Testing the Reasoning Power for NLI Models with Annotated Multi-perspective Entailment Dataset

Dong Yu[1], Lu Liu[1], Chen Yu[1], and Changliang Li[2]

[1] Beijing Language and Culture University, Beijing, China
[2] Kingsoft AI Lab
yudong@blcu.edu.cn, {luliu.nlp,yuchen7312}@gmail.com,
lichangliang@kingsoft.com

Abstract. Natural language inference (NLI) is a challenging task to determine the relationship between a pair of sentences. Existing Neural Network-based (NN-based) models have achieved prominent success. However, rare models are interpretable. In this paper, we propose a Multi-perspective Entailment Category Labeling System (METALs). It consists of three categories, ten sub-categories. We manually annotate 3,368 entailment items. The annotated data is used to explain the recognition ability of four NN-based models at a fine-grained level. The experimental results show that all the models have poor performance in the commonsense reasoning than in other entailment categories. The highest accuracy difference is 13.22%.

Keywords: Natural Language Inference · Multi-perspective Entailment Category Labeling System · Entailment Categories.

1 Introduction

NLI is an important subtopic in Natural Language Understanding. NLI is to identify the entailment relation between two sentences, commonly formalized as 3-way classification task. Previous works devote to the development of NLI datasets and models [9, 2, 8].

In 2015, the SNLI [4] corpus is released by Bowman. It provides more than 570K P-H sentence pairs. After that, other datasets have also emerged, such as MultiNLI [34], Dialogue NLI [33], and QA-NLI [11]. Those datasets have promoted the development of many NN-based NLI models, especially SNLI. Since the release of this dataset, the highest accuracy of it is constantly improved [22, 32, 8, 25, 26, 12], and the recognition ability of some models now have exceed human beings. However, those models are hard to interpret. It's often not clear why they work or how they exactly work. Consequently, some researches set out to explore the nature of NLI tasks in novel ways [23, 5, 16]. Some studies begin to do more fine-grained classification [28, 15, 2]. But the sentence pairs that contain multiple entailment categories are still unexplained. The categories of entailment are still not fine-grained enough.

In this paper, we propose METALs aiming at the entailment data (in 4.2). The METALs is set into three categories, ten sub-categories. We chose entailment because it is more regular than neutral, less studied than contradiction. We manually annotate 3368 entailment items in SNLI test set. The current version of our labeled corpus is freely available at https://github.com/blcunlp/Multi-perspective-Entailment-Dataset.

To interpret the recognition ability of NN-based models at a fine-grained and multi-perspective level, we conduct experiments based on four NLI models (in 5.1). The experimental results demonstrate that all the models are excellent in the entailment data due to the reduction of information or the change of syntactic structure. But they are inadequate in the entailment examples which contain inference (in 5.2).

2 Related Works

Since the NLI task is presented, many NLI datasets have released. An early attempt is Recognizing Textual Entailment (RTE) Challenge [10]. There are several larger benchmarks inspired by this work. The Sentences Involving Compositional Knowledge (SICK) [19] benchmark is proposed in 2014 and collect about 10k sentence pairs. In 2015, the SNLI [4] corpus is released by Bowman. It provides more than 570K P-H sentence pairs. After that, other datasets have appeared, such as MultiNLI [34], Dialogue NLI [33], and QA-NLI [11]. Those datasets promote the development of many NN-based NLI models, especially SNLI. And our work focus on SNLI models.

There are two main categories of the NN-based models on SNLI: one is encoding sentence embeddings and integrating the two sentence representations to predict [22, 18, 6, 8, 29, 20], and the other is cross sentence attention-based models which concern more about the interaction between each sentence pair [24, 32, 25, 14]. Recently, some pre-training models [12, 26, 35, 17] achieve good results in NLI tasks.

Although some models have got high accuracy, how good the model is, where it is, and why it is good are still inconclusive. Recently, some works devote to explain the performance of models. [28] proposes a detailed evaluation and sketches a process to generate its annotation. They label a subset of 210 examples and utilize a series of experiments to prove that these annotations are useful. [15] creates a new NLI test set that shows the deficiency of models in inferences which require lexical and world knowledge. [2] proposes a methodology for the creation of specialized data sets for NLI and experiments it over a sample of pairs taken from the RTE-5 data set [3].

Early studies have shown that the relationship between P and H is identified by analyzing and utilizing lexical, syntax, and world knowledge. [30] predicts the datasets completely based on the grammatical clues, and puts forward more than 50 grammatical tags. [7] extracts 100 entailment samples and classifies entail phenomena. The conclusion is that lexical and world knowledge is needed for RTE. [13] proposes an annotation method, ARTE, which extract 23 entailment relations. Subsequently, researches begin to focus on the specific phenomenon of NLI. [27] studies motion-space reasoning of entailment and constructs reasoning corpus. [1] shows how to improve classification accuracy by using the hypostatic relation.

Inspired by the previous work, we propose a Multi-perspective Entailment Category Labeling System (METALs) and experiment it over a sample of pairs from SNLI. We select several models with high classification accuracy and analyze them below.

3 Categories

In order to detect the recognition ability of the model more finely, we consider it from three aspects. It refers to widely accepted linguistic categories in the literature [13]: lexical, syntactic, and reasoning.

At the lexical level, We regard the ontological knowledge relation between words as the inference basis. The entailment is mainly based on language conversion, addition, deletion, and replacement at the level of syntactic structure. After lexical and syntactic level mining, there are still some sentence pairs that cannot be categorized. These pairs need additional world knowledge, so we sum up them as commonsense reasoning. In addition, some of the data is classified as discard due to they contain many errors, such as spelling errors, adding irrelevant ingredients, capital letter, and punctuation change, etc.

From the three levels, we divide the entailment types into three categories, ten sub-categories. The categories are not independent. They can be applied in combinations so that they can achieve a result as informative as possible. We briefly introduce the definitions and main ideas associated with each component of the architecture in this section.

3.1 Lexical and Phrasal Level (LP)

The lexical level is intended to capture basic lexical‘s inherent properties of the entailment phenomenon. It also called as ontological relations. Ontological involves three kinds of lexical ontological relations, which are drawn from WordNet [21]. We chose three relations that are typically associated with the notion of semantic similarity.

Hypernymy (Hyp) refers to two entail words linked by the is-a-kind-of relation. A hyponym is a word or phrase whose semantic field is more specific than its hypernym. “poodle-dog-animal” is a typical example of transitive relation in this category.

Synonymy (Syn) refers to words or phrase that have the same meaning or similar meaning in the sentence pair. These two words may have slight differences in using range, emotional color, or colloquial style. But their basic meanings are mostly same. “screaming” and “yelling” have a similar meaning. They are interchangeable within the context in which they appear, such as “A man is X at a camera”.

Meronymy (Mer) is often expressed as “part-of”. A meronym denotes a constituent part of, or a member of something, such as “tree” and “tree branch”.

3.2 Syntactic Structure(SS)

The syntactic structure in entailment is reflected in three aspects. The first one is changing the surface structure of the sentence and keeping the meaning in the same. The second is the extraction and reorganization of sentence components. The third is the ellipsis of syntactic components.

Syntactic Transformations (ST) is using different language forms to express synonymous meaning. There are four typical transformations: changes in active-passive voice, changes of word order, uses double negatives, changes between simple-compound sentences.

Extraction (Ext) refers to extract a certain component of a sentence and reassembling it into a new sentence. The new sentence is H, which contains part of semantic information of the original sentence. The extracted part can be a syntactic component or a complete clause. The existential sentence which often appears in the H sentence is one of the reorganizing sentence forms.

Ellipsis (Elli) refers to delete or omit some structures in a sentence. Language recursion makes the language structure layer upon layer without causing structural confusion. Therefore, the overall structure of the sentence remains unchanged and the semantic information is extracted, when some structures in a sentence are deleted or omitted. Reduction information is an universal entailment.

3.3 Commonsense Reasoning (CR)

Understand natural language requires not only linguistic knowledge but also human commonsense and social experience. Commonsense reasoning includes judgments about the physical properties, purpose, intentions, and behavior of people and objects as well as possible outcomes of their actions and interactions.

Spatial Reasoning (SR) refers to the inference obtained by judging the absolute location like “on the beach” and “in the sand” or relative location like “in front of” and “behind”. The judgment of the spatial relationship is complicated.

Quantities Reasoning (QR) includes not only the addition, subtraction, and numbers comparison but also the judgment of cardinal numbers with approximate numbers.

Emotion Reasoning (EmoR) is to infer emotions, emotional states, and psychological changes through the characters actions, facial expressions, and other external information. Emotion is one of the inherent affective phenomena which is intrinsic in human experience.

Event Reasoning (EveR) is the judgment of the attributes, purposes, intentions, and behaviors of people as well as the possible results of their interactions. We describe the event from the general event relationship, event framework and event environment, and background.

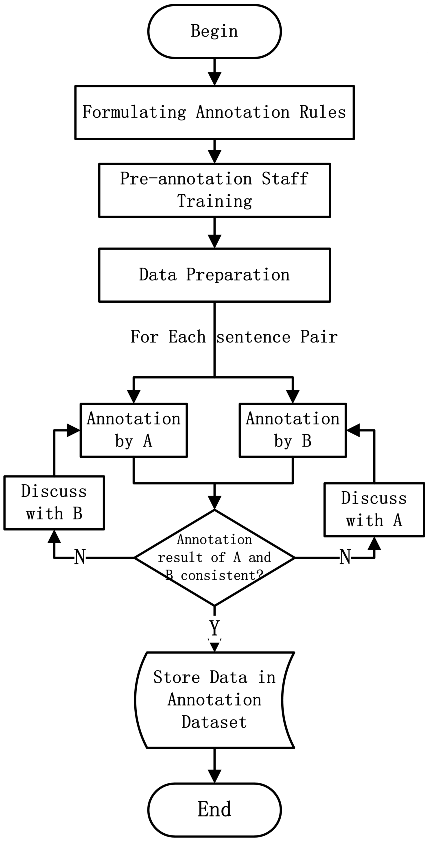
Begin
Formulating Annotation Rules
Pre-annotation Staff Training
Data Preparation
For Each sentence Pair
Annotation by A
Annotation by B
Discuss with B
Discuss with A
Annotation result of A and B consistent?
N
N
Y
Store Data in Annotation Dataset
End

Fig. 1. The annotation flow diagram.

Table 1. Percentages for all categories on pre-labeled one hundred data.

Category	Percentage
LP	2
SS	38
CR	17
AllSingle	57
Complex	**43**
Discard	1

Table 2. Percentages for all categories and sub-categories on annotated 3368 sentence pairs.

category	sub-category	num	percent
LP	Hyp	296	8.79
	Syn	608	18.05
	Mer	5	0.15
	LP-com	75	2.23
SS	Elli	1646	48.87
	Ext	361	10.72
	ST	78	2.32
	SS-com	144	4.28
CR	EmoR	22	0.65
	QR	115	3.41
	SR	441	13.09
	EveR	803	23.84
	CR-com	73	2.17
Diacard		86	2.55

4 Data Annotations

4.1 Data Selection

We select the open source dataset SNLI [4] as a data source. SNLI is a current mainstream NLI dataset. It consists of 570k pairs of sentences. Each sentence pair includes a P, an H, and a label describing the relationship between them (entailment, neutral, or contradiction). Its data are manually labeled through crowdsourcing.

The crowdsourcing workers obtain a description of the photo scene as a P. And they are asked to write different H according to the three requirements, corresponding to three inference labels.

SNLI test set consists of 10k sentence pairs. After removing the invalid data (The workers are inconsistent in the labeling of certain sentence pairs, and these sentences are ultimately unlabeled.) in the test set, we get 9824 samples. 3368 sentence pairs are labeled as entailment among them. Those are the data source for the annotation data in this paper.

4.2 Multi-perspective Entailment Category Labeling System

After analyzing 100 pre-labeled sentence pairs, we find that nearly half of P-H pairs have more than one cause of entailment relationship. The statistical results are shown in Table 1. "AllSingle" are sentence pairs with one and only one category. "Complex" are sentence pairs with more than two categories. Differing from the previous method of labeling only one category for each P-H pairs [23], we adopt a **METALs** to maintain the integrity of the entailment categories of each P-H pair. If a sentence pair belongs to both category A and category B, we label it on both A and B. METALs has three categories and ten sub-categories (in 3.1 to 3.3). The following is a specific example of an annotation.

Example 1. P: African woman walking through field. H: There is women navigating through the fields.

The example contains three perspectives of judgments, so we assign it three labels. Firstly, omitting the attribute modifier "Africa" is labeled as *Elli* in *SS*. Secondly, "Walking-navigating" with similar meanings is labeled as *Syn* in *LP*. Thirdly, "There be" is a sentence pattern transformation, we label it as *ST* in *SS*.

METALs avoids the influence of subjective judgment on objective results in single labeling. Our comprehensive, accurate and highly operable system can solve the ambiguous problems in classification, and make a more fine-grained evaluation of the causes of entailment.

4.3 Data Collection

We manually annotate all 3368 items selected from SNLI test set with METALs. We formulate annotation rules before labeling. In the labeling process, the annotator should first judge the category and find the entailment fragment in the sentence. Then they should match the entailment fragment to sub-categories. The annotation flow diagram is shown in Fig. 1.

In order to ensure the quality of the annotations, we recruit two graduate students with linguistic learning background to annotate. They begin to label the data after the training. In the labeling process, two annotators judge the entailment type under the guidance of the annotation specification. A consistency assessment is then performed. Sentence pairs that are consistent with the evaluation are identified as the final labeled data and stored in the final labeled dataset. Sentence pairs that assess inconsistencies would be discussed by two annotators and relabeled until consistent results are achieved.

The annotation process is not linear. Multiple iterations improve correctness and credibility of the annotation results.

Model	**Author-acc**	**Our-acc**
ESIM	88.0	88.0
GPT	89.9	89.8
BERT_base	-	89.3
MT-DNN_base	91.1	91.0

Table 3. Experimental results on four models. "Author-acc" is the accuracy given in the original paper. "Our-acc" is the accuracy of our retraining model.

Category	ESIM	GPT	BERT	MT-DNN
LP	82.76	85.52	85.52	**89.66**
SS	96.73	**97.16**	96.90	96.81
CR	84.09	**84.79**	83.68	83.54
LP&SS	94.06	94.44	95.40	**95.59**
LP&CR	**87.03**	**87.03**	85.95	**87.03**
SS&CR	87.44	88.65	**89.13**	87.92
LP&SS&CR	91.67	92.42	**93.18**	**93.18**
Discard	76.74	74.42	**80.23**	75.58

Table 4. The classification accuracy (%) of ESIM, GPT, BERT and MT-DNN in entailment categories.

4.4 Data Annotation Results

Table 2 exhibits the percentage of categories and sub-categories on total entailment samples. "*-com" is the complex sub-categories of category *.

In the annotation process, the entailment relationship of some sentence pairs contains a complex of two sub-categories under the same category. For example, sentence pair "P: A *younger* man *dressed* in *tribal attire*. H: A *young* man *wears clothing*. " contains two kinds of entailment relations, including sub-category *HH* (tribal attire-clothing) and sub-category *Syn* (younger-young and dressed-wears) below category *LP*. Sub-categories have many combinations. But the number of each combination is rare. We name the combination of the different sub-categories under the same category as "*-com" and analyze them as a whole.

5 Experiments

5.1 Experimental Setup

We concentrate on several classic and popular models on SNLI, which attain strong performance.

Enhanced Sequential Inference Model (ESIM)[25] is a strong benchmark on SNLI. It achieves 88.0% in accuracy on SNLI and exceeds the human performance (87.7%) for the first time.

Generative Pre-trained Transformer (GPT) [26] explores a semi-supervised approach for language understanding tasks using a combination of unsupervised pre-training and supervised fine-tuning. It trains a left-to-right Transformer Language Model [31].

Encoder Representations from Transformers (BERT) [12] addresses unidirectional constraints by proposing a new pre-training objective: the "masked language model". It bases on a multi-layer bidirectional Transformer encoder [31]. In this paper, we retraining the BERT_base model.

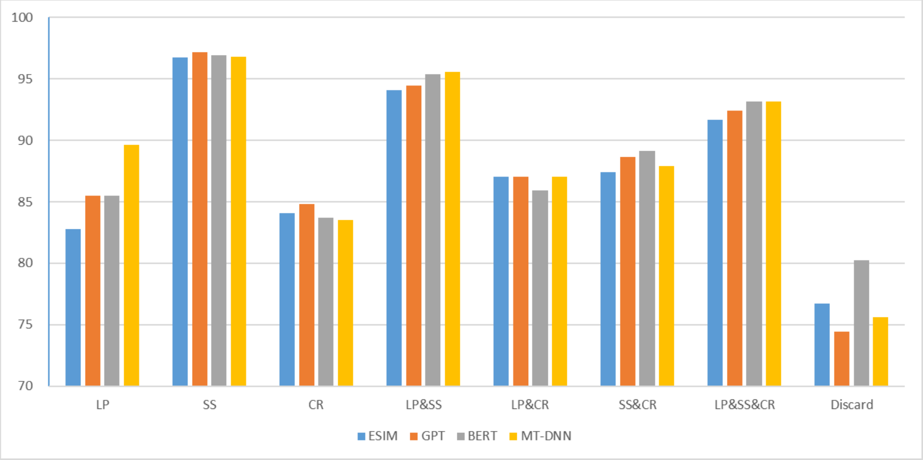

Fig. 2. The classification accuracy (%) of ESIM, GPT, BERT and MT-DNN in entailment categories.

Multi-Task Deep Neural Network (MT-DNN) [17] is a multitask deep neural network model for learning universal language embedding. MT-DNN integrates the advantages of multitask learning and BERT, and outperforms BERT in all ten natural language understanding tasks. MT-DNN_base are fine-tuned based on the pre-trained BERT_base.

We train above four models on SNLI. Our experimental results are shown in Table 3. The results obtained from our retraining almost all reach the results given by the author, which fully demonstrate that our analysis in the next chapter is credible.

5.2 Results and Analysis

In order to explore the performance of these models on different entailment categories, we test them on our annotated dataset using the reproductive models.

Performance on Coarse-grain The results of the four models in three categories and *Discard* are presented in Table 4 and Fig. 2. Since we use METALs, some sentence pairs may be labeled as two categories at the same time. The "*LP*&*SS*" means that the sentence pairs are entailment caused by *LP* and *SS* at the same time.

For sentences that contain only one category, we can easily observe from Fig. 2 that all models are excellent in *SS*. But models have poor performance in both *CR* and *Discard*. In Table 4, the classification accuracy of the four models is 19.99%, 22.74%, 16.67%, and 21.23% higher in *SS* than in *Discard*, and 12.63%, 12.37%, 13.22%, and 13.27% higher in *SS* than in *CR*. This stems from the model's stronger capability in recognizing information reduction and

Table 5. The classification accuracy (%) of ESIM, GPT, BERT and MT-DNN in entailment sub-categories. "*-com" is the complex sub-categories of category *.

category	sub-category	ESIM	GPT	BERT	mt-dnn
LP	Hyp	93.58	94.26	93.92	**94.59**
	Syn	89.64	90.13	90.79	**91.45**
	Mer	**100.00**	80.00	**100.00**	**100.00**
	LP-com	88.00	92.00	92.00	**96.00**
SS	Elli	93.74	94.71	**94.96**	94.65
	Ext	95.84	**96.40**	95.57	95.57
	ST	91.03	85.90	**92.31**	91.03
	SS-com	**95.14**	94.44	93.75	94.44
CR	EmoR	77.27	86.36	**95.45**	90.91
	QR	91.30	88.70	**93.04**	92.17
	SR	93.88	**95.01**	94.56	93.88
	EveR	81.07	**81.82**	80.57	80.57
	CR-com	89.04	**90.41**	87.67	89.04

semantic structural change in *SS*. On the contrary, the cases in *CR* are more complex, because they require additional commonsense knowledge for reasoning and imagination. The results of all models are not satisfactory in *LP*. And the classification accuracy of all pre-training models outperforms ESIM in this category. In particular, MT-DNN is 6.90% higher than ESIM. It shows that pre-training models bring a lot of knowledge about words and phrases.

The classification accuracy of these four models in *LP&SS* is significantly higher than that in *CR*&*SS*, *CR*&*LP*, and *CR*&*LP*&*SS*. This further indicates that the performance of the models on *CR* is poor, so the classification accuracy is still relatively low when combined with other categories.

The results for *Discard* are significantly lower than the other categories, and even the highest is only 80.23%. We illustrate this problem with a concrete example. Given a P: *A woman holds a newspaper that says "Real change"* and an H: *A woman* ***on a street*** *holding a newspaper that says "real change"*, the extra ingredients, like on a street, add in H are often hard to be inferred by models, even human. In essence, this kind of example should not be considered as entailment, but the noise mix into the dataset when it is set up. The noise results from the flexible annotation rules of SNLI.

Performance on Fine-grained The classification performance of all four models on all sub-categories is displayed in Table 5. We label sentence pairs that have multiple subcategories under the same category as complex entailment in that category.

In category *LP*, because the *Mer* only have five samples in the labeled test dataset, the results of four models in this sub-category valueless to analyze. MT-DNN outperforms other models in all sub-categories. This model is good at recognizing sentence pairs with word relations and phrase relations. And all three pre-training models perform better than ESIM. It indicates that the pre-training model has mastered more lexical and phrasal information in the process of pre-training.

In category *SS*, the classification accuracy of each model on *Ext* is more than 95%. In this sub-category, the H is part of the information proposed from the P, so it is easy for models to recognize the entailment relationship. All models have higher classification accuracy on almost all sub-categories of category SS than they do on all test sets.

In category *CR*, it is obvious that all pre-training models are excellent in *EmoR*, the classification accuracy of the three pre-training models is 9.09%, 18.18%, and 13.64% higher than ESIM. Pre-trained language models enable texts with similar emotions to have an associated representation, such as reasoning from "laugh/grin/smile" to "happy". Each model performs well on both *QR* and *SR*. This proves that the entailment of *QR* and *SR* are relatively easy to identify. The *EveR* makes up a big proportion of the total entailment samples, reaching 23.84 %. However, the classification accuracy of all models in this category is relatively low. This leads to a decrease in the classification accuracy of the model for the entire datasets. *EveR* is difficult for all models because the recognition of such entailment requires a strong knowledge, such as common sense knowledge. It is urgent to study how to improve the ability of models for *EveR* sentence pairs recognition.

From the above analysis, we can draw the conclusion that current models are better at recognizing the entailment relationship of *SS*. *CR* and *LP* are a bit difficult for the model, especially *CR*. Future work should focus on the representation and use of common sense knowledge in the training model. We have already seen the promotion of the pre-training model *LP*. In the future, we should also explore how to increase it further.

6 Conclusion

In this paper, we propose a complete entailment category annotation system (METALs). It has three categories, ten sub-categories. We manually annotate 3368 items of SNLI test set. To granularly examine the recognition ability of NN-based NLI models, We conduct experiments with four models on the labeled dataset for more fine-grained comparison of the strengths and weakness of each model. The experimental results demonstrate that all the models are excellent in *SS*. But results are not satisfactory in both *CR* and *LP*.

Acknowledgments

This work is funded by National Key R&D Program of China, "Cloud computing and big data" key projects (2018YFB1005105).

References

1. Akhmatova, E., Dras, M.: Using hypernymy acquisition to tackle (part of) textual entailment. In: Proceedings of the 2009 Workshop on Applied Textual Inference. pp. 52–60. Association for Computational Linguistics (2009)
2. Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo, D., Leggio, M.L., Magnini, B.: Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In: LREC. Citeseer (2010)
3. Bentivogli, L., Clark, P., Dagan, I., Giampiccolo, D.: The fifth pascal recognizing textual entailment challenge. In: TAC (2009)
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. Computer Science (2015)
5. Carmona, V.I.S., Mitchell, J., Riedel, S.: Behavior analysis of nli models: Uncovering the influence of three factors on robustness (2018)
6. Chen, Q., Zhu, X., Ling, Z.H., Wei, S., Jiang, H., Inkpen, D.: Recurrent neural network-based sentence encoder with gated attention for natural language inference. arXiv preprint arXiv:1708.01353 (2017)
7. Clark, P., Murray, W.R., Thompson, J., Harrison, P., Hobbs, J., Fellbaum, C.: On the role of lexical and world knowledge in rte3. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. pp. 54–59. Association for Computational Linguistics (2007)
8. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data (2017)
9. DAGAN, I.: The pascal recognising textual entailment challenge, p. 2. Machine Learning Challenges **3944**, 177–190 (2006)
10. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. In: Machine Learning Challenges Workshop. pp. 177–190. Springer (2005)
11. Demszky, D., Guu, K., Liang, P.: Transforming question answering datasets into natural language inference datasets (2018)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Garoufi, K.: Towards a better understanding of applied textual entailment. Ph.D. thesis, Citeseer (2007)
14. Ghaeini, R., Hasan, S.A., Datla, V., Liu, J., Lee, K., Qadir, A., Ling, Y., Prakash, A., Fern, X.Z., Farri, O.: Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. arXiv preprint arXiv:1802.05577 (2018)
15. Glockner, M., Shwartz, V., Goldberg, Y.: Breaking nli systems with sentences that require simple lexical inferences. arXiv preprint arXiv:1805.02266 (2018)
16. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data (2018)
17. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504 (2019)
18. Liu, Y., Sun, C., Lin, L., Wang, X.: Learning natural language inference using bidirectional lstm model and inner-attention. arXiv preprint arXiv:1605.09090 (2016)
19. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., Zamparelli, R.: Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). pp. 1–8 (2014)

20. McCann, B., Keskar, N.S., Xiong, C., Socher, R.: The natural language decathlon: Multitask learning as question answering. arXiv preprint arXiv:1806.08730 (2018)
21. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
22. Mou, L., Rui, M., Ge, L., Yan, X., Lu, Z., Rui, Y., Zhi, J.: Natural language inference by tree-based convolution and heuristic matching. In: Meeting of the Association for Computational Linguistics (2016)
23. Naik, A., Ravichander, A., Sadeh, N., Rose, C., Neubig, G.: Stress test evaluation for natural language inference (2018)
24. Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. arXiv preprint arXiv:1606.01933 (2016)
25. Qian, C., Zhu, X., Ling, Z.H., Si, W., Inkpen, D.: Enhanced lstm for natural language inference (2017)
26. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Tech. rep., Technical report, OpenAI (2018)
27. Roberts, K.: Building an annotated textual inference corpus for motion and space. In: Proceedings of the 2009 Workshop on Applied Textual Inference. pp. 48–51. Association for Computational Linguistics (2009)
28. Sammons, M., Vydiswaran, V., Roth, D.: Ask not what textual entailment can do for you... In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 1199–1208. Association for Computational Linguistics (2010)
29. Tay, Y., Tuan, L.A., Hui, S.C.: Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. arXiv preprint arXiv:1801.00102 (2017)
30. Vanderwende, L., Dolan, W.B.: What syntax can contribute in the entailment task. In: Machine Learning Challenges Workshop. pp. 205–216. Springer (2005)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
32. Wang, Z., Hamza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences (2017)
33. Welleck, S., Weston, J., Szlam, A., Cho, K.: Dialogue natural language inference (2018)
34. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference (2017)
35. Zhang, Z., Wu, Y., Li, Z., He, S., Zhao, H., Zhou, X., Zhou, X.: I know what you want: Semantic learning for text comprehension. arXiv preprint arXiv:1809.02794 (2018)