Textual Inference: getting logic from humans

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@nuance.com

Abstract

This paper describes a manual investigation of the SICK corpus, which is the proposed testing set for a new system for natural language inference. The system provides conceptual semantics for sentences, so that entailment-contradiction-neutrality relations between sentences can be identified. The investigation of the SICK corpus was a necessary task to check the quality of the testing data which is to be used as a golden standard for the new system. This checking also provides crucial insights for the implementation of the components of the system. The investigation showed that the human judgements used in the building of the SICK corpus can be erroneous, in this way deteriorating the quality of an otherwise useful resource. We also show that detecting the relationship between some pairs of the SICK corpus requires more than just lexical semantics, which provides us with guidelines and intuitions for our further implementation.

1 Introduction

This paper describes our manual investigation of the SICK corpus[1] by Marelli et al. (2014), which is to be used as a testing baseline for a new system for natural language inference (NLI). SICK is a corpus containing pairs annotated for their degree of similarity and for the inference relation between the sentences of each pair, i.e. entailment, contradiction, neutrality. The long-term goal is to be able to provide conceptual semantics for sentences so that entailment-contradiction-neutrality relations between sentences can be identified.

In order to evaluate this testing set (and future ones), we explore a preliminary pipeline that is open source and free to use and which we will expand by providing our own representations and implementation. We put different tools together and investigate which kinds of improvements are needed in each of those components so that the evaluation and verification of the corpus is efficient and successful. Investigating the testing data can give us insights on issues that need to be taken into account when we build the components of the new NLI system. By looking into a corpus such as SICK we can see what humans consider entailment-contradiction-neutrality and we can obtain clues on where a logic based pipeline might fail because of the lack of encyclopedic knowledge or the lack of higher reasoning mechanisms that humans possess. Additionally, investigating the test corpus can show us to what extent the off-the-self components can be used as-are or need improvements. Last but not least, by verifying the testing set we can be sure that it can be used as a golden standard for the new system and thus serve as a reliable baseline.

In what follows we briefly refer to the different components of this pipeline, describe the SICK corpus and then raise some issues on how to define textual entailment based on SICK. The main part of this work is a preliminary analysis of these problems, our suggestions for improving them and an investigation of other phenomena of the corpus, which are not discussed in Marelli et al. (2014).

[1] Available at http://clic.cimec.unitn.it/composes/sick.html.

2 The framework

The preliminary pipeline is based on dependency parsing provided by the Stanford Parser, in particular universal dependencies (UD). The Stanford parser can produce enhanced universal dependencies (Schuster and Manning (2016)), which are more semantic than conventional universal dependencies. Enhanced dependencies augment the information present in the basic dependency structure by adding more explicit syntactic relations and labels that facilitate many kinds of semantic transformations. Our pipeline also uses tools for lexical semantics. We make use of Princeton WordNet [2] as described in Fellbaum (1998) as a repository of word senses. For disambiguation of senses we use the JIGSAW algorithm[3] by Basile et al. (2007). Thirdly, we use the mappings from Wordnet to SUMO [4] by Niles and Pease (2001) which provide us with traditional knowledge representation concepts. These resources should give us a baseline knowledge representation semantics and we integrate them to see how much we can get back from this preliminary pipeline for our purposes of checking the SICK data.

For the ultimate system, we use the same tools to rewrite (syntactic representations of the) sentences into their semantic representations. Depending on the form of those rewritten conceptual semantics the last stage of the system will be implemented, which will be responsible for inference and reasoning, integrating some of the ideas from the PARC Bridge blueprint by Bobrow et al. (2007), but also combining state-of-the-art machine-learning techniques.

To test the quality of the preliminary pipeline and to see to what extent our lexical resources suffice or how and where they need to be improved, we need a simple, common-sense corpus, meaning a corpus which contains simple sentences about concrete, everyday, common events or activities. Since SICK was built to contain such common-sense sentences, it seemed ideal for our kind of testing. To verify the quality of the SICK data and to confirm our choice of corpus, we manually investigated it, checking what the human annotators considered inference relations and how they codified them. Before diving into the analysis, we define what we mean by entailment, contradiction and neutrality. We take entailment to be the semantic relation between sentence A and sentence B, where sentence A entails sentence B, if whenever A is true, then B must also be true. Contradiction is the semantic relation where sentence A contradicts sentence B if whenever A is true, then B cannot be true. If neither of those two relations holds for sentences A and B, then we say that sentence A is neutral with respect to B because there can be a world where both sentences hold or not.

3 The corpus

SICK (Sentences Involving Compositional Knowledge) by Marelli et al. (2014) is an English corpus, created to provide a benchmark for compositional extensions of Distributional Semantic Models (DSMs). DSMs approximate the meaning of words using vectors, which summarize the patterns of co-occurrence of words in corpora. SICK includes 9840 sentence pairs that are rich in the lexical, syntactic and semantic phenomena that compositional DSMs are expected to account for (e.g. lexical variation phenomena, impact of negation, etc.) but do not require dealing with other aspects of existing sentential data sets (e.g. named entities, temporal phenomena, etc.) that are not within the main scope of compositional distributional formal semantics. The curators of the corpus also made an effort to reduce the amount of encyclopedic world-knowledge needed to interpret the sentences.

The SICK corpus was created from captions of pictures talking about daily activities and non-abstract entities. Therefore, such pictures should "only" require concrete, common-sense concepts since they do not include many actions or actors or a too complicated description. With this setting, SICK becomes an ideal corpus for testing the off-the-shelf lexical resources.

The set of SICK pairs may seem large, but these sentences were "expanded" from a core set, which was normalized to restrict the linguistic phenomena and also to make sure that complete sentences and not

[2] Available at `http://wordnet.princeton.edu/`.

[3] Available at `https://github.com/pippokill/JIGSAW`.

[4] Available at `http://www.adampease.org/OP/`.

just caption-phrases were included. The SICK creators describe the process as follows: each normalized sentence was used to generate three new sentences based on a set of rules, such as adding passive or active voice, adding negations, etc. Each sentence was then paired with all of those three generated sentences. According to the authors, a native speaker eliminated odd and ungrammatical sentences. To repeat an example from Marelli et al. (2014), the caption *The turtle followed the fish* was normalized to the sentence *The turtle is following the fish* and expanded to the three sentences *The turtle is following the red fish*, *The turtle isn't following the fish* and *The fish is following the turtle*. After de-duplication, we have 6076 unique sentences combined in the different pairs. Also the lexical items involved are limited to less than two thousand lemmas of content words.

Each pair of sentences in the SICK corpus was annotated by Amazon Mechanical Turkers [5] to show their degree of similarity and their semantic relationship, namely entailment, contradiction and neutrality. The semantic relationships were annotated in both directions, meaning that annotators described the relation of sentence A with respect to sentence B and conversely, the relation of sentence B with respect to sentence A. Then, each pair was given one of the labels *ENTAILMENT*, *CONTRADICTION*, *NEUTRAL* based on the judgement of the majority of the annotators. However, the annotators were not told that the sentences came from captions and they were only given examples of the three kinds of inference relation as guides. Crowdsourcing techniques can be useful for such annotation tasks. However, the quality and consistency of these annotations can be poor. When looking at the corpus to investigate what humans considered entailment and contradiction, we realized that there were many troublesome annotations. Hence we decided to delve deeper into the corpus to find the reasons for those incorrect annotations and to manually correct the mistaken ones.

4 The problematic pairs

Contradictions in logic are symmetric; if proposition A is contradictory to B, then B must be contradictory to A. This is not what happens with the annotations of SICK. From our processing of the corpus [6] we have many asymmetrical 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.

In total 611 pairs out of 9840 are annotated in a way that logically does not make sense. These may seem few (6%) but since the sentences chosen ought to describe simple, common-sense situations and since these wrong annotations are not even self-consistent, this is a cause for concern. The ones where A entails B, but B contradicts A are the worst ones. In fact, we find surprising that the creators of the corpus decided to label pairs of this category as *ENTAILMENT*. If A entails B, but B contradicts A we have a logical contradiction, an absurd situation. The unidirectionality can work for $AeBBnA$, because such a relation is possible in logic: a sentence A can entail B and sentence B might be neutral with respect to A because it does not make any commitments about A. However, for the category $AeBBcA$, if sentence B contradicts A it can never be the case that A entails B; there must be a mistake somewhere in the annotation. Both other asymmetrical sets are also logically inconsistent and hence disturbing. Thus, it is important to verify why these occur. After all, the corpus is supposed to have been simplified and checked manually, to a large extent.

We manually looked into 108 of those 611 wrongly annotated pairs to discover what the mistakes were and see if those cases offer us insights for the task at hand. First off, we have the obvious case of mistaken annotations of different referents within the same pair, already discussed in Marelli et al. (2014). Since the annotators were not given information on where the sentences came from or what their

[5] Crowdsourcing platform avalaible at `https://www.mturk.com/mturk/welcome`.

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

frame of reference was, they did not have any kind of context to judge the sentences. Therefore, we see examples such as in the pair *A = An Asian woman in a crowd is not carrying a black bag. B = An Asian woman in a crowd is carrying a black bag.* The annotators decided that A contradicts B and that B is neutral with respect to A. It is clear that the same woman cannot at the same time be carrying a black bag and not carrying it, but there might be one woman carrying a bag and another one not carrying a bag. We might simply be talking about different women. This corpus design flaw seems to have created much confusion for the annotators. These mistakes seem to be the most common ones within the corpus, as 81 out of the 108 cases we looked at were of this nature. This lack of specific reference means that all contradiction pairs in which both sentences have indefinite determiners or in which one of the sentences has an indefinite determiner and the other one does not have a universal quantifier need to be checked as well. Contradictions need a common reference background, as already argued in Zaenen et al. (2005) and de Marneffe et al. (2008), and since there was none, it might be the case that we find more wrong annotations among the contradictions that we have not checked.

But not all problems are of this kind. We discovered pairs that contain an ungrammatical sentence, e.g., *The black and white dog isn't running and there is no person standing behind.* Although the grammatical errors are not dramatic, we can assume that each annotator mentally fixes the ungrammaticality 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 that decision the sentence pair might have a different relation. Moreover, there is the case of nonsensical sentences, e.g., *A motorcycle is riding standing up on the seat of the vehicle.* (did they forget to add *rider* after motorcycle?), over which it is hard to reason. Thus, it might be reasonable to exclude such sentences from the corpus.

Another common issue within the SICK annotations is the difference between contradictory and alternative concepts. When someone says *Three civilians died in the incident* and someone else says *No civilians died in the incident* and there is a single referent for the incident in question, you have a true contradiction. Both sentences cannot be true in any possible world. But when concepts are alternative to others, the annotator may have a problem deciding whether they are contradictory or not. For example, for the pair *A = The lady is cracking an egg into a bowl. B = The lady is cracking an egg into a dish.*, the annotators decided that A entails B, but B is contradictory to A. Clearly, a *bowl* is a sub-type of *dish*, as any bowl is a dish (a container), that is round and deep, so the entailment is easy to see. But why did the annotators decide that being a dish is contradictory to being a bowl? One reason could be that *bowl* and *dish* are considered alternatives (bowls are deep, dishes are flat) and therefore these were judged as contradictory.

Similar to this example, but harder to detect is the pair *A = The man is aiming a gun. B = The man is drawing a gun.*, for which the annotators decided that A entails B, but B is contradictory to A. The rationale seems to be that to aim a gun, you first need to draw it from the holder, so aiming a gun entails having drawn it beforehand. This is similar to the example that every bowl is primarily a dish and then a specific kind of dish. But *drawing the gun* does not contradict *aiming the gun*, as *drawing the gun* is a sub-concept or a precondition, we could say, of *aiming the gun*.

Another interesting case is the pair *A = There is no man on a bicycle riding on the beach. B = A person is riding a bicycle in the sand beside the ocean.* This seems to be about what the definition of a *beach* should be. The question might be whether a beach is some "sand beside the ocean" or not. There are beaches in seas, lakes and rivers too, for sure. There are stone beaches, as well. So, what do we take a beach to be?

Furthermore, there were sentences among the 611 pairs that were simply wrongly annotated. We could not tell what the reason for the wrong label was. 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. This cannot be the case because 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. It seems that other testsuites of the RTE (Recognizing Textual Entailment [7]) task had similar problems with pairs that can only be classified as "plain errors"

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

and no apparent reason for the mistakes can be found (Zaenen et al. (2005)).

Thus, to be able to use SICK as a real golden standard, we need to make sure that whatever is in the corpus is correct, as if humanly checked. In order to achieve this, we need a good understanding of what kinds of inferences are included in such a corpus and also of the kinds of mistakes that are found in there. Therefore, we started to check the corpus, by cleaning the entailment section included in it. We are making it available at `https://github.com/kkalouli/SICK-processing`.

We consider the category of pairs where A entails B and B is neutral with respect to A ($AeBBnA$). It is natural to start with this category as it is required to judge the others. The investigation of the wrong pairs showed us that it is easier to say if something entails something else, rather than to say that it contradicts it, because you need to check for more parameters in order to make the second decision. That is we need to decide whether the entities in the two sentences are co-referents or not, whether there can be a possible world where the two sentences can be true, etc. Our goal was to check all 9640 pairs and create our own "healthy" corpus, to begin with. For now, we simply investigate the 1513 pairs in the $AeBBnA$ class and try to collect the kinds of mistakes found, the reasons that could have led to them and also the kinds of inference involved in them.

5 Cleaned entailments

On total, we deemed 178 of the 1513 pairs as wrong, almost 12% of the pairs. Most of these mistakes were similar to the ones already discussed. The most common ones being the "plain" errors (148 can be categorized as such), the ungrammatical sentences (found in 8 pairs), the nonsensical sentences (found in 3 pairs) and the referents issue (found in 2 pairs). There are also sentences (found in 4 pairs) that do not really refer to common-sense concepts, as they were supposed to according to the design of the corpus, and this influences the annotations. For example, how do you really reason over and annotate a pair containing a sentence about *singing hamsters* since a singing hamster is not really in your every day experience?

Apart from those, we have the issue of compound nouns, in particular, of deverbal adjectives modifying nouns, such as in the pair *A = The microphone in front of the talking parrot is being muted. B = A parrot is speaking*. If the *talking parrot* is a parrot that is talking right now, then A entails B and B is neutral to A. But if the *talking parrot* refers to the parrot's general ability to talk (rather than that the parrot is talking right now), then A should be neutral with respect to B and B should be neutral with respect to A.

There is also the traditional problem of what Partee (2010) calls "privative adjectives" (4 examples found). These are adjectives that contradict the noun they modify, e.g. a fake gun is not a gun. Consider the pair *A = A cartoon airplane is landing. B = A plane is landing*. A cartoon airplane is not a proper airplane, the same way a toy car is not a car. Thus, A cannot entail B and should rather be neutral with respect to B as B is neutral with respect to A.

Additionally, there are still the expected issues with ambiguous sentences – at least, 5 pairs can be categorized as such – as in the pair *A = Two bikes are being ridden by two people. B = Two people are riding a bike*. Here, sentence B is underspecified in a way that does not allow us to judge if the two people are together riding one bike or if they are each riding their own bike. One might expect that such ambiguities will not be encountered in a corpus like SICK which ought to contain simple, common-sense sentences. Nevertheless, it seems that language cannot avoid ambiguity even in its simplest forms. This shows that even a basic, preliminary effort for an inference system will have to be able to deal with such basic ambiguities from the beginning.

We also observed that some definitions are cultural (observed in 2 pairs). A representative example is the pair *A = Different teams are playing football on the field. B = Two teams are playing soccer.* Depending on whether sentence A talks about American football or not, we can say that it entails B or not. If we are talking about American football, then sentence A does not entail sentence B.

We were also able to observe the kinds of inference intended by the corpus creators. Since the corpus was expanded from an initial core set of sentences, through semi-automatic transformations,

e.g. conversion of passive to active voice, synonyms of words, addition of adjective modifiers, etc., these kinds of inference are all there (Marelli et al. (2014) provide their complete list of expansions). Apart from those, however, we made a couple of other observations. Firstly, we found pairs where the entailment is based on more than basic lexical semantics. The pair *A = One man is turning on the microwave. B = The buttons of a microwave are being pushed by a man.* is an example. For this pair we need to infer that turning on the microwave requires the pushing of some buttons and that therefore some button must have been pushed. Such inferences are more than what Wordnet and SUMO can give us, off-the-shelf. Similar cases of explicit world-knowledge were already observed by Zaenen et al. (2005) and indicate that our proposed pipeline might have to find other ways to add some basic forms of world knowledge. Another interesting issue is related to agentive nouns, such as *swimmer*. Everyone who swims is a swimmer, but is everyone who poses for a photo a model? We found many pairs where this linguistic phenomenon seems to have an impact on human judgements. For example, the annotators may say *A man is wearing a purple shirt and black leather chaps and is posing for the camera* entails *A model is wearing a purple shirt and black leather chaps and is posing for the camera.* However, we do not think that a model is anyone who poses for a photo, but only the ones who are doing that intentionally (for money or not). Hence it should not be the case that A entails B, for any man. The dictionary we use, Wordnet, will not necessarily map *man* to *model*, as these annotators seem to have done. Since Wordnet is the resource we are using to link lexical knowledge to world knowledge, and since its first analysis contradicts some of the human judgements, we cannot take for granted that the whole analysis will be correctly done by Wordnet. For the example of *swimmer* we could use the Wordnet relation *derivated-by* that relates *swimmer* to *swim*, but in the case of *pose* and *model*, there is no explicit relation between these synsets. However, if we consider the glosses offered by Wordnet, we might be able to get that inference, see synset `http://compling.hss.ntu.edu.sg/omw/cgi-bin/wn-gridx.cgi?synset=10324560-n`, where *model* is described as "person who poses for a photographer or painter or sculptor". Even if Wordnet does offer both senses for *model*, it is unlikely that it would get the same sense as the annotators have because for other humans such as ourselves this is not an obvious relation.

6 Conclusions

Our manual investigation of the SICK corpus was an important preliminary step to check both the kinds of inference that we want to have and to detect the mistakes that can appear within the corpus design as well as the mistakes that humans can make when annotating inferences. On the one hand, the investigation helped us conceptualize the limits of lexical semantics. Even a simple, common-sense corpus as SICK contains sentences that cannot be captured by the semantics given by Wordnet and SUMO and therefore our preliminary pipeline may also need more than that. Moreover, world-knowledge has to be integrated in a way that does not conflict with lexical semantics. On the other hand, the wrong cases we found motivate us to think through the special challenges that semantics pose and whether and how these can be handled in an automatic system. We will need more sophisticated solutions for sentences whose interpretation needs more context, for deverbal and privative adjectives, for alternative vs. contradictory concepts, for the underspecificity of some of the definitions of concepts, etc. Our next step will be to provide a baseline corpus of basic entailments and contradictions, based on SICK, which can then be used as a reliable golden standard. For that we would like to use the strengths of our preliminary pipeline. Together with the completion of this task, we will focus on the other components of our system.

References

Basile, P., M. de Gemmis, A. L. Gentile, P. Lops, and G. Semeraro (2007, June). Uniba: Jigsaw algorithm for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, pp. 398–401. Association for Computational Linguistics.

Bobrow, D. G., B. Cheslow, C. Condoravdi, L. Karttunen, T. H. King, R. Nairn, V. de Paiva, C. Price, and A. Zaenen (2007). Parc's bridge and question answering system. In *Proceedings of the Grammar Engineering Across Frameworks Workshop (GEAF 2007)* .

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

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

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*.

Niles, I. and A. Pease (2001). Toward a Standard Upper Ontology. In C. Welty and B. Smith (Eds.), *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pp. 2–9.

Partee, B. H. (2010). Privative adjectives: Subsective plus coercion. In *Presuppositions and Discourse: Essays Offered to Hans Kamp*, pp. 273–285. Brill.

Schuster, S. and C. D. Manning (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

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.