System Explanations: A Cautionary Tale

Ellen M. Voorhees
National Institute of Standards
and Technology
Gaithersburg, MD 20899-8940,
USA
ellen.voorhees@nist.gov

Abstract
There are increasing calls for systems that are able to explain themselves to their end users to increase transparency and help engender trust. But, what should such explanations contain, and how should that information be presented? A pilot study of justifications produced by textual entailment systems serves as a cautionary tale that there are no general answers to these questions. Six different judges each acting in the role of surrogate end-user independently rated how comprehensible justifications of entailment decisions were on a five-point scale. Interrater agreement was low, with an intra-class correlation of about 0.4. More than half of the explanations received both one rating of 'Very Poor' or 'Poor' and one rating of 'Good' or 'Very Good'; and in 32 cases, the same explanation received all five possible ratings from 'Very Poor' through 'Very Good'.

Author Keywords
disagreement; explanation content; human ratings; textual entailment.

CCS Concepts
•Computing methodologies → Lexical semantics;
•Human-centered computing → Laboratory experiments;

There is a long history of interest in automatic systems creating a human-understandable explanation of its processing [1]. Sometimes the target of the explanations are system builders who examine the explanations as a way of validating and improving the system. More often the explanations are targeted at the end user of the system with the hope of engendering trust in it.

The desire to create systems that can explain their decisions to the target user base, intelligence analysts, was the motivation for an optional 'Justification' task of the 2007 Recognizing Textual Entailment challenge (RTE3)[1]. Textual entailment systems decide whether the content of one text fragment can be inferred from the content of another text fragment. The purpose of the RTE3 Justification task was to gather data on how effectively systems could explain their decisions to human judges acting as surrogate users of an entailment system. To this end, a broad sample of justification approaches was solicited from the RTE community, and a subset of the justifications were subsequently rated for comprehensibility by each of six judges[2].

An earlier paper [10] summarizes some key findings from the RTE3 Justification task, including the low level of agreement among the judges as to which justifications are effective. That paper focuses largely on how different ratings affects the ability to measure systems' explanatory power, while this paper provides a more detailed picture of the ratings themselves. The data are unequivocal: the beauty of an explanation is clearly in the eye of the beholder. Effective explanations need to be tailored not only for different categories of users [8], but for the specific person consuming the explanation as well.

[1] https://nlp.stanford.edu/RTE3-pilot/

[2] Each justification was assigned two ratings by each judge, one for 'Understandability' and the other for 'Correctness'. This paper focuses only on the Understandability ratings in the interest of space since both rating types had the same kinds of variability.

RTE

In the RTE entailment task, systems report whether a *hypothesis* is entailed by a *text* for a given pair of text passages [3]. Three-way entailment decisions involve deciding whether the hypothesis is entailed by the text (that is, the hypothesis must be true if the text is true), is contradicted by the text, or is neither entailed nor contradicted by the text. The RTE3 data set contains 800 entailment test pairs and a gold-standard answer key created for the test set through a process of adjudication [4]. The hypothesis is entailed for about half of the test pairs, is contradicted for about a tenth of the pairs, and has neither relationship for the remainder.

For the Justification challenge, a system returned a rationale for its decision for each of the 800 entailment pairs. A justification was defined as a collection of ASCII strings with no minimum or maximum size. This deliberate underspecification gave participants as much flexibility to define their own justification approach as possible, though it did not support interactive techniques such as drill-downs. The target audience of the explanations was users of an information system that employed entailment decisions—specifically, intelligence analysts. The task was explicitly *not* explaining the inner workings of the system to system builders.

The challenge received six submissions (called runs in the remainder of the paper) from five teams. Two example justifications returned for one entailment pair by different runs are shown in Figure 1. The top of the figure gives the text and hypothesis (where the correct answer is that the hypothesis is entailed). The response from a system is then

Text: Claude Chabrol (born June 24, 1930) is a French movie director and has become well-known in the 40 years since his first film, Le Beau Serge, for his chilling tales of murder, including Le Boucher.
Hypothesis: Le Boucher was made by a French movie director.

Run A: (entailment decision: entails)
It can also be assumed that a french movie director refers to director. From the fact that director exists, we can infer that direct exists. We can further infer that direct is Le Boucher was made by a french movie director. We conclude that Le Boucher was made by a french movie director.
Ratings: VPoor, VGood, Fair, Poor, Poor, Fair

Run C: (entailment decision: neither)
I don't know - the H sentence doesn't seem to follow from T, but doesn't seem to contradict it either.
Ratings: Good, VGood, VPoor, VGood, VGood, VGood

Figure 1: Example justifications for one entailment pair.

displayed, where a response is both the entailment decision and the justification. The bottom line of each run's entry shows the ratings assigned to that justification, which are discussed below.

The runs were submitted to the National Institute of Standards and Technology (NIST), one of the organizers of the challenge. NIST organizes a variety of different community evaluations of natural language processing (NLP) technology and employs experienced human judges to provide language annotations. These judges all have some sort of information analysis background, and most of them are retired intelligence analysts [11]. Six judges independently rated the comprehensibility of the justifications for a subset of 100 entailment pairs, with all six judges rating the same 100 pairs for each of the six runs, and using a five-point rating scale of 'Very Poor', 'Poor', 'Fair', 'Good', and 'Very Good'.

NIST selected the subset of 100 entailment pairs to be rated after looking at the runs to maximize the informativeness of the rated set. Each run had a very small number of templates used to generate its justifications, so the 100 pairs included a lot of redundancy with respect to the structure of the justifications.

Because we wanted the judges' individual reactions to the explanations, NIST provided minimal training guidance to the judges. They were told the motivation for doing the task was that the community was interested in how to develop systems that can explain themselves to their end users to thereby increase users' confidence and trust in the system. They were also told that the use of system details or technical jargon were acceptable reasons to penalize comprehension scores, and that the justification of a wrong decision could be comprehensible, but system assertions that were wildly wrong could hurt comprehension if the assertion prevented the judge from seeing any relationship between the justification and the decision.

Judges used a system that displayed the text, hypothesis, and gold standard answer for whether a pair was entailed, and then displayed each run's justification in turn. The order in which the runs' justifications were displayed was randomly selected for each pair; for a given pair, each judge saw the same order.

Recent work on evaluating explanations in AI systems suggests that the measuring process itself may disrupt a user's true behavior with a system and lead to invalid conclusions if the evaluation is based on proxy tasks and subjective measures [2]. The RTE Justifications task was not a proxy task, however. The envisioned work flow for the target users, intelligence analysts, included the ability for the analyst to directly request an explanation of the system's reasoning at any time. While a judgment of comprehensibility is indeed subjective, comprehensibility is (one of) the qualities of interest for the target users who will certainly be subjective when they use the system. The main aim of the study was to see if there is general agreement as to what kind of explanations are understandable given the inherent subjectiveness.

Interrater Agreement

Figure 1 shows the ratings assigned to the justifications for the example entailment pair. The ratings from the six judges are shown in the same order for each of the justifications. The judges clearly disagree on the rating to be assigned for the example entailment pair. Both of the justifications in the figure received both Very Poor and Very Good ratings, for example. While this is just a single example entailment pair, it is not an unusual case.

Pattern	Freq
..x.x	96
.xxxx	77
.xxx.	76
..xxx	59
xxxx.	55
...xx	40
xxxxx	32
xx.x.	25
.x.xx	22
xx...	17
.x..x	14
x.xxx	14
xxx..	13
x.xx.	9
.xx.x	8
..xx.	6
.x.x.	6
xx.xx	6
x...x	5
x..xx	5
....x	4
xxx.x	4
.xx..	2
x....	2
x.x..	2
x.x.x	1

Table 1: Number out of 600 justifications that received a given pattern of ratings. (See text for pattern meaning.)

Each of the 100 entailment pairs had justifications from six different runs, and each justification was independently rated by each of six judges. So, each judge had 600 ratings to assign, and each justification was expected to receive six ratings. A few justifications were mistakenly skipped, however, so 589/600 justifications had ratings from all six judges. Since the rating scheme is an ordinal scale, we have more than two judges per justification, and we have a fully-crossed design if we consider just the 589 cases with all six ratings, we can compute the intra-class correlation (ICC) on the set of 589 justifications as a measure of the inter-rater agreement [5]. In particular, we use ICC model 2 for single random raters as computed using the ICC function of the *psych* R package [9]. The ICC2 score ranges from 1.0 to -1.0 where the extreme points represent perfect agreement (1.0) or disagreement (-1.0), and 0.0 represents random agreement. Larger differences in ratings (for example, POOR vs. VGOOD) affect the agreement more than smaller differences (GOOD vs. VGOOD). The ICC2 score for the justification ratings data is 0.42, with a lower bound of 0.27 and an upper bound of 0.54. While this level of agreement is statistically different from random agreement ($p < 0.00001$), it is nonetheless generally considered poor agreement [6].

Table 1 provides a more detailed view of the conflict in ratings assigned to justifications. The table gives the number of justifications that were assigned a certain pattern of ratings. Rating patterns are five-bit strings where each binary digit says whether at least one rating of that type was assigned to the justification ('x') or none of that type was assigned ('.'). The leftmost position represents the Very Poor rating and the rightmost the Very Good rating. So, pattern 'x..xx' indicates a justification that received at least one rating for Very Poor, for Good, and for Very Good, and no ratings of Poor or Fair. Over all 600 justifications, 359 have at least one Very Poor or Poor rating as well as at least one Good or Very Good rating; 32 justifications had every possible rating assigned to it.

These explanations are almost 15 years old now, and significant research on explainability has occurred in the intervening years [7]. So, automatically produced explanations are very likely to be better if created now. However, the focus of this investigation is not on the quality of the justifications per se, but instead on how different the ratings assigned to the justifications are for different judges. Judges had very different ideas of what constitutes a good explanation. Some judges preferred explanations in precise logical notation that other judges hated; anecdotal evidence suggests that this varied with the amount of mathematical training in a judge's background. Despite the divide on logical notation, all judges valued conciseness and wanted the explanations to focus on the specifics of the current pair. Different judges had different tolerances for stilted or ungrammatical prose.

Whatever the representation, meeting a user's expectation with regard to the correct level or granularity of an explanation is key but difficult to do automatically. System justifications often gave too many system-internal details that were not meaningful to the judges. Linguistic jargon including 'polarity', 'adjunct', 'hyponym' and 'scope contraction' that were used frequently were more detrimental than helpful for the judges. Generic phrases such as "there is a relationship between" and "there is a match" (part of some systems' templates) were penalized as clutter with no expository value. System internals such as scores from different components were unhelpful because the judges didn't know what the components were (and didn't want to know) and they had no way to calibrate the scores.

By design, the judges received minimal training on how to rate a justification. Interrater agreement levels would certainly rise with more training. But of course, increasing interrater agreement through training does not solve the underlying problem that eventual end users still have different opinions as to what constitutes an effective explanation.

Conclusion

The need for systems that can explain themselves to their human users or partners is great. While in-roads have been made for some kinds of systems where the problem definition offers good candidates for the elements of an explanation, the decisions of other autonomous systems are equally as important to explain but provide greater challenges with respect to the form and content of effective explanations.

The data collected in this study demonstrate that there is a wide variance in what different users find acceptable in an explanation of textual entailment decisions. The variance is large despite the fact that both the problem domain, entailment of short text fragments, and the target of the explanation, intelligence analysts, were well-defined and fairly narrow.

Acknowledgement

Thanks to the anonymous referee who encouraged using ICC rather than Fleiss' kappa in reporting agreement rates.

REFERENCES

[1] Or Biran and Coutenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *Proceedings of IJCAI-17 Workshop on Explainable AI (XAI)*. 8–13.

[2] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM. DOI: http://dx.doi.org/10.1145/3377325.3377498

[3] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment.* Lecture Notes in Computer Science, Vol. 3944. Springer, 177–190.

[4] 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*. Association for Computational Linguistics, 1–9.

[5] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology* 8, 1 (2012), 23–34. DOI: http://dx.doi.org/10.20982/tqmp.08.1.p023

[6] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163. DOI: http://dx.doi.org/10.1016/j.jcm.2016.02.012

[7] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. arXiv:1706.07269[cs.AI]. (June 2017).

[8] P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, David A. Broniatowski, and Mark A. Przybocki. 2020. Four Principles of Explainable Artificial Intelligence. Draft NISTIR 8312. (August 2020). https://doi.org/10.6028/NIST.IR.8312-draft

[9] William Revelle. 2017. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych R package version 1.7.8.

[10] Ellen M. Voorhees. 2008a. Contradictions and Justifications: Extensions to the Textual Entailment Task. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, 63–71.

[11] Ellen M. Voorhees. 2008b. Overview of TREC 2007. In *Proceedings of The Sixteenth Text REtrieval Conference (TREC 2007)*. 1–16. NIST Special Publication SP 500-274. http://trec.nist.gov/pubs/trec16/t16_proceedings.html.