Ranking of Potential Questions

Luise Schricker
Department of Linguistics
University of Potsdam
Germany
luise.schricker@uni-potsdam.de

Tatjana Scheffler
Department of Linguistics
University of Potsdam
Germany
tatjana.scheffler@uni-potsdam.de

Abstract

Questions are an integral part of discourse. They provide structure and support the exchange of information. One linguistic theory, the Questions Under Discussion model, takes question structures as integral to the functioning of a coherent discourse. This theory has not been tested on the count of its validity for predicting observations in real dialogue data though. In the present study, a system for ranking explicit and implicit questions by their appropriateness in a dialogue is presented. This system implements constraints and principles put forward in the linguistic literature.

1 Introduction

Questions are important for keeping a dialogue flowing. Some linguistic theories of discourse structure, such as the Questions under Discussion model (Roberts, 2012, and others), view questions and their answers as the main structuring element in discourse. As not all questions are explicitly stated, the complete analysis of a discourse in this framework involves selecting adequate implicit questions from the set of questions that could potentially be asked at any given time. Correspondingly, a theory of discourse must provide constraints and principles by which these potential questions can be generated and ranked at each step in the progression of a discourse.

As a first move towards putting this linguistic model of discourse structure into practice, we implemented a ranking system for potential questions. Such a system might be used to investigate the validity of theoretic claims and to analyze data in order to enrich the theory with further insights.

The given task is also relevant for practical considerations. A system for ranking potential questions, i.e. questions that are triggered by some assertion and could be asked in a felicitous discourse, is a useful component for applications that generate dialogue, such as chatbots. At some point in a dialogue, several questions could be asked next and the most appropriate one has to be determined, for example by using a question ranker.

Q0: *What is the way things are?*
- Q1: **What did you eat for lunch?**
– A1: **I ate fries,**
– Q1.1: *How did you like the fries?*
— A1.1: **but I didn't like them at all!**
— Q1.1.1: **Why?**
—- A1.1.1: **They were too salty.**
— Q1.1.2: *What did you do?*
—- A1.1.1: **So I threw them away.**

Figure 1: Constructed example of a QUD annotated discourse. Explicit questions and answers are marked in bold typeface. Implicit questions are set in italic type.

2 Background

2.1 The Questions-Under-Discussion Model

In 1996[1], Roberts (2012) published a seminal paper describing a framework that models discourse as a game. This game allows two kinds of moves, questions and assertions. The questions that have been accepted by the participants, also referred to as *questions under discussion* (QUDs), provide the structure of a discourse. An example discourse annotated with a question-structure is shown in Figure 1. The overall goal of the game is to answer the big question of *how things are*. The question structure is given by explicit questions that are proffered and accepted by the participants and implicit questions that can be accommodated.

We follow the variant by Riester (2019), who developed the QUD framework further and formalized the model. Riester models the question

[1] Here, we cite the reissued 2012 version.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 143–148
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

structures as *QUD trees* and introduces the notion of assertions that trigger subsequent questions. Following Van Kuppevelt (1995), he refers to such assertions as *feeders*. Furthermore, Riester introduces three constraints on the formulation of implicit QUDs in coherent discourse. These constraints ensure that modeled discourses are well-formed. The first constraint, **Q-A-Congruence**, states that the assertions that are immediately dominated by a QUD must provide an answer for it. The second constraint, **Q-Givenness**, specifies that implicit QUDs can only consist of given or highly salient material. Finally, the third constraint, **Maximize-Q-Anaphoricity**, prescribes that as much given or salient material as possible should be used in the formulation of an implicit QUD. Implicit questions are therefore constrained by both the previous discourse and the following answer.

The notion of questions triggered by feeders was strengthened by Onea (2013) who introduces the concept of *potential questions* within a QUD discourse structure. This concept refers to questions that are licensed by some preceding discourse move. That move can be a question, but also an assertion. Depending on the context, some potential questions are more appropriate than others. In chapter 8, Onea addresses this observation by describing a number of generation and ordering principles, which are listed below. In this paper, we implement Riester's Q-Anaphoricity constraint and Onea's potential question principles as features for a question ranker, allowing us to test them on naturally occurring dialog.

2.2 Generation Principles

Follow formal hints Certain linguistic markers trigger the generation of potential questions, e.g. appositives, indefinite determiners and overanswers.

Unarticulated constituents Whenever constituents in an assertion are not articulated, questions about these constituents are generated.

Indexicals For every assertion, questions about unspecified indexicals are generated.

Rhetorical relations Any assertion licenses typical questions related to rhetorical relations, e.g. questions about the result, justification, elaboration, and explanation.

Parallelism and contrast For any question in the discourse, parallel or contrastive questions that are triggered by a following assertion should be generated as potential questions.

Animacy hierarchy Every time a human individual is introduced into the discourse, questions about this individual should be generated.

Mystery Questions about surprising objects or events that enter the discourse should be generated.

2.3 Ordering Principles

Strength Rule The Strength Rule states that more specific questions are generally better (i.e., more coherent) than less specific ones.

Normality Rule The Normality Rule predicts that a question triggered by a normal or common context is better than a question triggered by an unusual context.

2.4 Question Ranking

While the described work by Roberts, Riester, and Onea is purely theoretical, other research is practically concerned with the ranking of questions. This research does not consider the notion of potential questions though and can therefore not offer a direct point of comparison for the present study. Heilman and Smith (2010), for example, present a system for automatically generating questions from a given answering paragraph. The system overgenerates questions, which are subsequently ranked regarding the questions' acceptability given the answering text. In contrast to this, the system described in the present paper considers the assertion *preceding* the question, rather than the answer, when determining a question's felicity in discourse.

3 System

In order to investigate which role the linguistic constraints and principles play in practice, we implemented a ranking system based on the theoretical insights. The system takes an assertion and a set of potential questions triggered by this assertion as input and ranks the set of potential questions by appropriateness, given the preceding assertion.

3.1 Data

The task of implementing a system for ranking potential questions is difficult, as no datasets exist that fulfils the requirements of the input. To circumvent this problem, the required data was approximated by different data extraction schemes. We used two corpora: the test set was extracted from a small manually annotated corpus of interview fragments. The training set was mined from the Switchboard Dialog Act corpus.

The test corpus consists of eight short texts. The texts are copies of a segment of an interview with Edward Snowden[2] that was annotated with a QUD structure like the one in Figure 1 by students of the class *Questions and Models of Discourse*, held at the University of Potsdam in 2018.[3] Some preprocessing, manual and automatic, had to be done in order to ensure a consistent structure amongst the texts. The interviews were segmented into assertions and explicit and implicit questions. Assertions are often not complete sentences and the segmentation differs between the individual texts.

We extracted every assertion that was followed by a question, explicit or implicit, together with the following question. The three preceding and three next questions were saved as an approximation of the set of alternative potential questions.[4] We deemed this acceptable because it is likely that the immediately surrounding questions will be about similar topics as the assertion. The question immediately following the assertion was regarded as the correct label, i.e. the question that should be ranked highest by the system. Items that contained the same assertions were merged, which resulted in several correct labels, and a larger set of alternative potential questions per assertion.

As the test set was not sufficiently big to use for training in a machine learning setting, a second dataset was extracted from the Switchboard Dialog Act corpus (SWDA)[5] (Stolcke et al., 2000). The SWDA corpus contains spontaneous telephone conversations that are annotated with dialog acts. The reasoning behind using this corpus was that a question following an assertion in a dialog can be interpreted as the highest ranked potential question available at that point.

Similar to the extraction of the test set, assertions directly followed by a question were extracted along with the question. We considered only prototypical types of assertions and questions[6], excluding for example rhetorical questions, to avoid inconsistent items. For each item, three questions were randomly picked from the set of all questions in the corpus to arrive at a set of approximate alternative potential questions. The individual questions and assertions were cleaned from disfluency annotation. The resulting training set consists of 2777 items.

3.2 Feature Extraction

In this work, we implemented a subset of Onea's (2013) generation and ordering principles and Riester's (2019) QUD constraints as features for ranking a question following a preceding assertion. For linguistic processing spaCy (Honnibal and Montani, 2019) (e.g. dependency parsing, named entity recognition and POS tagging), NLTK (Bird et al., 2009) (wordnet, stopwords) and neuralcoref[7] (coreference resolution) were used. For features using word embeddings, a pretrained Word2vec model[8] (Mikolov et al., 2013a,b) was used, the model was handled via the gensim package (Řehůřek and Sojka, 2010). Below, the implemented features are described.

Indefinite Determiners This feature detects indefinite noun phrases in the assertion that are coreferent to some mention in the question.

Indexicals This feature analyzes whether the question is about a time or a place by searching for question phrases that inquire about a time or a place (e.g. *when*, *where* etc.).

Explanation Following Onea (2019), who draws parallels between certain patterns in discourse trees with question structures and rhetorical relations, the rhetorical relation *Explanation* is detected by searching for *why*-questions.

Elaboration The rhetorical relation *Elaboration* is linked by Onea (2019) to questions that ask

[2] https://edwardsnowden.com/2014/01/27/video-ard-interview-with-edward-snowden/

[3] The raw annotated files can be accessed under: https://github.com/QUD-comp/analysis-of-QUD-structures/tree/master/Snowden

[4] Incomplete datapoints from the start and end of a document, which were followed or preceded by fewer than three questions, were excluded.

[5] The version distributed by Christopher Potts (https://github.com/cgpotts/swda) was used, as well as the code he provided for better accessibility of the corpus.

[6] List of considered assertion tags: ['s', 'sd', 'sv'] (statements with or without opinions); list of considered question tags: ['qy', 'qw', '^d', 'qo', 'qr', 'qw^d'] (different syntactic sub-types of questions).

[7] https://github.com/huggingface/neuralcoref

[8] https://code.google.com/archive/p/word2vec/

about an explicit or unarticulated constituent in an assertion with a *wh*-question phrase. This is implemented by checking the question for *wh*-question phrases that enquire about properties of some NP (e.g. *which*, *what kind* etc.) and that are used in a non-embedded sentence.

Animacy This feature detects mentions of persons, i.e. named entities or words that belong to the Wordnet synset *person*, in the assertion and checks whether any of these are coreferent to mentions in the question.

Strength Rule I This method approximates the specificity of the question as the relation between the length of assertion and question. A question much shorter than the assertion is likely unspecific, a question much longer might talk about something else and therefore also lose specificity.

Strength Rule II Questions specific to an assertion are likely to be semantically similar to the assertion. Following this observation, the feature approximates specificity as the cosine similarity of the word vector representation of the assertion to the representation of the question. These representations are computed by adding the word vectors for the individual words.

Normality Rule This feature checks the normality of a context by first computing separately the average cosine similarities of the words within the question and within the assertion. Unexpected words in a sentence should have a lower similarity score than expected words when compared to the rest of the sentence. For example, the words *sandwich* and *ham* should have a higher similarity score than the words *sandwich* and *screws*, giving the phrase *a sandwich with ham* a higher normality score than the phrase *a sandwich with screws*. In a second step, a ratio of the normality scores of the assertion and the question is computed. If the assertion talks about an unnormal context it is normal for the question to relate to this.[9] Overall, the closer the score is to 1.0, the more normal the context of the question is, given the assertion.

Maximize Anaphoricity This method counts mentions in the question that are coreferent to something in the assertion and string matches between question and assertion that were not already counted as coreference mentions.

Assertion: ”It was the right thing to do.”
Potential questions:
”When was this your greatest fear?”
”But isn’t there anything you’re afraid of?”
”Why don’t you lose sleep?”
”Was it the right thing to do?”
”But are you afraid?”
”Mr. Snowden, did you sleep well the last couple of nights?”
”Is this quote still accurate?”

Figure 2: Example input for the potential question ranker from the test set. The correct following question is marked in italic.

3.3 Ranking Component

The ranking component takes an assertion and a list of potential questions as input (see Figure 2 for an example input), transforms every assertion-question pair into a feature representation, and ranks the questions based on this representation. Three modes of ranking are possible. The *Baseline* mode shuffles the questions randomly and returns them in this order.

The *Uniform* mode transforms every assertion-question pair into a scalar representation by adding up the individual features. All features based on Onea’s generation principles return either 0 or 1, depending on whether the feature is present or not. Strength Rule I and the Normality Rule should return a value as close to 1.0 as possible for a high ranking question. Therefore, the absolute distance of the return value from 1.0 is subtracted from the representation. Strength Rule II and the Maximize Anaphoricity feature return continuous values. These are also added to the scalar representation. The questions are sorted by the value of the feature representation.

The *ML* (short for machine learning) mode accumulates features for an assertion-question pair into vector representations which are fed into a Random Forest classifier. The choice of using a Random Forest classifier was motivated by the amount of available training data and by considerations about transparency. Decision Trees are usually a good option for small training datasets and it’s easy to analyze the patterns they learn by inspecting feature importance. Scikit-learn’s (Pedregosa et al., 2011) Random Forest implementation was used. A grid search was performed on a small set of parameters to arrive at an optimal con-

[9] Imagine the following conversation: A: *”I had a sandwich with screws yesterday.”* B: *”A sandwich with screws??”* (example adapted from (Onea, 2013)). In this context, it would be rather unnormal if B did not ask about the screws.

Mode	Top-1	Top-3	Top-5
Baseline	19.23	46.15	65.38
Uniform	38.46	61.54	**88.46**
ML	**50.00**	**73.08**	80.77

Table 1: Results on test set. *Top-N* signifies the stage of evaluation.

figuration, results for the best configuration[10] are detailed in Table 1.

4 Evaluation

The different ranking modes and classifier configs were evaluated on the test set extracted from the annotated Snowden interview. In order to get a more detailed insight into the performance of the ranking system, evaluations were done in three stages. To this end, the *Top-N accuracy* measure was used. As a result of the merging of items described in section 3.1, the average number of questions that are correct if ranked in first place is three per item, and the average number of potential questions available for ranking is 21 per item. Results are listed in Table 1.

Interestingly, the uniform mode works quite well, providing the best result in the easiest evaluation setting with an accuracy score of almost 90%. The overall best ranking system (ML mode) achieves an accuracy of 50% for ranking a correct label highest and a score of 73% for placing a correct label amongst the top three ranks. These numbers show improvements of over 30 and over 25 points compared to the random baseline.

It should be noted that the data in the training and test sets have different properties. While the training data is built from spontaneous dialogue, the test set contains QUD annotations that were added in hindsight and that are sometimes not phrased like natural speech. Training and test sets that are more similar might therefore provide better results. This experiment should be repeated in the future if a reasonably sized QUD-annotated corpus becomes available.

Furthermore, the random baseline is quite simple and might be too easy to beat. An anonymous reviewer suggested implementing a deep learning model trained for next sentence prediction as an additional baseline. While we agree that this would be worthwhile, we have to leave it for future work due to time constraints.

Type	Utterance
Assertion	*There are significant threats*
Question	*Are there significant threats?*
Assertion	*"The greatest fear I have," and I quote you, "regarding these disclosures is nothing will change."*
Question	*But isn't there anything you're afraid of?*

Table 2: Examples of questions incorrectly ranked in top place that the assertion already answers

An additional inspection of the best performing Random Forest model's features by importance showed the three ordering constraint features, Strength Rule II, the Normality Rule and Strength Rule I in top position. This confirms the theoretic background: the ordering principles should be more important for ranking potential questions than the generation principles.

4.1 Error Analysis

In order to better understand the failings of the ranking system, the best configuration was inspected more closely in an error analysis. The most prominent error by far is ranking a question that the assertion already answers highest, instead of one that is triggered by the assertion. Some examples of this type of error are listed in Table 2.

This can be explained by the nature of the training data. As alternative potential questions were sampled randomly from the data during training, they are more likely to be about a different topic than the assertion compared to the correct question, which would enhance the importance of similarity features like Strength Rule II. An answer to a question can be as similar to the question as the assertion directly preceding a question. In a real application, questions that are answered by the preceding assertion should not be part of the set of potential questions that are fed into the system, though.

5 Conclusion

Potential questions are a concept stemming from theories that organize discourse around questions. A ranking system[11] based on these theories was

[10] The best configuration has *min_samples_leaf* = 5, *max_depth* = 10, and *class_weight* = {0:0.5, 1:1}.

[11] The code and data presented here have been made available for public use under a GPL-3.0 license: https://github.com/QUD-comp/ranking-potential-questions.

able to improve rankings of a small test dataset by up to 30 percentage points compared to a random Baseline. This system is a first step towards an implementation of the until now theoretic but influential QUD discourse model. It might be of help for further evaluation and enrichment of these linguistic theories, but might also be useful in dialogue generation applications, e.g. for machine dialogue systems and chatbots.

Acknowledgements

We would like to thank the participants of the 2018 University of Potsdam class *Questions and Models of Discourse* for annotating the interview segments that were used as test data. Furthermore, we would like to thank Bonnie Webber for the detailed and extremely valuable mentoring feedback. Finally, we are grateful to the anonymous reviewers for their helpful comments.

References

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2019. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Edgar Onea. 2013. *Potential questions in discourse and grammar*. Habilitation thesis, University of Göttingen.

Edgar Onea. 2019. Underneath rhetorical relations. the case of result. In K. v. Heusinger, E. Onea, and M. Zimmermann, editors, *Questions in Discourse*, volume 2. Brill, Leiden.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Arndt Riester. 2019. Constructing QUD trees. In Klaus v. Heusinger, Edgar Onea, and Malte Zimmermann, editors, *Questions in Discourse*, volume 2. Brill, Leiden.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

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