Investigating the Performance of Transformer-Based NLI Models on Presuppositional Inferences

Jad Kabbara and **Jackie Chi Kit Cheung**[†]
School of Computer Science, McGill University, Montreal, QC, Canada
Montreal Institute for Learning Algorithms (Mila), Montreal, QC, Canada
{jad, jcheung}@cs.mcgill.ca
[†] Canada CIFAR AI Chair

Abstract

Presuppositions are assumptions that are taken for granted by an utterance, and identifying them is key to a pragmatic interpretation of language. In this paper, we investigate the capabilities of transformer models to perform NLI on cases involving presupposition. First, we present simple heuristics to create alternative "contrastive" test cases based on the ImpPres dataset and investigate the model performance on those test cases. Second, to better understand how the model is making its predictions, we analyze samples from sub-datasets of ImpPres and examine model performance on them. Overall, our findings suggest that NLI-trained transformer models seem to be exploiting specific structural and lexical cues as opposed to performing some kind of pragmatic reasoning.

1 Introduction

Natural language inference (NLI) is the task of predicting whether a sentence entails, contradicts or is neutral with respect to another sentence. While NLI is presented as a task with the goal of pushing the frontier of language understanding, one question that remains elusive is the extent to which learning models trained on NLI datasets acquire competence in pragmatic reasoning. For example, "Rose's drawing stunned the audience" presupposes that "Rose has a drawing." In terms of the NLI task, the relationship between these two statements is one of entailment. However, "Rose stunned the audience" does not trigger the same presupposition. If an NLI system relies on an incorrect heuristic to predict entailment, such as the fact that both statements have "Rose" in them, the system would fail on the modified version.

While there exists a plethora of NLI datasets (e.g. SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), HANS (McCoy et al., 2019)), issues related to pragmatics remain under-explored. One exception is the ImpPres dataset (Jeretic et al., 2020), which contains English NLI instances involving two pragmatic phenomena: implicature and presupposition.

In linguistics and philosophy of language, presuppositions are assumptions and beliefs that are shared and taken for granted by discourse participants without an explicit mention in the discourse context. Presuppositions are prevalent in language and understanding them facilitates smooth communication and is crucial for a proper understanding of the meaning being conveyed in a given context.

In this work, we investigate how well transformer models perform on cases from ImpPres involving presupposition. We ask the following question: In cases where these models seem to be doing well in terms of (NLI) accuracy performance, are they exploiting any patterns or correlations between certain words and decision labels which happen to lead to a good performance on ImpPres, but which might not generalize beyond it?

Our contribution is two-fold: First, we present simple heuristics to create alternative "contrastive" test cases based on the Presupposition sub-datasets in ImpPres and investigate the model performance on those test cases. Second, from those datasets, we draw a random set of testing samples and examine the performance of the models on these sets. In each of those cases, we identify specific patterns or cues that seem to be influencing the accuracy performance (in either the correct or incorrect predictions). Our findings suggest that transformer-based NLI models seem to be exploiting surface-level lexical and structural cues as opposed to performing some kind of pragmatic reasoning and that these cues are highly task-specific and do not generalize to adversarially perturbed versions of the input.

2 Related Work

Natural language inference (NLI) is the task of predicting whether one passage entails another. In

Type	Premise	Hypothesis	Label
All N	All six roses that bloomed died.	Exactly six roses bloomed.	E
Both	Both flowers that bloomed died.	Exactly two flowers bloomed.	E
Change of State	Rene might have hidden.	Rene hid.	N
Cleft Existence	It might be Becky who researched Jesus.	Someone researched Jesus.	E
Cleft Uniqueness	It is Joel who helps Diana.	Exactly one person helps Victoria.	N
Only	Susan only writes.	Susan doesn't write.	C
Possessed Definites	Rose's bird did alarm Peter.	Rose has a bird.	E
Question	Did Bill wonder when Omar hunted?	Omar didn't hunt.	C

Table 1: Examples showing the different presupposition types in the ImpPres dataset.

our work, we adopt the modern formulation[1] cast as a three-way classification task where two statements, a premise and a hypothesis, are bound by a relationship of entailment, contradiction or neither (the "neutral" case). Many datasets have been introduced for the task including SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), MPE (Lai et al., 2017), XNLI (Conneau et al., 2018), etc.

The ImpPres dataset (Jeretic et al., 2020) has been proposed to focus specifically on presupposition and implicature. ImpPres consists of 25.5k sentence pairs further divided into several sub-datasets, each focusing on a specific type of implicature or presupposition. In our work, we focus on the presupposition part of ImpPres. Table 1 presents one example from each sub-dataset.

Other work studied the same general issue of checking whether systems systematically changed their labels on perturbed samples in an expected way (Ribeiro et al., 2020; Emami et al., 2019; Sinha et al., 2021; Gardner et al., 2020; Niven and Kao, 2019). The HANS dataset (McCoy et al., 2019) in particular targets issues where NLI models fail because they rely on superficial syntactic heuristics. Unlike our work, the NLI instances they examine—and NLI in general—often require additional world knowledge, or linguistic knowledge that is highly lexicalized (e.g., a list of factive verbs). By contrast, the class of presuppositions that we examine is tied to specific linguistic constructions that systematically trigger them, such as definite articles and cleft constructions, which could potentially be easier. We nevertheless show that the models we tested fail to generalize systematically on presuppositional inferences, complementing previous results.

Also relevant to this work is the CommitmentBank dataset (De Marneffe et al., 2019) which focuses on issues involving presupposition projection in various environments. This work highlights that projection is hardly a binary affair, and that judgments about entailment (roughly corresponding to speakers' commitment in the language of CommitmentBank) are graded.

3 Methodology

Our study consists of fine-tuning a pre-trained transformer model (either BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019)), then evaluating its performance on the various presupposition sub-datasets of ImpPres in addition to perturbed versions of those sub-datasets introduced in this work. We fine-tune on MNLI, following Jeretic et al. (2020).

We conduct two studies in order to investigate the performance of these models and better understand their capabilities in making pragmatic inferences. In the first study, we present simple heuristics to create "contrastive" test cases. We then evaluate the finetuned model on those test cases quantitatively. In the second study, we draw from each sub-dataset of ImpPres a random set of 100 samples. We examine these sets for structural and lexical cues that might be influencing the performance of BERT.

3.1 Training details

Similar to previous work using BERT-based models for NLI, we concatenate the premise and hypothesis separated by the [SEP] token, with the special [CLS] token preceding them. For the BERT model, we use HuggingFace's bert-large-uncased implementation (Wolf et al., 2019) of a 24-layer, 1024-hidden, 16-heads, 336M-parameter version of the model that was trained on lower-cased English text. Similarly, for RoBERTa, we use the roberta-large implementation. All models are implemented in PyTorch (Paszke et al., 2019) and

[1] For earlier (two-way classification) formulations, the reader is referred to (Dagan et al., 2005) and (Manning, 2006)

trained to minimize the standard cross-entropy cost with Adam (Kingma and Ba, 2015) as the optimizer with all default parameters except for the learning rate. All model hyperparameters are kept as default except for the following: We follow the recommended ranges for fine-tuning hyperparameters in the BERT paper (Devlin et al., 2019) and find that the optimal performance on the dev set is reached for a batch size of 16, learning rate of 2e-5 and weight decay of 0.01 for BERT, and learning rate of 1e-5 and weight decay of 0.001 for BERT. The BERT model achieves a dev set accuracy of 85.14% (comparable to that reported in (Jeretic et al., 2020) and (Devlin et al., 2019)).

4 Study 1: Contrastive Probing

In this study, we present simple heuristics to create "contrastive" test cases for the various ImpPres sub-datasets and investigate model performance on those test cases. We divide our probes into two types: those involving contrastive sentences which are grammatical English sentences, and those with corrupt input which are ungrammatical.

4.1 Probing with Valid English Sentences

Removing Possessives (POSS): In the first experiment, we focus on the "Possessed Definites Existence" subset of ImpPres. Possessive noun phrases trigger a presupposition about the existence of the possessed noun, and its possession by the possessor. For example, "Alice's painting is amazing" (premise) presupposes "Alice has a painting" (hypothesis), which corresponds to a gold label of entailment. We create a contrastive example by removing the possessed noun (or noun phrase) in the premise: "Alice is amazing", while keeping the hypothesis the same, changing the correct label to neutral. Applying this rule to all samples in this sub-dataset, all labels become neutral. We sampled and manually checked 100 cases and found that this heuristic is correct for all 100 sampled cases.

Replacing Names (NAMES): We process the premises using spaCy's Named Entity Recognizer (Honnibal et al., 2020), identify all instances of names of people and replace each by a (different) randomly selected name from the 200 most common names in USA.[2] For samples where a change is made, as the premise and hypothesis are now about different people, the label becomes neutral.

Label	Original Accuracy	Contrastive Accuracy
Entailment	99.2 %	-
Contradiction	99.0 %	-
Neutral	31.5 %	62.4 %
Label	Original Gold Labels	Contrastive Gold Labels
Entailment	26.3 %	0 %
Contradiction	31.6 %	0 %
Neutral	42.1 %	100%
Label	Original Predictions	Contrastive Predictions
Entailment	37.2 %	0.4 %
Contradiction	49.5 %	37.2 %
Neutral	13.3 %	62.4 %

Table 2: Per-class accuracy, distribution of golden labels (in %), distribution of predictions in the original and contrastive dataset (in %) for the BERT model for the POSS experiment (which focuses only on the "Possessed Definites Existence" sub-dataset.

4.2 Contrastive Probing with Corrupt Input

We are interested in probing the model with test cases that are corrupt in a way that the model is expected to perform worse. For example, if the model is provided with less context in the input or if we were to randomize the word order, the model is expected to perform worse.

Providing Less Context in the Premise (1ST-HALF): In this experiment, instead of a full premise, we present the first half of the sentence. The hypothesis and labels are kept unchanged.

Randomizing Word Order (RNDMZ): Here, we shuffle randomly the order of tokens in the premise and hypothesis and keep the labels unchanged.

4.3 Results and Discussion

For the POSS experiment, the accuracy drops from 70.6% in the original sub-dataset to 62.4% in the contrastive dataset for the BERT model and from 64.8% to 35.2% for the RoBERTa model. The sharp drops suggest the model failed to pick up on the deletion of the possessive which altered the pragmatic context. We present additional statistics in Table 2 for the BERT case. While the model had near-perfect accuracy for the entailment and contradiction labels, it performed at less than chance for the neutral case. While the labels are relatively spread across the three labels, the model's original predictions of "neutral" were less than a fourth of the expected number which explains the low accuracy per class (31.5%). From Table 3, while the original predictions of "contradiction" should have been switched to neutral, the majority of these predictions did not change. This shows that the model

[2] www.ssa.gov/oact/babynames/decades/century.html

		Original Predictions		
		Ent: 707	Cont: 941	Neu: 252
Changed to	Entailment	4	3	0
	Contradiction	148	529	30
	Neutral	555	409	222

Table 3: Break down of how BERT's predictions changed from the original to the contrastive dataset in the (POSS) experiment. As a reminder, here in the contrastive set, all samples become neutral.

failed to translate the change in context to a correct change in prediction.

Tables 4 and 5 present the results of the three experiments NAMES, 1ST-HALF and RNDMZ for the two models BERT and RoBERTa.

NAMES. The most striking result is that changing names led to a significant drop in accuracy with a sharp decrease seen in several datasets (e.g., going from 70.6% to 26.3% (BERT) and 64.8% to 22.6% (RoBERTa)). For RoBERTa, accuracy numbers decreased across all datasets. For BERT, the drop was seen for the subdatasets with the highest performance but, interestingly, in the datasets where the performance was low, the new performance was actually higher. This could be explained by the fact that in those cases, the model was performing very poorly to start with so for these cases the results might be not suggestive.

1ST-HALF. Here, for both BERT and RoBERTa, we notice that in most cases (13/18 cases), the contrastive accuracy was higher. This observation is counter-intuitive because one would expect that providing less context to the model will impact the performance negatively. Indeed, with the exception of the Possessed Definites Existence dataset where the presupposition is consistently in the first half of the premise, the remaining sub-datasets are quite diversified in where the presupposition appears in the premise (i.e., first half vs second half). Moreover, upon inspection of samples from various datasets, we see how cutting off the second half of the premise often removes key parts of the context, thus making it virtually impossible to establish the (original) presupposition based on that (cut) premise.

RNDMZ. In the last experiment, we notice a drop in the performance in several datasets for both models. However, the drop is not as large as one might expect given that the input is no longer grammatical or coherent.

	Original	Contrastive Accuracy		
	Accuracy	NAMES	1ST-HALF	RNDMZ
Poss. Def. Existence	70.6	26.3	67.6	52.2
Question	66.4	28.6	50.2	52.3
Cleft Existence	63.0	15.8	65.8	53.7
Only	62.3	39.4	32.9	52.7
All n	43.5	44.7	46.7	40.7
Both	32.6	42.6	41.8	34.0
Change of state	30.4	36.8	37.3	27.8
Poss. Def. Uniqueness	23.3	46.3	27.0	36.4
Cleft Uniqueness	11.1	20.3	36.0	21.8

Table 4: Accuracy results for the BERT model.

	Original	Contrastive Accuracy		
	Accuracy	NAMES	1ST-HALF	RNDMZ
Poss. Def. Existence	64.8	22.6	66.1	50.6
Cleft Existence	63.3	18.2	65.1	57.1
Question	61.8	26.0	55.2	50.8
Poss. Def. Uniqueness	56.9	13.0	60.7	41.1
Only	55.0	35.4	47.3	46.7
All n	50.8	35.0	56.7	38.7
Both	49.1	45.6	55.9	39.4
Change of state	36.1	35.5	42.3	37.7
Cleft Uniqueness	26.7	12.4	41.1	27.4

Table 5: Accuracy results for the RoBERTa model.

5 Study 2: Sample Error Analysis

We examine BERT's performance on 100 samples randomly drawn from each sub-dataset. We see patterns repeating across several datasets—all about surface cues which should not be directly responsible for presupposition. We group our insights into themes.

5.1 Exploiting Lexical Cues

One pattern that we noticed across several datasets is that certain tokens and negation heavily affect performance. In the "All n" sub-dataset, we notice that the token "exactly" appears in the hypothesis in 76% of samples (e.g., "All six roses that bloomed died." / "Exactly six roses bloomed"). For these samples, the accuracy was at 48.68%. For the remaining 24% (not having "exactly"), the accuracy was at 16.67%. A closer look at those 76% samples shows that samples that have negation in the premise had an accuracy of 72% versus 44.61% for those without negation. Interestingly, these results are closely replicated in another sub-dataset on the similar "both" presupposition effect. In this dataset, we find that there are 69% of samples having "exactly" in their premise. For these samples, the accuracy is at 42.02%. For the remaining 31% (i.e. not having "exactly" in the premise), the accuracy is at 19.35%. Similarly, in samples that have negation in their premise, the accuracy is 66.67% versus 33.33% for samples with no negation.

In the Cleft Existence dataset (e.g., “It is Keith who stunned Christina.” / “Keith stunned Christina.”), the model properly learns with 100% accuracy the association between “no one” and the “contradiction” label (e.g. “It is Helen that talked about Cheryl.” / “No one talked about Cheryl.”). However, for the samples not having “no one”, the accuracy drops to 52.78%. This could suggest that the correct prediction of “contradiction” in these cases was likely based on a cue, the presence of “no one” as opposed to some pragmatic understanding of the presupposition at hand. Furthermore, while the model has learnt with perfect accuracy the connection between “no one” and a label of contradiction, it does less well with other samples where “someone” replaces the main subject in the premise (which is crucial for understanding the cleft presupposition). In those cases, the accuracy is at 68.75%, suggesting that the presence of “no one” had a stronger impact on the model’s performance.

In principle, such patterns shouldn’t occur with a model that can systemically reason about presuppositions; yet we found them in our analysis and they echo findings (Degen, 2015; Schuster et al., 2020) that negation and certain tokens can affect pragmatic inferences.

5.2 Exploiting Similarity Information

In the Change of State sub-dataset (e.g. “Rene might have hidden.” / “Rene hid.”), we notice cases where the same verb appears in both the premise and hypothesis or appears in a close morphological variation (e.g., hidden/hid, gotten/got). In 85% of those cases, the model predicted “entailment”—all incorrectly. This suggests that the model might be learning to associate “entailment” with some kind of similarity between the premise and hypothesis, which is incorrect in these cases. Previous work also noted how the similarity between two statements can affect inferences related to pragmatic phenomena (e.g., implicature (Degen, 2015)). In the “Only” sub-dataset (e.g., “Susan only writes.” / “Susan writes.”), samples differed in one main aspect: For some, the same subject appeared in both the premise and hypothesis. For others, there was no subject agreement. We found that 88% of the correctly predicted samples had the same subject with a majority of “entailment” predictions (55%). Furthermore we noticed that among the incorrectly predicted samples with same subjects, 75% were also predicted as “entailment”. This suggests that the model might have been exploiting cues on subject agreement to make its predictions.

In the Question sub-dataset (e.g. “Did Bill wonder when Omar hunted?” / “Omar hunted.”), in 51% of samples, word-to-word phrases/expressions appeared in both the premise and hypothesis. The model predicted entailment in 96% of these cases. However, the accuracy was lower—only 72.54%.

5.3 Exploiting Structural Information

In several datasets, a common pattern found in the premise is the structure “if ..., it’s okay” (e.g. Premise: “If Amanda had left, it’s okay.” / Hypothesis: “Amanda used to be here.”). In the Change of State dataset, in 81.25% of such samples, the model wrongly predicted “neutral”. Similarly, in the Possessed Definites dataset, in 85% of such samples, the model wrongly predicted “entailment”.

In the Cleft Uniqueness dataset, another common pattern is the structure “it is ... who ...” in the hypothesis (e.g.: Premise: “It is Sandra who disliked Veronica.” / Hypothesis: “Exactly one person disliked Veronica.”). For 76.9% of such samples, the model predicted entailment—all incorrectly. Similarly, samples where the hypothesis had “exactly one” were predicted as contradiction 70% of the time–again, all incorrect.

In all these cases, the model seemed to have learned incorrect associations between syntactic patterns and presuppositional entailment decisions.

6 Conclusion

We investigated BERT’s capabilities to perform NLI on cases involving presupposition. Our analysis suggests that NLI-trained BERT exploits lexical and structural cues to do so, and that these cues are highly task-specific and do not generalize to adversarially perturbed versions of the input.

Acknowledgements

The authors would like to thank the reviewers for their valuable comments. This work is partially supported by the Canada CIFAR AI Chair Program held by Jackie CK Cheung at Mila. Jad Kabbara is also supported by an IVADO PhD fellowship.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Judith Degen. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8:11–1.

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.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceeding of the 2015 International Conference on Learning Representation (ICLR 2015)*, San Diego, California.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher Manning. 2006. Local textual inference : It's hard to circumscribe , but you know it when you see it - and nlp needs it.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.