Toward Dialogue Modeling: A Semantic Annotation Scheme for Questions and Answers

María Andrea Cruz Blandón, Gosse Minnema, Aria Nourbakhsh
Maria Bortichev, Maxime Amblard
LORIA, UMR 7503, Université de Lorraine, CNRS, Inria
Nancy, France
{mariaandrea.cruzblandon, gosseminnema}@gmail.com
aria.nourbakhsh@outlook.com
{maria.boritchev, maxime.amblard}@univ-lorraine.fr

Abstract

The present study proposes an annotation scheme for classifying the content and discourse contribution of question-answer pairs. We propose detailed guidelines for using the scheme and apply them to dialogues in English, Spanish, and Dutch. Finally, we report on initial machine learning experiments for automatic annotation.

1 Introduction

Question-answer pair (QAP) labeling is the problem of characterizing the content and discourse contribution of questions and answers using a small but maximally informative tagset that can be consistently applied by both human annotators and NLP systems. QAP labeling has many potential use cases, for example as a preprocessing step for dialogue modeling systems or for chatbots. The problem is not new: in the NLP literature, different aspects of QAP tagging have been addressed in the context of question answering systems (Li and Roth, 2002), question generation systems (e.g. Graesser et al., 2008), and dialogue act classification (e.g. Allen and Core, 1997; Stolcke et al., 2000).

However, we see several gaps in the literature: existing approaches to QAP classification often do not cover the full range of questions and answers found in human dialogues and are limited in the types of semantic information that they cover. To address these issues, we propose a new annotation scheme that was developed based on corpora of natural conversations in several languages (English, Spanish, and Dutch) and provides several layers of annotations for QAPs. Notably, where applicable, we annotate the semantic role of the questioned constituent in questions and their corresponding answer (e.g. 'Does she live in *Paris or London?*' ⇒ LOCATION), which we believe is an informative, yet easy definable way of globally characterizing the content of a QAP.

Our paper has two main contributions: the annotation scheme itself (section 3) and two ways of applying it to real data. We developed detailed and explicit guidelines for human annotators, and tested these on corpus data (section 4.1). Additionally, we started experimenting with machine learning approaches for automating part of the annotation process (section 4.2).

2 Related Work

Our annotation scheme is related to two existing schemes in particular. The first of these is Freed (1994), which categorizes questions along an *information continuum* that ranges from questions purely asking for factual information to questions that convey, rather than request, (social) information. Within this continuum, questions are divided into classes that are defined based on a combination of formal (syntactic) and functional criteria. Both of these ideas are also used in our scheme: our question types are also distinguished by whether they ask or convey information ('phatic questions' and 'completion suggestions' fall into the latter category) and are defined as combinations of specific forms and functions.

Another related scheme is Stolcke et al. (2000), an adapted version of DAMSL ('Dialog Act Markup in Several Layers', Allen and Core 1997), an annotation scheme for dialogue acts (including QAPs). The scheme includes a set of eight different question types (e.g. *yes/no* questions, *wh*-questions, *rhetorical* questions) that has considerable overlap with our set of question types.

3 Annotation scheme

Annotated information is split between two main 'layers': *question/answer type* and *feature* (se-

Proceedings of the 13th Linguistic Annotation Workshop, pages 230–235
Florence, Italy, August 1, 2019.

mantic role). Every question or answer is assigned at least a type tag, and depending on the type, a feature tag.

3.1 Questions

The question tagset was designed in a corpus-driven way, starting with two basic types and expanding the tagset based on corpus data. Our starting assumption is that the corpora would contain at least two well-known and well-defined categories of questions: *yes/no* questions and *wh*-questions (Freed, 1994). In our opinion, both of these types are useful *a priori*, because they are each associated with a clear set of syntactic, semantic, and pragmatic characteristics (at least for the languages that are included in this study). Prototypical English *yes/no* questions are characterized by subject-auxiliary inversion and do-support (syntax), express a proposition that could be true or false (semantics), and their answers are expected to either confirm or deny this proposition (pragmatics). On the other hand, a prototypical English *wh*-question contains a fronted constituent that starts with a *wh*-word (syntax), expresses a proposition with missing information (semantics), and expects the answerer to supply this missing information (pragmatics) (Freed, 1994).

Next, we looked for questions in our corpora that did not correspond to either of the two prototypes and extended the scheme to fit them (see table 1 for the final scheme and examples). First, there are questions that are similar to *wh*-questions or *yes/no* questions but have a deviant form (e.g. *wh-in-situ* questions like 'You saw what?', or yes/no questions without inversion such as 'You saw him?'). We decided not to introduce new categories for these on the basis of their semantics and pragmatics.

A second group of questions has the syntactic characteristics of a *yes/no* question or a *wh*-question, but a different pragmatics and/or semantics. For example, the asker of the question suggests a way to complete the utterance of the previous speaker, and the expected answer would confirm or deny this suggestion. This is subtly different from a prototypical *yes/no* question because the asker of the question does not necessarily ask their interlocutor to confirm the truth value of the suggestion (e.g. *A: it includes heat and uhm, I think B: Water?*, SCoSE/Amy, line 746-747[1]). We call these types of questions *completion suggestions*.

Tag	Name	Tag	Name
YN	Yes/No question	WH	Wh-question
CS	Completion suggestion	PQ	Phatic question
DQ	Disjunctive question		

Table 1: Question types

The third group of questions appear to be a *yes/no* question or a *wh*-question, respectively, but their context and intonation make clear that the asker is not actually interested in the confirmation or denial of the proposition. Instead, such questions can have various so-called *phatic functions*, i.e. their semantic content is less important than their social and rhetorical functions (Freed, 1994; Senft, 2009). We call this type of questions *phatic questions* (e.g. *right? / oh yeah? / you know?*).[2]

Finally, some questions containing a disjunction (e.g. 'Do you go on Monday or on Tuesday?') are semantically and pragmatically similar to *wh*-questions, but are syntactically closer to *yes/no* questions. This kind of questions, like *yes/no* questions, exhibits subject-auxiliary inversion (at least in English), but does not ask for the confirmation or denial of the proposition that it expresses. Instead, it expects the answerer to provide some missing information with the set of options to choose from. We call this type of questions *disjunctive questions* (sometimes also called *alternative questions* in the literature).

3.2 Features

Wh- and *disjunctive* questions are always 'about' a particular constituent (e.g. '*Which man* is running?', 'Do you want *coffee or tea*?'). The *feature*, or semantic role of this constituent provides information about the content of the question and the expected answer (e.g. if the questioned constituent is an AGENT then it is likely that the answer will refer to a person). Detecting semantic roles requires semantically analyzing the sentence, but for *wh*-questions, *wh*-words often provide cues (e.g. 'where' for LOCATION). Our feature annotations follow the feature set (see table 2) and the

[1] See section 4.1.2 for information about our corpora.

[2] Note that our use of the term *phatic question* is somewhat broader than the *phatic information* question described in Freed (1994); for example, our definition also includes rethorical questions, while in Freed's scheme, these are not included.

mapping from (English) *wh*-words to features proposed in Boritchev (2017) (adapted from Jurafsky and Martin 2000).

Tag	Name	Tag	Name
TMP	Temporality	OW	Owner
LOC	Location	RE	Reason
AG	Agent	TH	Theme
CH	Characteristic		

Table 2: Features

3.3 Answers

The main intuition underlying our answer annotation scheme is that question types restrict their answers: for example, *yes/no* questions are prototypically answered by 'yes' or 'no', and *wh*-questions ask for a constituent with a particular feature. Table 3 summarizes our answer types and their corresponding question types. Among these types of answers, there may be overlaps. For example, a 'deny the assumption' answer can be thought of as a negative answer because it is possible that they share the same grammatical and semantic structure. Different factors including the context and prosody are relevant to decide overlapping tags.

Some questions are not followed by answer. We distinguish between two situations. First, there are questions that receive a reply that, while not providing the information asked for in the question, clearly do respond to it. For example, in the QAP *A: 'When will you guys get off?' / B: 'My last exam is like . . . I don't know'* (SCoSe/Amy, line 243-244), B's response does not answer A's question directly but does engage with it as there is a logical connection between finishing the exams and going on vacation. In such cases, the response is tagged as *unrelated topic (UT)* because it is about a different topic but still responds to the question. By contrast, when there is no response at all, no answer should be annotated.

4 Annotation Experiments

In this section, we discuss our experiments with applying the scheme manually (section 4.1) and using machine learning techniques (section 4.2).

4.1 Manual annotation

We have experimented with applying the scheme on real-world data. Our experiment consists of two parts: writing annotation guidelines to explicitly define the annotation process and annotating 701 questions across three languages, namely, English, Spanish, and Dutch.[3]

Tags	Name	Question Type
PA	Positive Answer	YN, CS
NA	Negative Answer	YN, CS
FA	Feature Answer	DQ, WH
PHA	Phatic Answer	YN, CS, DQ, WH, PQ
UA	Uncertainty Answers	YN, CS, DQ, WH, PQ
UT	Unrelated Topic	YN, CS, DQ, WH, PQ
DA	Deny the Assumption	YN, CS, DQ, WH, PQ

Table 3: Answers

4.1.1 Annotation guide

In order to help annotators apply the scheme consistently, we wrote annotation guidelines for English, which include examples and instructions for how to use the annotation software (ELAN 2017, Sloetjes and Wittenburg 2008). The annotation procedure guides the annotator in identifying questions, dealing with transcription errors, determining question types, and adding tags for additional information such as features, complexity, and indirectness.

Some question types have a very specific prototypical syntactic form (e.g. *wh*-questions), whereas other questions can have several different forms (e.g. *phatic* questions). We exploit this by defining a precedence order for question types, which serves as a filter for identifying questions. The precedence order lists question types from the most specific to the most general ones, i.e. from questions with easily identifiable characteristics to those that can have different forms as it is the case for the phatic questions. The precedence order is as follows: (1) *Wh-questions*, (2) *Disjunctive questions*, (3) *Yes/No questions*, (4) *Completion suggestions* (5) *Phatic questions.*

4.1.2 Corpora

We annotated several dialogues from three different corpora in three languages: the *Saarbrücken Corpus of Spoken English (SCoSE)* (Norrick, 2017), a corpus of face-to-face conversations; the *CallFriend* corpus (Spanish) (Canavan and Zipperlen, 1996), a corpus of phone conversations;

[3]Our guidelines and annotations are available in our repository at https://github.com/andrea08/question_answer_annotation.

Annotators	A_o	κ
Questions	0.73	0.63
Features	0.90	0.67
Answers	0.59	0.49

Table 4: Cohen's Kappa score (κ) and observed agreement (A_o) for gold standard dialogue.

and the *Spoken Dutch Corpus (CGN)* Oostdijk 2001, a corpus of phone conversations. The purpose of annotating these dialogues was to test the annotation scheme on different languages and produce annotated data.

We annotated all questions in a subset of 4,939 utterances from the SCoSE corpus. Of these, 3,578 utterances were used to build the 'gold standard' corpus (used for calculating agreement scores and training machine learning algorithms). The remainder of the corpus was used as a test set in the machine learning algorithms. Furthermore, we annotated questions and answers from 2,618 and 935 utterances of CallFriend and CGN corpora, respectively. We relied primarily on the transcriptions of the corpora; in case of doubt, we made use of the audio recordings as well.

4.1.3 Results

We annotated 701 questions (Q) and 483 answers (A), distributed as follows: 422 (Q) / 289 (A) in the ScoSE corpus; 87 (Q) / 72 (A) in the CGN corpus; and 192 (Q) / 122 (A) in the CallFriend corpus. A descriptive analysis of our annotations shows that *yes/no* questions are the most common type in the three corpora, 40% (Spanish), 42% (English) and 64% (Dutch).

To evaluate the annotations, inter-annotator agreement was calculated based on a subset of the gold standard corpus.[4] Table 4 illustrates the values of observed agreement (A_o) and Cohen's κ (Cohen, 1960) obtained for question, feature and answer annotation. The agreement values obtained for question types were over 0.6 (for all annotators combined). This would generally be considered to be a 'moderate' level of agreement (Landis and Koch, 1977). A large share of our disagreements came from phatic questions; distinguishing these from other question types sometimes relies on subtle pragmatic and semantic contextual judgements. Agreement for answer types is lower than for question types because question types restrict answer types and hence question type disagreements can cause answer type disagreements.

In order to improve the annotation guidelines, we systematically examined all of the disagreements, most of which fell into one of four categories: (1) Simple mistakes, such as missing a question or choosing an (obviously) wrong tag. (2) Disagreements as a consequence of a previous disagreement; e.g., *wh*-questions need feature annotations, but *phatic* questions do not. In this case, a disagreement about the question type can cause further disagreement about feature type. (3) Missing instructions in the annotation guidelines for handling particular situations, e.g. annotating utterances containing interruptions. (4) Utterances whose interpretation was ambiguous and depends on subtle intonational or contextual cues for which it is hard to formulate a general rule.

4.2 Machine learning

We also conducted preliminary machine learning experiments for automating the annotation process. For the moment, we focus only on question type classification for English dialogues. So far, the approach that shows the most promising results is a decision tree algorithm (Quinlan, 1986) that takes as input a set of hand-designed features representing formal characteristics of a question, such as its length, the presence of a *wh*-word, and the presence of words such as *really?* or *you know?* Our full feature set is given in Table 5. Note that these features are quite superficial and do not take into account the discourse context of a question. Still, the algorithm achieves an accuracy score of 0.73 and an F1-score of 0.58, outperforming our majority-class baseline algorithm by a wide margin ($acc. = 0.47$, $F1 = 0.31$).[5]

Analysing the effect of the features in the predictions of the decision tree, we found that the majority of the mistakes were associated with the length of the questions. From the questions that were misclassified and had a length less than 6 (26 questions), 50% were wrongly predicted as *phatic* questions. Particularly, as with manual annotations, *phatic* questions that contain *wh-words* were source of disagreement and misclassified. Table

[4] This subset consists of the 690 utterances jointly annotated by all three annotators.

[5] A global $F1$ score was calculated by macro-averaging the scores for individual classes.

6 shows the confusion matrix for all the question types.

Feature	Description	Value
has_wh	Contains a wh-constituent	True, False
has_or	Contains the word "or"	True, False
has_inversion	Verb before NP (based on shallow parse)	True, False
has_tag	Contains a tag ('isn't it', 'right')	True, False
last_utt_similar	Question shares $\geq$ 50% of its words with the previous utterance	True, False
last_utt_incomplete	Previous utterance is interrupted (marked with special transcription symbol)	True, False
has_cliche	Contains a phatic marker ('you know?', 'really?')	True, False
length	Number of words	Numerical

Table 5: Extracted features for the classification task

	YN	DQ	PQ	CS	WH	Support
YN	74	1	8	3	2	88
DQ	0	3	0	0	0	3
PQ	7	0	15	0	8	30
CS	1	0	0	0	0	1
WH	10	0	9	0	43	62

Table 6: Confusion matrix of decision tree prediction. Testing data set, 184 questions.

Furthermore, we experimented with two neural architectures, a bag-of-words (BOW) classifier and a recurrent neural network (RNN), to test what input representations are most informative. However, so far these models suffer from overfitting and perform worse than the decision tree model (BOW: $acc. = 0.76$, $F1 = 0.44$; RNN: $acc. = 0.54$, $F1 = 0.24$). We expect these models to perform better when more training data is available.

5 Conclusion

This paper introduced a new annotation scheme for question-answer pairs in natural conversation. The scheme defines five question types and seven answer types based on a mix of formal and functional criteria. An annotation guide was developed and multi-lingual corpora were annotated. Inter-annotator agreement scores were moderately high; a qualitative analysis of disagreements led to improvements to the annotation guidelines. Initial machine learning experiments show that a simple decision tree algorithm achieves above-baseline performance, but much work remains to be done for making automatic annotation practically feasible. For future work, we would also like to expand the multilingual component of our work by adding language-specific guidelines, annotating more corpora, and adapting our machine learning algorithms to different languages.

Acknowledgements

This paper was written while the first authors (Cruz Blandón, Minnema, Nourbakhsh) were enrolled in the European Master Program in Language and Communication Technologies (LCT) and were supported by the European Union Erasmus Mundus program.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/, accessed January 22, 2019.

Maria Boritchev. 2017. Approaching dialogue modeling in a dynamic framework. Master's thesis, Université de Lorraine.

Alexandra Canavan and George Zipperlen. 1996. CALLFRIEND, Spanish-Non-Caribbean Dialect (LDC Catalog Number: LDC96S58).

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

ELAN (version 5.2). 2017. The Language Archive, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. https://tla.mpi.nl/tools/tla-tools/elan/.

Alice F. Freed. 1994. The form and function of questions in informal dyadic conversation. *Journal of Pragmatics*, 21(6):621 – 644.

Art Graesser, Vasile Rus, and Zhiqiang Cai. 2008. Question classification schemes. In *Proceedings of the Workshop on Question Generation*.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall PTR, Upper Saddle River, NJ, USA.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING '02 Proceedings of the 19th international conference on computational linguistics*, pages 1–7.

Neal Norrick. 2017. SCoSE part 1: Complete conversations. English Linguistics, Department of English at Saarland University.

Nelleke Oostdijk. 2001. The design of the Spoken Dutch Corpus. *Language and Computers*, 36:105–112.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning*, 1(1):81–106.

Gunter Senft. 2009. Phatic communion. In Gunter Senft, Jan-Ola stman, and Jef Verschueren, editors, *Culture and language use*, pages 226–233. John Benjamins Publishing, Amsterdam/Philadelphia.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR. In *LREC*.

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.