Textual Entailment with Natural Language Explanations: The Italian e-RTE-3 Dataset

Andrea Zaninello[1,2], Sofia Brenna[1,2] and Bernardo Magnini[1]

[1]Fondazione Bruno Kessler, Trento (Italy)
[2]Free University of Bozen-Bolzano (Italy)

Abstract

We introduce the 'e-RTE-3-it' dataset, an enriched version of the Italian RTE-3 dataset, where each text-hypothesis pair, in addition to the 'entailment', 'contradiction', or 'neutrality' label, has been enriched with an explanation for the label itself. Moreover, the dataset includes the level of confidence with which the annotators could write the explanation, and in cases where the annotators did not agree with the original label, an alternative label, along with an explanation for the new label. This offers the opportunity to analyse cases of uncertainty in annotation and delve into different perspectives on language understanding.

Keywords
Explanations, recognizing textual entailment, lexical resources

1. Introduction

Recently, Large Language Models (LLMs) like T5 [1], GPT-3.5/4 [2], LLama-2 [3], It5 [4], and Camoscio [5] have demonstrated impressive performance across various natural language processing tasks. Despite their success, these LLMs also face limitations and risks, such as lack of factuality [6], hallucinations [7], and poor transparency [8]. As a result, there is a growing demand for "inherent explainability," which refers to the ability of models to provide human-like, natural language explanations for their predictions. Many studies have thus focused on natural language explanations, and numerous datasets have been created for this purpose, primarily in English [9]. However, there is a notable gap for non-English languages, including Italian.

To fill this void, this paper introduces the 'e-RTE-3-it' dataset, the first Italian dataset for natural language inference enriched with free-form, human-written explanations for the relationship between two sentences. Additionally, the dataset includes alternative labels and confidence scores from annotators to account for the variability in human judgments. This aspect of the annotation scheme enhances the 'e-RTE-3-it' dataset, making it a valuable resource for exploring subjectivity and variability in language understanding[1].

2. Background and Related Work

Recognising Textual Entailment (RTE) emerged as a task in 2005 [10], aiming to determine if two sentences have an entailment, contradiction, or neutrality relationship. An Italian version of the RTE-3 dataset was later developed to explore language comprehension and textual entailment [11].

The significance of free-form explanations in enhancing understanding and interpretability has led to the creation of various datasets. For example, the CODAH dataset presents commonsense reasoning problems with adversarially constructed explanations [12]. Similarly, the COPA-SSE dataset offers crowd-sourced explanations for commonsense reasoning tasks [13]. The COS-E dataset couples commonsense reasoning problems with explanations [14], providing valuable insights into human approaches to these tasks.

The e-SNLI dataset is a relevant resource, as an enriched version of the Stanford Natural Language Inference (SNLI) corpus, containing human-written explanations for entailment decisions [15]. However, this dataset, while valuable for tasks requiring extensive training data, is not manually curated and focuses exclusively on the English language.

3. Methodology

3.1. Annotation layers

For each text-hypothesis pair in the original Italian RTE dataset, annotators were asked to provide an explanation (<e>) for the given label and rate their confidence in providing that explanation on a 5-point Likert scale. We also encouraged diversity in perspectives by allowing

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

azaninello@fbk.eu (A. Zaninello); sbrenna@fbk.eu (S. Brenna); magnini@fbk.eu (B. Magnini)
0000-0001-9998-1942 (A. Zaninello); 0009-0001-3748-1448 (S. Brenna); 0000-0002-0740-5778 (B. Magnini)

[1] We make the e-RTE-3-it dataset available at the following link: https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/e-rte-3-ita

annotators to disagree with the original label.

In such instances, they provided an explanation for the original label as well as an alternative label (`<a>`) and a corresponding explanation for the new label, along with the level of confidence for the second explanation.

3.2. Data collection and guidelines

We recruited 40 annotators among students (from undergraduate to PhD level) at the University of Bologna. Annotators were native Italian speakers fluent in at least one other language, and each took at least one linguistics course, ensuring meta-linguistic proficiency as well as broader cultural understanding. Annotators were provided with 50 text-hypothesis pairs each labeled with an entailment relationship.

They were asked to write one free-form, natural language explanation in Italian explaining why the two sentences stood in that particular entailment, contradiction, or neutrality relationship. To ensure language variety as well as uniformity across labelers, the following guidelines were given:

- please write an explanation in the form of one or two self-contained sentences for each `<pair, label>`;
- you can refer back to, quote, or paraphrase chunks of both the text and hypothesis;
- use case marking and punctuation consistently with the original sentences;
- you can use metalanguage to refer back to the original sentences with phrases such as “in the text, it is stated that...”, “the hypothesis does not mention...”, etc.;
- please provide your level of confidence (i.e. how sure you are about the reasons provided in your explanation) on a scale from 1 to 5;
- if you (even partially) disagree with the given label, provide a new label for the pair, an explanation for the new label, and your level of confidence in the new explanation.

3.3. Post-editing of the explanations

Finally, two different linguistics experts post edited the explanations to proofread them, validate them and address any discrepancies, as well as ensure uniformity and coherence in spelling.

In some cases, the experts discarded some explanations because of logical errors, in cases when explanations only paraphrased the input texts or included information not originally conveyed by the input texts. For example:

```
<pair id="224" entailment="UNKNOWN" task="IR" length=
     "short">
<t>Basandosi su uno studio mondiale [...] gli
     epidemiologi [...] dimostrano che il fumo e' la
     causa principale degli incendi e delle morti
     per incendi nel mondo.</t>
<h>Gli incendi domestici sono una causa importante
     delle morti da incendio.</h>
<e confidence="4">Il fumo e' la causa principale
     degli incendi domestici.</e></pair>
```

In this case, the explanation was stating something that could not be inferred from the input sentences, and was re-written by another annotator in the following way.

```
<e confidence="3">Il fumo e' la causa principale
     degli incendi e delle morti per incendio, ma
     non e' specificato se un'altra causa importante
      di morti da incendio siano proprio gli incendi
      domestici.</e>
```

3.4. Original dataset correction

While editing the explanations, the experts also detected and corrected some errors in the original dataset. In few cases, these included missing information that made it impossible to infer the right label for t and h. For example, consider the following text-hypothesis pair from the test set (id: 52). `Text: Oscar Chisini (nato il 4 marzo 1889 a Bergamo, morto il 10 aprile **1967** a Milano) fu un matematico italiano. Lui introdusse la media Chisini nel 1929.`; `hypothesis: Oscar Chisini morì nel 1967.`; `entailment: "YES"`. The information within stars ** (the year of death) was missing from the Italian dataset but was present in the original English RTE-3 dataset. This information was essential to infer the entailment relationship, and was re-introduced by checking the original English version.

Moreover, the Italian RTE-3 dataset, as reported in the description[2] changed the original label (from ”YES”: entailment, to ”NO”: contradiction) in 15 pairs, creating a mismatch with the English dataset. To ensure comparability, we decided to restore the original label provided by the English dataset, as our annotators were still able to express an alternative label in case they did not agree with it. They did so only in the dev set, where they provided an alternative label ”NO” in pairs 51, 490, 549, and a label ”UNKNOWN” (neutrality) in pair 604. In all other cases, they agreed with the original label.

For these reasons, the e-RTE-3-it dataset can also be regarded as an emended, manually curated version of the original RTE-3-it dataset.

[2] The original RTE 3 Italian dataset description can be found at https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/rte-3-ita

	Total/Average	Entailment	Contradiction	Neutrality
Original label	1600	800	150	650
New labels in <a>	147	48	62	37
New labels from entailment	41	-	10	31
New labels from contradiction	6	0	-	6
New labels from neutrality	100	48	52	-
Confidence (mean) in <e>	4.00	4.17	4.03	3.81
Confidence (mean) in <e> w/o <a>	4.01	4.24	4.05	3.91
Confidence (mean) in <e> with <a>	3.14	2.78	3.33	3.28
Confidence (mean) in <a>	3.48	3.35	3.47	3.65

Table 1
Statistics for the enriched 'e-RTE-3-it' dataset. Number of labels are in absolute values, confidence values are averaged over pairs on a scale from 1 to 5. ”New labels from” indicate times when the original label was changed, and to which label.

4. Dataset Description

The final dataset comprises 1600 text-hypothesis pairs, divided into two dev/test splits of 800 pairs each. Each pair inherits the original dataset's attributes indicating the pair's ID, the entailment relation (yes, no, unknown), the original task for which the pair was collected, and whether the text is long or short. Each pair is complemented with one explanation and a confidence score. It also provides 147 alternative labels with their respective explanation and confidence score. In the following, we provide a snippet of the test set.

```
<?xml version="1.0" encoding="UTF-8"?>
<pair id="201" entailment="YES" task="IR" length="
      short">
<t>Berlino ha un nuovo punto di riferimento. Sopra le
      gru che ancora dominano l'orizzonte della
      nuova capitale dell'Europa adesso c'e' una
      cancelleria, dove vivra' il capo del governo
      Gerhard Schroeder e il governo tedesco terra' i
      suoi incontri regolari.</t>
<h>Nuovi edifici sono stati eretti a Berlino.</h>
<e confidence="4">La frase "sopra le gru... adesso c'
      e' una cancelleria" e' da intendersi in modo
      figurato, e indica che e' stato costruito un
      nuovo edificio dove ha sede la cancelleria.</e>
<a confidence="5" new_label="UNKNOWN">Il fatto che
      ora sopra le gru c'e' una cancelleria, non
      implica che nuovi edifici sono stati eretti a
      Berlino.</a>
</pair>
```

5. Data analysis

Annotators' Agreement with Original Labels. Table 1 reports a detailed description of the dataset. The original labels in the dataset exhibit a distribution of 50% 'entailment', 10% 'contradiction', and 40% 'neutrality'. A characteristic of our dataset is the allowance for annotators to disagree with the original labels and propose an alternative one. If we consider disagreements from the original label (147 pairs), we observe an increase of the contradiction relationship to 13% and a decrease to 37% of the neutrality label. As an example, consider the following 't-h' pair:
Text: ”Finora non ci sono segnalazioni di qualche parente che abbia reclamato i corpi dei quattro uomini delle forze armate che sono presumibilmente morti quando l'aereo si schiantò.”
Hypothesis: ”Quattro uomini delle forze armate morirono in uno schianto aereo.”
The original label in the Italian RTE-3 dataset is 'YES'. However, an annotator disagrees and assigns the alternative label 'NO', explaining: ”Affermando che quattro uomini delle forze armate sono presumibilmente morti quando l'aereo si schiantò, si manifesta una mancata certezza totale dell'episodio.” The annotator rated their confidence in this explanation as 4.

We observed that among the cases where annotators disagreed with the original label, the ”neutrality” label 'UNKNOWN' was most frequently revised to ”contradiction” ('NO', 52) and to ”entailment” ('YES', 48). Upon examining the explanations provided for these revised labels, a common theme emerged: they often stated that the interpretation of 'h' needed to assign a neutrality label was too narrow and did not match with commonsense reasoning and inferences often made in discourse.

For example, in a case when an annotator changed the label from neutrality to entailment, 't' and 'h' stated that
Text: *[...] Michael Howard non riuscì a scalzare il Governo Laburista, sebbene i Conservatori avessero guadagnato 33 seggi”*
Hypothesis: *i Conservatori ottennero 33 seggi.* Here, the usual interpretation would be that they obtained *at least*, and not *exactly* 33 seats, explaining that “Guadagnare in questo caso è sinonimo di ottenere”.

In a case when the annotator changed the label from ”UNKNOWN” to ”NO” with confidence 4, 't' and 'h' stated
Text: *I proprietari di Phinda, l'Ente per la Conservazione con base in Sud Africa, non avrebbero potuto pagare per*

una pubblicità migliore per la loro filosofia di tutela della natura: un approccio alla tutela basato sulle persone, che sta lentamente guadagnando terreno in Africa poiché le riserve di caccia sono sempre piü minacciate dalle popolazioni locali affamate, povere e arrabbiate
Hypothesis: *L'Ente per la Conservazione con base in Sud Africa minaccia la popolazione locale.* The explanation given was that "L'Ente per la Conservazione con base in Sud Africa basa il suo rapporto di tutuela sulle persone, sulle popolazioni povere e affamate, quindi aiutandole non minacciandole".

Cases like these underline the subtleties involved in the inference process, and how tightly it connects to the interpretation of words in context, which may also be influenced by some level of subjectivity, an observation that paves the way for further investigation.

Lexical variety. We were also interested in the lexical variety of both the original sentences and the collected explanations. We noted that while the type/token ratio for each sentence is very high, indicating that few words are repeated in the same sentence, if we look at the lexical overlap between the sentences, we noted a high overlap between the alternative label explanation and the hypothesis, even compared to the text. This seems to indicate that the alternative explanations may rely on the information in the hypothesis more than the explanations for the original label, or and that they may be more 'metalinguistic' in nature, with a tendency to repeat the whole hypothesis literally.

mean length	**t**	**h**	**e**	**a**
length (tokens)	34	9	22	23
length (types)	30	9	19	20
types/tokens ratio	0.9	0.99	0.88	0.89

lexical overlapping	**t**	**h**	**e**	**a**
t	1.	0.11	0.16	0.24
h	0.56	1.	0.61	0.95
e	0.22	0.17	1.	0.38
a	0.21	0.17	0.23	1.

Table 2
Lexical variety in the dataset. The lexical overlapping indicates the word types present in the field in the column laso present in the field in the row, divided by the field in the row.

Confidence in Explanations. As can be seen in Table 2, the confidence scores assigned by annotators to explanations were generally high, with a mean score of 'e' of 4 on a 5-point Likert scale and the highest score being given to the entailment label. However, when annotators disagreed with the original label (and no alternative label was given) the mean confidence score for <e> decreased to 3.14 and the entailment label became the label with the lowest score (2.78). The fact that overall confidence in 'a' is lower than that of 'e' seems to indicate that while annotators felt confident in their judgments when they agreed with the label, cases involving label revision posed more challenges and perhaps involved a higher degree of uncertainty.

6. Conclusion and future work

The insights derived from the 'e-RTE-3-it' dataset pave the way for multifaceted research directions. The provided explanations can serve as a gold standard for training models to generate human-like explanations. Further, the alternative labels and explanations open avenues for investigating the subjectivity in language understanding. The rich layers of the dataset also allow for the study of correlation between the original and alternative labels, the confidence score, and the degree of disagreement among annotators. Future work includes utilizing the data to develop models capable of providing explanations for their entailment decisions and conducting a deeper analysis into the dynamics of subjectivity in the entailment task.

Acknowledgments

This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and by the ANTIDOTE project (CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4- 0002).

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[2] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov,

P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[4] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[5] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[6] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, Y. Matias, True: Re-evaluating factual consistency evaluation, arXiv preprint arXiv:2204.04991 (2022).

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys (2022).

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93:1–93:42. URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.

[9] S. Wiegreffe, A. Marasović, Teach me to explain: A review of datasets for explainable nlp, in: Proceedings of NeurIPS, 2021. URL: https://arxiv.org/abs/2102.12060.

[10] I. Dagan, O. Glickman, B. Magnini, The pascal recognising textual entailment challenge, in: Machine learning challenges workshop, Springer, 2005, pp. 177–190.

[11] B. Magnini, A. Lavelli, S. Magnolini, Comparing machine learning and deep learning approaches on NLP tasks for the Italian language, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2110–2119. URL: https://aclanthology.org/2020.lrec-1.259.

[12] M. Chen, M. D'Arcy, A. Liu, J. Fernandez, D. Downey, Codah: An adversarially authored question-answer dataset for common sense, arXiv preprint arXiv:1904.04365 (2019).

[13] A. Brassard, B. Heinzerling, P. Kavumba, K. Inui, COPA-SSE: semi-structured explanations for commonsense reasoning, CoRR abs/2201.06777 (2022). URL: https://arxiv.org/abs/2201.06777. arXiv:2201.06777.

[14] N. F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, arXiv preprint arXiv:1906.02361 (2019).

[15] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).