Explaining Simple Natural Language Inference

Aikaterini-Lida Kalouli
University of Konstanz
first.last@uni-konstanz.de

Annebeth Buis
University of Colorado Boulder
anne.buis@colorado.edu

Livy Real
University of São Paulo
livyreal@gmail.com

Martha Palmer
University of Colorado Boulder
martha.palmer@colorado.edu

Valeria de Paiva
University of Birmingham
valeria.depaiva@gmail.com

Abstract

The vast amount of research introducing new corpora and techniques for (semi-)automatically annotating corpora shows the important role that datasets play in today's research, especially in the machine learning community. This rapid development raises concerns about the quality of the datasets created and consequently of the models trained, as recently discussed with respect to the Natural Language Inference (NLI) task. In this work we conduct an annotation experiment based on a small subset of the SICK corpus. The experiment reveals several problems in the annotation guidelines, and various challenges of the NLI task itself. Our quantitative evaluation of the experiment allows us to assign our empirical observations to specific linguistic phenomena and leads us to recommendations for future annotation tasks, for NLI and possibly for other tasks.

1 Introduction

In the era of big data and deep learning there is an increasing need for large annotated corpora that can be used as training and evaluation data for (semi-)supervised methods. This can be seen by the vast amount of work introducing new datasets and techniques for (semi-)automatically annotating corpora. Different NLP tasks require different kinds of datasets and annotations and provide us with different challenges. One task that has lately gained much attention in the community is the task of Natural Language Inference (NLI). NLI, also known as Recognizing Textual Entailment (RTE) (Dagan et al., 2006), is the task of defining the semantic relation between a premise text p and a conclusion text c. p can a) entail, b) contradict or c) be neutral to c. The premise p is taken to entail conclusion c when a human reading p would infer that c is most probably true (Dagan et al., 2006). This notion of "human reading" assumes human common sense and common background knowledge. This means that a successful automatic NLI system is a suitable evaluation measure for real natural language understanding, as discussed by Condoravdi et al. (2003) and others. It is also a necessary step towards reasoning as more recently discussed by Goldberg and Hirst (2017) and Nangia et al. (2017) who say that solving NLI perfectly means achieving human level understanding of language. Thus, there is an increasing effort to design high-performing NLI systems, which in turn leads to the creation of massive learning corpora. Early datasets, like FraCas (Consortium et al., 1996) or the seven RTE challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Dagan et al., 2010; Bentivogli et al., 2009b,a, 2011), contained a few hundred hand-annotated pairs. More recent sets have exploded from some thousand pairs (e.g., SICK, Marelli et al., 2014b) to some hundred thousand examples: SciTail (Khot et al., 2018), SNLI (Bowman et al., 2015), Multi-NLI (Williams et al., 2018). The latter two have been vastly used to train learning algorithms and achieve high performance. However, it was recently shown that this high performance can drop significantly by slightly modifying the training process (Poliak et al., 2017; Glockner et al., 2018). It was also shown that such training sets contain annotation artifacts that bias the learning (Gururangan et al., 2018; Naik et al., 2018). Other recent work (Kalouli et al., 2017b,a, 2018) discussed problematic annotations of the SICK corpus (Marelli et al., 2014b) and attempted to improve the annotations. All this work leads to the conclusion that corpus construction, including the annotation process, is much more important than what is often assumed and that bad corpora can falsely deliver promising results.

In this paper we take a closer look at the work

Proceedings of the 13th Linguistic Annotation Workshop, pages 132–143
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

by Kalouli et al. (2017b,a) and attempt to build on the two conclusions that arise from their work. The first conclusion is that the guidelines for the NLI annotation task need be improved, as it seems clear that human annotators often have opposing perspectives when annotating for inference. This can result in faulty and illogical annotations. The second conclusion concerns the annotation procedure: having an inference label is not enough; knowing *why* a human subject decides that an inference is an entailment or a contradiction is useful information that we should also be collecting, if we want to make sure that the corpus created adheres to the guidelines given. Specifically, in this work we discuss an experiment, realized at the University of Colorado Boulder (CU), which attempts to address both these issues: provide uncontroversial, clear guidelines and give the annotators the chance to justify their decisions. Our goal is to evaluate the guidelines based on the resulting agreement rates and gain insights into the NLI annotation task by collecting the annotators' comments on the annotations. Thus, in the current work we make three contributions: Firstly, we discover which linguistic phenomena are hard for humans to annotate and show that these do not always coincide with what is assumed to be difficult for automatic systems. Then, we propose aspects of NLI and of the annotation task itself that should be taken into account when designing future NLI corpora and annotation guidelines. Thirdly, we show that it is essential to include a justification method in similar annotation tasks as a suitable way of checking the guidelines and improving the training and evaluation processes of automatic systems towards explainable AI.

2 Background: the SICK corpus

To achieve these goals, we look at the SICK corpus (Marelli et al., 2014b). 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. The corpus was created from captions of pictures talking about daily activities and non-abstract entities. It was also further simplified in terms of the linguistic phenomena included, e.g. named entities and temporal phenomena were removed. Annotators were not given strict definitions as guidelines but instead one example for each type of label. They were also not told that the sentences came from pictures. This creation process caused much confusion as discussed in the original paper but also in Kalouli et al. (2017b,a). In particular, the process did not resolve event and entity coreference issues so that a pair like *A woman is carrying a bag* and *A woman is not carrying a bag* ended up labelled as neutral, instead of as a contradiction. This weakness was specifically targeted in the later corpora SNLI and Multi-NLI. In these corpora, in an attempt to provide premise examples grounded in specific scenarios, the annotators were given the freedom to write themselves a conclusion sentence for a given premise and they were informed that the premises come from captions of pictures.

3 The CU experiment

Our experiment was undertaken with the help of 12 Computer Science and Linguistics graduate students in a Computational Linguistics seminar. These annotators were not under the pressure of making hasty judgements for money and had a much smaller number of pairs to work with than an average 'Mechanical Turker'. The goal was to provide the students with clear, uncontroversial guidelines and ask them to annotate a small part of SICK. They were also asked to justify their decisions, in order for us to see whether the given guidelines solved some of the problems discussed in relevant literature (e.g. Marelli et al. (2014b); Bowman et al. (2015); Kalouli et al. (2017b,a)) and whether we could gain additional insights from the students' justifications. Apart from the inference relation and the justification, the students were also asked to give a score from 0-10 for what we would like to call "computational feasibility", i.e. their estimation of the likelihood of an automatic system getting the inference right.

The guidelines The guidelines for the CU experiment gave a detailed definition of NLI/RTE by using common literature definitions. The annotators were asked to imagine sentence *A* as a caption of a picture, describing whatever is on that picture – following the creators of SNLI and MultiNLI to deal with coreference issues. For each judgment, they were instructed to consider only the inference relation from *A* to *B* and not vice versa. They were also instructed to assume that sentence *A* represents everything they know about the world of the picture; *A* represents the truth based on which they have to judge sentence *B*. If *A* is talking about *a man in red pants walking* and *B* is also talking about *a man in red pants running*, they were told to assume that both sentences are talking about the same man and event. The guidelines also provided detailed examples of each inference relation, along with the kinds of justifications expected. Finally, special remarks were made for corner cases or cases that had already been shown in Kalouli et al. (2017b,a) to cause confusion. For example, they were told to ignore differences in determiners[1] and to use common-sense for matters that might seem subjective, e.g. *a huge stick* contradicts *a small stick*, even if a huge stick for a child might be a normal size stick for an adult, etc.

The annotation process For the current experiment, a total of 224 pairs was randomly chosen from SICK. The pairs were annotated for their inference relation in both directions, resulting in a total of 448 judgments. Each direction was annotated *separately* by 3 annotators. The annotators had to provide an inference label (*E, C, N* for entailment, contradiction, neutrality, or, if they could not decide at all, *DN* for "don't know"), a justification for their choice and the "computational feasibility" score discussed above. They could also note whether something was ungrammatical or nonsensical or if they had additional comments.[2] A set of 24 pairs (48 judgements) was given to all annotators at the beginning of the process for calibration. The annotators were instructed to use the same four labels described above (E, C, N, DN). In this set the three inference relations were almost equally represented: 16 entailments, 14 contradictions and 18 neutrals. For the set there was 75.8% overall inter-annotator agreement (IAA) with Cohen's κ at 0.68 ("allowing tentative conclusions" according to Carletta (1996)).[3] More concretely, there was 80% IAA for contradiction, 93% for entailment and 63% for neutrals. These agreement rates gave the preliminary impression that the guidelines were satisfactory.

4 Preliminary Observations

After collecting all annotations, we first calculated their IAA to compare it to the calibration set. Indeed, the overall average IAA was 73.25% with κ 64.25, comparable to the calibration set. κ is a standard metric in any similar task and here the high Kappa means that our guidelines work well enough to propose them for future tasks and allow us to make the annotated set available for further purposes. However, we decided to look deeper into the annotated data and examine whether this metric is indeed sufficient to ensure reliable annotations. After all, the goal of this work is to examine the annotation process in detail, especially observing the usefulness and need for the justifications we asked from the annotators. This goal was reinforced by our further finding that the annotations provided by our annotators were different from the original SICK annotations in 17% of the annotated cases! Assuming that our annotators are more reliable due to their training and better "working" conditions, this finding raises questions about the quality of the original SICK corpus, as already discussed by Kalouli et al. (2017a).

Detailed analysis of the data revealed different kinds of justifications. Firstly, there were the expected, less-informative justifications of the kind "no relation" or "sentences mean the same thing". Though allowed, such justifications do not offer a lot of insight into the annotation. Secondly, there were justifications describing the relation between the sentences and thus explaining the decision. For example, for the pair *A = A person is brushing a cat. B = Nobody is brushing a cat*, we got the justifications: "cat cannot be both brushed and not brushed", "cannot both brush and not brush a cat" and "someone != no one". Such justifications were the expected ones and what we hoped for when integrating the justification annotation.

Thirdly, the justifications and the annotations themselves indicated that there was much confusion about when a pair should be a contradiction or neutral. Annotators considered as contradiction pairs in which sentence *B* had nothing to do with *A*. In an attempt to find some relation between the sentences and without paying attention to the fact that contradictions can be defined only when entities/events are coreferent, the annotators found many contradictions. For example, the pair *A = Two sumo ringers are fighting. B = A man is riding a water toy in the water* was labeled as contradiction, with the justification "the subjects and activities are completely different". However, in what we considered clear guidelines, we had stated that "*A* represents everything you

[1] This "forced" equivalence of the determiners is suitable for this restricted annotation scenario, but would be unnatural for other contexts, e.g. consecutive sentences in a text.

[2] The guidelines and the re-annotated subcorpus are available under `https://github.com/kkalouli/SICK-processing`

[3] Label "DN" was also included for computing κ.

know about the world of the picture, *A* represents the truth based on which you have to judge sentence *B*" and that therefore in such an example, sentence *B* cannot be judged given *A*, hence the pair should be neutral. This observation is very interesting because it seems to concern other NLI corpora as well, e.g. in SNLI we find pairs like *A = A young boy in a field of flowers carrying a ball. B = dog in pool* also marked as contradiction, although it is clear that there is no coreference and thus it should be neutral. Conversely, we found many cases where there was an obvious coreference and contradictory events/entities but the annotators attempted to think of scenarios where both things could still co-occur. The pair, *A = A girl is getting a tattoo removed from her hand. B = A girl is getting a tattoo on her hand*, was correctly judged by two annotators as contradiction because "getting a tattoo contradicts tattoo removal" but the third one thought of it as neutral because "could be getting both at the same time".

Another more important observation was that the same pair had different agreement rates depending on its direction. Recall that the pairs were given in both directions but separately from each other. An example is the calibration pair *A = A light brown dog is sprinting in the water. B = A light brown dog is running in the water.* This direction of the pair (A → B) was unanimously annotated as entailment by 12 annotators. However, the opposite direction B → A got an agreement of 25% entailment and 75% neutrality. Here, some annotators gave justifications like "running and sprinting are kind of the same for every day situations" while others, following dictionaries more carefully, assumed that while sprinting is a kind of running, running does not entail sprinting. Only one direction of the pair is thus uncontroversial. This raises questions of whether one direction is indeed harder than the other and whether such directionality effects should be considered in the design and evaluation of NLI annotation tasks. To the best of our knowledge, this has so far not been taken into account for such datasets.

This observation is closely related to another: pairs involving what we would call "loose definitions/loose human inference" are also more prone to disagreements. Looking at the calibration pair *A = A white dog is standing on a hill covered by grass. B = A dog is standing on the side of a mountain*, the annotators have to decide whether *hill covered by grass* is the same as *mountain* and since definitions tend to be loose and subjective, such pairs get bad IAA (25% E, 33% C, 41% N). Interestingly, the opposite direction gets a slightly better agreement (17% C, 83% N), which again brings up the issue of directionality described above.Another good example is *A = A man is talking on the phone. B = A man is making a phone call.* Here, one annotator marked it as neutral as "talking on the phone does not entail that the man initiated the call", another marked it as contradiction because "making a phone call is an action that precludes talking on the phone", while the third one considered it an entailment because "talking on implies phone call". For tasks like NLI and for certain domains, we might need this kind of looseness that would allow the pair to be an entailment even though "talking on the phone" does not logically entail "making a phone call" (assuming that "making a phone call" contains the concept of in fact *initiating* the call, "talking on the phone" does not entail "initiating the call" and thus it also does not logically entail "making a phone call" (modus tollens)). But then, how do we define such corner cases? Could the annotation guidelines ever exactly *define* the concept of common sense, so that such cases are treated uniformly?

Another preliminary observation was the correlation of high "computational feasibility" scores (CF scores) with highly unambiguous pairs. The CF score was introduced in the annotation to check whether the annotators thought it was likely for an NLI system to get the inference label right. Since the score relied more on the annotators' intuition and less on objective annotation guidelines, we observed that the given answers varied widely with poor agreement. However, general observations can be made: high scores (above 8) were mainly given to pairs with direct, clear-cut negations like *A = Nobody is holding a hedgehog. B = Someone is holding a hedgehog.* or to entailments with only differences in determiners, such as *A = The person is peeling an onion. B = A person is peeling an onion.* or to entailment pairs with only one-word-difference, e.g. *A = A child in orange is playing outdoors with a snowball. B = A kid in orange is playing outside with a snowball*, where child = kid is an easy lexical entailment. These observations are not surprising: Kalouli et al. (2018) discuss such cases that can be easily solved solely based on WordNet (Fellbaum, 1998) and heuristics.

5 The experiment on the experiment

The previous observations lead us to two important conclusions: for one, the justifications the annotators provided were crucial to make us understand what was being annotated and what aspects of the guidelines were still unclear. Thus, if we are interested in annotated data that enables us to confirm the quality of the annotation task, similar justification fields are needed. Furthermore, the guidelines need to address aspects that can be controversial, e.g. they need to state explicitly and a priori that contradictions can occur if and only if coreference can be established. Such improvements will be discussed further in Section 6.1. The second conclusion is even more crucial: what if the previous observations are not merely random but can indeed be classified in phenomena and observed in other NLI data? While we know that many linguistic phenomena impose challenges for automatically detecting the inference relation between a pair of sentences, it is unclear which phenomena are also difficult for a human to annotate. For example, the passive/active voice distinction is a phenomenon that always receives attention when dealing with inference relations. However, this kind of phenomenon seems very easy for humans. On the other hand, dealing with loose definitions or coreference seems difficult even for humans. Since such phenomena repeatedly appeared in the justifications of the annotators, we decided to verify if the sentences that had lower agreement actually showed exactly these phenomena. We conjecture that these phenomena are measurable quantities that need to be considered in all future annotation tasks. If so, there should be a measurable correlation among the phenomena and the low IAA, so that these phenomena lead to statistically worse agreements. To investigate these questions, we conducted a second experiment based on the CU experiment: based on our observations of Section 4 and the previous literature on SICK, we defined six distinct categories according to which we ourselves meta-annotated all 224 pairs. Although this meta-annotation took place after making our preliminary observations on the data, the validity of this annotation is not influenced in any significant way: our preliminary observations were only that; observations and no real analysis of the data, also not an informal one. It was exactly this question that we seek to answer by this second experiment: can these abstract observations be quantified and analyzed in a formal way?

Specific Annotation Precisely, we meta-annotated the pairs for *coreference, directionality, loose definitions, atomicity, negation* and *quantification* phenomena. For the feature *coreference*, we marked whether a pair contains events or entities that are hard to assume coreferent (we annotated True for hard coreference and False for easy coreference). Coreference difficulty could lead to the first phenomenon described above; not being able to decide whether something is coreferent and thus contradictory, or neutral. In the category *directionality*, we marked for each pair direction whether this direction was harder, easier or equally difficult to annotate as the opposite direction. In the *loose definition* category, we checked whether the pair contains concepts that are "loose", subjective or vague to define (annotated as True) or not (annotated as False). The next category was inspired by the previous work of Kalouli et al. (2017a) on SICK: *atomicity* concerns the question of whether a sentence contains only one predicate-argument structure or more. This relates to the observation by Kalouli et al. (2017b) that marking the inference relation, and especially making events and entities coreferent, is easier to do when the pair only contains atomic sentences, i.e. sentences with one main verb. In non-atomic sentences, all parts of the sentence should be able to be made coreferent with the other sentence, something that often proves a challenge, especially if the other sentence is atomic. An example is the pair *A = The singer is playing the guitar at an acoustic concert for a woman. B = A person is playing a guitar and singing*. A is atomic but B is not (*playing and singing*), so that the question arises whether the *person singing* can be coreferent with the *singer*. We annotate each sentence of each pair with True or False, depending on whether they are atomic or not. *Negation* also contains the labels True or False: here we mark if each sentence of the pair contains a negation of any kind (verbal, pronominal, etc.). We do a similar task for *quantifiers*: we mark whether each sentence contains a quantifier or not.[4] We added these last two categories to quantitatively test our impression that negation and quantifiers also cause more annotation problems, just as coreference, loose definitions, etc.

[4] *a* is taken to be a determiner and not a quantifier

	IAA		CF score	
Phenomenon	**True**	**False**	**True**	**False**
A_is_atomic	72.06	79.41	6.81	6.68
B_is_atomic	72.60	76.81	6.83	6.59
A_is_negated	**88.88**	**71.46**	**7.66**	**6.68**
B_is_negated	**90.47**	**71.27**	**7.51**	**6.7**
A_has_quant	79.67	72.60	7.03	6.76
B_has_quant	80.48	72.50	7.05	6.75
hard_coref	**62.45**	**77.27**	**6.22**	**6.99**
loose_def	**59.60**	**77.19**	6.2	6.95

	Directionality		
Measure	**Easier**	**Harder**	**Equal**
IAA	**81.18**	**58.33**	**74.90**
CF score	6.57	6.58	6.88

Table 1: Overview of the average IAA (%) and CF score (1-10) for each condition of our experiment.

Results The overall goal of these meta-annotations was to check if the presence of these phenomena correlates with low IAA and low CF scores. In other words, we wanted to test whether the IAA and CF scores are statistically worse in pairs with such phenomena. To this end, we calculated the IAA and the CF score[5] for each pair and each of the six meta-annotations. We then computed the average IAA and CF score of the pairs in each condition of our meta-annotations. The results are shown in Table 1. We should note that we could conduct this kind of study only on the re-annotated SICK pairs of our CU experiment (Section 3) and not on the original SICK annotations because for those the exact IAAs are not available but only the final majority label. Thus, it would not be possible to quantify our findings over those annotations. However, we did investigate how the pairs that had been differently annotated by the original annotators and our annotators (17% of the cases, as explained above) showed these linguistic categories and we could retrace some of the findings discussed below: for example, among the pairs that were differently annotated by the original and our annotators there were significantly more pairs containing loose definitions (37% vs. 20%) and hard coreference (32% vs. 26%) than among the pairs that were annotated with the same label by the original and our annotators.

To test for the involved effects, we analyzed the IAA results using generalized additive mixed models (GAMMs) with the *ocat-linking* function for ordered categorical data (Wood, 2011, 2017). We chose this kind of modelling due to the nature of our dependent variable IAA.[6] The six meta-annotation categories were added as fixed factors with interactions and the pairs were entered as random smoothers. The fixed factors *coreference*, *loose definitions*, *atomicity of A*, *atomicity of B*, *negation of A*, *negation of B* and *quantification of A* and *quantification of B* were binary (True or False for each of them as described in 5) (cf. Table 1, top) and the effect *directionality* was a 3-level variable ("easier", "harder" and "equal") (cf. Table 1, bottom). Interaction, main effects and random smoothers were removed if they were not significant at α = 0.05 and the model was refitted.

Concerning the inter-annotator agreement, the results showed main effects of *coreference*, *directionality*, *loose definitions* and *negation*. For the *coreference* setting, there was statistically lower agreement in pairs with coreference marked as hard than in pairs with easy coreference, with p < 0.04. *Directionality* also showed a correlation with the agreement rates, with pairs in the "harder" direction having statistically lower IAA (p < 0.001) than pairs in the "easier" and "same" direction and pairs in the "same" direction having statistically lower agreements than pairs in the "easier" direction (p < 0.001). A similar observation can be made for the *loose definitions* effect: pairs not containing loose definitions showed a statistically better agreement than pairs with such definitions (p < 0.02). The three factors presented so far confirmed our preliminary observations that these phenomena are not random but are quantitatively depicted in the data. As far as negation is concerned, the results were counter-intuitive at first glance: pairs with negation in one of the sentences A or B had statistically higher IAA rates (p < 0.001) than pairs with no negation at all. However, after a closer look, this is not so puzzling: the pairs of our dataset containing negation are the kind of clear-cut, textbook types of negation with one sentence negating exactly what the other sentence is stating by the use of "not", "no" or "nobody", as *A = Nobody is holding a hedgehog. B = Someone is holding a hedgehog.*. Thus, this statistical result shows that it might in fact be easier to decide for such straight-forward pairs with clear-cut negation than for pairs that have no negation

[5] Calculated by averaging the scores of the 3 annotators.

[6] IAA normally ranges from 0 to 1 or from 0 to 100 but since we have four possible annotation labels (E, C, N, DN) and three annotators per pair there can only be distinct ordered agreements of 0.00, 33.33 or 100%

but contain hard coreference or loose definitions or generally some complex context. There was no main effect of quantification, i.e. there is no statistical difference between the agreement of annotators in pairs with and without quantifiers. This is probably expected given the very small number of quantifiers found in our data. Otherwise, it could indicate that quantifiers are not so hard for humans as they are assumed to be for machines. Last but not least, the effect of atomicity offers grounds for discussion: for one, annotating atomicity is not as clear cut as one could expect, e.g. there is the open question whether sentences with participles should count as atomic or not. In the example *A = A white dog is standing on a hill covered by grass. B = A white dog is standing on a grassy hillside*, it is not clear whether the participle *covered* should count as an additional predicate-argument structure. We decided to annotate such sentences as atomic (we considered non-atomic only sentences containing more than one *main* clause verbs). For another, we expected pairs with atomic sentences to be significantly easier to annotate for the inference relations compared to non-atomic sentences. This turns out not to be the case in our dataset: the atomicity of the sentences does not impact the agreement rates; the slightly higher agreement when A or B are non-atomic (condition False) is not statistically significant ($p > 0.08$). It is necessary to test this factor with more and more diverse data to see if the significance changes. No significant interactions could be established for this model.

To test for the involved effects in the CF scores results, we analyzed our results with a logistic mixed-effects regression model with CF score as dependent variable and the six meta-annotation categories as fixed factors (main effects and interactions) and the pairs as random effects, using the R-packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017). Then, the random and fixed effects were backward fitted, using the *step()*-function in lmerTest with the default α cut-off levels (0.1 for random effects and 0.05 for fixed effects). The best fitted model showed main effects of *coreference* and *negation*. Pairs involving hard coreference have statistically lower CF scores, i.e. they are considered harder for an automatic system to label. This correlation also shows that *coreference* is indeed an intuitively detectable factor of inference pairs that annotators "caught" by giving such pairs lower CF scores. Pairs with negation in A or B sentence have statistically higher CF scores, i.e. they are considered easier for an automatic system to label. Both these findings are consistent with our preliminary observations. As we observed in Section 4, high CF scores seem to correlate with pairs that are highly unambiguous. In our case, these are pairs with the kind of clear-cut, textbook negations like *A = A woman is slicing a tomato. B = There is no woman slicing a tomato.* or pairs containing easy entailments, e.g that a kid is a child or that a small boy is a boy. The fact that the CF scores are statistically higher when there is negation or when the coreference is clear, i.e. there is an easy entailment of the previous kind, confirms this observation. Nevertheless, as we noted for the inter-annotator agreement above, negation seems to be an easy case due to its nature in this dataset; it is expected that in more complex data, negation will play a different role. No significant interactions could be established for this model. Note that the small differences in the average CF scores shown in Table 1 result from the actual average scores used by the annotators for each pair ranging from a minimum of 3.54 to a maximum of 8.65.

In a small side experiment we also tested how the CF scores correlate with what is really hard for automatic systems. We chose the best performing system from the SemEval 2014 task (Marelli et al., 2014a) on SICK by Lai and Hockenmaier (2014) and extracted from their test data those pairs that were also included in our subcorpus. These 92 pairs were split into two groups: those where the label given by the automatic system was the same as the label given by our annotators and those where it was different, i.e. the system got it wrong. For each of those groups we calculated the average CF score. Both groups have an average CF between 6.2 and 6.8, which means that for our subcorpus and this NLI system there is no strong correlation between what our annotators considered hard for machines and what is indeed hard.

6 Discussion

The above results allow us to formulate three conclusions. Firstly, when certain linguistic phenomena are involved in NLI pairs, it is harder for humans to annotate the inference relation and the upper limit they can reach seems to be below the perfect 100% agreement that much research has assumed so far. Given this and the fact that our "ultimate goal" is indeed the human-level understanding, the NLI task should try to account for these cases: either create corpora without those phenomena and expect systems to achieve an almost perfect performance (as humans probably would, without these hard cases) or include the phenomena in the corpus but be aware of them and treat them differently in training and evaluation. Additionally, it seems that our findings strongly confirm our preliminary observations and these observations were possible due to the justifications of the annotators. Thus, enhancing similar annotation tasks with justifications (of some sort) might be a suitable way for building high quality corpora and gaining insights into a given task. Such practices might reduce the original benefits of crowdsourcing annotations, which lie in much data being gathered fast and cheaply: however, for tasks like NLI, having correctly annotated data might be more beneficial than having huge amounts of data.

6.1 Improvement of the NLI process

The experiment was conducted on a small subset of SICK, yet it was enough to show that even a small subset of a simple NLI dataset like SICK, contains linguistic phenomena that can cause much confusion among the annotators and lead to low inter-annotator agreement or even worse, to acceptable agreement rates but annotations that were not intended in the first place. On the one hand, we showed that *coreference*, *directionality* and *loose definitions* have a strong effect on the resulting agreement and thus these factors should be taken into account at different stages of the process. Some of these issues such as coreference can partly be addressed in the guidelines. Guidelines like the ones we proposed for the CU experiment or the ones from corpora such as SNLI fail to show annotators the difference between contradiction and neutrality. The suggestion of assuming a photo sounded promising but was still not able to avoid confusions. Other phenomena like loose definitions could also partly be treated by appropriate guidelines: the annotators could be motivated to judge the pairs strictly or leniently according to the needs of the corpus creators. To this end, they could be given specific examples like the one mentioned above with the dog sprinting/running and be told that in such situations they should assume double entailment, i.e. be lenient, or assume neutrality in the direction running → sprinting, i.e. be strict. They could alternatively be given dictionaries to adhere to. Still, those issues cannot be fully treated by guidelines and other aspects such as directionality can altogether neither be treated by guidelines nor be predicted during the corpus data creation/generation. This highlights the problem that has plagued the RTE task since its inception: the definition of entailment and contradiction in terms of likely human inference leaves a lot of room for interpretation and neither sufficient annotator training nor unambiguous guidelines can prevent that. However, accepting the fact that the task, though very useful, cannot be well-defined should not scare us but instead motivate us to deal with it in a more efficient way. We need to start devising corpora based on the notion of *human* inference which includes some inherent variability, and find appropriate methods to train our systems on such data and measure their performance on them. For example, NLI pairs could be labelled with the information about the specific kind of inference they are dealing with, similarly to what was already proposed by Zaenen et al. (2005). On the other hand, the systems could be adapted to consider these different labels: in the case of directionality, for example, we could post-hoc measure the IAAs of each pair in both directions and find the harder one. This feature can then be exploited by automatic systems to evaluate their performance on "harder" vs. "easier" cases. It can also be considered for the training process itself: pairs in the "easier" direction have a higher IAA, are more reliable and should have a stronger learning effect, e.g. have higher training weights, than pairs in the "harder", less-reliable direction. Moreover, we showed that phenomena that are considered "hard" for machines can be easy for humans, e.g. quantifiers, while other phenomena are not only considered hard for machines but are proven hard for humans too, e.g. coreference. But since our ultimate goal is human level understanding, certain machine weaknesses are to be expected.

6.2 Justifications for better tasks

The preliminary observations which led us to the quantitative experiment and revealed the impact of the discussed phenomena, were facilitated by the justifications of the annotators. Such justifications can firstly reveal, as in our case, whether the guidelines of the task are clear enough or whether there is confusion. In this way the corpus creators can check the quality of the annotation data. We have shown that the commonly used metric of simple inter-annotator agreement or Cohen's Kappa can be hiding crucial aspects of the annotation quality. Secondly, justifications can indicate other aspects of the task that need to be taken into account during the annotation task, similarly as in this experiment. However, the insights gained can also be exploited in the use of the corpus, i.e. in the training process of some supervised method. When the insights gained can be classified and quantified in clear patterns as in our case, these patterns can be used as additional features during training. This is common in active learning scenarios: the goal in active learning is to output annotations for an initially unlabelled corpus, in addition to linguistic insight (e.g., in the form of rules or deduced patterns). During the labeling stage of the learning loop, the user interacts with the algorithm by labeling an unannotated data instance, verifying a given annotation, providing an estimate of her confidence, and providing a justification for the decision. These justifications along with the annotations and the provided confidence are used to update the existing model in the form of updated or new rules and train the algorithm further (e.g. cf. Sevastjanova et al., 2018). Similarly, the produced justifications in such annotation tasks could be integrated in a "static" learning system in the form of additional rules, patterns or weights and thus lead to a more explainable model. Such justifications can be beneficial in annotations where there is a specific label or score to be chosen among other labels/scores, e.g. in NLI, in semantic similarity tasks, in sentiment analysis, in argument annotation, etc.

7 Relevant Work

Most relevant work on annotation focuses on issues of crowd-sourced annotations. Some work compares such annotations with expert-user annotations (Snow et al., 2008; Munro et al., 2010), while others recommend guidelines and other constraints to make the most of such annotations (Kittur et al., 2008; Aker et al., 2012; Sabou et al., 2014; Dligach et al., 2010). Some researchers propose ways to control and improve discrepancies in such data (Hovy et al., 2013; Tibshirani and Manning, 2014) and others try to point out the quality and ethical issues that arise from such practices (Fort et al., 2011). Considerably less research has been done in task-specific annotations. For NLI there is work discussing annotation challenges (de Marneffe et al., 2008; Kalouli et al., 2017b) and other focusing on improving crowd-sourced corpora (Kalouli et al., 2017a, 2018).

8 Conclusions

This work describes an experiment in which we re-annotated a small subset of the SICK corpus, a benchmark for the NLI task, to investigate how guidelines and specific linguistic phenomena influence annotation quality. Particularly, we discuss the benefits of justifications of the annotation decisions. Based on them, we were able to draw conclusions about aspects of NLI that are hard for humans and need special attention. With a quantitative experiment inspired by these justifications, we could measure the influence of these aspects and make proposals for future annotation tasks, in the NLI domain but also generally. Since NLI is defined based on common human understanding, being aware of the linguistic phenomena that make an inference complex for humans is a fundamental step towards a grounded expectation of what machines should do. We leave as future work to trace and quantify similar trends in other NLI data, e.g. in the SNLI corpus which has been largely used for training NLI systems but also seems to suffer from similar problems. Also, we would like to investigate better the category we called 'loose definitions', following the work of Zaenen et al. (2005). In addition, further research should focus on creating better guidelines for NLI, taking into account the findings of this experiment.

Acknowledgments

We would like to thank the Colorado students for their annotations, Bettina Braun and Katharina Zahner for useful feedback on the statistical analysis, as well as the anonymous reviewers for their constructive comments. We gratefully acknowledge the support of DTRA 1-IDTRAl - 16-1-0002/Project 1553695, eTASC - Empirical Evidence for a Theoretical Approach to Semantic Components and the Linguistics Department of the University of Colorado. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DTRA or the U.S. government.

References

Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour, and Udo Kruschwitz. 2012. Assessing crowdsourcing quality through objective tasks. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1456–1461, Istanbul, Turkey. European Language Resources Association (ELRA).

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, and Danilo Giampiccolo. 2006. The Second PASCAL Recognizing Textual Entailment Challenge. *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Douglas Bates, Martin Mchler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009a. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC)*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC)*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009b. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC)*.

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*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Cleo Condoravdi, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel G Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*, pages 38–45. Association for Computational Linguistics.

The Fracas Consortium, Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. The Fourth PASCAL Recognizing Textual Entailment Challenge. *Journal of Natural Language Engineering*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Dmitriy Dligach, Rodney D. Nielsen, and Martha Palmer. 2010. To annotate more accurately or to annotate more. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 64–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Karen Fort, G Adda, and Kevin Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 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.

Yoav Goldberg and Graeme Hirst. 2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Aikaterini-Lida Kalouli, Valeria de Paiva, and Livy Real. 2017a. Correcting contradictions. In *Proceedings of the Computing Natural Language Inference Workshop*.

Aikaterini-Lida Kalouli, Livy Real, and Valeria DePaiva. 2018. WordNet for Easy Textual Inferences. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Aikaterini-Lida Kalouli, Livy Real, and Valeria de Paiva. 2017b. Textual inference: getting logic from humans. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Tushar Khot, Ashutosh Sabharwal, and Peter Clark. 2018. SciTail: A Textual Entailment Dataset from Science Question Answering. In *AAAI*.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA. ACM.

Alexandra Kuznetsova, Per Brockhoff, and Rune Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, Articles*, 82(13):1–26.

Alice Lai and Julia Hockenmaier. 2014. Illinois-LH: A Denotational and Distributional Approach to Semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. 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)*, pages 1–8.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130, Los Angeles. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Adam Poliak, Pushpendre Rastogi, M. Patrick Martin, and Benjamin Van Durme. 2017. Efficient, Compositional, Order-sensitive n-gram Embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 503–508, Valencia, Spain. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines". In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rita Sevastjanova, Mennatallah El-Assady, Annette Hautli-Janisz, Aikaterini-Lida Kalouli, Rebecca Kehlbeck, Oliver Deussen, Daniel A. Keim, and Miriam Butt. 2018. Mixed-Initiative Active Learning for Generating Linguistic Insights in Question Classification. In *3rd Workshop on Data Systems for Interactive Analysis (DSIA) at IEEE VIS*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Julie Tibshirani and Christopher D. Manning. 2014. Robust Logistic Regression using Shift Parameters. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 124–129, Baltimore, Maryland. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Simon N. Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Simon N. Wood. 2017. *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.