Can a Gorilla Ride a Camel?
Learning Semantic Plausibility from Text

Ian Porada[1], Kaheer Suleman[2], and Jackie Chi Kit Cheung[1,3]

[1]Mila, McGill University
{ian.porada@mail, jcheung@cs}.mcgill.ca
[2]Microsoft Research Montreal
kasulema@microsoft.com
[3]Canada CIFAR AI Chair

Abstract

Modeling semantic plausibility requires commonsense knowledge about the world and has been used as a testbed for exploring various knowledge representations. Previous work has focused specifically on modeling physical plausibility and shown that distributional methods fail when tested in a supervised setting. At the same time, distributional models, namely large pretrained language models, have led to improved results for many natural language understanding tasks. In this work, we show that these pretrained language models are in fact effective at modeling physical plausibility in the supervised setting. We therefore present the more difficult problem of learning to model physical plausibility directly from text. We create a training set by extracting attested events from a large corpus, and we provide a baseline for training on these attested events in a self-supervised manner and testing on a physical plausibility task. We believe results could be further improved by injecting explicit commonsense knowledge into a distributional model.

1   Introduction

A person riding a camel is a common event, and one would expect the subject-verb-object (s-v-o) triple *person-ride-camel* to be attested in a large corpus. In contrast, *gorilla-ride-camel* is uncommon, likely unattested, and yet still semantically plausible. Modeling semantic plausibility then requires distinguishing these plausible events from the semantically nonsensical, e.g. *lake-ride-camel*.

Semantic plausibility is a necessary part of many natural language understanding (NLU) tasks including narrative interpolation (Bowman et al., 2016), story understanding (Mostafazadeh et al., 2016), paragraph reconstruction (Li and Jurafsky, 2017), and hard coreference resolution (Peng

Event	Plausible?
bird-construct-nest	✓
bottle-contain-elephant	✗
gorilla-ride-camel	✓
lake-fuse-tie	✗

Table 1: Example events from Wang et al. (2018)'s physical plausibility dataset.

et al., 2015). Furthermore, the problem of modeling semantic plausibility has itself been used as a testbed for exploring various knowledge representations.

In this work, we focus specifically on modeling physical plausibility as presented by Wang et al. (2018). This is the problem of determining if a given event, represented as an s-v-o triple, is physically plausible (Table 1). We show that in the original supervised setting a distributional model, namely a novel application of BERT (Devlin et al., 2019), significantly outperforms the best existing method which has access to manually labeled physical features (Wang et al., 2018).

Still, the generalization ability of supervised models is limited by the coverage of the training set. We therefore present the more difficult problem of learning physical plausibility directly from text. We create a training set by parsing and extracting attested s-v-o triples from English Wikipedia, and we provide a baseline for training on this dataset and evaluating on Wang et al. (2018)'s physical plausibility task. We also experiment training on a large set of s-v-o triples extracted from the web as part of the NELL project (Carlson et al., 2010), and find that Wikipedia triples result in better performance.

2 Related Work

Wang et al. (2018) present the semantic plausibility dataset that we use for evaluation in this work, and they show that distributional methods fail on this dataset. This conclusion aligns with other work showing that GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013) embeddings do not encode some salient features of objects (Li and Gauthier, 2017). More recent work has similarly concluded that large pretrained language models only learn attested physical knowledge (Forbes et al., 2019).

Other datasets which include plausibility ratings are smaller in size and missing atypical but plausible events (Keller and Lapata, 2003), or concern the more complicated problem of multi-event inference in natural language (Zhang et al., 2017; Sap et al., 2019).

Complementary to our work are methods of extracting physical features from a text corpus (Wang et al., 2017; Forbes and Choi, 2017; Bagherinezhad et al., 2016).

2.1 Distributional Models

Motivated by the distributional hypothesis that words in similar contexts have similar meanings (Harris, 1954), distributional methods learn the representation of a word based on the distribution of its context. The occurrence counts of bigrams in a corpus are correlated with human plausibility ratings (Lapata et al., 1999, 2001), so one might expect that with a large enough corpus, a distributional model would learn to distinguish plausible but atypical events from implausible ones. As a counterexample, Ó Séaghdha (2010) has shown that the subject-verb bigram *carrot-laugh* occurs 855 times in a web corpus, while *manservant-laugh* occurs zero.[1] Not everything that is physically plausible occurs, and not everything that occurs is attested due to reporting bias[2] (Gordon and Van Durme, 2013); therefore, modeling semantic plausibility requires systematic inference beyond a distributional cue.

We focus on the masked language model BERT as a distributional model. BERT has led to improved results across a variety of NLU bench-marks (Rajpurkar et al., 2018; Wang et al., 2019), including tasks that require explicit commonsense reasoning such as the Winograd Schema Challenge (Sakaguchi et al., 2019).

2.2 Selectional Preference

Closely related to semantic plausibility is selectional preference (Resnik, 1996) which concerns the semantic preference of a predicate for its arguments. Here, *preference* refers to the typicality of arguments: while it is plausible that a gorilla rides a camel, it is not preferred. Current approaches to selectional preference are distributional (Erk et al., 2010; Van de Cruys, 2014) and have shown limited performance in capturing semantic plausibility (Wang et al., 2018).

Ó Séaghdha and Korhonen (2012) have investigated combining a lexical hierarchy with a distributional approach, and there have been related attempts at grounding selectional preference in visual perception (Bergsma and Goebel, 2011; Shutova et al., 2015).

Models of selectional preference are either evaluated on a pseudo-disambiguation task, where attested predicate-argument tuples must be disambiguated from pseudo-negative random tuples, or evaluated on their correlation with human plausibility judgments. Selectional preference is one factor in plausibility and thus the two should correlate.

3 Task

Following existing work, we focus on the task of single-event, physical plausibility. This is the problem of determining if a given event, represented as an s-v-o triple, is physically plausible.

We use Wang et al. (2018)'s physical plausibility dataset for evaluation. This dataset consists of 3,062 s-v-o triples, built from a vocabulary of 150 verbs and 450 nouns, and containing a diverse combination of both typical and atypical events balanced between the plausible and implausible categories. The set of events and ground truth labels were manually curated.

3.1 Supervised

In the supervised setting, a model is trained and tested on labelled events from the same distribution. Therefore, both the training and test set capture typical and atypical plausibility. We follow the same evaluation procedure as previous work

---

[1] This point was made based on search engine results. Some, but not all, of the *carrot-laugh* bigrams are false positives.

[2] Reporting bias describes the discrepancy between what is frequent in text and what is likely in the world. This is in part because people do not describe the obvious.

Wikipedia	*male-have-income*
	village-have-population
	event-take-place
NELL	*login-post-comment*
	use-constitute-acceptance
	modules-have-options

Table 2: Most frequent s-v-o triples for each corpus.

and perform cross validation on the 3,062 labeled triples (Wang et al., 2018).

3.2 Learning from Text

We also present the problem of learning to model physical plausibility directly from text. In this new setting, a model is trained on events extracted from a large corpus and evaluated on a physical plausibility task. Therefore, only the test set covers both typical and atypical plausibility.

We create two training sets based on separate corpora: first, we parse English Wikipedia using the StanfordNLP neural pipeline (Qi et al., 2018) and extract attested s-v-o triples. Wikipedia has led to relatively good results for selectional preference (Zhang et al., 2019), and in total we extract 6 million unique triples with a cumulative 10 million occurrences. Second, we use the NELL (Carlson et al., 2010) dataset of 604 million s-v-o triples extracted from the dependency parsed ClueWeb09 dataset. For NELL, we filter out triples with non-alphabetic characters or less than 5 occurrences, resulting in a total 2.5 million unique triples with a cumulative 112 million occurrences.

For evaluation, we split Wang et al. (2018)'s 3,062 triples into equal sized validation and test sets. Each set thus consists of 1,531 triples.

4 Methods

4.1 NN

As a baseline, we consider the performance of a neural method for selectional preference (Van de Cruys, 2014). This method is a two-layer artificial neural network (NN) over static embeddings.

Supervised. We reproduce the results of Wang et al. (2018) using GloVe embeddings and the same hyperparameter settings.

Self-Supervised. We use this same method for *learning from text* (Subsection 3.2). To do so, we turn the training data into a self-supervised train-ing set: attested events are considered to be plausible, and pseudo-implausible events are created by sampling each word in an s-v-o triple independently by occurrence frequency. We do hyperparameter search on the validation set over learning rates in $\{1e-3, 1e-4, 1e-5, 2e-5\}$, batch sizes in $\{16, 32, 64, 128\}$, and epochs in $\{0.5, 1, 2\}$.

4.2 BERT

We use BERT for modeling semantic plausibility by simply treating this as a sequence classification task. We tokenize the input s-v-o triple and introduce new entity marker tokens to separate each word.[3] We then add a single layer NN to classify the input based on the final layer representation of the [CLS] token. We use BERT-large and fine-tune the entire model in training.[4]

Supervised. We do no hyperparameter search and simply use the default hyperparameter configuration which has been shown to work well for other commonsense reasoning tasks (Ruan et al., 2019). BERT-large sometimes fails to train on small datasets (Devlin et al., 2019; Niven and Kao, 2019); therefore, we restart training with a new random seed when the training loss fails to decrease more than 10%.

Self-Supervised. We perform *learning from text* (Subsection 3.2) by creating a self-supervised training set in exactly the same way as for the NN method. The hyperparameter configuration is determined by grid search on the validation set over learning rates in $\{1e-5, 2e-5, 3e-5\}$, batch sizes in $\{8, 16\}$, and epochs in $\{0.5, 1, 2\}$.

5 Results

5.1 Supervised

For the supervised setting, we follow the same evaluation procedure as Wang et al. (2018): we perform 10-fold cross validation on the dataset of 3,062 s-v-o triples, and report the mean accuracy of running this procedure 20 times all with the same model initialization (Table 3).

BERT outperforms existing methods by a large margin, including those with access to manually labeled physical features. We conclude from

---

[3]Our input to BERT is of the form: [CLS] [SUBJ] <subject> [/SUBJ] [VERB] <verb> [/VERB] [OBJ] <object> [/OBJ] [SEP].

[4]We use Hugging Face's PyTorch implementation of BERT, https://github.com/huggingface/pytorch-transformers.

Model	Accuracy
Random	0.50
NN (Van de Cruys, 2014)	0.68
NN+WK (Wang et al., 2018)	0.76
Fine-tuned BERT	**0.89**

Table 3: Mean accuracy of classifying plausible events for models trained in a supervised setting. NN+WK combines the NN approach with manually labeled world knowledge (WK) features describing both the subject and object.

Event	Plausible?	
	BERT	GT
dentist-capsize-canoe	✓	✓
stove-heat-air	✗	✓
sun-cool-water	✓	✗
chair-crush-water	✗	✗

Table 4: Interpreting log-likelihood as confidence, example events for which BERT was highly confident and either correct or incorrect with respect to the ground truth (GT) label.

these results that distributional data does provide a strong cue for semantic plausibility in the supervised setting of Wang et al. (2018).

Examples of positive and negative results for BERT are presented in Table 4. There is no immediately obvious pattern in the cases where BERT misclassifies an event. We therefore consider events for which BERT gave a consistent estimate across all 20 runs of cross-validation. Of these, we present the event for which BERT was most confident.

We note that due to the limited vocabulary size of the dataset, the training set always covers the test set vocabulary when performing 10-fold cross validation. That is to say that every word in the test set has been seen in a different triple in the training set. For example, every verb occurs within 20 triples; therefore, on average a verb in the test set has been seen 18 times in the training set.

Supervised performance is dependent on the coverage of the training set vocabulary (Moosavi and Strube, 2017), and it is intractable to have 18 plausibility labels for all verbs across English. Furthermore, supervised models are susceptible to annotation artifacts (Gururangan et al., 2018; Poliak et al., 2018) and do not necessarily even learn

Model	Wikipedia		NELL	
	Valid	Test	Valid	Test
Random	0.50	0.50	0.50	0.50
NN	0.53	0.52	0.50	0.51
BERT	**0.65**	**0.63**	**0.57**	**0.56**

Table 5: Accuracy of classifying plausible events for models trained on a corpus in a self-supervised manner.

the desired relation, or in fact any relation, between words (Levy et al., 2015).

This is our motivation for reframing semantic plausibility as a task to be learned directly from text, a new setting in which the training set vocabulary is independent of the test set.

5.2 Learning from Text

For *learning from text* (Subsection 3.2), we report both the validation and test accuracies of classifying physically plausible events (Table 5).

BERT fine-tuned on Wikipedia performs the best, although only partially captures semantic plausibility with a test set accuracy of 63%. Performance may benefit from injecting explicit commonsense knowledge into the model, an approach which has previously been used in the supervised setting (Wang et al., 2018).

Interestingly, BERT is biased towards labelling events as plausible. For the best performing model, for example, 78% of errors are false positives.

Models trained on Wikipedia events consistently outperform those trained on NELL which is consistent with our subjective assessment of the cleanliness of these datasets. The baseline NN method in particular seems to learn very little from training on the NELL dataset.

6 Conclusion

We show that large, pretrained language models are effective at modeling semantic plausibility in the supervised setting. Supervised models are limited by the coverage of the training set, however; thus, we reframe modeling semantic plausibility as a self-supervised task and present a baseline based on a novel application of BERT.

We believe that self-supervised results could be further improved by incorporating explicit commonsense knowledge, as well as further incidental signals (Roth, 2017) from text.

Acknowledgments

References

Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are elephants bigger than butterflies? reasoning about sizes of objects. In *AAAI*.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 399–405, Hissar, Bulgaria. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1306–1313. AAAI Press.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 26–35, Doha, Qatar. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects.

In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Maria Lapata, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgements. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 354–361, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maria Lapata, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 30–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Christian Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *HLT-NAACL*.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Lucy Li and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 435–444, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diarmuid Ó Séaghdha and Anna Korhonen. 2012. Modelling selectional preferences in a lexical hierarchy. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 170–179, Montréal, Canada. Association for Computational Linguistics.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819, Denver, Colorado. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Dan Roth. 2017. Incidental supervision: Moving beyond supervised learning. In *AAAI*.

Yu-Ping Ruan, Xiaodan Zhu, Zhen-Hua Ling, Zhan Shi, Quan Liu, and Si Wei. 2019. Exploring unsupervised pretraining and sentence structure modelling for winograd schema challenge. *CoRR*, abs/1904.09705.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *ArXiv*, abs/1907.10641.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Ekaterina Shutova, Niket Tandon, and Gerard de Melo. 2015. Perceptually grounded selectional preferences. 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)*, pages 950–960, Beijing, China. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Su Wang, Stephen Roller, and Katrin Erk. 2017. Distributional modeling on a diet: One-shot word learning from text only. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 204–213, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.