner. However, no model considered here closely matches expert performance at judging acceptability. Closing the gap to experts may yet require a more sophisticated language model.

Acknowledgments

Work supported by a Key Project of the National Social Science Foundation of China (Grant No.18AYY08) to Wang Jing, a Post-Doctoral Fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC) to M. A. Kelly, and a National Science Foundation grant (BCS-1734304) to David Reitter and M. A. Kelly.

References

Blache, P., Hemforth, B., & Rauzy, S. (2006). Acceptability prediction by means of grammaticality quantification. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 57–64). Sydney, Australia. doi: 10.3115/1220175.1220183

British National Corpus Consortium. (2007). *British national corpus* (version 3 XML ed.). Bodleian Libraries, University of Oxford. Retrieved from http://www.natcorp.ox.ac.uk/

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*, 113-124. doi: 10.1109/TIT.1956.1056813

Clark, A., Giorgolo, G., & Lappin, S. (2013a). Statistical representation of grammaticality judgements: The limits of n-gram models. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 28–36). Sofia, Bulgaria. Retrieved from https://www.aclweb.org/anthology/W13-2604

Clark, A., Giorgolo, G., & Lappin, S. (2013b). Towards a statistical model of grammaticality. In N. S. M. Knauff M. Pauen & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (p. 2064-2069). Austin, TX: Cognitive Science Society. Retrieved from https://escholarship.org/uc/item/0s47f6vj

Elman, J. (1998). Generalization, simple recurrent networks, and the emergence of structure. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates. Retrieved from https://langev.com/pdf/elman98generalizationSimple.pdf

Ferraro, F., Post, M., & Van Durme, B. (2012). Judging grammaticality with count-induced tree substitution grammars. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 116–121). Retrieved from https://www.aclweb.org/anthology/W12-2013

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 1195–1205). Association for Computational Linguistics. doi: 10.18653/v1/N18-1108

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., & Tetreault, J. (2014). Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 174–180). doi: 10.3115/v1/P14-2029

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2016). Experience as a free parameter in the cognitive model-

ing of language. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (p. 1325-1330). Austin, TX: Cognitive Science Society. Retrieved from https://mindmodeling.org/cogsci2016/papers/0397/paper0397.pdf

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1-37. doi: 10.1037/0033-295X.114.1.1

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Unpublished doctoral dissertation, University of Edinburgh. doi: 10.7282/T3GQ6WMS

Kelly, M. A., Reitter, D., & West, R. L. (2017). Degrees of separation in semantic and syntactic relationships. In *Proceedings of the 15th International Conference on Cognitive Modeling* (p. 199-204). Warwick, U.K.: University of Warwick. Retrieved from https://par.nsf.gov/servlets/purl/10067553

Lau, J. H., Clark, A., & Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (p. 821-826). Austin, TX: Cognitive Science Society. Retrieved from https://cogsci.mindmodeling.org/2014/papers/149/

Lau, J. H., Clark, A., & Lappin, S. (2015). Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 1618–1628). doi: 10.3115/v1/P15-1156

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, *122*(3), 485–515. doi: 10.1037/a0039267

McArthur, T. (Ed.). (1992). *The Oxford companion to the English language*. Oxford University Press.

Mikolov, T. (2012). Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*. Retrieved from http://www.fit.vutbr.cz/~imikolov/rnnlm/google.pdf

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 1st International Conference on Learning Representations.* Scottsdale, Arizona, USA. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association* (pp. 1045–1048). Makuhari, Chiba, Japan. Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1045.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D11-1024

Pauls, A., & Klein, D. (2012). Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 959–968). Retrieved from https://dl.acm.org/citation.cfm?id=2390654

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). doi: 10.3115/v1/D14-1162

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing* (p. 257-286). Retrieved from https://www.isca-speech.org/archive/archive_papers/icslp_2002/i02_0901.pdf

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, *41*(1), 102-136. doi: 10.1111/cogs.12330

Vertolli, M. O., Kelly, M. A., & Davies, J. (2018). Coherence in the visual imagination. *Cognitive Science*, *42*(3), 885-917. doi: 10.1111/cogs.12569

Wong, S.-M. J., & Dras, M. (2010). Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop 2010* (pp. 67–75). Retrieved from https://www.aclweb.org/anthology/U10-1011