[Proceedings of the Society for Computation in Linguistics](https://scholarworks.umass.edu/scil)

Volume 3 Article 43

2020

BLiMP: A Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt
New York University, alexwarstadt@gmail.com

Alicia Parrish
New York University, alicia.v.parrish@nyu.edu

Haokun Liu
New York University, haokunliu@nyu.edu

Anhad Mohananey
New York University, amohananey@ea.com

Wei Peng
New York University, weipeng@nyu.edu

See next page for additional authors

Follow this and additional works at: https://scholarworks.umass.edu/scil

Part of the Computational Linguistics Commons

Recommended Citation

Warstadt, Alex; Parrish, Alicia; Liu, Haokun; Mohananey, Anhad; Peng, Wei; Wang, Sheng-Fu; and Bowman, Samuel R. (2020) "BLiMP: A Benchmark of Linguistic Minimal Pairs for English," *Proceedings of the Society for Computation in Linguistics*: Vol. 3 , Article 43.
DOI: https://doi.org/10.7275/zejz-qs04
Available at: https://scholarworks.umass.edu/scil/vol3/iss1/43

This Abstract is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

BLiMP: A Benchmark of Linguistic Minimal Pairs for English

Authors

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman

This abstract is available in Proceedings of the Society for Computation in Linguistics: https://scholarworks.umass.edu/scil/vol3/iss1/43

BLiMP : A Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt,[1] **Alicia Parrish,**[1] **Haokun Liu,**[3] **Anhad Mohananey,**[3]
Wei Peng,[3] **Sheng-Fu Wang,**[1] **and Samuel R. Bowman**[1,2,3]

[1]NYU Linguistics [2]NYU Data Science [3]NYU Computer Science
Correspondence: warstadt@nyu.edu

Introduction & Prior Work We introduce BLiMP (The **B**enchmark of **Li**nguistic **M**inimal **P**airs, or), a large new benchmark dataset for the targeted evaluation of statistical language models' knowledge of linguistic phenomena. The benchmark consists of 67 datasets, each containing 1000 minimal pairs isolating a specific grammatical contrast and collectively offering broad coverage of major phenomena in English grammar. Like the GLUE benchmark for reusable sentence understanding models (Wang et al., 2018), assigns a single numerical score to a language model (LM) measuring its overall mastery of grammar, enabling straightforward comparison of LMs. The dataset is ideal for fine grained analysis of an LM's knowledge of different grammatical domains. For baselines, we evaluate four representative LMs from NLP literature. We find that is hard even for state-of-the-art models, though Transformers perform better than LSTM and n-gram LMs. Humans overwhelmingly agree with the generated minimal pair contrasts in .

A growing body of work evaluates LSTM LMs' knowledge of grammar by testing whether they prefer acceptable sentences over minimally different unacceptable ones (Linzen et al., 2016, a.o.). So far, results have been mixed, motivating the creation of this benchmark which scales up this kind of investigation to isolate dozens of grammatical contrasts within an otherwise-uniform controlled artificial dataset. Our results show that knowledge of grammar has increased as LM technology progressed from n-grams to LSTMs to Transformers. LSTMs and Transformers alike are very accurate in detecting morphological and agreement violations, but state-of-the-art Transformer LMs have an especially large advantage over LSTMs in contrasts where simple generalizations are difficult to find, such as NPI licensing and island effects.

Data consists of 67 datasets of 1000 minimal pairs each, grouped into twelve broader categories (Table 1). A minimal pair consists of two minimally different sentences where one is grammatically acceptable and the other is not. All minimal pairs in contain the same number of tokens and differ only in word order or the identity of one lexical item, following Marvin and Linzen (2018).

We include minimal pairs illustrating linguistic phenomena well known in morphology, syntax, and semantics. While this set is not exhaustive, it does cover a wide range of topics found in formal implementations of English grammar (e.g., HPSG; generative linguistics textbooks). To fully isolate the phenomena of interest, we use realistic artificially-generated sentences, following Marvin and Linzen, a.o. To generate text, we construct a vocabulary of over 3300 lexical items labeled with features reflecting morphology (e.g. singular/plural), syntax (e.g. transitive/intransitive), and semantics (e.g. animate/inanimate), and build a simple artificial grammar for each paradigm.

We validate the acceptability contrasts in the generated pairs with Mechanical Turk annotators, testing 5 randomly-selected pairs from each paradigm using the same forced-choice task models are presented with. Majority vote of 20 annotators agrees with on at least 4/5 examples from each paradigm and on 96.4% of pairs overall.

Baselines We evaluate 4 baselines: (1) An **n-gram** LM trained on the English Gigaword corpus (Graff et al., 2003), based on a modified Kneser Ney implementation by (Heafield, 2011), which considers up to 5-grams, restricting the model from learning dependencies spanning more than 5 words. (2) An **LSTM** recurrent neural network LM from Gulordava et al. (2018). (3) **Transformer-XL** (Dai et al., 2019), a transformer LM with additional features that enable it to model

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 437-438.
New Orleans, Louisiana, January 2-5, 2020

Phenomenon	N	Acceptable Example	Unacceptable Example
Anaphor agreement	2	*The cats licked themselves.*	*The cats licked itself.*
Argument structure	9	*The cat broke the lamp.*	*The cat vanished the lamp.*
Binding	7	*Bob thinks Ann saw herself.*	*Ann thinks Bob saw herself.*
Control/Raising	5	*The cat is likely to purr.*	*The cat is tough to purr.*
Determiner-Noun agr.	8	*Meg pets those cats.*	*Meg pets that cats.*
Ellipsis	2	*I have a black cat and you have two.*	*I have a cat and you have two black.*
Filler-Gap	7	*The cat noticed the mouse that slept.*	*The cat noticed what the mouse slept.*
Irregular forms	2	*The cat ate the mouse.*	*The cat eaten the mouse.*
Island effects	8	*Whose cat are you petting?*	*Whose are you petting cat?*
NPI licensing	7	*A man who can see Jan hasn't ever left.*	*A man who can't see Jan has ever left.*
Quantifiers	4	*No cat ate more than three treats.*	*No cat ate at least three treats.*
Subject-Verb agr.	6	*The cat that chased the mice sleeps.*	*The cat that chased the mice sleep.*

Table 1: Minimal pairs exemplifying each of the twelve linguistic phenomenon categories covered by . *N* is the number of 1000-example minimal pair paradigms within each category.

model	**Overall**	Ana. Agr	Arg. Str	Binding	Ctrl. Rais.	D-N Agr	Ellipsis	Filler. Gap	Irregular	Island	NPI	Quantifiers	S-V Agr
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	70.8	95.2	73.5	73.2	67.9	84.2	67.3	71.3	92.3	43.9	66.7	62.2	85.1
Transf.-XL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Table 2: Percentage accuracy of four baseline models and raw human performance on using a forced-choice task. A random guessing baseline would give expected accuracy of 50%.

long contiguous inputs of thousands of words during training. (4) **GPT-2** (Radford et al., 2019), a larger neural network LM based on a standard architecture, which is not recurrent and directly models long-distance dependencies.

Our primary evaluation is a forced choice task, in which we test whether a model assigns a higher probability to the acceptable sentence than unacceptable one in each pair. While probability may not correspond to grammaticality when comparing very different sentences, we expect this to be a viable proxy when comparing minimally different sentences as in our data. Additional metrics using word-level probabilities to more narrowly isolate model behavior yield broadly similar conclusions.

Results & Discussion We report model accuracy for the 12 broad categories (Table 2). Overall, the state-of-the-art GPT-2 achieves the highest score and the n-gram the lowest, though all models perform significantly below humans. We find that some phenomena are easier than others: determiner-noun agreement is easy for all models, while islands are quite difficult. We replicate Marvin and Linzen's finding that LSTMs succeed at subject-verb agreement and to some extent binding/anaphora, but largely fail at NPI licensing.

The n-gram model's poor overall performance confirms is not solvable from co-occurrence information alone. Rather, success at is driven by the more abstract (and less interpretable) features learned by neural networks. There are a few exceptions to this pattern: n-grams are mostly sufficient to capture irregular verb forms. Furthermore, SoTA models still show little improvement over n-grams on some phenomena, such as quantifier restrictions and, most strikingly, island effects.

Conclusion We have offered a human-solvable challenge set that covers a broad overview of major grammatical phenomena in English. is hard even for SotA models, though recent large-scale Transformers outperform simple baselines.

References

Z. Dai, Z. Yang, Y. Yang, W. Cohen, J. Carbonell, Q. Le, and R. Salakhutdinov. 2019. TransformerXL.

D. Graff, J. Kong, K. Chen, and K. Maeda. 2003. English gigaword.

K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically.

K. Heafield. 2011. KenLM.

T. Linzen, E. Dupoux, and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies.

R. Marvin and T. Linzen. 2018. Targeted syntactic evaluation of language models.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. GLUE.