Annotation guidelines

Compilation of the Slovenian SI-NLI dataset for Natural Language Inference
Call for proposals CLARIN.SI 2022
Matej Klemen, Aleš Žagar, Jaka Čibej, Marko Robnik Šikonja
Version: 1.0
Last update: 2022-07-08
CC-BY-SA 4.0

1 Introduction

Natural language inference (NLI) is a task designed to assess the ability of advanced machine models to understand natural language.

On the basis of the given pair of texts (premise and hypothesis), the aim is to establish:

- whether the meaning of the second text **is entailed** in the first text (*entailment*);
- whether the second text **contradicts** the first in meaning (*contradiction*);
- the semantic relation between the two texts **cannot be inferred** (neutrality).

Each pair of texts is annotated with exactly one of these three categories.

In the following example, we give three examples of hypotheses for the same premise (one for each semantic relation).

PREMISE:
Investigations during the trial showed that events had unfolded very differently.

—
HYPOTHESIS 1:
Investigators found that the reports of the incident were not true.

The hypothesis follows from the premise (*entailment*), because both the premise and the hypothesis mention the falsity of the (originally described) events.
—
HYPOTHESIS 2:
The autopsy showed a very different picture.

The relationship between the premise and the hypothesis cannot be concluded (*neutrality*), since it is not necessarily the autopsy that led to the realisation that the reports were not true.
—
HYPOTHESIS 3:
The story proved to be true after the investigations were carried out.

A hypothesis is a *contradiction* in meaning to the premise - the premise claims that events unfolded in a significantly different way, while the hypothesis claims that the events (the "story") happened as they were presented before the investigation.

2 Objectives and Task Description

The main purpose of the task is to prepare a training set of pairs of sentences for the presented semantic relations. The examples should be challenging/non-trivial, but natural.

The task gives two similar sentences (a **premise** and a **hypothesis**) and the aim is to determine which label (*entailment*, *contradiction* or *neutrality*) is appropriate for the hypothesis. Then, according to the instructions, two additional sentences need to be formed for the two missing relations (either by *modifying the hypothesis* or by *forming new sentences*). If the second sentence in the originally given pair is inadequate, sentences for all three semantic relations have to be formed.

WARNING! The premise itself (i.e. the first sentence of the source pair) should never be modified!

Be careful not to get caught up in overthinking whether the example represenets entailment, contradiction, or neutrality. As a general guideline, we should spend no more than 30 seconds on each example. If the case is problematic, this can be noted in the comments column and/or by sending a message to [project mailing list].

3 Sentence Formation Instructions

Below, we define the instructions for creating examples in more detail.

WARNING! Please note that these are guidelines, not strict rules. The guidelines have been tested, but in some cases it may not be possible to follow them. If in doubt, please contact the [project mailing list].

3.1 ❌ Inadequate Ways of Forming Sentences

It is possible to create (too) simple examples for each semantic relation; however, the model will not learn the more complex rules of language understanding from them. Such examples should be avoided.

3.1.1 General Instructions for Inadequate Examples

3.1.1.1 Excessive Overlap

If more than 70% of the hypothesis is identical to the assumption (e.g. contains almost identical words), the example is inadequate. At least one third of the sentence must be different.

[P] John bought apples at the grocery store.
[H] John purchased apples at the grocery store. ❌

3.1.1.2 Non-Standard Language

Do not introduce typos or non-standard language into the hypothesis (except in cases where non-standard language is already present in the source pair).

[P] John bought apples at the grocery store.
[H] John bougt aples at the store. ❌

3.1.1.3 Simple Substitution of Numbers, Named Entities, etc.

It is not enough to just modify a few numbers or names in the hypothesis compared to the premise.

[P] John bought 5 apples.
[H] John bought **10** apples. ❌
[H] **Mary** bought 5 apples. ❌

3.1.2 ❌ Inadequate Examples for Entailment

3.1.2.1 Significant Overlaps or Substrings

The hypothesis contains part of the premise (with minimal modifications) or reuses many of the words from the premise.

[P] We found some sort of a middle road.
[H] We took the middle road. ❌

3.1.2.2 The Hypothesis Is Just a Shortened Premise

A hypothesis is just a shortened version of the premise (this is also a common consequence of using substrings) or is always significantly shorter than the premise.

[P] John bought 5 apples to make an apple strudel.

[H] John bought 5 apples. ❌

3.1.3 ❌ Inadequate Examples of Neutrality

3.1.3.1 Significant Change of Topic

If the hypothesis is about a completely different topic than the premise, the example is inadequate. When forming a neutral hypothesis, we need to form a sentence that is still semantically related to the premise.

[P] John bought 5 apples.
[H] It's raining outside. ❌

3.1.3.2 Simple Additions of New Information

It is inadequate to make a hypothesis neutral simply by adding new information to the premise (e.g. listing previously unmentioned objects).

[P] John bought 5 apples.
[H] John bought 5 apples **and 7 pears**. ❌

A statement is neutral when it cannot be inferred from a premise - usually because its meaning is more specific (or narrower). The easiest way to achieve this is to add new information to the sentence, but in this case the model will only learn to recognise as neutral those sentences that contain something new compared to the premise. Our aim is to provide it with diverse examples: for example, we can make the meaning more detailed or narrower by using different words – e.g. in the example below, the hypothesis states that the person cut the onion; in the premise, it says that she chopped it, which is not necessarily the same thing (the size of the pieces is different; if the onion is chopped, it is cut into pieces; if it is cut into pieces, it is not necessarily chopped).

[P] Before lunch, you'd usually see her at the kitchen counter, **cutting** onions.

[H] She often **chopped** onions in the kitchen while preparing lunch. (neutrality)

3.1.3.3 The Hypothesis is an Extended Premise

It is inadequate for the hypothesis to be just an extended version of the premise (this is often the result of simply adding information). We should also be careful that the hypothesis is not always consistently longer than the premise, otherwise the model relies too heavily on sentence length alone to establish the meaning relationship.

[P] John bought 5 apples.
[H] John bought 5 apples to make an apple strudel. ❌

3.1.4 ❌ Inadequate Examples of Contradiction

3.1.4.1 Simple Negation of the Premise

A hypothesis is just a negated form of the premise, substituting for example "am" with "am not", "is" with "is not" or "isn't", and so on.

[P] I've handed in my homework.
[H] I didn't hand in my homework. ❌

3.1.4.2 Repeated Use of a Very Limited Set of Words

Be careful not to negate the hypothesis in exactly the same way across multiple examples, or to use the same words over and over again – the examples below are inadequate because we keep using the word "slept" to deny what the person was doing in the hypothesis. The only thing the model would learn from such examples is to pay attention to the word "slept".

[P] John played football this morning.
[H] John **slept** in the morning. ❌

[P] John ate breakfast.
[H] John **slept**. ❌

3.2 ✅ Adequate Ways of Forming Sentences

In this section, we provide some more examples of how to create suitable sentences. The combination of these patterns should serve as a starting point for you when creating examples.

WARNING! The list **is not exhaustive**: there are other adequate ways to create examples. Do not focus on just one category of sentence formation, as the aim of the task is to get as many different examples as possible.

3.2.1 Common Sense Reasoning

The hypothesis contains additional knowledge that is considered common sense (e.g. in the example below, we know that if a person pushes the brakes, the speed decreases, even though this is not explicitly stated in the hypothesis).

[P] John slammed on the brakes while driving.
[H] The speed decreased. ✅ (entailment)

3.2.2 Non-Trivial Reasoning Involving Numbers

Adequate examples can use inference about numbers in the hypothesis, but it should be non-trivial and it should require some additional knowledge. For instance, in the example below, the hypothesis implies that the person's birthday is three months after June, so most likely in September. To reach this conclusion, we do not only need to know how to count, but we also need additional knowledge about the months of the year, so the example is not trivial.

[P] He started planning in June, three months before her birthday.
[H] His birthday is in autumn, in September, to be precise. ✅ (entailment)

3.2.3 Integration of General Knowledge About the World

Adequate examples include (commonly known) facts about the world in the hypothesis. For instance, in the example below, we know from common knowledge that the "purple" team represents Maribor and the "green" team represents Olimpija.

[P] The match between the purples and the greens was a draw.
[H] A football match between Maribor and Olimpija was held. ✅ (entailment)

3.2.4 Temporal/Spatial Reasoning

Adequate examples may include our own reasoning about space and time in the hypothesis. In the example below, we learn from the hypothesis that the river is five meters wide. Since we know that people cannot jump that far, we can conclude that the hypothesis that someone jumped from one bank to the other bank contradicts the assumption.

[P] **A** five-meter-wide river flows between the two banks.
[H] I jumped from one bank to the other. ✅ (contradiction)

3.2.5 Phrases, Idioms, and Metaphors

Adequate examples may contain the use of metaphorical expressions and phrases (e.g. in the example below "to postpone" - "the plans fell through").

[P] The meeting was postponed due to unforeseen circumstances.
[H] The plans for the meeting fell through. ✅ (entailment)

3.3 ❕ Conditionally Adequate Ways of Forming Sentences

In this section, we list ways of forming sentences that are allowed and adequate, but you must be careful not to form sentences using only one of the listed methods (e.g. replacing only one word with a synonym). As a general guideline, at least one third of the sentence should be modified. At best, the sentences should be as varied as possible.

3.3.1 Use of Synonyms, Hypernyms, Antonyms

[P] When I still lived in Frankfurt, I used to take the train a lot.

[H] When I lived in Frankfurt, I often traveled by train. ❌

We have replaced "to take the train" with "to travel by train" and removed "still" - the change is not sufficient, the two sentences are too overlapping.

[H] In Germany, I frequently used public transport. ✅ (entailment)

The example includes the common knowledge that Frankfurt lies in Germany; "train" was replaced with the hypernym "public transport"; "often" was replaced with "frequently"; the hypothesis is not significantly shorter than the premise and is sufficiently different.

3.3.2 Word Order, Passive/Active Voice

Please make sure that the change of word order is not the only modification to the text. This applies both to the change of the word order and to the change of agent/patient of the action (subject – object; passive – active).

[P] All the forks and knives that had been taken from the sink were put in drawers in the cupboard.

[H] They put all the forks and knives that they had taken from the sink in the drawers in the cupboard. ❌

The modification is insufficient.

[H] The sink was full of cutlery, so they stacked it in the cupboard. ✅ (entailment)

We've changed the word order and used a synonym ("stacked") and a hypernym ("cutlery") at the same time.

[P] John gave Mary flowers.

[H] Mary was given flowers by John. ❌

Insufficient modification.

[H] Metka received a bouquet of flowers as a gift from John. ✅ (entailment)

We have used an expression that offers a different view of the action from the perspective of the patient (to give - to receive as a gift), we have used a synonymous expression ("a bouquet of flowers"), and we have reversed the order.

[P] Last week, perhaps on Wednesday, the postman delivered a letter to her in a heavily decorated envelope.

[H] The letter in a heavily decorated envelope was delivered last week, perhaps on Wednesday. ❌

The example is inadequate because it is only a change of the active voice to the passive voice.

[H] The letter in an ornate envelope was delivered to her last week around Wednesday. ❕

A conditionally acceptable example (passive - active; "perhaps on Wednesday" - "around Wednesday", "heavily decorated" - "ornate"), but there is room for improvement.

[H] The letter in a kitschy and over-the-top envelope was handed to her some time in the middle of the week. ✅

"delivered" - "handed over", "around Wednesday" - "some time in the middle of the week", active - passive.

3.3.3 Use of Morphological Features

For example, sensible changes in grammatical gender, number, etc. are adequate, but only in combination with other modifications. Also, be careful not to change the number and gender in a way that the makes the sentence describe someone else who was not included in the premise (e.g. "She played the guitar" -> "He played the guitar").

[P] We cycled to an old mill not far from where the two rivers meet.

[H] I cycled to the old mill not far from where the two rivers flow. ❌

We only mention one person instead of both (which is adequate - if both cycled to the mill, it entails that the individual person did that as well), and we make a slight modification to the last part ("the two rivers flow"), but there is still too much overlap.

[H] I took the bicycle all the way to the mill, which stands very close to the confluence of the two rivers. ✅ (entailment)

Sufficient change - synonyms ("to take the bicycle" - "to cycle"; "the place where the two rivers meet" - "the confluence"), meaningful change of grammatical number (we - I).

3.3.4 Use of Abbreviations/Acronyms or Their Extended Forms

Use of established abbreviations is permitted. Abbreviations can also be written in full form. However, abbreviations should not be introduced ad hoc (e.g. abbreviating "temperature" to "temp.").

[P] A state of emergency has been declared in the USA due to freezing temperatures.

[H] The United States declares a state of emergency amid freezing temperatures. ❌

The expanded abbreviation and the modification ("amid") are not sufficient modifications.

[H] The United States of America is in the grip of a severe cold snap and a state of emergency has been introduced. ✅ (entailment)

Full form of the abbreviation, synonymous terms ("low temperatures" - "severe cold snap", "declare" - "introduce")

[P] For more information on how to make elderflower syrup, see p. 5.

[H] Read more about the preparation of the syrup on page 5. ❌

Insufficient modification - only extended abbreviation (p. – page) and shortened term ('elderberry syrup' - 'syrup').

[H] The process of making syrup from elderflowers is described in more detail on page 5. ✅ (entailment)

Synonymous term ('elderflower syrup' - 'syrup from elderflowers'), full form of abbreviation ('p.' - 'pages'), restructured sentence ('read more about the preparation' - 'the process is described').

3.3.5 Anaphora and Coreference

The use of pronouns and other anaphoric devices is permitted, but should not be the only modification.

[P] After another long-winded and unnecessary speech, she decided not to listen to the Vice-President at any more meetings.

[H] She decided not to listen to him anymore. ❌

The example is inadequate. The hypothesis is significantly shorter than the premise; it is a shortened part of the original premise; only a pronoun is used instead of "Vice-President".

[H] He was again making completely unnecessary comments, and she came to the decision not to pay any attention to what he said during the meetings. ✅ (entailment)

The example is adequate: a heavily modified sentence, synonymous expressions ("decided" - "came to the decision", "to listen" - "to pay attention to"), anaphoric devices ("Vice-President" - "him").

3.4 Additional Illustrative Examples

Here are some more examples, with explanations why they are (in)adequate.

[P] My eyes got used to the twilight.

[H] A darkness fell over my eyes. ❌

An irrelevant example that cannot be linked to the premise at all - it has a metaphorical meaning that implies something completely different. It is seen as a significant change of subject.

[H] I started to see the outlines in the dark better. ✅ (entailment)

A good example of entailment - the sentence is not overlapping with the premise, it is not significantly shorter or longer, and it contains general knowledge (if our eyes get used to the twilight, we can better perceive the outlines in the dark).

[H] It was dark in the cave. ❕ (neutrality)

A conditionally acceptable case for neutrality - there is no (in)direct mention of a cave or anything like that in the hypothesis, so we have added some new information to the hypothesis. This is not the best way to form neutral sentences (see 3.1.3.2).

[H] I usually had no problems with my vision in the dark. ✅ (neutrality)

The assumption does not imply that a person does not normally have problems with their vision in the dark.

[H] I couldn't see anything in the dusk. ✅ (contradiction)

A good example of a contradiction. We have synonymous terms ("twilight" - "dusk"), the sentence is not a substring of the premise and is sufficiently restructured, and its meaning contradicts the premise (if our eyes have become accustomed to twilight, we are probably seeing something).

[P] He told us today, "At one point, our police officer asked me to come in for a short chat."

[H] On this day, he informed us of a police officer's summons for an interview. ✅ (entailment)

Relevant example - synonymous expressions ('today' - 'on this day', 'told' - 'informed', 'chat' - 'interview'), restructured sentence (change from direct quotes; verb structure to noun structure - 'asked me' - 'summons').

[H] He was never contacted by the men in blue on this matter. ✅ (contradiction)

Relevant example of contradiction. Related terms ("policeman" - "men in blue", "contact" - "invite") and semantic contradiction to the premise (if a person said that a policeman invited him to an interview, it cannot be assumed that he was never contacted by the men in blue).

[H] He spoke to a police officer about the incident. ✅ (neutrality)

A good example of neutrality - we don't know whether the person has already spoken to the police officer or not.

4 Frequently Asked Questions

The premise contains non-standard language (e.g. slang) or a typo (e.g. "they're" -> "there"). Can I correct it?

No, the premise should never be modified. If the premise contains non-standard language, it is best (but not necessary) if the generated hypothesis also contains non-standard language.