Beyond the Tip of the Iceberg: Assessing Coherence of Text Classifiers

Shane Storks and **Joyce Chai**
Computer Science and Engineering Division
University of Michigan
Ann Arbor, MI 48109, USA
{sstorks, chaijy}@umich.edu

Abstract

As large-scale, pre-trained language models achieve human-level and superhuman accuracy on existing language understanding tasks, statistical bias in benchmark data and probing studies have recently called into question their true capabilities. For a more informative evaluation than accuracy on text classification tasks can offer, we propose evaluating systems through a novel measure of prediction coherence. We apply our framework to two existing language understanding benchmarks with different properties to demonstrate its versatility. Our experimental results show that this evaluation framework, although simple in ideas and implementation, is a quick, effective, and versatile measure to provide insight into the coherence of machines' predictions.

1 Introduction

Large-scale, pre-trained contextual language representations (Devlin et al., 2018; Radford et al., 2018; Raffel et al., 2020; Brown et al., 2020) have approached or exceeded human performance on many existing language understanding benchmarks. However, due to increasing complexity and concerns of statistical bias enabling artificially high performance (Schwartz et al., 2017; Poliak et al., 2018b; Niven and Kao, 2019; Min et al., 2020), the coherence of these state-of-the-art systems and their alignment to humans is not well understood.

This is perhaps because benchmarks geared toward language understanding only cover the tip of the iceberg, typically focusing on a high-level end task rather than diving deeper into the kind of coherent, robust understanding that takes place in humans. Language understanding in machines is often boiled down to text classification, where a classifier is tasked with recognizing whether a text contains a particular semantic class, e.g., textual entailment (Dagan et al., 2005; Bowman et al., 2015), commonsense implausibility (Roemmele et al., 2011; Mostafazadeh et al., 2016; Bisk et al., 2020), or combinations of several phenomena meant to serve as comprehensive diagnostics (Poliak et al., 2018a; Wang et al., 2018, 2019). Without regard to the underlying evidence used to reach a conclusion, systems are rewarded for correct predictions on the task without "showing their work."

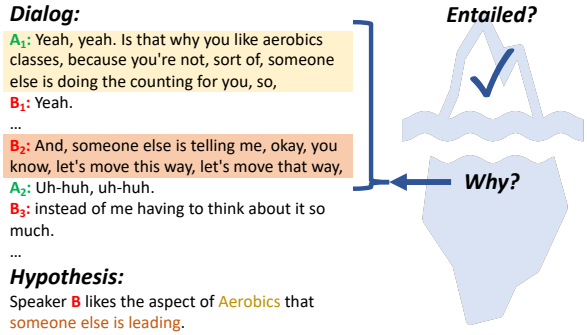

Figure 1: In Conversational Entailment (Zhang and Chai, 2010), systems only predict whether a hypothesis is entailed by a dialog, while ignoring the underlying evidence in the discourse toward this conclusion.

To make meaningful improvement on machine language understanding, it is important to have more informative performance measures. To address this issue, the contribution of this paper is to introduce a novel model- and task-agnostic evaluation framework that allows a quick assessment of text classifiers' ability in terms of the coherence of their predictions. We apply our framework to two existing language understanding benchmarks of different genres to demonstrate its versatility. Our results support recent findings of spurious behaviors in fine-tuned large LMs, and show that our framework, although simple in ideas and implementation, is effective as a quick measure to provide insight into the coherence of machines' predictions.

2 Related Work

In the face of data bias and uninterpretability of large LMs, past work has proposed methods to robustly interpret and evaluate them for various tasks

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3169–3177

and domains. Some work has sought to probe contextual language representations through various means to better understand what knowledge they hold and their correspondence to syntactic and semantic patterns (Tenney et al., 2018; Hewitt and Manning, 2019; Jawahar et al., 2019; Tenney et al., 2019). Meanwhile, behavior testing approaches have also been applied to understand model capabilities, from automatically removing words in language inputs and examining model performance as the input becomes malformed or insufficient for prediction (Li et al., 2016; Murdoch et al., 2018; Hewitt and Manning, 2019), to curating fine-grained testing data to measure performance on interesting phenomena (Zhou et al., 2019; Ribeiro et al., 2020). Similar work has used specialized natural language inference tasks (Welleck et al., 2019; Uppal et al., 2020), logic rules (Li et al., 2019; Asai and Hajishirzi, 2020), and annotated explanations (DeYoung et al., 2020; Jhamtani and Clark, 2020) to support and evaluate consistency and coherence of inference in these models. Other works have studied coherence of discourse through the proxy task of sentence re-ordering (Lapata, 2003; Logeswaran et al., 2018). Different from these previous works that focus only on specific tasks or methods, or require heavy annotation, this paper introduces an easily-accessed, versatile evaluation of machine coherence from a small amount of additional annotation.

3 Coherent Text Classification

For any text classification task requiring reasoning over a discourse, a coherent classifier should use the same evidence as humans do in reaching a conclusion. For any positive example, we expect that there are specific regions of the text which contain the semantic class of interest and thus directly contribute to the positive label. Conversely, for any negative example, there should be no such regions of the text. At a high level, we will propose a coherence measure that captures whether classifiers can give consistent and human-aligned predictions on these regions to support the end task conclusion.

Depending on specific tasks, this measure can have different implementations while maintaining the same high-level goal. In the following sections, we will use two example benchmark datasets, Conversational Entailment (**CE**) from Zhang and Chai (2010) and Abductive Reasoning in narrative Text (**ART**) from Bhagavatula et al. (2020), to illustrate how the coherence measure can be applied. We intentionally chose these two distinctive benchmark datasets for our investigation. CE is formulated as a textual entailment task, while ART is a multiple-choice text plausibility classification task. CE is small-scale, created over ten years ago before the era of deep learning, while ART is a large-scale (∼171k examples) dataset created more recently. Through these two different datasets, we aim to demonstrate the versatility of this framework.

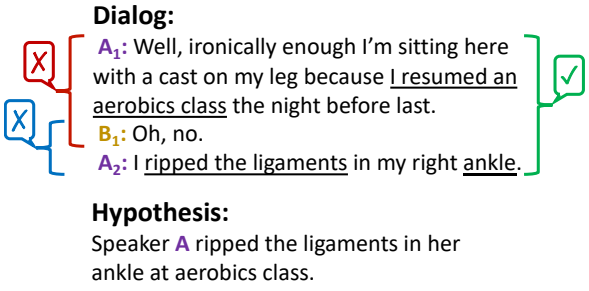

Figure 2: In CE, we label each sub-span of dialog with whether it entails the hypothesis (✓ for yes, ✗ for no).

3.1 Coherence in Textual Entailment

CE poses a textual entailment task where context is given as several turns of a natural language dialog, and we must determine whether the dialog entails a hypothesis sentence. All required information is explicitly given in the dialog. In each positive example, only some dialog turns directly contribute to the entailment, while others are irrelevant to the hypothesis. For example, as shown in Figure 1, turns A_1 and B_2 together entail the hypothesis, while others are not necessary for entailment.

As shown in Figure 2 for CE, we can label individual spans of a discourse that entails a hypothesis with whether or not consecutive sub-spans of the discourse also entail the hypothesis. Here, while the entire dialog from A_1 through A_2 entails the hypothesis, the spans from A_1 through B_1 and B_1 through A_2 do not, as they omit details required by the hypothesis. Given an example of length N,[1] we can decompose it into $N + \binom{N}{2}$ possible consecutive sub-spans[2] to label with human judgements.

For a correctly classified example, we can then perform inference on all sub-spans. If the system

[1]Length can be defined in units of dialog turns, sentences, paragraphs, or other appropriate units of the text. Text should be decomposed such that individual sub-spans are not malformed or fragmented, so token- and character-level sub-spans will typically be inappropriate for this evaluation.

[2]There are $\binom{N}{2}$ combinations of starting and ending points for multi-sentence sub-spans, plus N individual sentences.

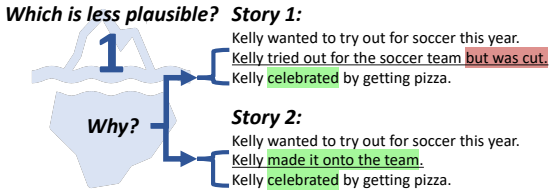

Figure 3: In Abductive Reasoning in narrative Texts (Bhagavatula et al., 2020), systems only compare two texts by their commonsense plausibility, ignoring which parts of the stories support this conclusion.

additionally classifies all of them correctly, we consider the prediction to be coherent. We then calculate **coherence** on the task as the percentage of examples coherently classified. Extremely simple to compute, this provides valuable insight beyond the surface of end task accuracy, measuring how well the classifier's perceived evidence toward the conclusion aligns with that of humans. Alternatively, the average sub-span accuracy may be considered as a more lenient measure.

3.2 Coherence in Plausibility Classification

ART, meanwhile, is a multiple-choice text classification benchmark for commonsense plausibility recognition. The task is to determine which of two candidate sentences most plausibly fits between two given context sentences when considering commonsense constraints on the world. This translates naturally into a choice between two three-sentence stories (differing only by the second sentence), one of which has some implausibility (the positive choice). For example, as shown in Figure 3, Story 1 is implausible because while the second sentence describes a negative event, the third sentence indicates celebration. Meanwhile, in Story 2, the agent is celebrating a positive event.

Multiple-choice tasks. To account for multiple-choice tasks like ART, where we identify one of two texts to be semantically implausible, we must adjust this setup. We still consider sub-spans of the context, breaking down each pair of texts into $N+\binom{N}{2}$ pairs of sub-spans. Intuitively, the model's choice on each pair should again align with that of humans. However, there is a possibility that none of the texts contain the positive class. In such cases, the classifier should not make a confident prediction, and instead believe the texts are equally likely. Confidence should be defined based on the classifier's internal model of the probability distribution over all possible class labels, i.e., text choices (typically calculated by applying softmax over the activations of several neural network branches). This is conceptually visualized in Figure 4, where a classifier should only become confident that Story B is implausible once both the second and third sentence are present, as *the trash* is less likely to end up on *the floor* with a *hole in the top* of the bag.

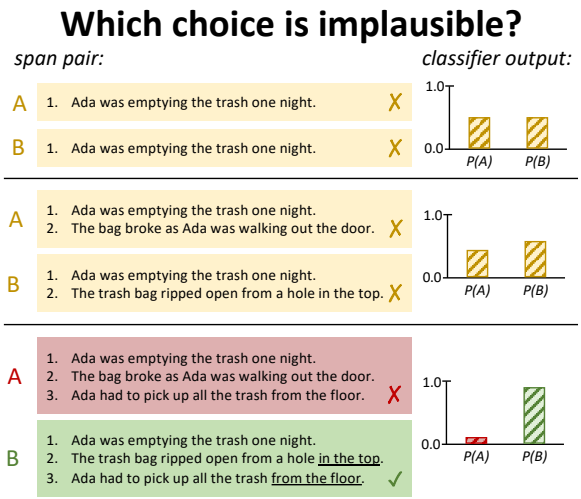

Figure 4: In ART, a multiple-choice text classification problem, we can label sub-spans with the least plausible choice, although in some cases, both choices are plausible. To address this, we consider the classifier's posterior probability for each choice; it is ideal if the classifier has low confidence in such instances.

Generally, let $T_{a:b}$ represent the consecutive subsequence of text T from unit a through b, e.g., sentences a through b of text T. Consider a set $S_{1:N}$ of M texts of length N such that $S = \{T^1_{1:N}, T^2_{1:N}, \cdots, T^M_{1:N}\}$, and a classifier f such that $f(S_{1:M}) \in [1, M]$.[3] When classifying a set $S_{a:b}$, let $f(S_{a:b}) = c^*$ be considered a *confident* prediction if $\max_{c \in [1,c^*) \vee (c^*,M]} (p(c^*) - p(c)) \geq \rho$, where $p(c)$ refers to probability of class c under the classifier's output distribution, and ρ is a confidence threshold. Where there is no positive text within $S_{a:b}$, then the desired outcome (ground truth) is for $f(S_{a:b})$ to be a non-confident prediction. This should be reflected in the calculation of coherence.

4 Coherence of SOTA Classifiers

Using our framework, we next establish baseline measures of coherence on the two benchmarks. The source code and data for our empirical study are shared with the community on GitHub.[4]

[3] While text choices may be different lengths, this can be trivially resolved by padding.

[4] https://github.com/sled-group/Verifiable-Coherent-NLU

4.1 Enabling Coherence Evaluation

To enable the type of evaluation described in Section 3 for our benchmarks, additional annotation is required. CE contains 50 unique dialog sources from the Switchboard corpus (Godfrey and Holliman, 1997). We randomly selected 10 testing sources to form the test set and left all remaining sources for training and validation, creating an 80%/20% split for training and validation (703 examples) versus testing (178 examples). We annotated the positive examples in the test set with the range of dialog turns entailing the hypothesis, allowing us to generate ground truth labels for the coherence measurement. Examples were labeled by two separate annotators and cross-verified with a near-perfect Cohen's κ (Cohen, 1960) of 0.91, then a third annotator resolved any disagreements.

To transfer ART to our framework, we annotated 200 random examples from the public validation set (1532 examples) with the evidence for implausibility. There are 3 possible cases in implausible story choices: 1) the second sentence conflicts with the first and/or third sentence, 2) the second sentence is malformed or nonsense, presumably due to annotation error or adversarial filtering (Zellers et al., 2018), and 3) the first and third sentence conflict with each other by default, and the second sentence does not resolve this. These cases are labeled by two annotators then merged with a fair Cohen's κ of 0.30 (perhaps lower due to subjectivity of commonsense-based problems), and a third annotator again resolving disagreements. 11 examples were discarded as two annotators agreed that both story choices were entirely plausible, presumably due to annotation error in ART.

4.2 Empirical Results

We evaluate three state-of-the-art, transformer-based language models from recent years: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021).[5] On CE, we additionally apply transfer learning from MultiNLI (Williams et al., 2017), a large-scale textual entailment dataset with some dialog-based problems. We measure both the *accuracy*, i.e., the proportion of instances where the end task prediction is correct, and *coherence* of models on respective evaluation sets. Specifically, we consider two kinds of coherence: strict and lenient. Given a set of evaluation instances, *strict coherence* refers to the proportion of instances where the end task prediction is not only correct, but also coherent as described in Section 3. While strict coherence only rewards systems for examples where all sub-span predictions are correct, *lenient coherence* averages the sub-span accuracy over all examples for a less rigid reward. We include this alternate form of coherence to accommodate some disagreement with our annotations (which can be subjective based on measured inter-annotator agreement) without severe penalty.

Training details. Following common practice, systems are trained with cross-entropy loss toward the end task of text classification, maximizing accuracy on the validation set for model selection. On CE, we used 8-fold cross-validation split by dialog sources, then re-trained the model with the highest average validation accuracy on all folds.

Pre-trained model parameters and implementations come from HuggingFace `transformers` (Wolf et al., 2020),[6] each trained with the AdamW optimizer (Loshchilov and Hutter, 2018). We performed a grid search over a wide range of learning rates and a maximum of 10 epochs. Training batch sizes are fixed based on available GPU memory. Selected hyperparameters can be found in Appendix A.

Discussion of results. Results on the test set of CE and public validation set of ART are listed in Table 1. All results show a statistically significant drop in performance from classification accuracy to strict coherence under a McNemar test (McNemar, 1947) with $p <$ 1e-5, some dropping below majority-class accuracy. While lenient coherence is slightly higher for both tasks, we still see large drops from accuracy. This demonstrates that while our text classifiers can achieve high classification accuracy on CE and ART, they do not deeply understand the tasks. Much of their performance is supported by incoherent intermediate predictions. Although pre-training on MultiNLI improves the end task accuracy on CE, it still suffers from comparably significant drops to the coherence measures. On ART, while all models see significant performance drops, DeBERTa, the state-of-the-art system for the task, achieves the best accuracy and coherence measures, as well as the highest chosen ρ values, which generally indicates more confident

[5] We use the "large" configuration of all models, which have 24 hidden layers and 16 attention heads.

[6] https://huggingface.co/transformers/

CE, *test*:

Model	**Accuracy** (%)	**Strict Coherence** (Δ; %)	**Lenient Coherence** (Δ; %)
majority	57.8	–	–
BERT	55.8	28.5 (-27.3)	35.7 (-20.1)
RoBERTa	70.9	39.0 (-31.9)	47.5 (-23.4)
↪ + MNLI	78.5	50.6 (-27.9)	58.2 (-20.3)
DeBERTa	67.4	37.2 (-30.2)	45.2 (-22.2)

ART, *validation*:

Model	**Accuracy** (%)	**Strict Coherence** (Δ; %)	ρ	**Lenient Coherence** (Δ; %)	ρ
majority	55.0 (50.1)	–	–	–	
BERT	66.7 (66.7)	42.3 (-24.4)	0.15	43.7 (-23.0)	0.85
RoBERTa	87.8 (84.2)	55.0 (-32.8)	0.1	59.3 (-28.5)	0.05
DeBERTa	88.4 (85.7)	59.8 (-28.6)	0.85	61.8 (-26.6)	0.95

Table 1: Accuracy, strict coherence, and lenient coherence on CE and ART for state-of-the-art text classifiers. Δ is the total performance drop from the classification accuracy to each coherence measure, and each ρ is the confidence threshold achieving the highest coherence. For ART, accuracy on the full validation set is given in parentheses.

predictions. Even though it only marginally outperforms RoBERTa in accuracy, we see larger improvements in coherence measures and the chosen ρ, suggesting DeBERTa is more robust.

5 Conclusion

In this work, we proposed a simple and versatile method to evaluate the coherence of text classifiers, particularly targeting the problem where end task prediction depends on a discourse rather than a single sentence. *By annotating a small amount of data in a benchmark, this method supports a quick assessment on whether machines' end task performance is supported by coherent intermediate evidence.* Future work driven by benchmarks should consider similar examination beyond the end task accuracy, whether this be through our proposed coherence measures or other appropriate means. As we showed, such effort is quite straightforward, and can drive progress toward more powerful classifiers that can support human-aligned reasoning.

Acknowledgements

This work was supported in part by IIS-1949634 from the National Science Foundation. We thank Bri Epstein and Haoyi Qiu for their diligent annotation work. We also thank the anonymous reviewers for their helpful comments and suggestions.

References

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Inteligence (AAAI-20)*, New York, NY, USA. AAAI Press.

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 (EMNLP 2015)*, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv: 2005.14165*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944, pages 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. 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 (NAACL HLT 2019)*, Minneapolis, MN, USA.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online. Association for Computational Linguistics.

John J. Godfrey and Edward Holliman. 1997. Switchboard-1 release 2. Linguistic Data Consortium.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654*.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, Minneapolis, MN, USA. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online. Association for Computational Linguistics.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv:1612.08220*.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

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: 1907.11692*.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, LA, USA.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online. 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 Framework 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 (NAACL HLT 2016)*, San Diego, CA, USA. Association for Computational Linguistics.

W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada.

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 (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, LA, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding with Unsupervised Learning. *OpenAI*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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 (ACL 2020)*, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford, CA, USA.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. 2017. The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In *Proceedings of the 21st Conference on Computational Natural Language (CoNLL 2017)*, Vancouver, BC, Canada. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada.

Shagun Uppal, Vivek Gupta, Avinash Swaminathan, Haimin Zhang, Debanjan Mahata, Rakesh Gosangi, Rajiv Ratn Shah, and Amanda Stent. 2020. Two-step classification using recasted data for low resource settings. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, BC, Canada. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. 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 (NAACL HLT 2018)*, New Orleans, LA, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): System Demonstrations*, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Brussels, Belgium. Association for Computational Linguistics.

Chen Zhang and Joyce Y. Chai. 2010. Towards Conversation Entailment: An Empirical Investigation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, Cambridge, MA, USA. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a Vacation" takes longer than "Going for a Walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, Hong Kong, China. Association for Computational Linguistics.

Task	Model	Batch Size	Learning Rate	Ep.
CE	BERT	32	7.5e-6	8
CE	RoBERTa	32	7.5e-6	10
CE	RoBERTa+MNLI	32	7.5e-6	8
CE	DeBERTa	16	1e-5	10
ART	BERT	64	5e-6	9
ART	RoBERTa	64	2.5e-6	5
ART	DeBERTa	32	1e-6	9

Table 2: Training hyperparameters (batch size, learning rate, epochs) for probed models.

A Model Training Details

The selected hyperparameters for each model presented in the paper are listed in Table 2.