Annotation Guidelines for Self-Contradictions in the CAPTURE corpus

April 15, 2024

Contents

1 Introduction

This document summarizes the guidelines for annotating locution pairs for whether they contain a self-contradiction. The aim is to create a corpus composed of locutions from ”real-world” discussions that presents contradicting and non-contradicting locution pairs from the same author. Since the aim is to model the properties of self-contradiction, criteria for this category will be described in broader detail and following the definition presented by [1], while any locution pair that does not meet these criteria will be annotated as non-contradictory. It is assured that every locution pair the annotators will receive is uttered by the same speaker, which rules out any necessity of annotations for speaker and ensures that the presented criteria can be applied under the assumption that the speaker of the two presented locutions is identical.

2 Annotating Self-Contradiction

As described in the introduction, this section presents the main criteria for the annotation. After presenting and discussing the strictly logical definition of contradiction in the light of self-contradiction, the definition by [1] will be introduced as the basic definition of contradiction to be adopted in the annotation.

2.1 Logical Definition of Contradiction

Speaking from a strictly logical point of view, a contradiction in a logical formula occurs, if and only if there is no possible world where the propositions A and B would be true at the same time.

[1] presents a variation to this definition that is looser in a strict logical sense, but appeals more to a pragmatic approach to contradiction detection as the CAPTURE corpus aims to capture. This definition assumes that a statement A and B are contradictory in case there is a very low probability that they are uttered in the same context and both true.

2.2 Semantic Conceptions of Contradiction

This idea can be found in the definition given in [2] as well, where contradictoriness is defined as a sentence, where a part (e.g. the subject) has a meaning that contains information that is incompatible with the meaning contained in another part of the sentence (e.g. the verb). This is illustrated in example 1. Contradictoriness is given here since the meaning of the subject *Dolphin* incompatible with the meaning that is attributed to it in the predication; in this case being a *fish*.

(1) A Dolphin is a fish.

To adapt this to dialogue, adding the definition of inconsistency from [2] is helpful. This phenomenon is presented as a comparison of two sentences that are both together neither true or false if they refer to the same event or individual, but rather one of them has to be true and the other needs to be false. This is illustrated in example 2, where only one of the two statements can be true in case they refer to the same cat.

(2) a. The cat is dead.
b. The cat is alive.

2.3 Applied Definition of Self-Contradiction for Annotation

This section contains criteria for annotating Self-Contradictions in the CAPTURE. The applied labels will be ”contradiction” and ”non-contradiction”. The data is provided to the annotators as a pair of locutions together with a pair of corresponding reconstructed propositions. The selection of the data is designed in a way that ensures that the speaker is the same for both locutions. Providing the propositions shall compensate for a lack of context that may hinder a correct annotation of the data. The annotators are asked to annotate as much data as possible without using the help of the propositions and annotate with a second label indicating whether the context was needed for a correct annotation. Here, the applied labels are ”True” and ”False” with ”True” indicating that the context was needed for a successful annotation.

The following sections will first present criteria that can be applied for the identification of self-contradictions in the corpus. The whole definition has to be seen under the assumption that the locutions are uttered in the same (discourse) context by the same speaker. Annotation should happen in a rather conservative manner and in doubt, the locution-pair should be rather annotated as non-contradiction. In the second subsection, examples will be presented that will be annotated as ”non-contradiction”.

2.3.1 Annotating Self-Contradictions

The definition of self-contradiction applied in the context of CAPTURE combines the concepts of contradictoriness and inconsistency as dialogue components with the definition of contradiction presented by [1]. In the scope of the annotation for the CAPTURE, a self-contradiction is defined by the following conditions. A locution pair is considered as containing a self-contradiction if the joint truth value of both locutions in a conjunction is improbable to be true, no matter if this fact arises from bare lexical contradictoriness or marks an inconsistency in the dialogue. This evaluation happens under the assumption that they were uttered by the same same speaker in a similar context. To help identifying this abstract criterion, the following indicators can be used:

- ... the locutions contain a lexical or semantic incompatibility that makes it improbable that the joint truth value of both locutions is true. This criterion refers to the definition of contradiction by [1], as well as to the concept of semantic contradictoriness in [2]. Please consider example 3 to illustrate this. It would be annotated as ”contradiction” as there is a mismatch between the implication of ”fluff and targeted rhetoric” and the term ”specifics”. Therefore it would be very improbable that the two locutions are both true in case the pronoun ”she” would be resolved to ”Hillary” as it is visible in the propositions.

(3) a. Locution 1: Hillary actually has specifics, names names, etc.

b. Locution 2: Most of what she said has been fluff and targeted rhetoric.

c. Proposition 1: Hillary Clinton actually has specifics, names names, etc.

d. Proposition 2: Most of what Hillary Clinton said has been fluff and targeted rhetoric.

- ... the locutions present positions that are unlikely to be compatible within the world view of the same speaker in the frame of the related discourse. This includes a deliberate change of position in the discourse as well since such an instance still marks a case of inconsistency even though this may be resolved by explanation in the further discourse. This criterion applies the concept of inconsistency as presented by [2] to the domain of dialogue and positions in a discourse. Please consider 4 as an example for this phenomenon. In the discourse around abortion, the positions ”pro-life” and ”pro-choice” are opposite positions and it is very unlikely that the same speaker has these two positions in the same discussion without having had a strong change of heart.

(4) a. Locution 1: I am pro-life

b. Locution 2: very pro- choice

c. Proposition 1: TRUMP is pro-life

d. Proposition 2: TRUMP is very pro- choice

- ... there is a factive or numeric mismatch between two locutions and is an interpretation of the contradictoriness presented by [2] applied to the domain of dialogue. Please consider 5 as an example for this criterion. In addition, it is a good example for a sample that does not profit from additional information given in the propositions.

(5) a. Locution 1: We’ve actually only created 21,400 green jobs in Scotland with the powers that we already have.

b. Locution 2: The Scottish Government promised that we would have 130,000 green jobs by 2020.

c. Proposition 1: We’ve actually only created 21,400 green jobs in Scotland with the powers that we already have.

d. Proposition 2: The Scottish Government promised that we would have 130,000 green jobs by 2020.

- ... there is an emotional mismatch or conflict between the locutions. This criterion is an interpretation of inconsistency taken to the level of emotional involvement. In example 6, such a mismatch is presented in the form of Boris Johnson stating that he knows that things went wrong, but also expresses that he would have the Downing Street parties again.

(6) a. Locution 1: I would do the Downing Street parties again

b. Locution 2: I know things went wrong

c. Proposition 1: Boris Johnson would do the Downing Street parties again

d. Proposition 2: Boris Johnson knows things went wrong

- ... a mismatch is uttered between ideals (of the speaker) and the reality of actions and relates partly to the emotional mismatch, but in this case not between two actions but between ideals and actions. Example 7 illustrates this case and expresses a mismatch between the wish of people to own a house and the situation in reality that there are not enough houses built in order to fulfil this need or wish. Locution pairs that contrast two options would rather not fall under this criterion.

(7) a. Locution 1: The reality is we don't build enough houses

b. Locution 2: Many of us want the opportunity to own our home

c. Proposition 1: the reality is we don't build enough houses

d. Proposition 2: many of us want the opportunity to own our home

- ... one locution denies the truth of the other. This is possibly the criterion that is closest to a very pure case of contradiction. Cases where the speaker obviously corrects their statement would still be considered under this criterion even though the intention might be different. This differs from human perception of contradiction, but is necessary to model contradiction for a machine-learning model which is not capable to understand pragmatic moves like the intention of a speaker. Please consider 8 as an example for a very clear case of denying the content of another locution.

(8) a. Locution 1: Women are pigs , slobs and dogs , and pregnancy is an inconvenience to employers.

b. Locution 2: I never said that.

c. Proposition 1: Women are pigs , slobs and dogs , and pregnancy is an inconvenience to employers.

d. Proposition 2: TRUMP never said that.

The whole definition has to be seen under the assumption that the locutions are uttered in the same (discourse) context by the same speaker. A self-contradiction does not need to fulfill every criterion listed above and there is no hard line, how many criteria have to be fulfilled since their perception may differ also with the semantic content that is conveyed in the locution. In doubt, a helpful question to ask can be: Would I notice and probably tell another person that there is a contradiction in their statements if I was in a conversation with them and they uttered the locutions in question?

2.3.2 Annotating Non-Contradiction

The annotations for "non-contradiction" labels is mostly defined by the insufficient or non-application of the criteria for annotating a "contradiction" label. Generally, a more generous annotation should happen for this category so that the corpus is able to capture a harder and more concise definition of "contradiction". The following will illustrate examples for non-contradiction labels with a short explanation.

The example 9 is rather straight forward and presents two locutions that present a point of view that is not contradictory or inconsistent. In this case, the context provided from the propositions is helpful to determine that the context of the two statements is actually the same so that the pronoun "it" can be resolved to the same entity in the discourse.

(9) a. Locution 1: It's a benevolent decision

b. Locution 2: it's great that they did

c. Proposition 1: Removing words from children's books is not a benevolent decision

d. Proposition 2: It's great that the publishers removed words from children's books

Example 10 illustrates a rephrase that widens the statement in the first locution, which does not not produce a contradiction. This also contrasts from the case of mismatch between ideals and reality 7, where there is a clear mismatch between an ideal and the reality, while this example asks for broadening the approach to solve a problem.

(10) a. Locution 1: that time is used to get on top of test, trace and isolate

b. Locution 2: what we have to do now is to make sure that we not only have that circuit breaker

c. Proposition 1: that time is used to get on top of test, trace and isolate

d. Proposition 2: what we have to do now is to make sure that we not only have that circuit breaker

Example 11 illustrates a more fuzzy case. Here, there could be a potential contradiction in the fact that political courage is attested to someone while disagreement with this person is expressed in the second locution. The propositions help to identify that the pronoun can be resolved to the same individual. Nevertheless, it is not necessarily contradictory to admire a person for their political courage while disagreeing with them on a specific position.

(11) a. Locution 1: You can't say he doesn't have political courage.

b. Locution 2: I tend to disagree with him on this particular judgment call.

c. Proposition 1: you can't say Douglas Ross doesn't have political courage.

d. Proposition 2: Iain Anderson tends to disagree with Douglas Ross on this particular judgment call.

Example 12 illustrates a case where the context from the proposition does not help to determine whether the statements are truly contradictory since there is still a lack of context.

An example like this, would be annotated as ”non-contradiction” and the second label would be marked as ”False” to indicate that the context was necessary to make this decision even though it was not helpful to the decision.

(12) a. Locution 1: A blind man on a galloping horse would see this for what it is

b. Locution 2: I try hard not to be cynical about these things

c. Proposition 1: A blind man on a galloping horse would see this for what it is.

d. Proposition 2: Naomi Long tries hard not to be cynical about these things, like parliamentary recess.

References

[1] M.-C. de Marneffe, A. N. Rafferty, and C. D. Manning. Finding contradictions in text. In J. D. Moore, S. Teufel, J. Allan, and S. Furui, editors, *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[2] J. J. Katz. *Semantic theory.* Studies in language. New York London Harper and Row, 1972.