Background

We will analyze some English texts. We assume that a discourse unit (a sentence, or a larger textual span formed by more than one sentence) has a topic, which can be identified by the question the discourse unit tries to answer. The question can be explicit, when there is a question in the text:

Sentence 1: Of all the ethnic tensions in America, which is the most troublesome right now?

Sentence 2: A good bet would be the tension between blacks and Jews in New York City. [wsj 2369]. *--- (example 1)*

For written texts, most of the time the question is implicit. To identify the topic of a discourse unit, you need to infer the question, and the discourse unit is considered to provide an answer to the implicit question:

1: Yesterday a jury of investigation came to the conclusion that the 31 casualties of the fire in the King's Cross London underground station died as the result of an accident and not as the result of negligence.

2: Relatives of the victims rejected it.

3: They are of the opinion that the jury did not do their job well.

4: Further prosecution of the officials of London Regional Transport is ruled out.

If implicit questions that each sentence answers are added:

F_1: *Yesterday a jury of investigation came to the conclusion that the 31 casualties of the fire in the King's Cross London underground station died as the result of an accident and not as the result of negligence.*

Q1: How did people react to the outcome of the investigation?

A1: Relatives of the victims rejected it.

Q2: Why?

A2: They are of the opinion that the jury did not do their job well.

Q3: What is the consequence of the outcome?

A3: Further prosecution of the officials of London Regional Transport is ruled out.

---(example 2)

The role of F_1 (Feeder):

Explicit and implicit questions do not arise without a cause. A feeder gives rise to the contextual induction of a question, such as F_1 above. As a feeder does not have any context, we do not infer a question for the feeder and put “What is the way things are?” as the question (Roberts 2012; Riester 2019).

Structural organization:

In example 2, it can be seen that Q2 is asked as a supplementary question to Q1, requesting more information. This means that Q1 is not answered satisfactorily yet at this step, and Q1 is the parent of Q2. After the answer A2 is given, no more questions are asked with respect to Q2 or its parent Q1 in the text, which suggests that the two questions have been answered satisfactorily. Therefore, the topics constituted by Q2 and Q1 are closed. Q3 forms a topic shift from discussion of people’s reaction to the consequence of the outcome of the investigation. Therefore, we can obtain a structural representation of the text:

The overall discourse topic Q0 is something like “How did people react to the outcome of the jury investigation of the casualties of the fire in the King’s Cross London underground station and what is the consequence of the outcome?” [1-4] (start sentence id-end sentence id).

Under Q0, two major topics exist: Q1 and Q3. Under Q1, a subtopic exists: Q2. If we reorganize the numeric reference of questions to reveal the hierarchical structure, the output is like:

Q1: [1-4]

Feeder: *Question: What is the way things are?*

Answer: [1] *Yesterday a jury of investigation came to the conclusion that the 31 casualties of the fire in the King’s Cross London underground station died as the result of an accident and not as the result of negligence.*

Q1.1 Question: *What is people’s reaction to the outcome of the investigation?*

Answer: [2-3]

Q1.1.1 Question: *How did people react to the outcome of the investigation?*

Answer: [2] *Relatives of the victims rejected it.*

Q1.1.1.1 Question: *Why?*

Answer: [3] *They are of the opinion that the jury did not do their job well.*

Q1.2 Question*: What is the consequence of the outcome?*

Answer: [4] *Further prosecution of the officials of London Regional Transport is ruled out.*

We distinguish two principles which guide the process of questioning:

A. Principle of recency: This principle indicates the order in which subquestions are contextually induced. It means that a subquestion Q_p is asked as the result of an answer A_{p-n}, which is the most recent unsatisfactory answer to a preceding question Q_{p-n}. In example 2, Q2 is asked because of A1, which is the most recent unsatisfactory answer to the preceding question Q1.

B. Principle of topic termination: A topic continues as long as subquestions are asked with respect to it. In other words, the occurrence of a subquestion indicates that a preceding question has not been answered satisfactorily yet. In example 2, Q2 is asked as a subquestion of Q1, which suggests that Q1 is not answered satisfactorily yet. When Q3 is raised and it is not a subquestion of Q2, it means that Q2 has been answered satisfactorily; and it is not a subquestion of Q1, it means that Q1 has been answered satisfactorily as well.

A method to test if a question has been answered satisfactorily---subordination test

We add a sentence that indicates the closure of the preceding topic, and if the following question becomes inappropriate, it means that the preceding topic is not closed yet, and the following question is subordinate to a previous question. The test sentence is added immediately after the answer to a question, and it typically has the following form: *I now understand (without discrepancy)...*

(a)

F_1: *Yesterday a jury of investigation came to the conclusion that the 31 casualties of the fire in the King's Cross London underground station died as the result of an accident and not as the result of negligence.*

Q1: How did people react to the outcome of the investigation?

A1: Relatives of the victims rejected it.

S: I now understand how people reacted to the outcome of the investigation.

Q2: Why?

A2: They are of the opinion that the jury did not do their job well.

Q3: What is the consequence of the outcome?

A3: Further prosecution of the officials of London Regional Transport is ruled out.

The test sentence *S* shows that Q1 has been fully addressed at that point. However, Q2 is raised, which indicates that this is not true and more information is required. Therefore, Q2 is subordinate to Q1.

(b)

F_1: *Yesterday a jury of investigation came to the conclusion that the 31 casualties of the fire in the King's Cross London underground station died as the result of an accident and not as the result of negligence.*

Q1: How did people react to the outcome of the investigation?

A1: Relatives of the victims rejected it.

Q2: Why?

A2: They are of the opinion that the jury did not do their job well.

S: I now understand how people reacted to the outcome of the investigation.

Q3: What is the consequence of the outcome?

A3: Further prosecution of the officials of London Regional Transport is ruled out.

When the test sentence *S* is added, the following question Q3 is still proper because Q3 and A3 do not talk about people's reaction anymore. Therefore, Q3 is not subordinate to Q2.

An example in dialogue:

(a) *F1* *A: Tomorrow is Harry's Birthday.*

Q1 *B: What would be a suitable birthday present for him?*

A1 *A: A monkey-wrench.*

S *B: I now understand what would be a suitable birthday present for Harry.*

****Q2*** ***B: What's a monkey-wrench?***

A2 *A: That's some kind of tool with which one can loosen or tighten nuts and bolts of various sizes.*

Q3 *B: Why would that be a suitable birthday present for him?*

A3 *A: He recently came to borrow one from me.*

When the test sentence *S* is added, it means that B has sufficient knowledge about Q1. However, Q2 is raised, and it shows that B is still not clear about some information that contributes to the clarification of Q1. Therefore, Q2 is subordinate to Q1.

(b) *F1* *A: Tomorrow is Harry's Birthday.*

Q1 *B: What would be a suitable birthday present for him?*

A1 *A: A monkey-wrench.*

Q2 *B. What's that?*

A2 *A: That's some kind of tool with which one can loosen or tighten nuts and bolts of various sizes.*

S *B: I now understand what would be a suitable birthday present for him.*

Q3** **B: Why would a monkey-wrench be a suitable birthday present for him?

A3 *A: He recently came to borrow one from me.*

Similarly, when the test sentence *S* is added, it means that B has sufficient knowledge about Q1. However, Q3 is raised, and it shows that B is still asking questions to get more information about Q1. Therefore, Q3 is subordinate to Q1.

(c) *F1* *A: Tomorrow is Harry's Birthday.*

Q1 *B: What would be a suitable birthday present for him?*

A1 *A: A monkey-wrench.*

Q2 *B. What's that?*

A2 *A: That's some kind of tool with which one can loosen or tighten nuts and bolts of various sizes.*

S *B: I now understand what a monkey-wrench is.*

Q3** **B: Why would a monkey-wrench be a suitable birthday present for him?

A3 *A: He recently came to borrow one from me.*

When the test sentence *S* is added, it means that B has sufficient knowledge about Q2. The following question Q3 is not about Q2, and given the test sentence, Q3 is still proper. This suggest that Q3 is not subordinate to Q2.

Task overview

We start with reading a text, and then identify what question the text is trying to answer (Q1) in a broad sense, which is called a discourse topic (Q1). Under Q1, a text may be divided into some sections, with each section comprised by one or more sentences working together to answer a question, which is more specific than Q1, and therefore, this set of questions are denoted as Q1.1, Q1.2, Q1.3 etc.. As indicated, under each section, there may be one or more sentences. When more sentences are involved, it is possible that these sentences can be further grouped when they work together to answer a question under Q1.1, for instance. This set of questions is denoted as Q1.1.1, Q1.1.2...Q1.2.1, Q1.2.2 and so on.

Example: (wsj_0652)

1 Wilfred American Educational Corp. said a federal grand jury in Boston indicted the operator of cosmetology and business schools for mail fraud.

2 The charges in the 12-count indictment, which stem from events that allegedly occurred in late 1984 and early 1985, involve enrollment procedures of six students and the preparation of certain reports, Wilfred said.

3 No individuals were charged in the indictment.

4 Wilfred American said it will "vigorously defend" itself against the charges and added that the charges relate to procedures that it has since changed.

5 Eight admissions representatives at two of Wilfred's former Massachusetts schools previously pleaded guilty to charges of aiding, abetting and counseling students to submit false financial-aid applications.

6 Wilfred closed its Massachusetts schools earlier this year.

7 In New York Stock Exchange composite trading Friday, Wilfred fell 6.25 cents to 93.75 cents a share.

Reference analysis:

Q1: What legal issues is Wilfred American Educational Corp. facing and what actions are being taken in response? [1, 7]

├── Q1.1: What are the specifics and implications of the indictment? [1, 3]

│ ├── Q1.1.1: What is the way things are? [1]

│ ├── Q1.1.2: What are the specifics of the indictment? [2]

│ └── Q1.1.3: Are there any people involved in the indictment? [3]

├── Q1.2: How does the company plan to handle the charges? [4]

├── Q1.3: What prior legal charges and operational changes occurred that relate to the indictment? [5, 6]

│ ├── Q1.3.1: What are previous legal issues at Wilfred American? [5]

│ └── Q1.3.2: What decision did Wilfred make regarding its Massachusetts schools? [6]

└── Q1.4: What was the financial repercussion of the indictment? [7]

Criteria to meet:

1. Sentences under the same topic should be adjacent. This means that for a topic, the start sentence id and end sentence id cover the whole span, for instance, a span [1-3] means all the sentences from 1 to 3 work together to answer the same question.
2. Implicit questions for each sentence should be inferred **only based on the preceding context**. Therefore, questions cannot be a simple conversion of the sentences they dominate or leak words in the answer sentences. For instance, the question for sentence 7 CANNOT be “What is the stock price of Wilfred in New York Stock Exchange composite trading Friday?”, because “New York Stock Exchange composite trading Friday” and “stock price” are not given in the preceding context.
3. An implicit question should be answerable by the sentence it dominates. For sentence 6 in the example, a question that **violates** this criterion would be something like “How did these admissions representatives defend themselves against these charges?”. Although this question is derived from the previous context, sentence 6 does not provide an answer to this question.
4. An implicit question should contain as more information of the preceding context as possible. For example, the question identified for sentence 6 in the above example is better in its current state than “What did Wilfred do?”.

A longer text: (wsj_1399)

1 Texaco Inc. has purchased an oil-producing company in Texas for $476.5 million, its first major acquisition since its legal brawl with Pennzoil Co. began more than four years ago.

2 The White Plains, N.Y., oil company said Friday that it had acquired Tana Production Corp., a subsidiary of TRT Energy Holdings Inc., for $95.1 million in cash, with the rest to be paid in shares of a new, non-voting issue of preferred stock.

3 Tana, which holds properties in 17 oil and gas fields in south Texas, will provide Texaco with mostly gas reserves.

4 The fields contain recoverable reserves of 435 billion cubic feet of natural gas and four million barrels of oil.

5 "This acquisition is another indication of Texaco's commitment to increase the company's reserve base," said Chief Executive Officer James W. Kinnear.

6 Texaco has also been attempting to sell oil properties.

7 At least two years ago, the company put 60 million barrels of oil reserves on the block.

8 They were either too small or uneconomic to maintain, the company said.

9 Not all of those parcels have yet been sold.

10 Texaco acquired Tana before it completed those sales because Tana's properties are high quality and near other fields Texaco already owns, a company spokeswoman said.

11 Texaco, like many other oil companies, has been struggling to replace its falling oil and gas reserves.

12 Texaco's situation had become particularly complex because much of its effort had for years been focused on its brawl with Pennzoil and then on New York investor Carl C. Icahn's attempt to take over the company.

13 Pennzoil had sued Texaco for improperly interfering with its acquisition of a portion of Getty Oil Co.

14 Eventually, Texaco, which was forced into bankruptcy proceedings by that litigation, settled its fight with Pennzoil for $3 billion in 1986.

15 Mr. Icahn, who played a key role in the settlement and attempted subsequently to take control of the company, sold his stake in Texaco just last summer.

16 Completion of Texaco's acquisition of Tana is subject to government approval under the Hart-Scott-Rodino Antitrust Improvements Act.

Reference analysis:

Q1: What are the details of Texaco Inc.'s acquisition of an oil-producing company and why does it happen? [1, 15]

├── Q1.1: What does the acquisition involve? [1, 4]

│ ├── Q1.1.1: What happens with the acquisition? [1, 2]

│ │ ├── Q1.1.1.1: What is the way things are? [1]

│ │ └── Q1.1.1.2: How will the acquisition be implemented? [2]

│ └── Q1.1.2: What is the situation with the acquired company? [3, 4]

│ ├── Q1.1.2.1: What is the general information of the acquired company? [3]

│ └── Q1.1.2.2: What do these fields do? [4]

├── Q1.2: What benefits does the acquisition bring to Texaco? [5, 10]

│ ├── Q1.2.1: What is the implication of the acquisition? [5]

│ ├── Q1.2.2: Why is the acquisition important? [6, 10]

│ │ ├── Q1.2.2.1: What business does Texaco have that requires increasing the reserve base? [6, 9]

│ │ │ ├── Q1.2.2.1.1: What is the background of Texaco in this business? [6]

│ │ │ └── Q1.2.2.1.2: What difficulties does Texaco face? [7, 9]

│ │ │ ├── Q1.2.2.1.2.1: What did Texaco do for this business? [7]

│ │ │ ├── Q1.2.2.1.2.2: What is the result? [8]

│ │ │ └── Q1.2.2.1.2.3: What is the current state of the barrels of oil reserves? [9]

│ │ └── Q1.2.2.2: What kind of benefits does the acquisition of Tana bring to Texaco? [10]

├── Q1.3: What particular difficulties does Texaco face? [11, 15]

│ ├── Q1.3.1: Why does Texaco need Tana's properties in particular? [11, 12]

│ │ ├── Q1.3.1.1: Why does Texaco need Tana's properties? [11]

│ │ └── Q1.3.1.2: What other reasons make Texaco's situation difficult? [12-15]

├──Q1.3.1.2.1: What other difficulties did Texaco face? [12]

│ │ ├── Q1.3.1.2.2: What happened with Texaco’s brawl with Pennzoil? [13, 14]

│ │ │ ├── Q1.3.1.2.1.1: What was the brawl with Pennzoil about? [13]

│ │ │ └── Q1.3.1.2.1.2: What was the result? [14]

│ │ └── Q1.3.1.2.3: What happened with Carl C. Icahn's attempt to take over the company? [15]

└── Q1.4: What is left to be done with the acquisition procedure? [16]

Annotation format:

You can choose any form in your annotation, as long as the structure is clear. A reference template:

1-7: ******?

1-3: ******?
1: *****?
2: *****? (3 subordinate to 2)
3: *****?
4-7: ******?
4: *****?
5: *****?
6-7: ****? (6 and 7 are coordinated questions)
6: ****?
7: ****?

There are mainly two types of relations between questions (Van Kuppevelt, 1996):
----subordinating (shown as 3 to 2 above): question 3 is derived from question 2, or question 3 is a subquestion of question 2. One question is more important or salient than the other.
----coordinating (shown as 6 and 7 above): questions are parallel to each other, such as question 1-3 and question 4-7 above. The questions are equally salient.

However, in some very rare cases, one segment may tell a part about a question, and another segment focuses on another part of the same question. It is difficult to formulate a different question for segment 2 other than using words such as “What about other….”. In this case, you can show the annotation as (Riester, 2019):
2-3: *****?
2:
3:

Difference between higher-level questions and sentence-level questions:

As you may notice during the annotation process, questions for sentences can be derived from the preceding context most of the time, but you need to know the content of a span to drive higher-level questions. For higher-level questions, you do not need to follow the rule that “implicit questions should be inferred only based on the preceding context”.

References:

Nicholas Asher and Alex Lascarides. Logics of Conversation. Cambridge University Press, 2003.

Riester, A. (2019). Constructing QUD trees. In *Questions in discourse* (pp. 164-193). Brill.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, *5*, 6-1.

Van Kuppevelt, J. (1995). Discourse structure, topicality and questioning. *Journal of linguistics*, *31*(1), 109-147.

Van Kuppevelt, J. (1996). Directionality in discourse: Prominence differences in subordination relations1. *Journal of semantics*, *13*(4), 363-395.