NLI Disagreement Taxnonmy Annotation Guidelines

Nan-Jiang Jiang
Department of Linguistics
The Ohio State University
jiang.1879@osu.edu

Natural language inference (NLI) is the task of whether the Hypothesis (H) is true/false/neither, assuming that the Premise (P) is true. In the following example, assuming P is true, H1 is definitely true (Entailment), H3 is definitely false (Contradiction), H2 may or may not be true (Neutral).

- P: The story remains to be told.
- H1: The story has not been told yet. → Entailment (if P is true, H is definitely true)
- H2: The story will be told tomorrow. → Neutral (if P is true, not sure if H is true)
- H3: The story has been told a million times. → Contradiction (if P is true, H is definitely false)

However, when multiple crowdsourced annotators see the same P/H pair, they sometimes disagree on which of the three labels applies. The goal of this annotation task is to identify linguistic phenomena that can potentially lead to disagreement in the NLI labels. There are 10 categories in total, falling into three high-level classes (inspired by Aroyo and Welty 2015).

- Uncertainty in sentence meaning
- Underspecification in task guidelines
- Annotator biases

Here we define the 10 categories. Certain items can have more than one category applied.

Uncertainty in Sentence Meaning

There are 5 categories under this high-level class. They involve linguistic phenomena with uncertainty in meaning, which then lead to uncertainty in the NLI labels.

1. Lexical In this category, disagreement stems from the uncertainty in meaning of a few lexical items. There are two kinds of examples that fall into this category.

For the first kind, P/H mostly mean the same thing except for a particular pair of lexical items, where there might be disagreement on whether they are hypernyms.

(1) **P:** It vibrated under his hand.
H: It hummed quietly in his hand.

It is uncertain whether "vibrating" entails "humming quietly". If one takes "vibrating" as entailing "humming quietly", the item is labeled as entailment. But one could also annotate the item as neutral.

(2) **P:** If the United States had used full conventional power.
H: If the United States had maximized their potential.

Here it is uncertain whether "used full conventional power" is the same as "maximized their potential": it is thus disagreeing over entailment vs. neutral. The notion of lexical item can be a bit more relaxed here, as this example concerns a pair of VPs.

Another kind of examples in this category have P/H where the decision of whether P entails H boils down to a single lexical item or a phrase in one of P or H, where the lexical item potentially has multiple meanings and the different meanings lead to different truths of the Hypothesis. The lexical item can also require certain arguments or contextual parameters that remain underspecified.

(3) **P:** Why isn't a lookalike good enough for them?'
H: The look alike is plenty good.

The argument "good for whom" is underspecified in H. If the argument is "them" as in "the look alike is plenty good (for them)", H is a contradiction because P is a wh-question, which presupposes the existence of an answer. In particular, *why Q* presupposes Q (Lawler, 1971). Therefore, P carries the presupposition that "a lookalike isn't good enough for them", which H would thus contradict. If the argument is the speaker of P, as in "the look alike is plenty good (for me)", then P can be viewed as either entailing H or being neutral with respect to H, since P (a negated question) carries the expectation that the speaker considers a look alike

good enough for themselves. Therefore, this example has three categories: Lexical, Presupposition, Probabilistic Enrichment

(4) **P:** On the days I go to my office, I wear a flannel shirt with no necktie if the weather is cool.
H: On hot days, I wear a flannel shirt to the office.

This example could be neutral or contradiction: contradiction when the "if" in "if the weather is cool" is interpreted as "if and only if", neutral otherwise (speaker could wear flannel shirts on hot and cool days).

2. Probabilistic Enrichment Items in this category tend to be disagreeing on whether the items get labeled neutral vs. contradiction or entailment. The hypothesis is an enriched inference of the premise in that it has some probability of being true, but is not definitely true given the premise. This also takes into account prior probabilities – how plausible the hypothesis is out of the blue.

(5) **P:** It's absurd but I can't help it. Sir James nodded again.
H: Sir James thinks it's absurd.

H is true if we assume Sir James nodding means to agree with the speaker that it's absurd. If we don't make such an assumption, H is neutral. This assumption may have some probability of being true, but is not definitely true.

Sometimes the enriched inference is the negation of the hypothesis. Therefore there is disagreement over whether the item is neutral vs. contradiction.

(6) **P:** The word itself, tapa, is translated as lid and derives from the old custom of offering a bite of food along with a drink, the food being served on a saucer sitting on top of the glass like a lid.
H: Tapas are large portions and are a very filling meal.

Here the second conjunct "Tapas are a very filling meal" involves enrichment: one might infer that tapas are NOT a very filling meal, because a tapa is "a bite of food", therefore contradiction. However the enriched inference is not definitely true of the premise, since the premise does not mention fillingness at all.

3. Presupposition The hypothesis involves a presupposition of the premise. See Beaver (1997) for a list of presupposition triggers.

Note: It is not enough that the premise contains presupposition triggers. For an item to be in this category, the truth of the hypothesis needs to depend on the presupposition coming from the premise.

(7) **P:** : Adrin's Third Lesson
H: Adrin had three lessons.

The possessive "Adrin's third lesson" is considered a presupposition trigger. It triggers the presupposition that Adrin's third lesson exists, and therefore Adrin had three lessons. There might be disagreement over whether the presupposition actually holds, therefore leading to disagreement over whether it is entailed.

4. Implicature The premise or hypothesis involving an implicature, and whether or not the implicature is canceled gives different NLI labels. Davis 2019 contains common forms of implicatures. See Rett (2020) for how to identify manner implicature. The Table 4 in Appendix in Jeretic et al. (2020) contains a list of triggers for scalar implicature.

(8) **P:** Today it is possible to buy cheap papyrus printed with gaudy Egyptian scenes in almost every souvenir shop in the country, but some of the most authentic are sold at The Pharaonic Village in Cairo where the papyrus is grown, processed, and handpainted on site.
H: The Pharaonic Village in Cairo is the only place where one can buy authentic papyrus.

(8) involves a scalar implicature, triggered by *some*: *some of the most authentic papyrus (are sold in The Pharaonic Village)* gives rise to the scalar implicature *but not all of the most authentic papyrus*, making the hypothesis false since it asserts that authentic papyrus are only sold in The Pharaonic Village. However the scalar implicature "some, but not all" can be cancelled. In that case either a neutral label or an entailment label are possible.

When an implicature is triggered by a specific lexical item, the category Lexical should also be used.

(9) **P:** Isn't a woman's body her most personal property?
H: Women's bodies belong to themselves, they should decide what to do with it.

The negated question in P implies that the speaker believes the prejacent "a woman's body is her most personal property" to be true. From this implication, we can infer probabilistically that "they should decide what to do with it". So this example would have both labels Probabilistic Enrichment and Implicature.

There is no consensus in the literature for

whether this implication from the negated question is an implicature or presupposition. Romero and Han (2004) considered it an implicature. van Rooij and Safarova (2003) considered it a "weak presupposition". Here we follow Romero and Han (2004) and say it's implicature.

Note: When in doubt, look up the literature for whether certain inferences are considered as implicatures, and include the link to the literature in the spreadsheet.

Similar to Presupposition, an item to be in the Implicature category, it is not enough that P/H contains the triggers. The judgment for whether H is entailed/a contradiction needs to depend on the implicature.

5. Imperfection Use this category when P/H do not express coherent propositions, have typos, are truncated sentences or stuttered speech transcripts, making it hard to understand what is being said. The disagreement results from annotators not sure about how to interpret the sentences and potentially choosing randomly or making up meaning of their own.

(10) **P:** profit rather
H: Our profit has not been good.

Underspecification in Task Guidelines

There are 3 categories under this high-level class "underspecification in task guidelines".

Although these categories also involve uncertainty in meaning, the resolution of such uncertainty should ideally be part of the NLI task guidelines.

6. Coreference Items in this category need a coreference assumption to be judged as entailment or contradiction. It could be:

- coreference of entities: whether P and H are talking about the same entity or different entities. If they are about different entities the label is often neutral.
- coreference of events: whether P and H describe the same event or events that could both happen but in different places/time.

The distinction between the two can be hard to make. So we only use the category "coreference" without distinguishing between the two types of coreference.

(11) **P:** The original wax models of the river gods are on display in the Civic Museum.
H: They have models made out of clay.

If we assume P and H are talking about the same models, H is a contradiction. If P and H are talking about different models, H is neutral, since the museum can have both wax models and clay models. Note that this would not fall under Lexical, because people would most likely consider wax and clay to be exclusive – if something is made of wax it is not made of clay. There is most likely no uncertainty in the meaning of the words "wax" and "clay".

(12) **P:** China's civil war sent distressing echoes to Hong Kong.
H: Japan fought a civil war.

This is an example of event coreference. If you assume that the "civil war" in the two sentences are referring to the same event, H is a contradiction. If you do not assume that they are the same event, it is neutral.

7. Temporal Reference Items in this category have different NLI labels when evaluating the truth of the hypothesis at different time points (temporal referents).

(13) **P:** However, co-requesters cannot approve additional co-requesters or restrict the timing of the release of the product after it is issued.
H: They cannot restrict timing of the release of the product.

Here it is not clear whether H is talking about restricting timing before or after it is issued. H is entailed if H only concerns restricting timing after it is issued. H is a neutral if H is taken to mean that they cannot restrict timing at all/at any point, since we don't know if they can restrict timing before it is issued.

8. Interrogative hypothesis The hypothesis is an interrogative clause. The NLI task has no definition of how to evaluate the truth of a question, therefore the task is undefined for such instances. In these cases, annotators tend to choose entailment for H that is a paraphrased question of P, and neutral/contradiction if P and H seem unrelated or opposite in polarity.

Note: The premise can be an interrogative clause, but that is not what is being targeted in this category. If the premise is interrogative and the hypothesis is declarative, this category does not apply.

(14) **P:** yeah but uh do you have small kids
H: Do you have any children?

Some annotators judge this as entailment, possibly because H is a paraphrase of P. But some annotators mark it as neutral, possibly because they do not know how to evaluate the truth of a question.

Annotator Biases

There are 2 categories in Annotator Biases. These categories involve a subset of annotators systematically behaving in certain ways.

9. Minimally added underspecified content When P and H mostly talk about the same thing, but H adds something that isn't talked about in P. The added content could be a clause or a modifier and is non-at-issue, easy to ignore.

(15) **P:** I had rejected it as absurd, nevertheless it persisted.
H: I rejected it as absurd but it persisted out of protest.

Here "out of protest" is not specified in the premise and the item should be neutral. However, many annotators tend to judge this as entailment, ignoring the underspecified part.

10. High Overlap Items in this category have P/H having mostly the same words, or read mostly the same when skimming through, but H is actually not inferred when reading closely.

(16) **P:** Yet, in the mouths of the white townsfolk of Salisbury, N.C., it sounds convincing.
H: White townsfolk in Salisbury, N.C. think it sounds convincing.

Many annotators judge this as entailment, as H has high lexical overlap with P. However H effectively expresses something different from P and should be neutral. H is saying "White townfolk think it sounds convincing", whereas P is saying "it sounds convincing" without mentioning who thinks it is convincing.

Note: The clues for non-entailment can be rather subtle, so pay close attention to things like tense and argument structure.

11. N/A If you feel that none of the above categories apply, put N/A.

Annotate with Multiple Labels

Each item may be annotated with several categories. For example, (17) has **Implicature, Temporal Reference, and Lexical** labels.

(17) **P:** The park was established in 1935 and was given Corbett's name after India became independent.
H: The park used to be named after Corbett.

The premise does not suggest that the park changed name, while the hypothesis does so with the implicature triggered by *used to*. Therefore, if we evaluate the truth of the hypothesis now, there can be disagreement between Neutral and Contradiction. If we evaluate the truth of the hypothesis in or before 1935, the hypothesis is entailed because the park was named after Corbett at some point. Also, given that the implicature is triggered by a specific lexical item (in contrast to non-conventional conversational implicatures), the category Lexical applies too.

References

David Ian Beaver. Chapter 17 - presupposition. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 939–1008. North-Holland, Amsterdam, 1997. ISBN 978-0-444-81714-3. doi: https://doi.org/10.1016/B978-044481714-3/50022-9. URL https://www.sciencedirect.com/science/article/pii/B9780444817143500229.

Wayne Davis. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.768. URL https://aclanthology.org/2020.acl-main.768.

John Lawler. any questions? *Papers from the 7th Regional Meeting, Chicago Linguistic Society (CLS 7)*, 1971.

Jessica Rett. Manner implicatures and how to spot them. *International Review of Pragmatics*, 12 (1):44–79, 2020.

Maribel Romero and Chung-hye Han. On negative yes/no questions. *Linguistics and Philosophy*, 27:609–658, 2004.

Robert van Rooij and M. Safarova. On polar questions. *Semantics and Linguistic Theory*, 13:292–309, 2003.