The Influence of Context on Sentence Acceptability Judgements

Jean-Philippe Bernardy
University of Gothenburg
jean-philippe.bernardy@gu.se

Shalom Lappin
University of Gothenburg
shalom.lappin@gu.se

Jey Han Lau
IBM Research Australia
jeyhan.lau@gmail.com

Abstract

We investigate the influence that document context exerts on human acceptability judgements for English sentences, via two sets of experiments. The first compares ratings for sentences presented on their own with ratings for the same set of sentences given in their document contexts. The second assesses the accuracy with which two types of neural models — one that incorporates context during training and one that does not — predict these judgements. Our results indicate that: (1) context improves acceptability ratings for ill-formed sentences, but also reduces them for well-formed sentences; and (2) context helps unsupervised systems to model acceptability.[1]

1 Introduction

Sentence acceptability is defined as the extent to which a sentence is well formed or natural to native speakers of a language. It encompasses semantic, syntactic and pragmatic plausibility and other non-linguistic factors such as memory limitation. Grammaticality, by contrast, is the syntactic well-formedness of a sentence. Grammaticality as characterised by formal linguists is a theoretical concept that is difficult to elicit from non-expert assessors. In the research presented here we are interested in predicting acceptability judgements.[2]

Lau et al. (2015, 2016) present unsupervised probabilistic methods to predict sentence acceptability, where sentences were judged independently of context. In this paper we extend this research to investigate the impact of context on human acceptability judgements, where context is defined as the full document environment surrounding a sentence. We also test the accuracy of more sophisticated language models — one which incorporates document context during training — to predict human acceptability judgements.

We believe that understanding how context influences acceptability is crucial to success in modelling human acceptability judgements. It has implications for tasks such as style/coherence assessment and language generation. Showing a strong correlation between unsupervised language model sentence probability and acceptability supports the view that linguistic knowledge can be represented as a probabilistic system. This result addresses foundational questions concerning the nature of grammatical knowledge (Lau et al., 2016).

Our work is guided by 3 hypotheses:

H_1: Document context boosts sentence acceptability judgements.

H_2: Document context helps language models to model acceptability.

H_3: A language model predicts acceptability more accurately when it is tested on sentences within document context than when it is tested on the sentences alone.

We sample sentences and their document contexts from English Wikipedia articles. We perform round-trip machine translation to generate sentences of varying degrees of well-formedness and ask crowdsourced workers to judge the acceptability of these sentences, presenting the sentences with and without their document environments. We describe this experiment and address H_1 in Section 2.

In Section 3, we experiment with two types of language models to predict acceptability: a standard language model and a topically-driven model. The latter extends the language model by incorporating document context as a conditioning

[1] Annotated data (with acceptability ratings) is available at: https://github.com/GU-CLASP/BLL2018.

[2] See Lau et al. (2016) for a detailed discussion of the relationship between acceptability and grammaticality. They provide motivation for measuring acceptability rather than grammaticality in their crowd source surveys and modelling experiments.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 456–461
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

variable. The model comparison allows us to understand the impact of incorporating context during training for acceptability prediction. We also experiment with adding context as input at test time for both models. These experiments collectively address H_2, by investigating the impact of using context during training and testing for modelling acceptability. We evaluate the models against crowd-sourced annotated sentences judged both in context and out of context. This tests H_3.

In Section 4 we briefly consider related work. We indicate the issues to be addressed in future research and summarise our conclusions in Section 5.

2 The Influence of Document Context on Acceptability Ratings

Our goal is to construct a dataset of sentences annotated with acceptability ratings, judged with and without document context. To obtain sentences and their document context, we extracted 100 random articles from the English Wikipedia and sampled a sentence from each article. To generate a set of sentences with varying degrees of acceptability we used the Moses MT system (Koehn et al., 2007) to translate each sentence from English to 4 target languages — Czech, Spanish, German and French — and then back to English.[3] We chose these 4 languages because preliminary experiments found that they produce sentences with different sorts of grammatical, semantic, and lexical infelicities. Note that we only translate the sentences; the document context is not modified.

To gather acceptability judgements we used Amazon Mechanical Turk and asked workers to judge acceptability using a 4-point scale.[4] We ran the annotation task twice: first where we presented sentences without context, and second within their document context. For the in-context experiment, the target sentence was highlighted in boldface, with one preceding and one succeeding sentence included as additional context. Workers had the option of revealing the full document context by clicking on the preceding and succeeding sentences. We did not check whether subjects viewed the full context when recording their ratings.

Henceforth human judgements made without context are denoted as $\mathtt{h}^-$ and judgements with context as $\mathtt{h}^+$. We collected 20 judgements per sentence, giving us a total of a 20,000 annotations (100 sentences × 5 languages × 2 presentations × 20 judgements).

To ensure annotation reliability, sentences were presented in groups of five, one from the original English set, and four from the round-trip translations, one per target language, with no sentence type (English original or its translated variant) appearing more than once in a HIT.[5] We assume that the original English sentences are generally acceptable, and we filtered out workers who fail to consistently rate these sentences as such.[6] Post-filtering, we aggregate the multiple ratings and compute the mean.

We first look at the correlation between without-context ($\mathtt{h}^-$) and with-context ($\mathtt{h}^+$) mean ratings. Figure 1 is a scatter plot of this relation. We found a strong correlation of Pearson's $r = 0.80$ between the two sets of ratings.

We see that adding context generally improves acceptability (evidenced by points above the diagonal), but the pattern reverses as acceptability increases, suggesting that context boosts sentence ratings most for ill-formed sentences. The trend persists throughout the whole range of acceptability, so that for the most acceptable sentences, adding context actually diminishes their rated acceptability. We can see this trend clearly in Figure 1, where the average difference between $\mathtt{h}^-$ and $\mathtt{h}^+$ is represented by the distance between the linear regression and the diagonal. These lines cross at $\mathtt{h}^+ = \mathtt{h}^- = 3.28$, the point where context no longer boosts acceptability.

To understand the spread of individual judgements on a sentence, we compute the standard deviation of ratings for each sentence and then take the mean over all sentences. We found a small difference: 0.71 for $\mathtt{h}^-$ and 0.76 for $\mathtt{h}^+$. We also calculate one-vs-rest correlation, where for each

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

[4] We ask workers to judge how "natural" they find a sentence. For more details on the AMT protocol and our use of a four category naturalness rating system, see Lau et al. (2015, 2016).

[5] A HIT is a "human intelligence task". It constitutes a unit of work for crowdworkers.

[6] Control sentence rating threshold = 3. Minimum accuracy for control sentences = 0.70. To prevent workers from gaming this system (by giving all perfect ratings), we also removed workers whose average rating $\geq$ 3.5. Using these rules we filtered out on average, for each sentence, 7.5125 answers for $\mathtt{h}^+$ and 3.9725 for $\mathtt{h}^-$. This gave us approximately 13 and 16 annotators for each $\mathtt{h}^+$ and $\mathtt{h}^-$ sentence respectively.

Language	Sentence	h$^-$	h$^+$
—	david acker, harry's son, became the president of sleepy's in 2001.	3.47	3.38
Czech	david acker harry' with son has become president of the sleepy' with in 2001.	1.75	2.08
German	david field, harry' the son was the president of " in 2001.	1.63	3.00
Spanish	david acker, harry' his son, became president of the sleeping' in 2001.	2.19	2.62
French	david acker, harry' son, the president of the sleepy' in 2001.	1.47	2.46

Table 1: A sample of sentences with their without-context (h$^-$) and with-context (h$^+$) ratings. The "Language" column denotes the intermediate translation language. The original English sentence is marked with "—".

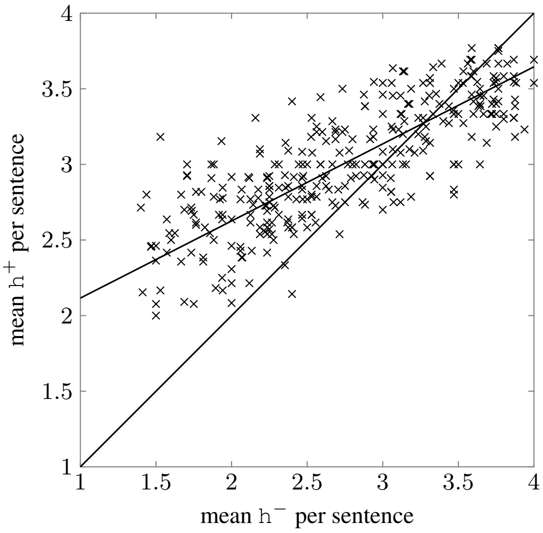

Figure 1: With-context (h$^+$) against without-context (h$^-$) ratings. Points above the full diagonal represent sentences which are judged more acceptable when presented with context. The total least-square linear regression is shown as the second line.

sentence we randomly single out an annotator rating and compute the Pearson correlation between these judgements against the mean ratings for the rest of the annotators.[7] This number can be interpreted as a performance upper bound on a single annotator for predicting the mean acceptability of a group of annotators.

We found a big gap in the one-vs-rest correlations: 0.628 for h$^-$ and 0.293 for h$^+$. We were initially surprised as to why the correlation is so different, even though the standard deviation is similar. Further investigation reveals that this difference is explained by the pattern shown in Figure 1. Adding context "compressess" the distribution of (mean) ratings, pushing the extremes to the middle (i.e. very ill/well-formed sentences are now less ill/well-formed). The net effect is that it lowers correlation, as the good and bad sentences are now less separable.

One possible explanation for this compression is that workers focus more on global semantic and pragmatic coherence when context is supplied. If this is the case, then the syntactic mistakes introduced by MT have less effect on ratings than for the out-of-context sentences, where global coherence is not a factor.

To give a sense how context influences ratings, we present a sample of sentences with their without-context (h$^-$) and with-context (h$^+$) ratings in Table 1.

3 Modelling Sentence Acceptability with Enriched LMs

Lau et al. (2015, 2016) explored a number of unsupervised models for predicting acceptability, including n-gram language models, Bayesian HMMs, LDA-based models, and a simple recurrent network language model. They found that the neural model outperforms the others consistently over multiple domains, in several languages. In light of this, we experiment with neural models in this paper. We use: (1) a LSTM language model (lstm: Hochreiter and Schmidhuber (1997); Mikolov et al. (2010)), and (2) a topically driven neural language model (tdlm: Lau et al. (2017)).[8]

lstm is a standard LSTM language model, trained over a corpus to predict word sequences.

[7] Trials are repeated 1000 times and the average correlation is computed, to insure that we obtain robust results and avoid outlier ratings skewing our Pearson coefficient value. See Lau et al. (2016) for the details of this and an alternative method for simulating an individual annotator.

[8] We use the following tdlm implementation: https://github.com/jhlau/topically-driven-language-model.

Acc. Measure	Equation
LogProb	$\log P_m(s,c)$
Mean LP	$\frac{\log P_m(s,c)}{\lvert s \rvert}$
Norm LP (Div)	$-\frac{\log P_m(s,c)}{\log P_u(s)}$
Norm LP (Sub)	$\log P_m(s,c) - \log P_u(s)$
SLOR	$\frac{\log P_m(s,c) - \log P_u(s)}{\lvert s \rvert}$

Table 2: Acceptability measures for predicting the acceptability of a sentence. s is the sentence ($|s|$ is the sentence length); c is the document context (only used by `lstm`$^+$ and `tdlm`$^+$); $P_m(s,c)$ is the probability of the sentence given by a model; $P_u(s)$ is the unigram probability of the sentence.

Rtg	Model	*LP*	*Mean*	*NrmD*	*NrmS*	*SLOR*
h$^-$	lstm$^-$	0.151	0.487	**0.586**	0.342	0.584
	lstm$^+$	0.161	0.529	0.618	0.351	**0.633**
	tdlm$^-$	0.147	0.515	0.634	0.359	**0.640**
	tdlm$^+$	0.165	0.541	0.645	0.373	**0.653**
h$^+$	lstm$^-$	0.153	0.421	0.494	0.293	**0.503**
	lstm$^+$	0.168	0.459	0.522	0.310	**0.546**
	tdlm$^-$	0.153	0.450	0.541	0.313	**0.557**
	tdlm$^+$	0.169	0.473	0.552	0.325	**0.568**

Table 3: Pearson's r of acceptability measures and human ratings. "Rtg" = "Rating", "LP" = *LogProb*, "Mean" = *Mean LP*, "NrmD" = *Norm LP (Div)* and "NrmS" = *Norm LP (Sub)*. Boldface indicates optimal performance in each row.

`tdlm` is a joint model of topic and language. The topic model component produces topics by processing documents through a convolutional layer and aligning it with trainable topic embeddings. The language model component incorporates context by combining its topic vector (produced by the topic model component) with the LSTM's hidden state, to generate the probability distribution for the next word.

After training, given a sentence both `lstm` and `tdlm` produce a sentence probability (aggregated using the sequence of conditional word probabilities). In our case, we also have the document context, information which both models can leverage. Therefore we have 4 variants at **test time**: models that use only the sentence as input, `lstm`$^-$ and `tdlm`$^-$, and models that use both sentence and context, `lstm`$^+$ and `tdlm`$^+$.[9] `lstm`$^+$ incorporates context by feeding it to the LSTM network and taking its final state[10] as the initial state for the current sentence. `tdlm`$^-$ ignores the context by converting the topic vector into a vector of zeros.

To map sentence probability to acceptability, we compute several *acceptability measures* (Lau et al., 2016), which are designed to normalise sentence length and word frequency. These are given in Table 2.

We train `tdlm` and `lstm` on a sample of 100K English Wikipedia articles, which has no overlap with the 100 documents used for the annotation described in Section 2. The training data has approximately 40M tokens and a vocabulary size of 66K.[11] Training details and all model hyperparameter settings are detailed in the supplementary material.

To assess the performance of the acceptability measures, we compute Pearson's r against mean human ratings (Table 3). We also experimented with Spearman's rank correlation, but found similar trends and so present only the Pearson results.

The first observation is that we replicate the performance of the original experiment setting (Lau et al., 2015). We achieved a correlation of 0.584 when we compared `lstm`$^-$ against `h`$^-$, which is similar to the previously reported performance (0.570).[12] *SLOR* outperforms all other measures, which is consistent with the findings in Lau et al. (2015). We will focus on *SLOR* for the remainder of the discussion.

Across all models (`lstm` and `tdlm`) and human ratings (`h`$^-$ and `h`$^+$), using context at test time improves model performance. This suggests that taking context into account helps in modelling acceptability, regardless of whether it is tested against judgements made with (`h`$^+$) or without context (`h`$^-$).[13] We also see that `tdlm` consis-

[9] There are only two trained models: `lstm` and `tdlm`. The four variants are generated by varying the type of input provided at test time when computing the sentence probability.

[10] The final state is the hidden state produced by the last word of the context.

[11] We filter word types that occur less than 10 times, lowercase all words, and use a special unkown token to represent unseen words.

[12] We note two differences. First, we use a different set of Wikipedia training and testing articles. Second, we employ a LSTM instead of a simple RNN for the language model.

[13] We believe incorporating context at test time for `lstm` improves performance because context puts the starting state of the current sentence in the right "semantic" space when predicting its words. Without context, the initial state for the current sentence is defaulted to a vector of zeros, and the

tently outperforms lstm over both types of human ratings and test input variants, showing that tdlm is a better model at predicting acceptability. In fact, if we look at tdlm$^-$ vs. lstm$^+$ (h$^-$: 0.640 vs. 0.633; h$^+$: 0.557 vs. 0.546), tdlm still performs better without context than lstm with context. These observations confirm that context helps in the modelling of acceptability, whether it is incorporated during training (lstm vs. tdlm) or at test time (lstm$^-$/tdlm$^-$ vs. lstm$^+$/tdlm$^+$).

Interestingly, we see a lower correlation when we are predicting sentence acceptability that is judged with context. The *SLOR* correlation of lstm$^+$/tdlm$^+$ vs. h$^+$ (0.546/568) is lower than that of lstm$^-$/tdlm$^-$ vs. h$^-$ (0.584/0.640). This result corresponds to the low one-vs-rest human performance of h$^+$ compared to h$^-$ (0.299 vs. 0.636, see Section 2). It suggests that h$^+$ ratings are more difficult to predict than h$^-$. With human performance taken into account, both models substantially outperform the average single-annotator correlation, which is encouraging for the prospect of accurate model prediction on this task.

4 Related Work

Nagata (1988) reports a small scale experiment with 12 Japanese speakers on the effect of repetition of sentences, and embedding them in context. He notes that both repetition and context cause acceptability judgements for ill formed sentences to be more lenient. Gradience in acceptability judgements are studied in the works of Sorace and Keller (2005) and Sprouse (2007).

There is an extensive literature on automatic detection of grammatical errors (Atwell, 1987; Chodorow and Leacock, 2000; Bigert and Knutsson, 2002; Sjöbergh, 2005; Wagner et al., 2007), but limited work on acceptability prediction. Heilman et al. (2014) trained a linear regression model that uses features such as spelling errors, sentence scores from n-gram models and parsers. Lau et al. (2015, 2016) experimented with unsupervised learners and found that a simple RNN was the best performing model. Both works predict acceptability independently of any contextual factors outside the target sentence.

5 Future Work and Conclusions

We found that (i) context positively influences acceptability, particularly for ill-formed sentences, but it also has the reverse effect for well-formed sentences (H_1); (ii) incorporating context (during training or testing) when modelling acceptability improves model performance (H_2); and (iii) prediction performance declines when tested on judgements collected with context, overturning our original hypothesis (H_3). We discovered that human agreement decreases when context is introduced, suggesting that ratings are less predictable in this case.

While it is intuitive that context should improve acceptability for ill-formed sentences, it is less obvious why it reduces acceptability for well-formed sentences. We will investigate this question in future work. We will also experiment with a wider range of models, including sentence embedding methodologies such as Skip-Thought (Kiros et al., 2015).

Acknowledgments

We are grateful to the anonymous ACL reviewers for their helpful comments. Versions of this paper were presented to the Queen Mary University of London NLP Seminar in March 2018, and the Saarland University SFB Language Sciences Colloquium in May 2018. We are grateful to the audiences of both forums, and to colleagues at CLASP for useful discussion of the ideas presented here.

The research of the first two authors was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability in FLoV at the University of Gothenburg.

References

E.S. Atwell. 1987. How to detect grammatical errors in a text without parsing it. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics Morristown, NJ, USA, pages 38–45.

J. Bigert and O. Knutsson. 2002. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *Proc. 2nd Workshop Robust Methods in Analysis of Natural language Data (ROMAND'02), Frascati, Italy.* pages 10–19.

model has no information as to what words will be relevant.

M. Chodorow and C. Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pages 140–147.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland, pages 174–180.

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

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. Montreal, Canada, pages 3294–3302.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic, pages 177–180.

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

Jey Han Lau, Alexander Clark, and Shalom Lappin. 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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1618–1628.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2016. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* pages 1–40.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. Makuhari, Japan, pages 1045–1048.

Hiroshi Nagata. 1988. The relativity of linguistic intuition: The effect of repetition on grammaticality. *Journal of Psycholinguistic Research* 17:1–17.

J. Sjöbergh. 2005. Chunking: an unsupervised method to find errors in text. *NODALIDA2005* page 180.

A. Sorace and F. Keller. 2005. Gradience in linguistic data. *Lingua* 115(11):1497–1524.

J. Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1:123–134.

J. Wagner, J. Foster, and J. Van Genabith. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. *Proceedings of EMNLP-CoNLL-2007* .