Breaking NLI Systems with Sentences that Require Simple Lexical Inferences

Max Glockner[1], **Vered Shwartz**[2] **and Yoav Goldberg**[2]

[1]Computer Science Department, TU Darmstadt, Germany
[2]Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel
{maxg216,vered1986,yoav.goldberg}@gmail.com

Abstract

We create a new NLI test set that shows the deficiency of state-of-the-art models in inferences that require lexical and world knowledge. The new examples are simpler than the SNLI test set, containing sentences that differ by at most one word from sentences in the training set. Yet, the performance on the new test set is substantially worse across systems trained on SNLI, demonstrating that these systems are limited in their generalization ability, failing to capture many simple inferences.

1 Introduction

Recognizing textual entailment (RTE) (Dagan et al., 2013), recently framed as natural language inference (NLI) (Bowman et al., 2015) is a task concerned with identifying whether a *premise* sentence entails, contradicts or is neutral with the *hypothesis* sentence. Following the release of the large-scale SNLI dataset (Bowman et al., 2015), many end-to-end neural models have been developed for the task, achieving high accuracy on the test set. As opposed to previous-generation methods, which relied heavily on lexical resources, neural models only make use of pre-trained word embeddings. The few efforts to incorporate external lexical knowledge resulted in negligible performance gain (Chen et al., 2018). This raises the question whether (1) neural methods are inherently stronger, obviating the need of external lexical knowledge; (2) large-scale training data allows for implicit learning of previously explicit lexical knowledge; or (3) the NLI datasets are simpler than early RTE datasets, requiring less knowledge.

[1]The contradiction example follows the assumption in Bowman et al. (2015) that the premise contains the most prominent information in the event, hence the premise can't describe the event of a man holding both instruments.

Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

Table 1: Examples from the new test set.

In this paper we show that state-of-the-art NLI systems are limited in their generalization ability, and fail to capture many simple inferences that require lexical and world knowledge. Inspired by the work of Jia and Liang (2017) on reading comprehension, we create a new NLI test set with examples that capture various kinds of lexical knowledge (Table 1). For example, that *champagne* is a type of *wine* (hypernymy), and that *saxophone* and *electric guitar* are different musical instruments (co-hyponyms). To isolate lexical knowledge aspects, our constructed examples contain only words that appear both in the training set and in pre-trained embeddings, and differ by a single word from sentences in the training set.

The performance on the new test set is substantially worse across systems, demonstrating that the SNLI test set alone is not a sufficient measure of language understanding capabilities. Our results are in line with Gururangan et al. (2018) and Poliak et al. (2018), who showed that the label can be identified by looking only at the hypothesis and exploiting annotation artifacts such as word choice and sentence length.

Further investigation shows that what mostly affects the systems' ability to correctly predict a test example is the amount of similar examples found in the training set. Given that training data will always be limited, this is a rather inefficient way to learn lexical inferences, stressing the need to develop methods that do this more

effectively. Our test set can be used to evaluate such models' ability to recognize lexical inferences, and it is available at https://github.com/BIU-NLP/Breaking_NLI.

2 Background

NLI Datasets. The SNLI dataset (Stanford Natural Language Inference, Bowman et al., 2015) consists of 570k sentence-pairs manually labeled as entailment, contradiction, and neutral. Premises are image captions from Young et al. (2014), while hypotheses were generated by crowd-sourced workers who were shown a premise and asked to generate entailing, contradicting, and neutral sentences. Workers were instructed to judge the relation between sentences *given that they describe the same event*. Hence, sentences that differ by a single mutually-exclusive term should be considered contradicting, as in "The president visited Alabama" and "The president visited Mississippi". This differs from traditional RTE datasets, which do not assume event coreference, and in which such sentence-pairs would be considered neutral.

Following criticism on the simplicity of the dataset, stemming mostly from its narrow domain, two additional datasets have been collected. The MultiNLI dataset (Multi-Genre Natural Language Inference, Williams et al., 2018) was collected similarly to SNLI, though covering a wider range of genres, and supporting a cross-genre evaluation. The SciTail dataset (Khot et al., 2018), created from science exams, is somewhat different from the two datasets, being smaller (27,026 examples), and labeled only as entailment or neutral. The domain makes this dataset different in nature from the other two datasets, and it consists of more factual sentences rather than scene descriptions.

Neural Approaches for NLI. Following the release of SNLI, there has been tremendous interest in the task, and many end-to-end neural models were developed, achieving promising results.[2] Methods are divided into two main approaches. Sentence-encoding models (e.g. Bowman et al., 2015, 2016; Nie and Bansal, 2017; Shen et al., 2018) encode the premise and hypothesis individually, while attention-based models align words in the premise with similar words in the hypothesis, encoding the two sentences together (e.g. Rocktäschel et al., 2016; Chen et al., 2017).

[2] See the SNLI leaderboard for a comprehensive list: https://nlp.stanford.edu/projects/snli/.

External Lexical Knowledge. Traditional RTE methods typically relied on resources such as WordNet (Fellbaum, 1998) to identify lexical inferences. Conversely, neural methods rely solely on pre-trained word embeddings, yet, they achieve high accuracy on SNLI.

The only neural model to date that incorporates external lexical knowledge (from WordNet) is KIM (Chen et al., 2018), however, gaining only a small addition of 0.6 points in accuracy on the SNLI test set. This raises the question whether the small performance gap is a result of the model not capturing lexical knowledge well, or the SNLI test set not requiring this knowledge in the first place.

3 Data Collection

We construct a test set with the goal of evaluating the ability of state-of-the-art NLI models to make inferences that require simple lexical knowledge. We automatically generate sentence pairs (§3.1) which are then manually verified (§3.2).

3.1 Generating Adversarial Examples

In order to isolate the lexical knowledge aspects, the premises are taken from the SNLI training set. For each premise we generate several hypotheses by replacing a single word within the premise by a different word. We also allow some multi-word noun phrases ("electric guitar") and adapt determiners and prepositions when needed.

We focus on generating only *entailment* and *contradiction* examples, while *neutral* examples may be generated as a by-product. *Entailment* examples are generated by replacing a word with its synonym or hypernym, while *contradiction* examples are created by replacing words with mutually exclusive co-hyponyms and antonyms (see Table 1). The generation steps are detailed below.

Replacement Words. We collected the replacement words using online resources for English learning.[3] The newly introduced words are all present in the SNLI training set: from occurrence in a single training example ("Portugal") up to 248,051 examples ("man"), with a mean of 3,663.1 and a median of 149.5. The words are also available in the pre-trained embeddings vocabulary. The goal of this constraint is to isolate lexical knowledge aspects, and evaluate the models' ability to generalize and make new inferences for known words.

[3] www.enchantedlearning.com, www.smart-words.org

	SNLI Test	New Test
Instances:		
contradiction	3,236	7,164
entailment	3,364	982
neutral	3,215	47
Overall	9,815	8,193
Fleiss κ:		
contradiction	0.77	0.61
entailment	0.69	0.90
Overall	0.67	0.61
Estimated human performance:		
	87.7%	94.1%

Table 2: Statistics of the test sets. 9,815 is the number of samples with majority agreement in the SNLI test set, whose full size is 9,824.

Replacement words are divided into topical categories detailed in Table 4. In several categories we applied additional processing to ensure that examples are indeed mutually-exclusive, topically-similar, and interchangeable in context. We included WordNet antonyms with the same part-of-speech and with a cosine similarity score above a threshold, using GloVe (Pennington et al., 2014). In *nationalities* and *countries* we focused on countries which are related geographically *(Japan, China)* or culturally *(Argentina, Spain)*.

Sentence-Pairs. To avoid introducing new information not present in the training data, we sampled premises from the SNLI training set that contain words from our lists, and generated hypotheses by replacing the selected word with its replacement. Some of the generated sentences may be ungrammatical or nonsensical, for instance, when replacing *Jordan* with *Syria* in sentences discussing *Michael Jordan*. We used Wikipedia bigrams[4] to discard sentences in which the replaced word created a bigram with less than 10 occurrences.

3.2 Manual Verification

We manually verify the correctness of the automatically constructed examples using crowdsourced workers in Amazon Mechanical Turk. To ensure the quality of workers, we applied a qualification test and required a 99% approval rate for at least 1,000 prior tasks. We assigned each annotation to 3 workers.

Following the SNLI guidelines, we instructed the workers to consider the sentences as describing the same event, but we simplified the annotation process into answering 3 simple yes/no questions:

1. Do the sentences describe the same event?
2. Does the new sentence (hypothesis) add new information to the original sentence (premise)?
3. Is the new sentence incorrect/ungrammatical?

We then discarded any sentence-pair in which at least one worker answered the third question positively. If the answer to the first question was negative, we considered the label as *contradiction*. Otherwise, we considered the label as *entailment* if the answer to the second question was negative and *neutral* if it was positive. We used the majority vote to determine the gold label.

The annotations yielded substantial agreement, with Fleiss' Kappa $\kappa = 0.61$ (Landis and Koch, 1977). We estimate human performance to 94.1%, using the method described in Gong et al. (2018), showing that the new test set is substantially easier to humans than SNLI. Table 2 provides additional statistics on the test set.[5]

4 Evaluation

4.1 Models

Without External Knowledge. We chose 3 representative models in different approaches (sentence encoding and/or attention): RESIDUAL-STACKED-ENCODER (Nie and Bansal, 2017) is a biLSTM-based single sentence-encoding model without attention. As opposed to traditional multi-layer biLSTMs, the input to each next layer is the concatenation of the word embedding and the summation of outputs from previous layers. ESIM (Enhanced Sequential Inference Model, Chen et al., 2017) is a hybrid TreeLSTM-based and biLSTM-based model. We use the biLSTM model, which uses an inter-sentence attention mechanism to align words across sentences. Finally, DECOMPOSABLE ATTENTION (Parikh et al., 2016) performs soft alignment of words from the premise to words in the hypothesis using attention mechanism, and decomposes the task into comparison of aligned words. Lexical-level decisions are merged to produce the final classification. We use the AllenNLP re-implementation,[6] which does not implement the optional intra-sentence attention, and achieves an accuracy of 84.7% on the SNLI test set, comparable to 86.3% by the original system.

[4] github.com/rmaestre/Wikipedia-Bigram-Open-Datasets

[5] We note that due to its bias towards *contradiction*, the new test set can neither be used for training, nor serve as a main evaluation set for NLI. Instead, we suggest to use it in addition to the original test set in order to test a model's ability to handle lexical inferences.

[6] http://allennlp.org/models

Model	Train set	SNLI test set	New test set	Δ
Decomposable Attention (Parikh et al., 2016)	SNLI	84.7%	51.9%	-32.8
	MultiNLI + SNLI	84.9%	65.8%	-19.1
	SciTail + SNLI	85.0%	49.0%	-36.0
ESIM (Chen et al., 2017)	SNLI	87.9%	65.6%	-22.3
	MultiNLI + SNLI	86.3%	74.9%	-11.4
	SciTail + SNLI	88.3%	67.7%	-20.6
Residual-Stacked-Encoder (Nie and Bansal, 2017)	SNLI	86.0%	62.2%	-23.8
	MultiNLI + SNLI	84.6%	68.2%	-16.8
	SciTail + SNLI	85.0%	60.1%	-24.9
WordNet Baseline	-	-	85.8%	-
KIM (Chen et al., 2018)	SNLI	88.6%	83.5%	-5.1

Table 3: Accuracy of various models trained on SNLI or a union of SNLI with another dataset (MultiNLI, SciTail), and tested on the original SNLI test set and the new test set.

We chose models which are amongst the best performing within their approaches (excluding ensembles) and have available code. All models are based on pre-trained GloVe embeddings (Pennington et al., 2014), which are either fine-tuned during training (RESIDUAL-STACKED-ENCODER and ESIM) or stay fixed (DECOMPOSABLE ATTENTION). All models predict the label using a concatenation of features derived from the sentence representations (e.g. maximum, mean), for example as in Mou et al. (2016). We use the recommended hyper-parameters for each model, as they appear in the provided code.

With External Knowledge. We provide a simple WORDNET BASELINE, in which we classify a sentence-pair according to the WordNet relation that holds between the original word w_p and the replaced word w_h. We predict *entailment* if w_p is a hyponym of w_h or if they are synonyms, *neutral* if w_p is a hypernym of w_h, and *contradiction* if w_p and w_h are antonyms or if they share a common hypernym ancestor (up to 2 edges). Word pairs with no WordNet relations are classified as *other*.

We also report the performance of KIM (Knowledge-based Inference Model, Chen et al., 2018), an extension of ESIM with external knowledge from WordNet, which was kindly provided to us by Qian Chen. KIM improves the attention mechanism by taking into account the existence of WordNet relations between the words. The lexical inference component, operating over pairs of aligned words, is enriched with a vector encoding the specific WordNet relations between the words.

4.2 Experimental Settings

We trained each model on 3 different datasets: (1) SNLI train set, (2) a union of the SNLI train set and the MultiNLI train set, and (3) a union of the SNLI train set and the SciTail train set. The motivation is that while SNLI might lack the training data needed to learn the required lexical knowledge, it may be available in the other datasets, which are presumably richer.

4.3 Results

Table 3 displays the results for all the models on the original SNLI test set and the new test set. Despite the task being considerably simpler, the drop in performance is substantial, ranging from 11 to 33 points in accuracy. Adding MultiNLI to the training data somewhat mitigates this drop in accuracy, thanks to almost doubling the amount of training data. We note that adding SciTail to the training data did not similarly improve the performance; we conjecture that this stems from the differences between the datasets.

KIM substantially outperforms the other neural models, demonstrating that lexical knowledge is the only requirement for good performance on the new test set, and stressing the inability of the other models to learn it. Both WordNet-informed models leave room for improvement: possibly due to limited WordNet coverage and the implications of applying lexical inferences within context.

5 Analysis

We take a deeper look into the predictions of the models that don't employ external knowledge, focusing on the models trained on SNLI.

5.1 Accuracy by Category

Table 4 displays the accuracy of each model per replacement-word category. The neural models tend to perform well on categories which are frequent in the training set, such as *colors*, and badly

Dominant Label	Category	Instances	Example Words	Decomposable Attention	ESIM	Residual Encoders	WordNet Baseline	KIM
Cont.	antonyms	1,147	*loves - dislikes*	41.6%	70.4%	58.2%	95.5%	86.5%
	cardinals	759	*five - seven*	53.5%	75.5%	53.1%	98.6%	93.4%
	nationalities	755	*Greek - Italian*	37.5%	35.9%	70.9%	78.5%	73.5%
	drinks	731	*lemonade - beer*	52.9%	63.7%	52.0%	94.8%	96.6%
	antonyms (WN)	706	*sitting - standing*	55.1%	74.6%	67.9%	94.5%	78.8%
	colors	699	*red - blue*	85.0%	96.1%	87.0%	98.7%	98.3%
	ordinals	663	*fifth - 16th*	2.1%	21.0%	5.4%	40.7%	56.6%
	countries	613	*Mexico - Peru*	15.2%	25.4%	66.2%	100.0%	70.8%
	rooms	595	*kitchen - bathroom*	59.2%	69.4%	63.4%	89.9%	77.6%
	materials	397	*stone - glass*	65.2%	89.7%	79.9%	75.3%	98.7%
	vegetables	109	*tomato -potato*	43.1%	31.2%	37.6%	86.2%	79.8%
	instruments	65	*harmonica - harp*	96.9%	90.8%	96.9%	67.7%	96.9%
	planets	60	*Mars - Venus*	31.7%	3.3%	21.7%	100.0%	5.0%
Ent.	synonyms	894	*happy - joyful*	97.5%	99.7%	86.1%	70.5%	92.1%
	total	8,193		51.9%	65.6%	62.2%	85.8%	83.5%

Table 4: The number of instances and accuracy per category achieved by each model.

on categories such as *planets*, which rarely occur in SNLI. These models perform better than the WordNet baseline on entailment examples (*synonyms*), suggesting that they do so due to high lexical overlap between the premise and the hypothesis rather than recognizing synonymy. We therefore focus the rest of the discussion on contradiction examples.

5.2 Accuracy by Word Similarity

The accuracies for *ordinals*, *nationalities* and *countries* are especially low. We conjecture that this stems from the proximity of the contradicting words in the embedding space. Indeed, the Decomposable Attention model—which does not update its embeddings during training—seems to suffer the most.

Grouping its prediction accuracy by the cosine similarity between the contradicting words reveals a clear trend that the model errs more on contradicting pairs with similar pre-trained vectors:[7]

Similarity	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Accuracy	46.2%	42.3%	37.5%	29.7%	20.2%

5.3 Accuracy by Frequency in Training

Models that fine-tune the word embeddings may benefit from training examples consisting of test replacement pairs. Namely, for a given replacement pair (w_p, w_h), if many training examples labeled as contradiction contain w_p in the premise and w_h in the hypothesis, the model may update their embeddings to optimize predicting contradiction. Indeed, we show that the ESIM accuracy on test pairs increases with the frequency in which their replacement words appear in contradiction examples in the training data:

Frequency	0	1-4	5-9	10-49	50-99	100+
Accuracy	40.2%	70.6%	91.4%	92.1%	97.5%	98.5%

This demonstrates that the model is capable of learning lexical knowledge when sufficient training data is given, but relying on explicit training examples is a very inefficient way of obtaining simple lexical knowledge.

6 Conclusion

We created a new NLI test set with the goal of evaluating systems' ability to make inferences that require simple lexical knowledge. Although the test set is constructed to be much simpler than SNLI, and does not introduce new vocabulary, the state-of-the-art systems perform poorly on it, suggesting that they are limited in their generalization ability. The test set can be used in the future to assess the lexical inference abilities of NLI systems and to tease apart the performance of otherwise very similarly-performing systems.

Acknowledgments

We would like to thank Qian Chen for evaluating KIM on our test set. This work was supported in part by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1), an Intel ICRI-CI grant, Theo Hoffenberg, and the Israel Science Foundation grants 1951/17 and 1555/15. Vered is also supported by the Clore Scholars Programme (2017), and the AI2 Key Scientific Challenges Program (2017).

[7] We ignore multi-word replacements in §5.2 and §5.3.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and D. Christopher Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 632–642. https://doi.org/10.18653/v1/D15-1075.

Samuel R Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1466–1477.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1657–1668.

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.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *International Conference on Learning Representations (ICLR)*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. New Orleans, Louisiana.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2021–2031. https://www.aclweb.org/anthology/D17-1215.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*. New Orleans, Louisiana.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* pages 159–174.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 130–136.

Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312* .

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2249–2255. https://aclweb.org/anthology/D16-1244.

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)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Joint Conference on Lexical and Computational Semantics (StarSem)*.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296* .

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. New Orleans, Louisiana.

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.