Improving the precision of natural textual entailment problem datasets

Jean-Philippe Bernardy, Stergios Chatzikyriakidis
University of Gothenburg
Centre for Linguistic Theory and Studies in Probability
jean-philippe.bernardy,stergios.chatzikyriakidis@gu.se

Abstract

In this paper, we propose a method to modify natural textual entailment problem datasets so that they better reflect a more precise notion of entailment. We apply this method to a subset of the Recognizing Textual Entailment datasets. We thus obtain a new corpus of entailment problems, which has the following three characteristics: 1. it is precise (does not leave out implicit hypotheses) 2. it is based on "real-world" texts (i.e. most of the premises were written for purposes other than testing textual entailment). 3. its size is 150. Broadly, the method that we employ is to make any missing hypotheses explicit using a crowd of experts. We discuss the relevance of our method in improving existing NLI datasets to be more fit for precise reasoning and we argue that this corpus can be the basis a first step towards wide-coverage testing of precise natural-language inference systems.

Keywords: Natural Language Inference, Textual Entailment, RTE, SNLI, SICK

1. Introduction and Background

Reasoning is part of our every day routine: we hear Natural Language (NL) sentences, we participate in dialogues, we read books or legal documents. Successfully understanding, participating or communicating with others in these situations presupposes some form of reasoning: about individual sentences, whole paragraphs of legal documents, small or bigger pieces of dialogue and so on. Depending on the domain, and in general the situation, it appears that the reasoning performed can be more or less precise. Consider the following example:

(1) Three representatives are needed.

If a human reasoner with expert knowledge was to interpret the above utterance in a legal context, s/he will most probably judge that a situation where more than three references are provided could be compatible with the semantics of the utterance. To the contrary, if the same reasoner was to interpret the above as part of a casual, everyday conversation, then *three* would most likely be interpreted as *exactly three*, making the same situation incompatible with the utterance. We aim to examine to what extent existing corpora for NLI capture precise reasoning. We call "precise reasoning", inference which is either performed by experts or normal people after taking some time to consider the inferences that follow or not from a set of premises.

There have been claims that large corpora ($>$1000 inference problems) like SNLI and MultiNLI (Bowman et al., 2015; Williams et al., 2017) are not suited to test systems of this type. Only the FraCaS test suite is intended to capture such reasoning, but it only involves 346 hand-constructed problems.

The next candidates for precise larger-scale corpora would be the Recognizing Textual Entailment (RTE) datasets. While the SICK dataset (Marelli et al., 2014) can be also thought to be fit for testing precise reasoning, to a certain extent, but SICK has been originally designed to specifically test distributional compositional semantics approaches (Section 3.3.). Therefore, in this paper, we consider the RTE dataset, as it represents the closest candidate to our ideal dataset.

To shine a clear light on the strengths and weaknesses of the RTE and the rationale behind choosing it instead of the FraCaS, we consider both dataset in somewhat more detail.

1.1. The FraCaS test suite

The FraCaS test suite[1] is an NLI data set consisting of 346 inference problems. Each problem contains one or more premises followed by one yes/no-question.[2] There is a three way classification: YES, NO or UNK (unknown, see example (2) for an example from FraCaS). The FraCaS test suite was later on turned into machine-readable format by Bill McCartney[3]

Expansions of FraCaS include: a) MultiFraCaS, in effect a multilingual FraCaS[4], and b) JSem, the Japanese counterpart to FraCaS, which expands the original FraCaS in a number of ways.[5]

The FraCaS test suite covers a wide range of NLI cases and is, at least to some extent, multilingual. It is for the most part precise. Except for a few quirky cases (dubbed as undefined in Bill MacCartney's XML version), well agreed-upon reasoning rules clearly define if there is entailment or

[1] `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`

[2] Yet FraCaS exhibits some formal problems. For example, four problems are not formulated as a question.

[3] `www-nlp.stanford.edu/~wcmac/downloads/fracas.xml`. There, the conclusion is presented in two forms: one in the original question format and one as a declarative statement following from the premises.

[4] `www.ling.gu.se/~cooper/multifracas/`

[5] More info on the suite and its innovations compared to the original FraCaS can be found here: `http://researchmap.jp/community-inf/JSeM/?lang=english`.

not.

(2) An UNK example from the FraCaS test suite.

P1 A Scandinavian won the Nobel Prize.

P2 Every Swede is Scandinavian.

H. Did a Swede win the Nobel prize?

H. A Swede won the Nobel prize.

Label UNK [FraCaS 065]

Despite its qualities, the FraCaS test suite suffers from two major drawbacks:

- It consists almost exclusively of artificial examples;
- It contains only 346 examples

1.2. Recognizing Textual Entailment

The Recognizing Textual Entailment (RTE) challenges constitute a yearly series of NLI tasks which appeared between 2004 and 2011. Their aim is to test textual entailment, i.e. relations between a premise text and a hypothesis text (3).

(3) An entailment example from RTE1.

P. Budapest again became the focus of national political drama in the late 1980s, when Hungary led the reform movement in eastern Europe that broke the communist monopoly on political power and ushered in the possibility of multiparty politics.

H. In the late 1980s Budapest became the center of the reform movement.

Label Entailment [RTE702]

In contrast to the FraCaS test suite, the RTE challenges use naturally occurring data as premises. The hypothesis text is then constructed based on this premise text. There is either a binary or a tripartite classification of entailment, depending on the issue of RTE. The first two RTE challenges follow the former scheme and make a binary classification of entailment (entailed or not entailed). Tripartite classification (entailment, negation of the hypothesis entailment or no entailment) is added in the later datasets. This classification is a strict refinement of the binary one; and so one can use it as such if one wishes. Seven RTE challenges have been created altogether.

As such, RTE addresses the drawbacks of the FraCaS test suite: it uses natural text and offers an order of magnitude more examples. However, as we demonstrate in this paper, RTE is not suitable for testing precise reasoning. Additionally, the abundance of real-world concepts is problematic (Section 3.2.).

This paper makes two contributions:

- We test the hypothesis (H0) that the RTE data-sets is insufficiently precise to capture logical reasoning or precise reasoning tasks.

  This hypothesis is perhaps not quite surprising. Indeed, the creators of RTE had in mind a more loose definition of inference where both a precise and an imprecise definition of entailment would be at play. (Dagan et al., 2010; Sammons et al., 2012) mention that "our applied notion of textual entailment is also related, of course, to classical semantic entailment in the linguistics literature... a common definition of entailment specifies that a text t entails another text h (hypothesis, in our terminology) if h is true in every circumstance (possible world) in which t is true." This is close to what we want to capture in this paper. But, at the same time, (Dagan et al., 2010) also mention that "however, our applied definition allows for cases in which the truth of the hypothesis is highly plausible, for most practical purposes, rather than certain".

  The novelty lies in the way that we test the hypothesis (H0). We do so by appealing to a panel of linguistic and logic experts; invoking their definition of "precise reasoning" as the gold standard.

- Having the RTE problem-sets as our starting point, we propose a method of collecting entailment pairs which combine real-world texts and precise reasoning. The idea is to make explicit the supporting missing/hidden premises which are used in justifying or not an entailment pattern.

We present our method in detail in Section 2. In Section 3., we show how we come to conclude (H0) to be validated. We also provide a detailed analysis of the discrepancies between RTE and the judgements that we collected. We discuss and conclude in Section 4.

2. Method

We have randomly selected 150 problems out of the RTE 3 (i.e. from 2007) development corpus which were marked as "YES" (i.e. entailment holds). The problems were not further selected nor doctored by us. The problems were then re-rated by experts in logic and/or linguistics. For each problem, three experts were consulted, and each expert rated 30 problems. More precisely, the experts were instructed to re-consider each problem and be especially wary of missing hypotheses. If they considered the entailment to hold, we gave the instruction to optionally mention any additional implicit hypothesis that they would be using. Similarly, if they considered that there was no entailment in the problem, they were suggested to (optionally) give an argument for their judgement — thereby also indirectly indicating missing hypotheses.

In order to facilitate data collection, the experts were chosen from the network of contacts of the authors. Despite focusing on reliable people, the process of data collection took nearly six months. The authors themselves were put to contribution in the data-collection process (taking one set of 30 problems each) in order to complete the survey.

Additionally, we consolidated all the inputs received and, using our best judgement, we have put together a test set of 150 problems comprised of the original problems, a new

Type	Count	Ratio
Yes, with no missing hypothesis	223	0.49
Yes, with missing hypotheses	146	0.33
No, with no explanation	33	0.07
No, with explanation	47	0.10
Total of doubtful entailment	226	0.50
Total of any type	449	1

Table 1: Number of responses by type

judgement ("yes" or "no"), and added missing hypotheses (if "yes" is a reasonable option). [6]

3. Results

In the process, we have gathered a total of 449 expert judgements (one expert failed to answer a given problem), 146 missing hypotheses and 47 explanations for negative judgements. The entailment judgements are found in Table 1.
Despite all original problems being classified as "yes" by the creators of the RTE3 testsuite — we find here that one average, one expert in two is likely to cast a doubt over this "yes". Here, we count as a doubt either a response of "no" or "yes" with missing hypotheses.

"Yes if ..." vs "No because ..."? We elected to group those categories in our summaries, because the classification between "yes" with missing hypotheses and "no" is a tenuous one. Indeed, experts often find the same missing hypotheses but classify the problems differently (as "yes" or "no"). Consider the following example:
Example: (Problem 672)
P: Philip Morris the US food and tobacco group that makes Marlboro, the worlds best-selling cigarette, shrugged off strong anti-smoking sentiment in the US.
H: Philip Morris owns the Marlboro brand.
We got the following answers:

A1 Yes, if making involves owning the brand

A2 Yes, if making something implies owning the brand

A3 No, because making the product does not imply owning the brand

It is clear for all experts, the same premise is missing, but some will consider it acceptable to add, others will not.
Therefore the doubtful/certain classification appears to be a more enlightening one. This common-sense analysis is confirmed statistically: the agreement factor (Fleiss' Kappa) is higher when grouping answers in the doubtful/certain categories (κ=0.33) than when grouping answers in the yes/no categories (κ=0.21).
Thus, another way to look at the data is counting the number of experts casting doubt on an entailment problem. In Fig. 1, we show the distribution of number of experts casting doubt on entailment, over all problems, as an histogram. Due to the limited number of respondents for each problem we can only draw preliminary conclusions about RTE in general, but we can make the following observations in addition of the Fleiss' Kappa:

[6] the data is available as part of the "share your LRs" initiative, and as appendix to this document.

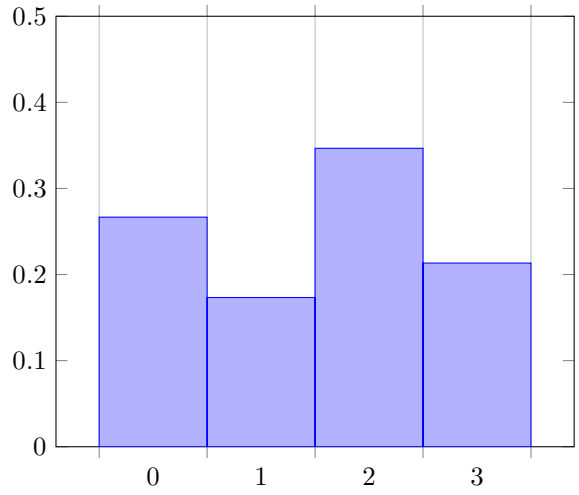

Figure 1: Distribution of the number of doubtful subjects.

1. Perfect agreement (0 or 3 doubts) occur in 48 percent of cases.

2. The probability of having a single doubt being cast is the lowest.

We find this level of agreement indicative of a good level of reliability. Additionally, with three experts per problem, we are likely to discover most missing hypotheses and incorrect entailments.
To obtain a final judgement on RTE problems, we have consolidated the experts judgements and their information of missing hypotheses, to our best judgement. In our compilation, we have marked 42 problems as straight "No", 64 as "Yes" with missing implicit hypotheses and "44" as plain "Yes". This means that, we expect, in our opinion, 28% of problems to be incorrectly labelled in RTE3 for precise reasoning, *even assuming reasonable world knowledge*. An additional 42% of problems require additional (yet reasonable to assume) hypotheses for entailment to hold formally, as prescribed by RTE3. This leaves only 30% of problems acceptable as such. The reason that the amount of doubt is larger than in the average numbers quoted above is that, for many problems, certain missing hypotheses and/or error were not detected by a majority experts, but, after careful inspection, we judge that the minority report is justified (the indicated missing hypothesis is compelling).
In order to further analyse the discrepancies between RTE and our data, we have additionally tagged each missing hypothesis according to the following classification:

1. Linguistic subtleties (Labelled "Language"; Example: "ownership in the past is enough to justify the possessive in the present"; 9 occurrences in our sample)

2. Lexical meaning, sometimes specific to the context of the problem; (Labelled "Lexicon"; Example: "buying entails selling"; 15 occurrences in our sample)

3. World knowledge (Labelled "World"; Example: "Increased amounts of CO2 and other greenhouse gases cause Greenhouse effect."; 13 occurrences in our sample)

4. Other missing hypothesis, see below for further details.

3.1. Analysis of reported missing hypotheses and incorrect labelling in RTE3

Pragmatic Strengthening A general theme explaining wrong conclusions is a pragmatic strengthening of the premises. (This occurs in many problems, at least 50, 51, 176, 278, 454, 643, 722, 740 in our sample). Indeed, in our experts judgements, we have found cases where problems were marked as "yes", and justified with a hypothesis which is, taken in isolation, false, but which could make sense in the context of the premises.
Thus, the problem is not "is there entailment", but rather becomes "does the questioner intend entailment". This semantic shift can be problematic in the context of testing a precise entailment system.
Example: (Problem 454)
P: On Aug. 6, 1945, an atomic bomb was exploded on Hiroshima with an estimated equivalent explosive force of 12,500 tons of TNT, followed three days later by a second, more powerful, bomb on Nagasaki.
H: In 1945, an atomic bomb was dropped on Hiroshima.
(Bombs can explode without being dropped.)
To conclude entailment, one must at the minimum assume that exploding a bomb implies dropping a bomb (according to our expert panel) — but this is false in general.

Mistaking claims for truth The largest specific source of incorrect labelling, found in 12 problems in our sample (750, 754, 756, 757, 51, 66, 178, 225, 294, 588, 643, 659) out of 42 errors, is mistaking claims for truth, as in the following example.
Example: (Problem 294)
P: Mental health problems in children and adolescents are on the rise, the British Medical Association has warned, and services are ill-equipped to cope.
H: Mental health problems increase in the young.
As it should be obvious with a instant's thought, the above should entail only if the word of the British Medical Association can be taken for fact. While it may be safe to behave as such in many situations in the real world, one cannot do so when reasoning precisely.

Mistaking intentions and facts Another source of common mistakes is the confusion of intentions and facts, found in 8 problems in our sample (33, 148, 191, 396, 420, 59, 121, 166).
Example: (Problem 191)
P: Though Wilkins and his family settled quickly in Italy, it wasn't a successful era for Milan, and Wilkins was allowed to leave in 1987 to join French outfit Paris Saint-Germain.
H: Wilkins departed Milan in 1987.
(Even though Wilkins was allowed to leave, it does not mean he actually left.)

Mistaking the past for the present Rather obviously, events described in the past cannot be taken to hold currently. Yet this error is still found in 6 problems in our sample (255, 230, 118, 308, 454, 175).
Example: (Problem 308)
P: On 29 June the Dutch right-wing coalition government collapsed. It was made up of the Christian-democrats (CDA) led by Prime Minister Jan Peter Balkenende, the right wing liberal party (VVD) and the so-called 'left-liberal' D66.
H: Three parties form a Dutch coalition government.
(The coalition may have collapsed at the time of solving the problem)

Incorrect application of monotonicity The final common source of errors that we identify is incorrect application of monotonicity reasoning. This error is a bit more subtle than the others — and thus, one might expect, easier to make and in turn more frequent — but it is found only in five occurrences in our sample (59, 202, 221, 231, 463). One explanation for its relatively low frequency is that RTE subjects paid special attention to it, perhaps because most RTE problems feature some kind of monotonicity reasoning, and thus it is present in the mind of the subjects at all time.
Example: (Problem 463)
P: Qin Shi Huang, personal name Zheng, was king of the Chinese State of Qin from 247 BCE to 221 BCE, and then the first emperor of a unified China from 221 BCE to 210 BCE, ruling under the name First Emperor.
H: Qin Shi Huang was the first China Emperor.
In this example, for entailment to hold, monotonicity reasoning dictates that "China" should to be at least as specific than "unified China" (in the context). Our panel of experts judge this not to be the case — following usual rules for common noun phrases.

3.2. Use of world-knowledge

We find that entailment problems which depend on any non-trivial amount of world knowledge are problematic from the point of view of training and testing systems for entailment. Indeed, in the presence of a large number of arbitrary facts, the conclusion can come solely from such knowledge, completely ignoring the premise. Such a situation occurs in a many cases in RTE. For example, in problem 191 shown above, it is public knowledge that *Wilkins departed Milan in 1987*. Likewise, it is common knowledge that a *in 1945, an atomic bomb was dropped on Hiroshima* in example 454.
The rules of logic tell us that if the conclusion holds, then the entailment holds as a whole. Yet, subjects often consider that the entailment does not hold *as such*. What seems to happen is a kind of *pragmatic weakening* of their knowledge. That is, in the context of judging entailment, subjects will treat the problems as fictive situations, suspend disbelief, and reject the use of some — but not all! — world knowledge in direct relation with the problem. Effectively, the subjects are second-guessing the tests: rather than answering spontaneously they act as they imagine the test is expecting of them.
If one wishes to define entailment by example, this is a problematic situation, because where to draw the line between what should be (temporarily) ignored and what should be still accepted is a purely subjective matter. We find that our approach (asking subjects to list the knowledge they use) resolves the problem satisfactorily: when a subject makes a judgement about entailment, they will

list the knowledge that they used (even if this "knowledge" turns out to be actually wrong in the real world). They are relieved from the burden of guessing which facts should be ignored, and there is no longer any point (or even opportunity) of second-guessing the test.
(Zaenen et al., 2005) analyse the same issue from a different perspective: they distinguish entailment from inference. According to them, entailment depends only on the text itself, not on the objects that it refers to. Because RTE refers heavily to real-world objects, they conclude that it can only be used to circumscribe inference. Unfortunately, our analysis shows that it is hard to distinguish between purely lexical knowledge and world knowledge. Following (Zaenen et al., 2005), "murdering" *entails* "killing" — but not the other way around. However this kind of knowledge may not be accessible, say, to someone with little legal background.
Thus we prefer not to make the same distinction, and simply let subjects inform us about any missing hypothesis, it being of lexical or of other nature.

3.3. Applicability to other corpora

SICK SICK (Marelli et al., 2014) was constructed to test compositional distributional semantics (DS) models (DCSM). It contains examples pertaining to logical inference (negation, conjunction, disjunction, apposition, relative clauses, etc.). It focuses on distributional semantic approaches and, thus, it normalises several cases that DS is not expected to account for. The dataset consists of approximately 10k test pairs annotated for (tripartite) inference and relatedness. According to (Kalouli et al., 2017a; Kalouli et al., 2017b) the SICK dataset is known to be problematic in its annotation. As evidence, they report that a number of contradictions in SICK are asymmetric, while according to usual logical rules, contradiction is symmetric: if a sentence A contradicts B, then sentence B should also contradict A. However, this is not always the case in SICK. **Example:** (Problem SICK221)
P: The blond girl is dancing behind the sound equipment.
H: The blond girl is dancing in front of the sound equipment.
For the above example, one finds the annotation A_contradicts_B, but also B_neutral_A in the dataset, thereby violating symmetricity of contradiction.
Like us (Kalouli et al., 2017a) aim to improve the precision of NLI dataset, and propose several ways to improve SICK towards this goal. We believe that our approach is a reasonable way supplement to those.

SNLI, MultiNLI SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) are probably the most standard datasets used today to train and test NLI systems. Both datasets have been constructed using crowd-sourcing (Amazon Mechanical Turk). More specifically, SNLI is constructed as follows: subjects are given a caption of a picture and then are asked to provide: a) an alternate true caption, b) an alternate possibly true caption, and c) an alternate false caption. In a subsequent stage, subjects annotate the above sentence pairs as entailment, contradiction or neutral. The dataset resulting out of this process contains 570k inference pairs. This makes SNLI orders of magnitude bigger than datasets like FraCaS or RTE. MultiNLI is modelled on SNLI but, contrary to SNLI, uses data from a variety of genres. More specifically, ten different genres are included from both written and spoken English. The dataset consists of 433k sentence pairs. Recent work by (Camburu et al., 2018) extends SNLI with an additional layer of human-annotated natural language explanations of the NLI relations. The constructed dataset is shown to be less prone to the annotation artefact problem reported for SNLI in (Gururangan et al., 2018). The general idea in (Camburu et al., 2018) is close in spirit to what we have been doing here. However, while the approach of (Camburu et al., 2018), can recover NL explanations it does not guarantee to recover hidden premises, if these exist. One could envisage a further extension that will contain the hidden premise information as part of the NL explanation.
Additionally, (Camburu et al., 2018) do not propose to overrule the annotations.
We believe that SNLI is problematic as a platform for precise textual entailment. Indeed, the use of captions means that there is an underlying —but unknown to annotators— image which constitutes ground truth for the sentences. Thus, rather than judging entailment on the basis of the text alone, we propose that subjects attempt to reconstitute this ground truth from the premise and check the compatibility of the conclusion with that reconstruction, *as a caption*.

(4)

P. A black race car starts up in front of a crowd of people.

H. A man is driving down a lonely road.

Label Contradiction

In the above example, the (P.) and (H.) are not contradictory using usual logical rules, since they can refer to two different events. But (H.) is a contradictory caption for a picture that uses (P.) as a caption.
In sum, the effect of second-guessing, that we observe with RTE (yielding pragmatic weakening/strengthening) is even more pervasive for SNLI, making it a particularly weak benchmark for precise reasoning. We believe that asking subjects to list their assumptions would greatly enhance the precision of annotations in the construction of an SNLI-like dataset.

4. Conclusion and Future work

We find that folklore is vindicated: RTE is not suitable as such to test a precise NLI system, because, for entailment to hold as tagged in RTE, much world-knowledge is required and many missing hypotheses are omitted.
Fortunately, along the way, we have made several discoveries which we believe useful.
First, by using a crowd of experts to repair the missing hypotheses, we have effectively constructed a dataset of 150 precise entailment problems, based on text found in real-world corpora. Even though the dataset is on the small size, it is, to the best of our knowledge, the first of this kind.
Second, we have uncovered several possible avenues to improve the precision of RTE-like problems sets, at construction time.

- We found that many of the imprecisions in the classifications were due to just a handful of reasons. In the future, when constructing RTE-like dataset, it will be possible to instruct annotators directly to pay specific attention to these and thereby construct a dataset which is suitable for precise reasoning.

- Should precise reasoning be employed, the large majority of problems in RTE do not correspond to entailment — remember that in this paper we only considered "yes" cases in RTE. We believe that when annotating, subjects tend to balance each category, thereby relaxing the precision of entailment.[7] Thus, when constructing datasets for entailment, one should take into account that the balance of yes/no in the set will influence the precision of reasoning.

- One should strive to avoid testing the competence of subjects in world-knowledge rather than textual entailment, or, worse test the competence of subjects in guessing *how much* world-knowledge they should assume to recreate gold annotations. As discussed above, offering to list used hypotheses is one flexible way to address the issue. Another approach would be rename all named-entities (proper names) with some generic name from the same class. We believe this to be within reach of state-of-the-art NLP tools.

Regardless, an issue with the method that we used to construct our new dataset of entailment problems, essentially by discovering missing hypotheses, is that it is difficult to scale: it demands several minutes of precious expert work per constructed problem. Our plan is to investigate the possibility to gamify the process, so that lots of people can participate in the construction of precise entailment problems, as a form of entertainment. We leave any detail to further work, but this would be an asymmetric game, where one one player tries construct watertight entailment problems, and the opposing player would try and refute such entailment problems (say, by giving counter-examples). Identifying the possible kind of mistakes, as we have done here, will help prompting the players about things to look for — in either of the possible roles.

We find it striking that many mistakes committed in RTE are falling for classical fallacies (appealing to authority accounts for more than a quarter of errors). Thus, we believe that even if bare entailment annotation are already useful for the construction of NLI systems, the added hypotheses (or moves taken in our hypothetical game) should be of interest to the linguistic community at large.

Acknowledgements

We are grateful to the crowd of experts that performed the hard work of precisely annotating problems. Most of them chose to remain anonymous.

The others were, in alphabetical order: Rasmus Blank, Robin Cooper, Matthew Gotham, Julian Hough and Aarne Talman. The authors are supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

[7] It is known that the presentation of scales in a survey can influence the result (Cabooter et al., 2016) and we suspect that similar effects occur with data balancing.

5. References

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Cabooter, E., Weijters, B., Geuens, M., and Vermeirc, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research*, 69:2574–2584.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. (2018). e-snli: natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2010). Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Kalouli, A.-L., de Paiva, V., and Real, L. (2017a). Correcting contradictions. In *Proceedings of the Computing Natural Language Inference Workshop*.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017b). Textual inference: getting logic from humans. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Sammons, M., Vydiswaran, V., and Roth, D. (2012). Recognizing textual entailment. *Multilingual Natural Language Applications: From Theory to Practice*, pages 209–258.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Zaenen, A., Karttunen, L., and Crouch, R. (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. Association for Computational Linguistics.