Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation

Adam Poliak[1] Aparajita Haldar[1,2] Rachel Rudinger[1] J. Edward Hu[1]
Ellie Pavlick[3] Aaron Steven White[4] Benjamin Van Durme[1]
[1]Johns Hopkins University, [2] BITS Pilani, Goa Campus, India
[3]Brown University, [4]University of Rochester

Abstract

We present a large scale collection of diverse natural language inference (NLI) datasets that help provide insight into how well a sentence representation encoded by a neural network captures distinct types of reasoning. The collection results from recasting 13 existing datasets from 7 semantic phenomena into a common NLI structure, resulting in over half a million labeled context-hypothesis pairs in total. Our collection of diverse datasets is available at `http://www.decomp.net/`, and will grow over time as additional resources are recast and added from novel sources.

1 Introduction

A plethora of new natural language inference (NLI)[1] datasets has been created in recent years (Bowman et al., 2015; Williams et al., 2017; Lai et al., 2017; Khot et al., 2018). However, these datasets do not provide clear insight into what type of reasoning or inference a model may be performing. For example, these datasets cannot be used to evaluate whether competitive NLI models can determine if an event occurred, correctly differentiate between figurative and literal language, or accurately identify and categorize named entities. Consequently, these datasets cannot answer how well sentence representation learning models capture distinct semantic phenomena necessary for general natural language understanding (NLU).

To answer these questions, we introduce the **D**iverse **N**LI **C**ollection (DNC), a large-scale NLI dataset that tests a model's ability to perform diverse types of reasoning. DNC is a collection of NLI problems, each requiring a model to perform a unique type of reasoning. Each NLI dataset contains labeled context-hypothesis pairs that we recast from semantic annotations for specific structured prediction tasks. Table 1 includes a sample of NLI pairs that test specific types of reasoning.

Event Factuality	▶ Find him before he finds the dog food The finding did not happen	✓
	▶ I'll need to ponder The pondering happened	✗
Relation Extraction	▶ Ward joined Tom in their native Perth Ward was born in Perth	✓
	▶ Stefan had visited his son in Bulgaria Stefan was born in Bulgaria	✗
Puns	▶ Kim heard masks have no face value Kim heard a pun	✓
	▶ Tod heard that thrift is better than annuity Tod heard a pun	✗

Table 1: Example sentence pairs for different semantic phenomena. ▶ indicates the line is a context and the following line is its corresponding hypothesis. ✓ and ✗ respectively indicate that the context entails, or does not entail the hypothesis.

We extend various prior works on challenge NLI datasets (Zhang et al., 2017), and define recasting as leveraging existing datasets to create NLI examples (Glickman, 2006; White et al., 2017). We recast annotations from a total of 13 datasets across 7 NLP tasks into labeled NLI examples. The tasks include event factuality, named entity recognition, gendered anaphora resolution, sentiment analysis, relationship extraction, pun detection, and lexicosyntactic inference (Table 2). Currently, DNC contains over half a million labeled examples that can be used to probe a model's ability to capture different types of semantic reasoning necessary for general NLU. In short, this work answers a recent plea to the community to test "more kinds of inference" than in previous challenge sets (Chatzikyriakidis et al., 2017).

[1]The task of determining if a hypothesis would likely be inferred from a context, or premise; also known as Recognizing Textual Entailment (RTE) (Dagan et al., 2006, 2013).

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 337–340
Brussels, Belgium, November 1, 2018.

2 Motivation & Background

Compared to eliciting NLI datasets directly, i.e. asking humans to author contexts and/or hypothesis sentences, recasting can 1) help determine whether an NLU model performs distinct types of reasoning; 2) limit types of biases observed in previous NLI data; and 3) generate examples cheaply, potentially at large scales.

NLU Insights Popular NLI datasets, e.g. Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and its successor Multi-NLI (Williams et al., 2017), were created by eliciting hypotheses from humans. Crowd-source workers were tasked with writing one sentence each that is entailed, neutral, and contradicted by a caption extracted from the Flickr30k corpus (Young et al., 2014). Although these datasets are widely used to train and evaluate sentence representations, a high accuracy is not indicative of what types of reasoning NLI models perform. Workers were free to create any type of hypothesis for each context and label. Such datasets cannot be used to determine how well an NLI model captures many desired capabilities of language understanding systems, e.g. paraphrastic inference, complex anaphora resolution (White et al., 2017), or compositionality (Pavlick and Callison-Burch, 2016; Dasgupta et al., 2018). By converting prior annotation of a specific phenomenon into NLI examples, recasting allows us to create a diverse NLI benchmark that tests a model's ability to perform distinct types of reasoning.

Limit Biases Studies indicate that many NLI datasets contain significant biases. Examples in the early Pascal RTE datasets could be correctly predicted based on syntax alone (Vanderwende and Dolan, 2006; Vanderwende et al., 2006). Statistical irregularities, and annotation artifacts, within class labels allow a hypothesis-only model to significantly outperform the majority baseline on at least six recent NLI datasets (Poliak et al., 2018). Class label biases may be attributed to the human-elicited protocol. Moreover, examples in such NLI datasets may contain racial and gendered stereotypes (Rudinger et al., 2017).

We limit some biases by not relying on humans to generate hypotheses. Recast NLI datasets may still contain some biases, e.g. non-uniform distributions over NLI labels caused by the distribution of labels in the original dataset that we recast.[2] Experimental results using hypothesis-only models (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018) can indicate to what degree the recast datasets retain some biases that may be present in the original semantic datasets.

Phenomena	Dataset
Event Factuality	Decomp (Rudinger et al., 2018b) UW (Lee et al., 2015) MeanTime (Minard et al., 2016)
Named Entity Recognition	Groningen (Bos et al., 2017) CoNLL (Tjong Kim Sang and De Meulder, 2003)
Gendered Anaphora	Winogender (Rudinger et al., 2018a)
Lexicosyntactic Inference	VerbCorner (Hartshorne et al., 2013) MegaVeridicality (White and Rawlins, 2018) VerbNet (Schuler, 2005)
Puns	(Yang et al., 2015) SemEval 2017 Task 7 (Miller et al., 2017)
Relationship Extraction	FACC1 (Gabrilovich et al., 2013)
Sentiment Analysis	(Kotzias et al., 2015)

Table 2: List of each type of semantic phenomena paired with its corresponding dataset(s) we recast.

NLI Examples at Large-scale Generating NLI datasets from scratch is costly. Humans must be paid to generate or label natural language text. This linearly scales costs as the amount of generated NLI-pairs increases. Existing annotations for a wide array of semantic NLP tasks are freely available. By leveraging existing semantic annotations already invested in by the community we can generate and label NLI pairs at little cost and create large NLI datasets to train data hungry models.

Why These Semantic Phenomena? A long term goal is to develop NLU systems that can achieve human levels of understanding and reasoning. Investigating how different architectures and training corpora can help a system perform human-level general NLU is an important step in this direction. DNC contains recast NLI pairs that are easily understandable by humans and can be used to evaluate different sentence encoders and NLU systems. These semantic phenomena cover distinct types of reasoning that an NLU system may often encounter in the wild. While higher performance on these benchmarks might not be conclusive proof of a system achieving human-level reasoning, a system that does poorly should not be viewed as performing human-level NLU. We argue that these semantic phenomena play integral roles in NLU. There exist more semantic phenomena which are integral to NLU (Allen, 1995) and we plan to include them in future versions of DNC.

[2] In a corpus with part-of-speech tags, the distribution of labels for the word "the" will likely peak at the `Det` tag.

References

James Allen. 1995. *Natural language understanding*. Pearson.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In *Handbook of Linguistic Annotation*, pages 463–496. Springer.

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)*. Association for Computational Linguistics.

Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. 2017. An overview of natural language inference data collection: The way forward? In *Proceedings of the Computing Natural Language Inference Workshop*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

I. Dasgupta, D. Guo, A. Stuhlmüller, S. J. Gershman, and N. D. Goodman. 2018. Evaluating Compositionality in Sentence Embeddings. *ArXiv e-prints*.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). *Note: http://lemurproject.org/clueweb09/FACC1/Cited by*, 5.

Oren Glickman. 2006. *Applied textual entailment*. Ph.D. thesis, Bar Ilan University.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL*.

Joshua K Hartshorne, Claire Bonial, and Martha Palmer. 2013. The verbcorner project: Toward an empirically-based semantic decomposition of verbs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1438–1442.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606. ACM.

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.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. Semeval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In *Language Resources and Evaluation Conference (LREC)*.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multitask video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1273–1283.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173. 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 *The Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018a. Gender bias in coreference resolution. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018b. Neural models of factuality. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *11th International Conference on Language Resources and Evaluation (LREC2018)*.

Lucy Vanderwende and William B Dolan. 2006. What syntax can contribute in the entailment task. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 205–216. Springer.

Lucy Vanderwende, Arul Menezes, and Rion Snow. 2006. Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In *Second PASCAL Challenges Workshop*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, page to appear, Amherst, MA. GLSA Publications.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association of Computational Linguistics*, 5(1):379–395.