Evaluating Implicit Knowledge Sentences Generated by Language Models

Annotation Manual

In texts and everyday communication a lot of information stays implicit. This information is often very obvious and therefore not expressed explicitly, but is important for understanding the connection between two sentences -- especially for computational systems, that can't infer implicit knowledge as easily as humans can do.

For an illustration, look at the following two sentences:

> Sentence 1: *The car of my brother is brand new.*
> Sentence 2: *It is a very expensive vehicle.*

Here, the implicit knowledge which connects the two sentences is that *cars are vehicles*.

We tried to reconstruct such implicit information that helps to understand the connection between two sentences using neural language models that are tasked to generate sentences that verbalize implicit knowledge connecting two (usually contiguous) sentences.
Your task is now to evaluate the output of these language models in terms of grammaticality, coherence, and content.

You are given two sentences (in the following called ***source sentences***), and some sentences produced by language models which try to explain the connection between the two sentences (in the following called ***generated sentences***).

The source sentences can support or attack each other (agree or disagree about a topic), or can be contradictory. Here, the generated sentence should give an explanation why the sentences are contradictory. For example given the two source sentences

> Sentence 1: *The girl is swimming in the lake*.
> Sentence 2: *The girl is inside*.

Here the generated sentence should express something like *A girl can't be inside and in a lake at the same time*, or *Lakes are usually not inside*.

You are also given a ***reference sentence*** (not produced by a language model), which should explain the connection (but be careful, sometimes this sentence might not provide useful information!)

Your task is now first to answer some questions **only** about the generated sentences **without** looking at the reference sentence, and in a second step you are supposed to compare the generated sentences to the reference sentence.
At the end of this file, you will find some example annotations.

Some hints before you start:

- Even if the generated sentences have some (smaller) grammatical errors, they might still be good explanations of the connection between the two given sentences. It is therefore important that you judge grammaticality and coherence/content independent from each other!
- The source sentences come from different genres such as argumentative texts, newspaper texts, technical/domain-specific texts, blogs etc. So there are some differences between the instances, as you will notice.
- The source sentences are taken from longer texts, so sometimes they might seem to be a bit "out of context", for example due to discourse markers which only make sense in a larger context. This is for example the case with the following source sentence pair, where the discourse marker "furthermore" might be out of place:
  Furthermore, the production of meat is comparably resource-intensive.
  We should thus all limit our consumption of meat rigorously to protect the environment.
  In these cases, just ignore the discourse markers (or other words) and focus on the content of the sentences!
- To understand the sentences better, sometimes it can be helpful to read them in reverse order (second sentence first, then first sentence).

Step 1: Evaluation of the generated sentences in relation to the two source sentences

You are given two source sentences, and some sentences generated by our language models which should represent the missing information between them.
Please answer the following questions (just use the given options listed in the column "Answer in the Sheet"):

Examples given for generated sentences refer to the following pair of source sentences:

Source sentence 1: *Furthermore, the production of meat is comparably resource-intensive.*
Source sentence 2: *We should thus all limit our consumption of meat rigorously to protect the environment.*

[At the end of this file, you will find some more example annotations.]

Dimension	Question	Choices	Answer in the Sheet	Examples for generated sentences that would be labelled accordingly (source/target sentence see above)
Grammaticality	Is the generated sentence grammatically correct?	**Yes,** fully *(don't mind lower cases at the beginning of a sentence, or missing punctuations.)*	**YES**	*The production of meat is resource- intensive, which is bad for the environment.*
		Almost, only minor grammatical errors	**ALMOST**	*All animal must be killed and disposed of in a manner that is both ethically and economically feasible.*
		There are **some bigger grammatical errors**	**ERRORS**	*The production of a meat is danger.*
		Not at all, the grammar is so bad or the sentence is so cryptic that it is almost impossible to understand its meaning → in this case, skip the other questions!	**NO**	*ethically and economically feasible.*
Coherence	Is the generated sentence coherent (that means, logically and semantically consistent) with respect to the two source sentences?	**Yes,** the generated sentence is coherent with the source sentences	**YES**	*The production of meat is resource- intensive, which is bad for the environment.*
		Parts of the generated sentence are coherent with the source sentences, but not the complete generated sentence; or coherence is given with one of the two source sentences only	**PARTLY**	*The production of meat is resource-intensive, and many people like to go to restaurants on the weekends.*
		No, the generated sentence is incoherent with the source sentences	**NO**	*Buying new clothes every week is bad and unethical.*
		The generated sentence does not make any sense at all → in this case, skip the other questions!	**NONSENSE**	*Either it is a production, or it is not.*
Content	(A) Are the two source sentences implicitly connected by some (unexpressed) piece of knowledge?	**1: Yes,** there exists some implicit knowledge that links the two source sentences **2: No,** *everything* that is needed to understand the connection between the two source sentences is explicitly expressed in them	**YES/NO**	*YES*

	[One answer for all candidates]			
	(B) Does the generated sentence provide implicit knowledge that is not explicitly mentioned in the two source sentences?	**Yes,** it expresses some implicit knowledge that is not mentioned in the two source sentences (regardless of the content/quality of the generated sentence, which we will ask for in the next question)	**YES**	*the production of meat is resource- intensive, which is bad for the environment.*
		No, it does not add anything new.	**NO**	*The production of meat is resource-intensive.*
	(C) Does the generated sentence give an explanation of the connection between the two source sentences?	**Yes,** it gives a meaningful explanation of the connection between the two source sentences	**YES**	*the production of meat is resource- intensive, which is bad for the environment.*
		2: Neutral: The generated sentence is thematically and semantically related to the source sentences, but not in a clear logical relation and does not explain the connection between the two source sentences.	**NEUTRAL**	*Production of meat is a relatively large share of the total animal protein produced in the United States per unit length of meat produced per annum*
		3: No: The generated sentence is **misleading** or **contradictory** in the context of the source sentences	**NO**	*meat is a good thing.*
	(D) Which of the generated sentences explains the connection between the two source sentences best? [One answer for all candidates]	**[Choose one sentence**, if two (or more) are equally good, list all of them (by referring to their given numbers, e.g. “i” or “i;iii”)**]**	**i/ ii/ iii ...**	

Step 2: Comparing the generated sentences to the reference sentence

You are now allowed to look at the reference sentence and are supposed to compare it to the sentence generated by the language models. Please answer the following questions (again, just use the given options listed in the column “Answer in the Sheet”):

Examples given for generated sentences refer to the following pair of source sentences and the given reference sentence:

Source sentence 1: *Furthermore, the production of meat is comparably resource-intensive.*
Source sentence 2: *We should thus all limit our consumption of meat rigorously to protect the environment.*

Reference sentence: *Resource-intensive production of meat is bad for the environment.*

Dimension	**Question**	**Choices**	**Answer in the Sheer**	**Examples for generated sentences that would be labelled accordingly** (source/target and reference sentence see above)
Comparison	(A) Is the generated sentence similar in meaning with the reference sentence?	**1: Yes,** the generated sentence expresses the same or very similar information	**YES**	*The production of meat is resource-intensive, which is bad for the environment.*
		2: Parts of the expressed information from the generated sentence and the reference sentence are comparable/overlap	**PARTLY**	*All animals must be killed and disposed of in a manner that is both ethically and economically feasible.*
		3: No, the information expressed in the generated sentence is completely different from the reference sentence	**NO**	*Meat is a good thing.*
	(B) Which sentence is a more meaningful explanation of the implicit knowledge that connects the given source sentences: the reference sentence or	**1: The generated sentence** provides a more meaningful explanation of how the two sentences are (implicitly) related	**GS**	*/*
		2: The reference sentence provides more meaningful information which connects the sentences (see above?)	**RS**	*Production of meat is a relatively large share of the total animal protein produced in the United States per unit length of meat produced per annum*

	the generated sentence?	**3: Both** are equally meaningful explanations of how the sentences are related and express the same amount of information	**BOTH**	*The production of meat is resource-intensive, which is bad for the environment.*
	(C) Which of the generated sentences is semantically closest to the reference sentence? [One answer for all candidates]	**[Choose one sentence** (by referring to their given numbers, e.g. "i" or "ii")**]**	**i/ ii/ iii ...**	

EXAMPLE 1

Source sentence 1: There should be much higher fines for dog dirt left on pavements.
Source sentence 2: Besides, the fines for public urination by a human are higher than for leaving behind dog dirt on the pavement, which is disproportionate.

Reference sentence: Dog dirt is a greater offense than public urination.

Dimensions:	Grammaticality	Coherence	Content A (impl. connection?)	Content B (new content?)	Content C (explanation?)	Content D (which is best?)	Comparison A (similar meaning)	Comparison B (better ex-planation?)	Comparison C (closest?)
Options:	YES/ ALMOST/ERRORS/NO	YES/ PARTLY/ NO/ NONSENSE	YES/ NO	YES/ NO	YES/ NEUTRAL/NO	i, ii...	YES/ PARTLY/ NO	GS/ RS/ BOTH	i, ii...
(i) a dog dirt is a greater offense than public urination.	ALMOST	YES	YES	YES	YES	(i)	YES	BOTH	(i)
(ii) dog dirt is not the same as public urination by a human, therefore the fines are not equal to each other.	YES	YES		YES	YES		PARTLY	RS	
(iii) Public urination is a misdemeanor, and therefore a violation of the law.	YES	YES		YES	NEUTRAL		NO	RS	

(iv) the people are either leaving behind dog dirt on the pavement or they are not.	YES	YES		NO	NEUTRAL		NO	RS	
(v) greater offense than public urination.	NO	-		-	-		-	-	

EXAMPLE 2

Source sentence 1: But one can make occasional mistakes in the Bachelor thesis.
Source sentence 2: If this carries more weight, many other individual assignments done during the course of study may be obscured.

Reference Sentence: The Bachelor thesis is one of the individual assignments of the course of study.

Dimensions:	Grammaticality	Coherence	Content A (impl. connection?)	Content B (new content?)	Content C (explanation?)	Content D (which is best?)	Comparison A (similar meaning)	Comparison B (better ex-planation?)	Comparison C (closest?)
Options:	YES/ ALMOST/ERRORS/NO	YES/ PARTLY/ NO/ NONSENSE	YES/ NO	YES/ NO	YES/ NEUTRAL/NO	i, ii...	YES/ PARTLY/ NO	GS/ RS/ BOTH	i, ii...
(i) a bachelor thesis is a type of individual assignment.	YES	YES	YES	YES	YES	(i)	YES	BOTH	(i)
(ii) one can not make an individual assignment and a thesis at the same time and be doing both simultaneously.	YES	PARTLY		NO	NEUTRAL		NO	RS	
(iii) a Bachelor thesis is a type of student.	YES	NONSENSE		-	-		-	-	
(iv) a Bachelor thesis is a type of student work.	YES	YES		YES	YES		PARTLY	RS	
(v) Individual assignments are due at the end of the class.	YES	PARTLY		YES	NEUTRAL		NO	RS	