Correcting Contradictions

Aikaterini-Lida Kalouli
University of Konstanz
aikaterini-lida.kalouli@uni-konstanz.de

Livy Real
University of São Paulo
livyreal@gmail.com

Valeria de Paiva
Nuance Communications
valeria.depaiva@gmail.com

Abstract

This paper describes a manual investigation of the contradiction asymmetries of the SICK corpus, which is the intended testing set for a new system for natural language inference. Any system providing conceptual semantics for sentences, so that entailment-contradiction-neutrality relations between sentences can be identified, needs a baseline test set. The investigation of this test set, a part of the SICK corpus, was necessary to check the quality of our testing data and to ensure that the set is logically valid. This checking showed us that the lack of a specific context or reference for the sentence pairs and the presence of indefinite determiners have made the task of annotating very hard, leading to those contradiction asymmetries. We propose a way of correcting these annotations, which solves some of the issues but also makes some compromises.

1 Motivation

This paper describes our continuing work to analyse and improve the SICK corpus of English sentences created by Marelli et al. (2014). We aim to use SICK as a golden corpus for our studies on inference detection. SICK (Sentences Involving Compositional Knowledge) provides a benchmark for compositional distributional semantic models. The corpus consists of sentence pairs that are rich in the lexical, syntactic and semantic phenomena that distributional semantics are expected to account for. However, the corpus is simplified in other aspects: there are no named entities, the tenses have been simplified to the progressive only, there are few modifiers, few complex verb expressions, etc[1].

The data set consists of 9840 English sentence pairs, generated from existing sets of captions of pictures. The authors of SICK selected a subset of the caption sources and applied a 3-step generation process to obtain their pairs. This data was then sent to Amazon Turkers who annotated them for semantic similarity and for inference relations, i.e. for entailment, contradiction and neutral stances. Since SICK was created from captions of pictures, it contains literal, non-abstract, common-sense concepts and is thus considered a simple corpus for inference.

Given its intended simplicity the SICK corpus is a good dataset to test a collection of approaches for obtaining deeper semantic representations and to test text entailment evaluation methods. While the corpus seems to have been mostly used, so far, for distributional comparisons (Bowman et al. (2015); Beltagy et al. (2015)), we decided to try a more logic-based approach. Thus, we started investigating the corpus to see on the one hand, what non-expert annotators consider entailment-contradiction-neutrality relations and on the other hand what kinds of inferences are included in such a simple corpus. This should give us an idea of where a logic based pipeline might fail because of the lack of encyclopedic knowledge or the lack of higher reasoning mechanisms that humans possess.

[1]These characteristics are indeed true. However, as with any corpus, there are some issues.

2 The Problem: Contradictions are symmetric

Looking into the data, however, we realized that the corpus construction and annotation process produced a very surprising result. Contradictions in logic are symmetric but the human annotators produced several cases where contradictions are asymmetric. Since each annotator annotated each pair only in one direction, the corpus ended up with 611 pairs where the first of the sentences is deemed contradictory to the second, but the second sentence is neutral or – in a few cases – even entails the other. Since we want a logically valid corpus, we first needed to correct such annotations that we deemed wrong and this led us to analyze the possible reasons for those mistakes.

In our previous work (Kalouli et al. (2017)), we manually looked at and corrected all 1513 pairs of one-sided entailments, notated in the corpus as $AeBBnA$ (meaning pairs where sentence A entails sentence B and sentence B is neutral with respect to A). We also analyzed some of the reasons for the wrong annotations. In the current work, we look at contradictions, since detecting conflicting or contradictory statements is a fundamental text understanding task within many applications, see Condoravdi et al. (2003). We started by looking the contradiction annotations which are asymmetric.

If proposition A is contradictory to B, then B must be contradictory to A. This is not what happens with these 611 pairs of SICK. From our processing of the corpus[2] we have the following asymmetric pairs:

- 8 pairs $AeBBcA$, meaning that A entails B, but B contradicts A;
- 327 pairs $AcBBnA$, meaning that A contradicts B, but B is neutral with respect to A;
- 276 pairs $AnBBcA$, meaning that A is neutral with respect to B, but B contradicts A.

These 611 asymmetric, problematic pairs, out of 9840 may seem few (around 6%) but given that the sentences were chosen to describe simple, non-abstract, common-sense situations, the number raises questions. We looked into these pairs and tried to analyze the reasons they were annotated as such and correct them, re-annotating them in a way that would make them symmetric and logical.

3 Observed mistakes

First of all, we discovered that some pairs contain ungrammatical sentences, e.g., *The black and white dog isn't running and there is no person standing behind.* Although the grammatical errors can be considered small, we can assume that each annotator mentally fixes the grammar in a different way, thus creating different relations and annotations. In this example, we could add an *it* at the end of the sentence or remove *behind* altogether and depending on this decision the sentence may have a different implicative behavior. There is also the case of nonsensical sentences in the corpus, e.g., *A motorcycle is riding standing up on the seat of the vehicle*, (did they just forget the word *driver* after *motorcycle*?) over which it is hard to reason – for the annotators and for us.

Additionally, there are sentences among the 611 pairs that simply seem wrongly annotated. We could not tell what was the reason for the wrong label. The pair *A = The blond girl is dancing behind the sound equipment. B = The blond girl is dancing in front of the sound equipment*, was marked as A contradicts B and B is neutral with respect to A. We should be talking about the same blond girl which is either in front or behind the sound equipment, for the same observer. Thus, B should also contradict A. A possible explanation here might be the vagueness of the notion of entailment for a "lay person". The RTE (Recognizing Textual Entailment[3]) task defines entailment as "given two text fragments called "Text" (T) and "Hypothesis" (H), it is said that T entails H (T $\rightarrow$ H) if, typically, a human reading T would infer that H is most likely true" (Dagan et al. (2006)). The problem with the "a human reading" clause is that it leaves too much choice for annotators. Annotators might have been overcompensating for the "human reading", assuming that – for the example above – different observers were at play.

[2] Available at `https://github.com/kkalouli/SICK-processing/tree/master/pairs`.

[3] `https://www.aclweb.org/aclwiki/index.php?title=Recognizing_Textual_Entailment`.

Similarly a pair like *A = A black and white dog is carrying a small stick on the green grass. B = A black and white dog is carrying a huge stick on the green grass.* Assuming that we are talking about the same dog, it is clear that a stick cannot be small and huge at the same time for the same observer, but notions of smallness will vary according to observers, so maybe the annotators decided that the size of the stick was not important enough to establish a contradiction between sentences.

We were able to deal with these kinds of mistakes by excluding them altogether from our cleaned version of the corpus. However, the task was not as easy for the other 520 cases out of the 611, for which some design flaws in the corpus construction created much confusion for the annotators and for us. As Marelli et al. (2014) mention in their paper, within corpora dedicated to inference, there is a high proportion of neutral annotations because most sentences are not related through inference. Also there are many neutral stances because the sentences of many pairs contain indefinite determiners and the usual interpretation of indefinite determiners creates a new referent, every time one indefinite determiner is present. Thus we discovered two main flaws in the original corpus design and construction. Firstly, when a picture is captioned most people use indefinites to introduce and describe the objects in the discourse. But this was not taken into consideration during the normalization process of the SICK sentences. This led to sentences with indefinites being paired with other sentences with indefinites. Such pairs are of course problematic due to the lack of a common reference. Secondly, the annotators were not given any guidelines on referents to judge the pairs and thus they tended to mark such pairs as neutral as the indefinite determiners were not binding to each other. An example is the pair: *A= A fearful little boy is on a climbing wall. B= A fearful little boy is on the ground*. Since there was no reason to assume that the indefinite little boys were the same, the annotators assumed that A and B are neutral with respect to each other because the two sentences can be talking about different *little boys*. If the sentences were talking about the same boy, they could not be true at the same time, so one would have a contradiction. But since the determiners are indefinite and there is no common referent, it is assumed that there were different boys.

Our investigation showed us that this phenomenon of "indefiniteness" and the lack of a specific context or reference created many other faulty annotations and not only the neutral ones already mentioned by Marelli et al. (2014).

4 Possible solutions

The previous literature in detecting and classifying contradictions in text is not very extensive. Previous works (Zaenen et al. (2005); de Marneffe et al. (2008)) argue that the events or entities of a sentence pair need to be coreferent if the pair is to be deemed a contradiction. Since the coreference of entities and events in the SICK pairs is not guaranteed, we could agree to take the sentences with indefinite determiners as they are and re-annotate them accordingly: the indefinite determiners in the example above would mean that the two boys are/can be different boys and therefore that the sentences should be marked neutral to each other. Although such an approach might feel natural and straightforward, it does not make a very useful contribution to the correction of the asymmetric annotations. This is because most pairs will get to be marked neutral by assuming independence of referents and events. This is due to the fact that if there is no binding referent, contradictions can only work if sentence A uses an indefinite determiner and B ultimately negates this with a universal quantifier, like *there is no*, *nobody*, *nothing*, etc. Other than that, the sentences ought to be reasonably similar. An example is the pair *A = There is no man chopping a log with an axe. B = A man is chopping a tree trunk with an axe*. But such constructions are not very common. In all other cases the pairs would have to be classified as neutral. We also did not consider good practice changing the determiners of those problematic sentences to definite ones. While this would avoid the existential quantification of the indefinites, this way we would be interfering with the corpus itself, changing it into something that is not the SICK corpus anymore.

Therefore, another kind of solution was necessary to correct these asymmetric annotations to something useful and reasonable. After considering different options, we concluded that the best way to correct the pairs would be to make a hard compromise: to try to *assume* that the pairs are talking about

the same event and entities no matter what definite or indefinite markers are involved, with some restrictions, though. To assume that the entities and the events are coreferent is not so easy, especially in cases where the two sentences are very distant from each other in meaning. Therefore, we could only correct using this method the sentences of the pairs that are of an *atomic* nature, meaning the ones with only one predicate-argument structure. Those are easier to compare, if we assume coreference. If we try to correct the pair *A = There is no man singing and playing the guitar. B = A man is playing a guitar.* we have the problem that A is not atomic (there are two predicates) and therefore A is not analogous to B where we just have one predicate. In other words, even if we assume that the two men are the same, we still cannot fully compare the two sentences because if there is no man *singing AND playing the guitar* it does not mean that there is also no man just *playing the guitar*. Thus, the two sentences would get to be neutral in the sense that they are incomparable according to our criteria. On the contrary, if we have the pair *A = A man is getting off the car. B= A man is getting into a car.*, we just have *atomic* sentences and – assuming coreference – we can compare them easily to one another; we can make the man and the event coreferent and mark the pair as a contradiction since the same man cannot be getting in and off the car at the same time.

Apart from this first constraint – **i.** "one predicate-argument structure constraint" – there is another one we have to take into account, namely that **ii.** even sentences with *atomic* structures have to be close enough in meaning. It is one thing to make the – linguistically bad – compromise that we do not consider the semantic contribution of definite and indefinite determiners, that we say we only care to mark the entities and predicates present in the sentences. But it is another thing to also make the – logically and semantically bad – compromise that all entities corefer *no matter what is said about them*. If the two sentences are not as close in meaning, it is hard to assume coreference and therefore the sentences are incomparable for our purposes. If we take the pair *A = A man is eating a banana. B= The man is not eating a banana by a tree.*, it is hard to assume that the men are the same, although both sentences are *atomic*, because the man in B is *by a tree* which means that the sentence is not comparable with A where we do not know where the man is. On the contrary, if we have the pair *A= A large green ball is missing a potato. B= A large green ball is hitting a potato.*, we can easily assume that the large green ball is the same in both sentences and that it can either be missing or hitting the same potato at the same time, so the pair is a contradiction. Of course, this notion of "close enough meaning" is very hard to pin down and might vary from annotator to annotator. For our purposes of calibrating our annotations, we first commonly corrected the same pairs of the 611 in order to be sure that our way of seeing "closeness" is similar enough.

Although this method of correcting the asymmetrical annotations does make some serious compromises, it seemed to us the best one for the task at hand. No other method would give us a satisfactory result without altering the corpus too much or turning all asymmetrical pairs to neutral stances. Making all pairs neutral would mean that the useful pairs of the corpus would shrink even more – they are already less than half of all pairs. Of course, with our approach we also lose some of the pairs because they get to be *incomparable* but there seems to be no way out. Despite our linguistic compromises, our approach follows de Marneffe et al. (2008) who say that "contradiction occurs when two sentences are extremely unlikely to be true simultaneously" , a definition that attempts to "more closely match human intuitions". It seems plausible to assume that if an annotator has a pair of sentences that could be understood as talking about the same referents and about the very same event, that annotator would assume the coreference.

5 Taxonomy of mistakes

All in all, we find four kinds of mistakes within the asymmetrical pairs. First, there are ungrammatical sentences and we can assume that those mistakes come from the generation process of the pairs; some of the transformations did not work as expected. Secondly, we have the non-sensical pairs which again must have arisen during some of the automatic transformations. Both these mistakes have probably led the annotators to mentally fix them in different ways, thus creating asymmetrical annotations. Thirdly,

we have discovered asymmetrical annotations that are plain errors for which we cannot say much about the source of the mistakes. Lastly, there are mistaken annotations due to the lack of a common reference and the presence of indefinites. The flaws in the corpus design and construction could probably account for those mistakes. We would like to stress at this point that we cannot give any quantitative measures of these mistakes as many sentences belong to more than one of these categories and thus we cannot classify the mistakes absolutely.

6 Conclusions

What we concluded is that for our purpose of creating a textual inference system but also more generally, for the purpose of creating inference corpora such as SICK, the explicitation of the referents of a sentence plays an important role, especially when dealing with contradictions. This explicitation is called the grounding of a sentence in Bowman et al. (2015), for instance. Whether creating a corpus or building an inference system, the creators of a corpus should make sure that there are specific referents on which the sentences are judged or grounded. On the one hand, for the creation of a corpus we suggest two such control mechanisms. Firstly, the creators could make the referents explicit by showing to the annotators the pictures from which the captions come and instructing them to judge the pairs according to the pictures and only them. However, such methods with pictures restrict the kind of language and information that can be conveyed and detected to concrete and literal actors and events. Other types of reference assigning, e.g. using Skolem constants as in Bobrow et al. (2007), can be tried out as well. Secondly, the corpus creators could employ other data collection methods (e.g. everyday conversations about concrete people) apart from captions' descriptions as these are bound to such problems as it was explained before. If it is deemed necessary to stick to pictures, the design of the collection could still be improved by providing the picture and asking the subjects to only finish off the captions which already introduce the referents of discourse so that no multiple referents are introduced by the subjects, e.g. *The woman is....* Moreover, if captions are the chosen way, a normalization step could make sure that such phenomena of indefiniteness disappear. On the other hand, inference systems should contain an extra pre-processing step where the sentences are compared for *predicate-argument structure similarity* or, in other words, for frames of coreference, so problematic cases like the ones presented above would not enter the inference engine at all.

A manual investigation and correction of these asymmetric pairs seems an important step in "healing" SICK so that we can use it for testing a semantic pipeline. Any pipeline that aims at extracting semantic representations from sentences and using those representations for entailment and contradiction detection in a logic-enforcing way needs to have symmetric contradictions. The entailments detected might be relaxed later on, to hold up to probabilities or up to likelihood only, but uncontroversial cases should be made explicit to begin with.

7 Future work

We hope to devise less expensive mechanisms for verifying the rest of the corpus, making use of lexical and knowledge resources. In particular we started a small experiment where we automatically filter out the pairs that are only one-word apart. This will allow us to see how many of these are easy entailments (e.g. *A = The snowboarder is leaping fearlessly over white snow. B = The snowboarder is leaping over white snow*), how many are easy contradictions (e.g. *A= A deer is jumping over a fence. B=A deer isn't jumping over the fence.*), how many we need to declare neutral. The hope is to come up with mechanisms for the construction of logic-based corpora that can be easily extended via composition, to serve as stringent baselines for hybrid logic-linguistic and machine learning systems.

References

Beltagy, I., S. Roller, P. Cheng, K. Erk, and R. J. Mooney (2015). Representing meaning with a combination of logical form and vectors. *CoRR abs/1505.06816*.

Bobrow, D. G., C. Condoravdi, R. Crouch, V. de Paiva, L. Karttunen, T. H. King, R. Nairn, L. Price, and A. Zaenen (2007). Precision-focused textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 16–21. Association for Computational Linguistics.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Condoravdi, C., D. Crouch, V. De Paiva, R. Stolle, and D. G. Bobrow (2003). Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*, pp. 38–45. Association for Computational Linguistics.

Dagan, I., O. Glickman, and B. Magnini (2006). The PASCAL recognizing textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop Recognizing Textual Entailment.*

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

Kalouli, A.-L., L. Real, and V. de Paiva (to appear, 2017). Textual inference: getting logic from humans. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

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

Zaenen, A., L. Karttunen, and R. 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*, pp. 31–36. Association for Computational Linguistics.