Annotation Difficulties in Natural Language Inference

Aikaterini-Lida Kalouli[1], Livy Real[2], Annebeth Buis[3], Martha Palmer[3], Valeria de Paiva[4]

[1]Center for Information and Language Processing – LMU
Munich – Germany

[2]americanas s.a. d.lab – São Paulo, Brazil

[3]Department of Linguistics – University of Colorado at Boulder
Boulder, Colorado – USA

[4]Topos Institute
Berkeley, California – USA

kalouli@cis.lmu.de, livyreal@gmail.com

Abstract. *State-of-the-art models have obtained high accuracy on mainstream Natural Language Inference (NLI) datasets. However, recent research has suggested that the task is far from solved. Current models struggle to generalize and fail to consider the inherent human disagreements in tasks such as NLI. In this work, we conduct an experiment based on a small subset of the NLI corpora such as SNLI and SICK. It reveals that some inference cases are inherently harder to annotate than others, although good-quality guidelines can reduce this difficulty to some extent. We propose adding a Difficulty Score to NLI datasets, to capture the human difficulty level of agreement.*

1. Introduction

Natural Language Inference (NLI), the task of determining whether a premise (P) entails, contradicts or is neutral to a hypothesis (H), has recently seen tremendous progress. The increasing availability of huge datasets has facilitated the training of massive models, pushing the state-of-the-art (SOTA) to high levels of accuracy, with some papers even claiming to reach human performance [Liu et al. 2019, Zhang et al. 2019]. With these results, one might consider NLI a solved task. To date, two different strands of research attempt to show that this seemingly perfect performance does not make the task solved. One strand focuses on detecting bias or artifacts in the training sets [Gururangan et al. 2018, Poliak et al. 2018] and creating challenging datasets that expose the generalization difficulties of the models [Glockner et al. 2018, Nie et al. 2018, McCoy et al. 2019, inter alia]. Another strand of research shows that due to inherent disagreements in tasks such as NLI, current SOTA models cannot claim to capture human level inference capabilities [de Marneffe et al. 2018, Palomaki et al. 2018, Pavlick and Kwiatkowski 2019].

This paper addresses both these directions. Specifically, in this work we discuss an experiment realized at X University, which investigates whether some NLI pairs are inherently more difficult and controversial to annotate than others, leading to significantly lower Inter-Annotator-Agreement (IAA). Parallel to that, we investigate how such difficult pairs can be detected, as well as to what extent different annotation guidelines might be able to solve some of the inherent complexity. For the latter investigation, we build

on previous work by [Kalouli et al. 2017] and [Kalouli et al. 2019], who list difficult phenomena that become sources of disagreement and also propose improved guidelines for the task. Thus, in this work, our contributions are three-fold. First, we quantitatively show that both corpora under investigation contain inherently controversial pairs that lead to significantly higher disagreements than other pairs. Secondly, we propose the augmentation of the NLI annotation task with a *Difficulty Score*. Such a score can contribute to a better training process as well as capture and avoid some of the aforementioned artifacts and bias. Finally, we show that the quality and thoroughness of guidelines and of the corpus construction itself play an important role in the amount of disagreement.

2. The Experiment

Our experiment was undertaken with the help of nine Computer Science and Linguistics graduate students in a Computational Linguistics seminar. These annotators were not under time pressure, they did not have a financial motive and had a much smaller number of pairs to work with than an average crowdworker. The students were split in three teams of three, each receiving the same set of 100 NLI pairs but with different annotation guidelines. The goal was to observe whether the different guidelines lead to different amounts of disagreement or whether some pairs have consistently lower IAA scores across guidelines. The students were asked to provide a label for each pair, but also to justify their decision in a short comment. Additionally, they had to give a *Difficulty Score* from 1 to 5 for how hard the annotation of each pair was for them; 1 being very easy and 5 very hard. Thus, the *Difficulty Score* is a subjective human rating score, similar to the common semantic similarity score given to inference pairs (cf. [Marelli et al. 2014]).

The dataset The dataset used for this experiment was constructed with pairs originating from the SICK [Marelli et al. 2014] and the SNLI [Bowman et al. 2015] corpora. SICK is an English corpus of almost 10,000 pairs, annotated for their degree of similarity and for the inference relation between the sentences of each pair. SNLI is an English corpus of over 550,000 pairs, annotated for inference. Both corpora were created from captions of pictures, talking about daily activities and non-abstract entities. SICK was also further simplified in terms of linguistic phenomena included, e.g., named entities and temporal phenomena were removed. From each corpus, we selected 50 pairs: 20 were originally annotated as contradictions (C), 20 as neutrals (N) and 10 as entailments (E). We chose this distribution of labels because controversies are most common in pairs labelled as contradictions and neutrals [de Marneffe et al. 2008, McCoy et al. 2019, Kalouli et al. 2019] and thus it makes sense to include more of these examples to test our hypothesis. From these 50 pairs, 25 were pairs we considered *clear-cut*, i.e., easy, unambiguous inferences where we expect people to agree on (10 Cs, 10 Ns and 5 Es), and 25 were *controversial*, i.e., some ambiguity within the pair could potentially lead to different annotations (10 Cs, 10 Ns and 5 Es). To capture the notion of controversy, we use the findings of [Kalouli et al. 2019]: pairs with *directionality, coreference* and *loose definitions* phenomena are considered controversial. Examples of each type are given in Table 1. At this point, we should clarify that this distinction is *annotation-based*. This means that the *clear-cut vs. controversial* distinction concerns the difficulty of the annotation of the pair and not the linguistic complexity of the pair. This is important because recent work [Glockner et al. 2018, Nie et al. 2018, Dasgupta et al. 2018, McCoy et al. 2019, in-

	Clear-cut	**Controversial**
E	P: A man is riding a horse on the beach. H: A guy is riding a horse.	P: A woman is making a clay pot. H: An artist is sculpting with clay. *is everyone who is making something with clay an artist?*
C	P: A woman holding a boombox. H: A man holding a boombox.	P: A man is holding a small animal in one hand. H: A man is holding a big animal in one hand. *depending on the judge, an animal might be small or big*
N	P: A woman is running a marathon in a park. H: The woman is running fast.	P: A woman is being kissed by a man. H: A lady is being kissed by a man. *is lady a synonym to woman or not?*

Table 1. Examples of *clear-cut* and *controversial* pairs.

ter alia] has shown how there can also be *easy* and *hard* inferences, based on the complexity of the linguistic phenomena involved in the sentences. For example, sentences with modals, passives, word-order scrambling, implicative and factive verbs, etc., are considered hard because models struggle with them. This distinction between *linguistic phenomena* and *annotation* difficulty is essential in our experiment.

The Guidelines As mentioned above, each group of annotators was given different guidelines to deal with the task.[1] The goal behind this strategy is to see if different guidelines lead to more or less controversy and whether there are pairs that are inherently more ambiguous across guidelines. Group 1 received the original SNLI guidelines. These guidelines provide a caption of a picture as the premise and ask the crowd workers to write a sentence that is a definitely-true/definitely-false/might-be-true description of that picture/caption. Since now we already have the P-H pairs, we reformulated these guidelines: the annotators were asked to judge whether H was a definitely-true/definitely-false/might-be-true description of P, as the SNLI creators also did in their validation stage. Group 2 received the improved guidelines proposed by [Kalouli et al. 2019][2] (KAL guidelines). These guidelines attempt to address issues found in the original SICK and SNLI guidelines, e.g., tackling coreference phenomena. The annotators are asked to imagine P as a caption of a picture, describing whatever is on that picture; P represents the truth based on which they have to judge H. Finally, Group 3 was not given any guidelines.[3]

3. Results and Discussion

The set-up of our experiment allows for different kinds of observations. Here, we mainly focus on the most prominent results. The overall goal of the experiment was to test

[1]The exact guidelines will become available after publication.

[2]Available under `https://github.com/kkalouli/SICK-processing`.

[3]They were only given the following: *You get two pieces of text: a premise and a hypothesis. For some examples, the hypothesis follows from the premise ("entailment"). In other cases, the text and the hypothesis are contradictory ("contradiction") and in some others the hypothesis neither follows from nor contradicts the text ("neutral"). Annotate each pair with an inference label, i.e., E for entailment, C for contradiction or N for neutral.*

Group: Guidelines	Min	Max	Mean	St.Dev.	# Clear-cut	# Controversial
Group 1: SNLI guidelines	1	4	2.14	0.71	80	20
Group 2: KAL guidelines	1	4	1.95	0.75	86	14
Group 3: No guidelines	1	2.6	1.5	0.48	71	29

Table 2. The min, max, mean and standard deviation of the *Difficulty score* for the pairs, as labeled by the annotators, and the number of *clear-cut* vs. *controversial* pairs based on the z-score computation.

Group: Guidelines	Own Classification		Annotators' Classification	
	Clear-cut	Controversial	Clear-cut	Controversial
Group 1: SNLI guidelines	72.1	51.9	72.1	20.6
Group 2: KAL guidelines	76.7	52.5	70.5	35.3
Group 3: No guidelines	50.8	38.9	51.5	23.3

Table 3. IAA between *clear-cut* and *controversial* pairs, across groups and guidelines, based on both classification schemes.

whether the controversial pairs correlate with low IAA and high *Difficulty Scores*. In other words, we wanted to test whether the IAA is statistically worse in controversial pairs. At the same time, we wanted to test what effect different guidelines have on this aspect. To this end, we split the pairs into *clear-cut* vs. *controversial* in two different ways: on the one hand, we relied on our own initial classification of the pairs into *clear-cut* vs. *controversial* (cf. Section 2). On the other hand, we used the annotators' *Difficulty Score* to get a notion of ambiguity and controversy: for each group, annotator and pair, we applied z-score normalization of the *Difficulty Score* to account for different raters using the scale differently; then, if the z-score of a given pair was greater than 1, the pair was considered *controversial* and, if it was equal or less than 1, it was considered *clear-cut*. The minimum, maximum, mean and standard deviation of *Difficulty Score* for each group, i.e. for each set of guidelines, is shown in Table 2. We also show the number of pairs classified as *clear-cut* or *controversial* with this process (the exact pairs with their classifications will be available after publication). Based on these two classifications of the pairs, we calculated the IAA between *clear-cut* and *controversial* pairs, across the three sets of guidelines. Results are shown in Table 3.

Despite the small-scale of this experiment, the results lead to enlightening observations. First, we observe that the clear-cut pairs have a significantly higher IAA ($p < 0.05$) than the controversial pairs, both in ours and in the annotators' classification. Interestingly, the significance level is higher for the annotators' classification, i.e., the agreement for the controversial pairs as defined by the annotators themselves is much worse than the agreement for the controversial pairs as defined by our classification. This finding is in line with [Pavlick and Kwiatkowski 2019], who show that disagreements are not to be dismissed as annotation noise, but rather persist as more ratings are collected and as the amount of context provided to raters increases. Similarly, our experiment shows that some disagreements are persistent, no matter the guidelines and the task definition; those are inherent disagreements in the way humans perceive semantic notions and deal with world knowledge. They can also not be solved by using a graded scale of annotation rather than distinct labels: the extent of disagreement will always persist. However,

this does not mean that the task is unsolvable, but rather that a different perspective is required.

One solution is proposed by [Pavlick and Callison-Burch 2016], who show that current SOTA models do not capture the same distribution over inference labels as that of the human judgments. Thus, they argue that NLI evaluation should explicitly incentivize models to predict distributions over human judgments. This solution is useful but does not tackle the issues presented above: the artifacts of the datasets and the generalization difficulties of the models. Concerning the artifacts, current datasets have been shown to have strong correlations between labels and words [Gururangan et al. 2018, Poliak et al. 2018], e.g. contradictions have a strong correlation with the words *no, not, nobody, sleep*, etc., so that models only pick up such statistical patterns rather than the reasoning rules behind the sentences. As far as the generalization difficulty is concerned, current models struggle with complex linguistic phenomena such as compositionality, negation, modals, passives, factives and implicatives, quantifiers, etc. [Nie et al. 2018, Dasgupta et al. 2018, McCoy et al. 2019, Richardson et al. 2019]. However, such *linguistically-hard* phenomena can be *annotation-clear-cut* and unambiguous for humans. For example, a pair like *P: The man chased the cat. H: The cat was chased by the man* or *P: The man walked the dog. H: The dog walked the man* is very easy for humans to annotate but contains compositionality rules which make it difficult for models to get right. Thus, our proposal attempts to also address these challenges: the NLI annotation should be complemented by a *Difficulty Score* like the one introduced here. This score can then serve two roles. First, it can be exploited during the training process: pairs that are *clear-cut* are more reliable for training and should thus have a stronger learning effect, e.g., have higher training weights, than *controversial* pairs with lower IAA. The score can also be exploited during the evaluation process by measuring performance on *clear-cut* vs. *controversial* pairs: it is expected that current SOTA models will fail on many of what might be *annotation-clear-cut* pairs (for humans) but *linguistically-hard* pairs (for machines), and thus this will reflect better the real reasoning power of these models. Second, our proposal can help reduce the artifacts of the datasets. For example, if pairs containing the word *sleep* in H are always judged as contradictory and *clear-cut*, no matter the complexity of P (due to the artifact that sleeping is used to contradict any other action), they can be recognized as artifacts and thus be removed or dealt with otherwise. Of course, consistent human explanations would be even better than the proposed score. However, these tend to be very expensive (if at all possible) during and after the corpus construction [Camburu et al. 2018]. Thus, a score seems a suitable way to mitigate the issues with annotation unevenness, if one is embarking on an annotation project. The alternative ways of generating explanations e.g. LIME (Local Interpretable Model-agnostic Explanations [Ribeiro et al. 2016]), NILE (Natural-language Inference over Label-specific Explanations [Kumar and Talukdar 2020]) or REXC (Rationale-inspired Explanations with Commonsense [Majumder et al. 2021]) from already annotated corpora, seem less direct and less trustworthy. Thus, the proposed difficulty score is one of the main contributions of this paper that we expect to impact how future annotation efforts and NLI datasets are conceived and executed.

Additionally, despite the small scale of this experiment and the inconclusive picture we get, we do observe that guidelines play an important role: Group 1, which was given the SNLI guidelines, has the lowest IAA for the *controversial* pairs in the classifi-

cation done by the annotators' themselves (20.6 vs. 35.4 and 23.3). On the other hand, Group 2 which is given the KAL guidelines, has the fewest *controversial* pairs (14 vs. 20 and 29) and the best IAA for these controversial pairs in both classification schemes (in the annotators' classification the difference is even statistically significant). Group 3 has the most *controversial* pairs and the lowest IAAs both for *clear-cut* and *controversial* pairs in both classification schemes – except for the *controversial* in the annotators' classification. From these findings, we can conclude that a) the nature of the SNLI guidelines leaves much room for interpretation and does not avoid some of the controversy which seems avoidable given the KAL guidelines, b) the SNLI guidelines do not address harder cases and thus there is high disagreement for the annotation of the controversial pairs, c) no guidelines whatsoever (Group 3) do lead people to think that the annotations are easy, presumably because they are not given "restrictions" based on which they should judge the pair (lowest mean Difficulty Score), but no guidelines clearly lead to poor IAA and thus the claim that people should be annotating as naturally as possible [Manning 2006] cannot be justified.

4. Conclusions

In this work we presented an NLI annotation experiment, aimed at investigating whether specific NLI pairs are inherently more difficult to annotate and thus lead to lower IAA. The experiment considered this question given different guidelines, to test to what extent the annotation difficulty is due to the guidelines quality. The results of this work show the value of augmenting the NLI annotation task with a *Difficulty Score* and the ways in which this score can be beneficial. Future work will seek to scale up these findings with a larger-scale experiment and confirm further preliminary findings of the current experiment. We also aim to reproduce the experiment with languages different from English, with corpora as [Fonseca et al. 2016] and [Real et al. 2018] for Portuguese.

References

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. (2018). e-SNLI: Natural language inference with natural language explanations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., and Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. *CoRR*, abs/1802.04302.

de Marneffe, M.-C., Rafferty, A. N., and Manning, C. D. (2008). Finding contradictions in text. In *Proceedings of ACL-08*.

de Marneffe, M.-C., Simons, M., and Tonhauser, J. (2018). Factivity in doubt: Clause-embedding predicates in naturally occurring discourse (poster). *Sinn und Bedeutung 23*.

Fonseca, E. R., dos Santos, L. B., Criscuolo, M., and Aluísio, S. M. (2016). Assin: Avaliação de similaridade semântica e inferência textual. In *Proceedings of PROPOR*, pages 1–8.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655. Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Kalouli, A.-L., Buis, A., Real, L., Palmer, M., and de Paiva, V. (2019). Explaining simple natural language inference. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143, Florence, Italy. Association for Computational Linguistics.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017). Correcting Contradictions. In *Proceedings of Computing Natural Language Inference (CONLI) Workshop, 19 September 2017*.

Kumar, S. and Talukdar, P. (2020). NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504.

Majumder, B. P., Camburu, O., Lukasiewicz, T., and McAuley, J. J. (2021). Rationale-inspired natural language explanations with commonsense. *CoRR*, abs/2106.13876.

Manning, C. D. (2006). Local textual inference: it's hard to circumscribe, but you know it when you see it–and nlp needs it. *https://nlp.stanford.edu/manning/papers/TextualInference.pdf*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*.

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

Nie, Y., Wang, Y., and Bansal, M. (2018). Analyzing compositionality-sensitivity of NLI models. *CoRR*, abs/1811.07033.

Palomaki, J., Rhinehart, O., and Tseng, M. (2018). A case for a range of acceptable annotations. In *SAD/CrowdBias@ HCOMP*, pages 19–31.

Pavlick, E. and Callison-Burch, C. (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, Berlin, Germany. Association for Computational Linguistics.

Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Real, L., Rodrigues, A., e Silva, A. V., Albiero, B., Guide, B., Thalenberg, B., Silva, C., Câmara, I. C. S., de Oliveira Lima, G., Souza, R., Stanojevic, M., and de Paiva, V. (2018). Sick-br: a portuguese corpus for inference. In *PROPOR 2018*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Richardson, K., Hu, H., Moss, L. S., and Sabharwal, A. (2019). Probing natural language inference models through semantic fragments. *CoRR*, abs/1909.07521.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., and Zhou, X. (2019). Semantics-aware BERT for language understanding.