Annotation Guide for QUD-Annotations

November 13, 2020

1 Preliminaries

This manual indicates how the driving reports are to be annotated w.r.t. rhetorical relations between text spans and question-under-discussion (QUD) related annotations. The latter comprises the annotations of the QUDs themselves and the information structural decompositions that can be derived from the QUD, i.e. the distinction between focus and background, the assigmnent of topics (and comments as the complementary parts), and non-at-issue components.

2 Technical requirements

The QUD-tree structures are stored in XML format with custom tags. The XML files are available on the file sharing platform Sciebo at `https://ruhr-uni-bochum.sciebo.de/s/6GUKmcj2EZewCW6`. We will use an online tool for editing, hosted on our own server and secured by login. For the time being, we edit offline using Microsoft Visual Studio Code with the "XML language support" extension by Red Hat. The VS Code extension checks the tree structures using DTD files linked in the file header in each XML file.

3 QUD-based annotations

There are different views in the literature which roles different constituents play in a discourse. However, central to all is that—in a QUD approach—there are parts of each proposition that answer the QUD while other parts are already given in the QUD. For instance, *What did John eat?* is asking what his meal was, but it already presupposes that the what he was doing is eating. The verb *eat* is part of the QUD. So when the answer is something like *John ate salmon*, we can say that only salmon is information not given in the QUD; *John* and *eating* are part of the QUD. Our QUD annotation is based on the guidelines proposed in Riester et al. (2017). Their key assumption is that the constituent which answers the QUD is *in focus*, i.e. it is the prominent information in the sentence (also potentially marked as such by accent). This means that all parts of a sentence (or clause) that are not the focus of the sentence must be presupposed in the QUD. If a sentence has more than one focus, we must ask are they (1) different answers to different QUDs or (2) all answers to the same QUD. In the XMLs, QUD tree structures are made up of the following tags:

- one single <ROOT> QUD at the top of the tree structure

- <QUD> tags

The leaf nodes of the QUD-tree are text spans enclosed in tags. Each is a constituent with a specific discourse role, i.e. one of the following tags:

- <AT> assertion topic
- <F> focus
- <CMT> comment
- <CON> context / background
- <CT> contrastive topic/focus
- <NAI> non-at-issue content

and <RES>, a catch-all that does not clearly fit one of the other roles. Looking at the tree structure top-down, QUDs thus embed other QUDs or the following tags: <DT>, <AT>, <F>, <CMT>, <CON>, <CT>, <NAI> (and <RES>).

The focussed constituent <F> is the answer to the QUD. It is what is said, stated about the topic <AT>. The topic is set by the QUD, and so the focus <F> can also be understood as a statement about or a comment <CMT> on the topic, where the information presupposed by the QUD is the background or context in which the new information (<F> / <CMT>) is understood. We can also think about the distinctions in terms of old versus new information: The QUD contain given—i.e. presupposed—information, which may be echoed by the answer, but primarily the focusses on conveying new information. We can also talk about this dichotomy as theme versus rheme: the QUD sets the theme, the topic, the background, while the answer to the QUD supplies the rheme, the focus, and comments on the theme. One technical detail between the different terminology concerns the status of discourse markers such as *sondern*: On the focus view, *sondern* is part of the focus phrase, while on the comment view, *sondern* is not part of the comment.

Q Was ist mit dem Bentley Flying Spur?

A_1 [Der wuchtige Bentley Flying Spur]$_{CON/AT}$ [ist kein Monument der Beharrung]$_F$, [sondern [die schnellste Limousine der Welt]$_{CMT}$]$_F$.

A_2 [Der wuchtige Bentley Flying Spur]$_{CON/AT}$ [ist kein Monument der Beharrung]$_F$, [(sondern) [(der wuchtige Bentley Flying Spur ist) die schnellste Limousine der Welt]$_{CMT}$]$_F$.

In our corpus, we should isolate discourse markers such as *sondern* in their own to simplify search queries later on.

We also recognise a discourse topic <DT>. By discourse topics we mean that a driving report is typically divided into sections (e.g., an introduction section, a section about technical details, one about comfort and driving experience, one about available accessories or different models of the vehicle, etc.). Each section would then be assigned a discourse topic, i.e. what that section is about, and have a corresponding QUD (e.g., the section contain technical information about a car would have its own <DT> and a QUD, e.g., *What about the technical specs of the car?*). The idea here is that different sections of the text may have their own structural principles (e.g., how information is structure in an introduction follows different rules than in which order technical specifications are given, and

those are again different from how to report on the driving experience during the test drive).

Non-at-issue content <NAI> is information which could be omitted without violating the QUD requirement.

Q Was ist mit dem Bentley Flying Spur?
A Der [wuchtige]$_{\text{NAI}}$ Bentley Flying Spur ist kein Monument der Beharrung, sondern die schnellste Limousine der Welt.

Text spans which are not embedded in a <NAI> tag are at-issue. Contrasts are a special discourse structure with a parallel syntax.

(1) a. [John]$_{\text{CT}}$ ate [the salmon]$_{\text{CF}}$ (but) [Lisa]$_{\text{CT}}$ didn't (eat [the salmon]$_{\text{CF}}$).
b. [John]$_{\text{CT}}$ ate [the salmon]$_{\text{CF}}$ (but) [Lisa]$_{\text{CT}}$ ate [the eggplant]$_{\text{CF}}$.

Depending on whether the contrasted constituents serve the discourse role of topic or focus they are called contrastive topic or contrastive focus (cf. Büring, 2003). In (1) we have two contrastive topics, *John* and *Lisa*, and two contrastive focuses, *the salmon* and *the egglpant*. For each focus we would need one QUD: *What did John eat?* and *What did Lisa eat?* But due to the fact that the topics also differ, we also need a super-QUD *Who ate what?* which has two unfilled variables: who and what.

(1) c. [John]$_{\text{T}}$ ate [the salmon]$_{\text{CF}}$ (but) ([John]$_{\text{T}}$) didn't eat / not [the eggplant]$_{\text{CF}}$.

When we have the same topic, as in (1c), we still have a super-QUD *Who ate what?* where what is filled by *John*. Then there are different approaches how to deal with the two focuses: (i) We consider them a list of foods which collectively answer a single QUD *What did John eat?* (ii) We consider a more complex subtree where the QUD *What did John eat?* splits into two QUDs What did John eat first?, What did John eat second? Approach (ii) may not seem as plausible as approach (i) when the second thing John ate is negated, but the approach is viable in the positive case. The super-QUD *Who ate what?* contains the verb *eat*, so we need another more abstract super-super-QUD above it which does not presuppose the verb, e.g., *Who did what?* or *What is the way things are?* (cf. Riester et al., 2017).

In the driving reports contrast serves an important argumentative function where benefits of a vehicle are often contrasted with its deficits. Discourse markers such as *but/aber* which (among other discourse relations) can mark a contrast may or may not occur in contrast constructions. Contrasts may also be marked between larger, super-sentential units. Disourse markers such as aber, allerdings, jedoch, etc. should be isolated in their own tag so that we can systematically search for these discourse relations in our corpus.

Q_1 Was ist mit dem Bentley Flying Spur?
Q_2 Was ist der Bentley Flying Spur?
Q_3 In wie fern ist der Bentley Flying Spur (widererwartend) kein Monument der Beharrung (verglichen mit anderen Limousinen)?
A Der wuchtige Bentley Flying Spur ist kein Monument der Beharrung sondern die schnellste Limousine der Welt.

According to Riester et al. (2017) the focussed constituent in a sentence is the answer to the sentence's QUD. Since we initially took a top-down approach in annotating QUDs, a lot of them are of the form in Q_1. However, in order to conform to the guidelines by Riester et al. (2017) we need to flesh out the QUD tree structure to go from the more abstract Q_1 to the more concrete Q_2. Q_2 conforms to the Riester guidelines and is an accurate QUD-representation of the focus structure, but Q_3 is perhaps a more accurate, abstract super-QUD which captures three nuances: (i) Contrary to a prior expectation the Bentley is not a *Monument der Beharrung*. Perhaps judging by the size, weight, and looks of the limousine, one would expect it to less mobile and fast. (ii) The discourse marker *sondern* marks the first counterargument against this prior expectation. Its speed is the one counterargument. (iii) The superlative *schnellste* marks an implicit speed comparison of the Flying Spur to all other limousines. None of the three nuances would have to be part of an answer to Q_2; an answer to Q_2 could simply be *Der Bentley Flying Spur ist die neueste Limnousine von Bentley*. So while Q_2 is a good QUD approximation of the answer's predicate subcategorization, Q_2 does not determine the actual content of the answer. In order to map to database queries the QUD needs to recognize that the answer should make a statement about the Bentley's speed, while the other nuances must come from a super-QUD such as Q_3. We could also say that although the sentence answers Q_2, the speed information is couched in discourse markers which signal a more complex QUD tree above Q_2. And it is the other branches of this more complex QUD tree which speak to the other nuances by way of modifying the surface realization of the answer with discourse markers.

4 Annotation procedure

For the text generator we need one XML QUD structure for each vehicle. However, we have annotated each driving report by two annotators in order to mitigate personal views. We therefore need to

1. make sure that annotators were consistent in their own annotation style,
2. compare the QUD tree structures for each report, and figure out where differences are due to different annotation styles or where they are due to true disagreement between annotators as to what the right QUD structure is (record true disagreement between annotators, regularize other differences), and
3. find one QUD structure for each report that we agree is the best compromise to be the input for the text generator.

In regularizing consistent annotation styles we need to keep in mind the following points (list in order of priority):

1. The focussed constituent <F> is the answer to a QUD. So there should be one QUD for each <F> and one <F> for every QUD. So when there are multiple <F> we need to decide whether they answer *different* QUDs or whether they are a *list* of things that together answer one QUD.
2. Make sure that tags were used consistently. For instance, make sure that <CON> always means background/contextual information and is not mistakenly used for CONjunction.

3. Check that pairings of discourse roles are consistent. Topics <AT> / background / context <CON> may not be verbalised, i.e. they may be ellided, but focus <F> / comment <CMT> should always present in each clause.
4. The distinction between <AT> and <CON>, or <F> and <CMT> may be irrelevant in most situations (where they simply represent different frameworks in the linguistic literature), but in the case of repetition information which has already been established in the text thus far may nevertheless be focussed material again. What the focussed constituents are is likely the easier decision. It's likely that there will be more nuance w.r.t. to non-focus material since it may have been established by a QUD higher up in the tree structure, but it may also be background information that goes beyond the text entirely (e.g., information about the manufacture or the history of a brand or model series) or be seemingly irrelevant (e.g., evoking the setting where the test drive took place, and immersing the reader in that atmosphere while not being relevant to any of the specific QUDs of the text).
5. When topic <AT> marks the same constituent as comment <CMT> would, and when focus <F> marks the same constituent as background <CON> would, use the following embedded structure:

```
<QUD>
      <AT>
           <CON>
                <SEGMENT></SEGMENT>
           </CON>
      </AT>
      <F>
          <CMT>
               <SEGMENT></SEGMENT>
          </CMT>
      </F>
<QUD>
```

6. Discourse markers such as allerdings, aber, jedoch, sondern, etc. and conjunctions such as und should be isolated in their own using the following structure:

```
<QUD>
      ⋮
      <F>
          <CMT>
               <SEGMENT>... (1) ... aber ... (2) ... </SEGMENT>
          </CMT>
      </F>
<QUD>
```

should look like this

<QUD>
⋮
<F>
<CMT>
... (1) ...
aber
... (2) ...
</CMT>
</F>
<QUD>

7. Check if non-at-issue content is correctly identified and marked with <NAI>.
8. Check if contrastive topic/focus is correctly identified and marked with <CT>.