page 5  
  
after Example 2.1.1, seventh line: ... we can conclude that background knowledge was involved -> sounds dangerous and could lead to philosophical discussions. ... Can a machine 'know'? ... what about the ML techniques? Does the machine really use background knowledeg, or just a huge amount of data? Try to reformulate in a way that it is not attackable  
  
  
Just before Example 2.1.2: You need to explain here how the results are evaluated by google hits. I guess that it follows the PMI method, i.e. the relative amount of hits a question with each answer receives. Probably you can find information on that in wikipedia. Actually, I have the impression that only one version of the example is google proof, but not the twin sentence. Let's compute the statistical result with the Example 2.1.2. My result was as follows:  
  
#(books+foolish) / (#books + #foolish) =44.600.000 / (4.230.000.000+69.500.000) = 0.01  
  
#(mistakes+foolish) / (#mistakes + #foolish) = 35.100.000 / (340.000.000+69.500.000) = 0.086  
  
0.086 > 0.01, thus foolish mistakes is more likely than foolish books.  
  
#(books+worthless) / (#books + #worthless) = 42.400.000 / (4.230.000.000+68.000.000) = 0.0098  
  
#(mistakes+worthless) / (#mistakes + #worthless) = 26.100.000 / (340.000.000+68.000.000) = 0.064  
  
mh, here the correct answer cannot be solved with google, because worthless books is not more likely than worthless mistakes (but that should be the answer, right?)  
  
Please verify the numbers and why/ which part of this example is really not google proof.

second paragraph: uses Transformer based architecture -> why did you write Transformer in capital letter?

-because it’s a name of that specific architecture  
  
page 13  
  
second sentence after stage 4: ... the precision, recall and F1 score ... why do you mention them here? if you do, you should tell what they are, and which are the scores wrt the different approaches

-because this is the only paper that mentions them and this is how they did their evaluation

page 14  
  
Table 2.1 is a nice overview, but I think you could make it more readable, by splitting the normal dataset and \* in two different columns. You could save space by leaving out the vertical lines and make the distances between the columns smaller. This table is a very important and helpful summary and should also be part of your project presentation  
  
page 15  
why did you underline 'the steps'?

-because it’s a reference to the steps in the previous section, you can click and it takes you back to the steps  
  
3.1. Example 3.1.1.: Isnt it in Q, who was \*too\* weak?

-no it’s just Who was weak

The different colors in the figures (and also later for the annotators i problematic, because normally people print the paper out in black/ white. Maybe you can think of another way of presentation? gray, thick lines, italic, ...  
  
  
  
page 18, 3.2  
  
why is Reasoning  Algorithm in capital letters?

-because I introduce here the shortcut for it (RA)  
  
Phase 1: To explain the 'merge' you could show this merge wrt Figure 3.1 and 3.2?  
  
I do not understand the second step of 2. Probably if you would have a merged figure, you could refer to these examples in that figure? and then finally for step 5, show the resulting graph, which you might want to refer to in the steps in Phase 2  
  
  
  
  
page 20, Observations  
  
I thought that by now, the WSC problems and their corresponding knowledge types are now publicly available, arent they?

-not really… I found them on the git hub of Sharma, but I don’t think this counts as publicly available because I was really looking for them.  
  
Btw. do you have the email that Sharma sent you? What exactly did he answer you, when you told him that this 'crucial' relation did not have any effect?

-should I include in this report that I contacted him and what was the discussion???

page 21, Chapter 4  
  
Different than in Sharma et al, our categories rely on the content and not on the grammatical/ causal structure of the sentences. I think that is something which you should emphazise more

-this is in the discussion, should I also include it here in the description of the cat?  
  
4.1  
  
About these five categories: Are they excluding each other while the annotators did them?

-No, later when the numbers from the annotations results are discussed I mention this  
  
fifth line from below: The annotation was based on the additional knowledge -> I would rather say, the annotation was based on the least possible (relevant) knowledge that is necessary in order to solve the problem. Possibly, there are different relevant knowledge about a certain problem necessary, which do not need to be known together, but either one of that knowledge is necessary. does that make sense?

- I don’t understand why the least possible. I will leave it as it is in the text now  
  
  
page 23  
  
fifth line from below: this sentence seems wrong, it should be the other way around. Actually, sth only can fit into sth else if that sth else is a container. That is additional (implicit) knowledge, which we might assume for granted, as one of the two answers has to be necessarily correct (-> that is what i meant when i wrote::: the least knowledge necessary to solve the task; much more properties exist about these objects, but only a very small part is crucial, some other properties can be assumed, given the experimental setup, i.e. that one answer has to be correct, i.e. one of the two objects necessarily has to be a container...)  
  
page 24  
  
  
Here in Example 4.2 we would actually need to know, either  
  
1. if someone lived in 1765, then that person is (normally) dead now (and thus s/he cannot travel)  
  
or  
  
2. if someone has a grave, then that person is (normally) dead now  
  
  
It seems to me that only one of the two pieces of information is necessary to get the correct information.  
  
Evaluation, Cohen's kappa, you can actually do it by hand (in excel, no need to refer to an implementation)  
  
I tried to compute p\_o and p\_e, please verify if they are correct and include them in your report:  
  
p\_e = 0.1165 and p\_o = 0.6911 ??  
  
In Table 4.2, annotator 1 annotated 151 whereas annotator 2 annotated 150 items. annotator 1 has a sum of 101%. Is that because of rounding?

-yes, it’s because of rounding

page 26

last sentence of 4.1: How do you come to that conclusion? The annotators were forced to choose one category right?

page 27

second paragraph, seventh line: Small object(s) could not fit **other objects** (?) shouldn't that 'fit' relation not be specified as a binary relation?

Code 4.2 -> What do you refer to in the caption, for S2?

-I don’t understand the question

page 32, Dav16: Is there a better reference for this paper than CoRR, ? Similar question for [LJL+16a]  [LJL+16b] and for [MCCD13] on page 33 and [TEC+18]

-these are from arxiv.