Linguistic Appropriateness and Pedagogic Usefulness of Reading Comprehension Questions

Andrea Horbach*, Itziar Aldabe◊, Marie Bexte*, Oier Lopez de Lacalle◊, Montse Maritxalar◊
* Language Technology Lab, University Duisburg-Essen, Germany
◊ University of the Basque Country (UPV/EHU), Spain
* {andrea.horbach,marie.bexte}@uni-due.de
◊ {itziar.aldabe,oier.lopezdelacalle,montse.maritxalar}@ehu.eus

Abstract

Automatic generation of reading comprehension questions is a topic receiving growing interest in the NLP community, but there is currently no consensus on evaluation metrics and many approaches focus on linguistic quality only while ignoring the pedagogic value and appropriateness of questions. This paper overcomes such weaknesses by a new evaluation scheme where questions from the questionnaire are structured in a hierarchical way to avoid confronting human annotators with evaluation measures that do not make sense for a certain question. We show through an annotation study that our scheme can be applied, but that expert annotators with some level of expertise are needed. We also created and evaluated two new evaluation data sets from the biology domain for Basque and German, composed of questions written by people with an educational background, which will be publicly released. Results show that manually generated questions are in general both of higher linguistic as well as pedagogic quality and that among the human generated questions, teacher-generated ones tend to be most useful.

Keywords: question generation, reading comprehension, evaluation guidelines

1. Introduction/Motivation

The automatic generation of reading comprehension questions for educational purposes has been a topic receiving considerable attention in the NLP community (Du et al., 2017; Heilman, 2011; Mazidi and Nielsen, 2014). A variety of approaches has been proposed and evaluated. However, a recent review of such systems and their evaluations, e.g., (Amidei et al., 2018) showed that many of these evaluation measures focus on linguistic quality of the produced questions only and often completely ignore the educational appropriateness of a question. A question might be linguistically correct and well-formed, but may not be helpful in pedagogical terms.

Consider Figure 1. showing an example from the SQuAD data set (Rajpurkar et al., 2016), one of the most frequently used data sets for training automatic current question generation systems. One can see that the questions for this passage are factoid questions whose answer can be found directly in one place in the text. While this was a design decision for the data set, originally created for the task of machine reading, this can be problematic when used to automatically generate questions in an educational scenario, where a teacher might feel the need to go beyond such relatively simple questions. This might also be the reason why evaluation measures for such questions target linguistic quality only.

In this paper we provide a novel annotation scheme for evaluating linguistic appropriateness and pedagogic usefulness of reading comprehension questions that should make it easy for the evaluators to follow the scheme. We assume that a major contributing factor for the low inter-annotator agreement (IAA) reported in the literature is that annotators are forced by standard annotation schemes to annotate measures for questions where they do not apply or do not make much sense. E.g. can we judge the ambiguity of a question

READING TEXT: *Most of the enlargement of the primate brain comes from a massive expansion of the cerebral cortex, especially the prefrontal cortex and the parts of the cortex involved in vision. The visual processing network of primates includes at least 30 distinguishable brain areas, with a complex web of interconnections. It has been estimated that visual processing areas occupy more than half of the total surface of the primate neocortex. The prefrontal cortex carries out functions that include planning, working memory, motivation, attention, and executive control. It takes up a much larger proportion of the brain for primates than for other species, and an especially large fraction of the human brain.*

READING COMPREHENSION QUESTIONS:

- *Primates have a visual processing network of how many brain areas?*
- *The visual processing areas occupy how much of the surface of the neocortex or primates?*
- *Planning, motivation, and attention are controlled by what area?*
- *The prefrontal cortex is the largest in what animals?*

Figure 1: A reading text and crowd-sourced reading comprehension questions about that text from SQuAD.

that is so garbled we cannot even understand what it might ask? Even if studies choose to include a not-applicable option in the annotations, whether actually to select that option is a subjective decision. In our annotation scheme, the decision whether a certain category is applicable for a question which has already received certain annotations on a lower level is pre-formulated and integrated into the annotation mechanism itself.

In order to evaluate our annotation scheme, we apply the annotation scheme to both manually crafted and automat-

ically generated reading comprehension questions in English. We also collect small evaluation data sets for Basque and German by having teachers or students about to become teachers manually generate questions and have them annotated according to the same guidelines.

In this paper, we ask the following research questions:

RQ1: *How reliably can the variables in our annotation scheme be annotated?* To answer this question, we apply our annotation scheme to all manually crafted and automatically generated reading comprehension questions and compare inter-annotator agreement for the different annotation categories.

As a follow-up to that question, we ask

RQ2: *How is the inter-annotator agreement of annotators with an NLP and teaching background in comparison to crowdworkers?* We compare our English initial annotations to a second set of annotation by crowdworkers from Amazon Mechanical Turk. We observe a pronounced agreement loss here.

RQ3: *How 'good' (according to our annotation scheme) are automatically generated questions from state-of-the-art systems compared to manually crafted questions?* In terms of measures of linguistic well-formedness, we expect automatically generated questions to perform below manually crafted ones. However, we hypothesize that in terms of educational value, both manually crafted data sets as well as automatically generated data leave room for improvement. I.e. we assume that neither gold standard data nor system output currently produces questions going beyond very simple literal question requiring only lower cognitive processing skills according to pedagogical criteria typically applied to reading comprehension questions in the educational domain.

In existing data sets, hand-crafted questions have mainly been crowd-sourced and only few data from real educational contexts are available. To assess the importance of questions from real educational contexts, we use data from the English LearningQ data set coming from either students or educators in addition to crowd-sourced data.

This is to answer **RQ4:** *How are crowd-sourced questions evaluated differently than questions from real educational contexts?* In order to assess this question more thoroughly, we also analyse the data sets for Basque and German. These small-scale data collections are a first step towards pedagogically motivated data sets, where we know that questions are not only linguistically well-formed, but also pedagogically adequate.

The research results show that our annotation scheme can be applied obtaining better results when expert annotators with some level of expertise are recruited. We also observe that teacher-generated questions tend to be the most useful in general.

2. Previous Work

In this section, we first present related work on the evaluation of generated questions. As the quality of such questions, especially when generated using neural networks, crucially depends on the availability of suitable training data, we provide next a brief review of existing English data sets that can be used for question generation.

2.1. Evaluation of Generated Questions

Ozuru et al. (2013) compare the nature of text comprehension as measured by multiple-choice questions on the one hand and open-ended questions on the other hand. They state that open-ended questions require active generation of information, while multiple-choice questions are often not strongly related to the active processing of reading comprehension of texts. Thus, the present work focuses on open-ended questions, and, specifically on the evaluation of the quality of open-ended questions created to test the reader about the comprehension of a text. However, there is a lack of a common framework to evaluate the quality of the questions. Amidei et al. (2018) present a review of papers from the ACL anthology between 2013 and 2018, and conclude that there is a lack of standardised approaches. They state that "given the ever-increasing number of publications in automatically generated questions (AQG), a common framework for testing the performance of generation systems is urgently needed".

Amidei et al. (2018) also report different studies regarding extrinsic and intrinsic evaluations. The final goal of extrinsic methods would be to evaluate the ability reached by the final users to accomplish the task, in our case the students. Nevertheless, the present work uses an intrinsic approach to measure the quality of the questions, hand-crafted as well as automatically generated. Concretely, we will ask humans to annotate criteria regarding the quality of the questions, without testing them in a real task context.

Prevailing intrinsic methods use human evaluation and/or automatic evaluation metrics. Some authors (Callison-Burch et al., 2006; Ananthakrishnan et al., 2007) criticise the automatic metrics used in the literature to evaluate automatically generated questions, because they are mainly word-overlap based metrics. That is why Nema and Khapra (2018) propose a metric related to the answerability of the question, that considers key words and entities. From our point of view, those metrics are not enough to evaluate reading comprehension questions, automatically generated or hand-crafted. We claim that, in both cases, human evaluation is necessary to elicit high-quality judgements. Thus, the evaluation guidelines proposed in this paper ask annotators about grammatical and pedagogical criteria in order to judge the quality of the questions. The main difference compared to most related work is that our guidelines do not use the same rating scale for all criteria, the most common rating scale in our guidelines is the binary scale in order to force annotators to make a clear decision, and, moreover, we apply evaluation criteria only for questions where they are appropriate based on previous criteria.

2.2. Existing Reading Comprehension Datasets

There are many existing English data sets containing questions that refer to texts. An overview of these can be found in Table 1. Most of them were developed with the task of machine comprehension in mind, such as the popular SQuAD (Rajpurkar et al., 2016) or the more recently released Natural Questions (Kwiatkowski et al., 2019) data set. Some of those data sets specifically aim at creating questions not easily answerable by computers, for example, questions that require script knowledge to be an-

swered (Ostermann et al., 2018) or cannot be answered at all (Kwiatkowski et al., 2019; Rajpurkar et al., 2016). However, questions that are hard to answer for computers are not necessarily difficult for humans. In many of the existing data sets, answers to the questions are in fact spans in the text they are based on Richardson et al. (2013; Kwiatkowski et al. (2019; Trischler et al. (2016; Rajpurkar et al. (2016; Joshi et al. (2017), meaning that they fall into the first category of literal understanding described by Day and Park (2005) and are thus likely easily answered by humans. There are some data sets that explicitly try to also include questions that require higher levels of understanding, for example by creating questions where information from more than one part of the text is needed to answer them (Richardson et al., 2013; Khashabi et al., 2018; Khashabi et al., 2017). One factor that might limit the depth of questions is that they are frequently crafted by crowdworkers, which are not experts on the given topic. LearningQ (Chen et al., 2018) is an exception in that it is a data set that not only contains questions created by students enquiring about a topic, but also by teachers who created questions about their lessons. Therefore, we use the LearningQ data as a source of teacher-generated questions and the SQuAD data set, the most-used one for AQG.

3. Evaluation Guidelines

As the aim is to evaluate questions for reading comprehension, we considered to include measures focusing on linguistic and pedagogical appropriateness. For that, the systems and measures analysed in (Amidei et al., 2018; Le et al., 2014) and the reading comprehension taxonomy by Day and Park (Day and Park, 2005) were examined. The general criteria for evaluating the adequacy of each question were set in terms of fluency, ambiguity, answerability, pedagogical relevance and comprehension type.

As some criteria are not applicable given some questions[3] , we assume that it does not make sense to evaluate a question that is not understandable according to any of the other criteria. If, for example a question is not understandable at all, it would not be easy to determine the usefulness of it. Such an approach also has the goal to reduce the overall annotation load.

In order to evaluate reading comprehension questions according to different criteria, we divided the evaluation task into four different groups, where at the end of each group the annotation might stop. The following list gives the exact wording of each evaluation question together with the multiple choice answer options:

Group 1

- **Understandable** Could you understand what the question is asking? (Yes/No)

Group 2

- **DomainRelated** Is the question related to the Biology domain? (Yes/No)
- **Grammatical** Is the question grammatically well-formed, i.e. is it free of language errors? (Yes/No)
- **Clear** Is it clear what the question asks for? (Yes/More or less/No)

Group 3

- **Rephrase** Could you rephrase the question to make it clearer and/or error-free? (Yes/No)
- **Answerable** Are students probably able to answer the question? (Yes/No)

Group 4

- **InformationNeeded** Which kind of information is needed to answer the question?
  (a) Information presented directly and in one place only in the text
  (b) Information presented in different parts of the text
  (c) A combination of information from the text with external knowledge
  (d) General knowledge about the topic (not from the text)
  (e) The reader's feelings/judgements/... about the text
- **Central** Do you think being able to answer the question is important to work on the topic covered by the text? (Yes/No)
- **WouldYouUseIt** If you were a teacher working with that text in class, do you think you would use this question or your rephrasal of the question? (Yes/Maybe/No)

Each group of evaluation criteria defines one stopping point. That means that depending on the answer given to some of the criteria from that group, it is determined whether the question can be judged or not according to the rest of the measures and the evaluation of a particular question might stop there.

In group 1, the **understandable** option checks if it is possible to understand what the question is asking. If it is incomprehensible, we consider that no further measures apply to the question and the evaluation ends.

Group 2 includes three evaluation criteria related to the domain and linguistic appropriateness of the question: **domainRelated, grammatical** and **clear**. If it is not clear what the question asks for (**clear**), the evaluation process ends. Otherwise, the process continues. Group 3 presents one (**answerable**) or two (**rephrase, answerable**) evaluation criteria based on group 2's answers. If the question is more or less clear (group 2 - clear) or the question is not grammatical (group 2 - grammatical), the evaluator should rephrase the question to make it clearer and/or error-free (**rephrase**). Otherwise, this evaluation criterion is not presented. In all cases, it is asked if students are probably able to answer the question (**answerable**). If not, the evaluation ends. If yes, the evaluators are asked to provide the answer to the question in a text field.

[3]We arrive to this conclusion after discussing with teachers a version of the guidelines with no hierarchy

Name	# of questions	Texts	Source	Level
AI2 Biology Corpus [1]	378	undergraduate textbook	domain expert	answers are spans
bAbI QA tasks (Weston et al., 2015)	20*10000	2-4 simple sentences per question	system	designed to challenge computers, but easy for humans
HotpotQA (Yang et al., 2018)	112779	2 wikipedia paragraphs per question	crowd	open & MC questions requiring multi-hop reasoning
LearningQ (Chen et al., 2018)	231470	10841 educational articles & transcriptions	experts, crowd	factoid, but also understanding, applying and analyzing
MCScript (Ostermann et al., 2018)	14074	2100 narrative texts	crowd	MC questions, 27% requiring script knowledge
MCTest (Richardson et al., 2013)	2640	660 fictional stories	crowd	MC questions, answers are spans, 50% require information from >1 sentence
MS MARCO (Bajaj et al., 2016)	1010916	∅ 10 passages per question	queries	factoid, mostly descriptions or numeric
MultiRC (Khashabi et al., 2018)	9872	871 paragraphs	crowd	MC questions, 60% require multi-sentence reasoning
NarrativeQA (Kočiskỳ et al., 2018)	46765	1572 scripts and summaries	crowd	> 50% of answers are not spans in texts
Natural Questions (Kwiatkowski et al., 2019)	323045	wikipedia articles	queries	only 50% with answer, answers are spans
NewsQA (Trischler et al., 2016)	100000	12744 CNN articles	crowd	answers are spans, 33% word matching, 27% paraphrasing, 20% synthesis
OpenBookQA (Mihaylov et al., 2018)	5957	1326 science facts, 1 per question	crowd	MC questions requiring transfer of world knowledge
QA-SRL Bank 2.0 (FitzGerald et al., 2018)	265140	76397 sentences, 1 per question	crowd	factoid, 1 question per predicate-argument pair in sentence, easily answerable by humans
Question Answer data set (Smith et al., 2008)	1209	40 wikipedia articles	crowd, system	factoid
SciQ (direct-answer) (Welbl et al., 2017)	12252	short passages from science study books	crowd	answers are trivial when questions are presented with MC options
SearchQA (Dunn et al., 2017)	140461	∅ 49.6 search result snippets per question	crawled	factoid, answers are at most 3 tokens long
SQuAD 1.1 (Rajpurkar et al., 2016)	107785	paragraphs from 536 wikipedia articles	crowd	answers are spans, questions mostly lexical & syntactic variations of text
TREC 1999-2007 [2]	>3500	newspaper articles & blog posts	crowd, queries	factoid
TQA (text questions) (Kembhavi et al., 2017)	13693	lessons from science textbooks	crawled	MC questions, 40% need 1, 40% >1 sentence in the text to be answered
TriviaQA (Joshi et al., 2017)	95956	6 wikipedia articles & search results per question	crawled	answers are spans, questions mostly lexical & syntactic variations, 40% require multiple sentences to answer
WikiQA (Yang et al., 2015)	3047	summary sections of wikipedia articles	queries	factoid, 2/3 not answerable using the articles

Table 1: Overview of existing English question-answer data sets.

Finally, group 4 contains the pedagogically oriented criteria. The **informationNeeded** measure requires the annotator to identify which kind of information is needed to answer the question. The five answer options are part of the reading comprehension taxonomy by Day and Park. Similarly, the last two criteria (**central,wouldYouUseIt**) ask for the evaluators' opinion concerning the importance and usefulness of the question, with the **wouldYouUseIt** measure being the most subjective one.

4. Annotation studies

We test the afore-described evaluation guidelines in an annotation study which is detailed in this section. We first select the reading texts to be used. For these texts, we either select existing manually generated reading comprehension questions (for the English texts, which are part of existing data sets) or have new questions generated (in the case of Basque and German). Additionally, we generate questions for the English texts using both a rule-based and a neural question generation system.

language	**EN**	**EU**	**DE**
#texts	24	18	21
#questions (all)	299	152	95
#questions (generated)	235	-	-
#questions (manual)	64	152	95

Table 2: Overview of the annotation data for English (EN), Basque (EU) and German (DE)

4.1. Selection of Textual Material

For the selection of material for our annotation study we use the following criteria. We want to test the evaluation of both automatically generated as well as manually-crafted questions. Therefore, we examine existing data sets to obtain texts together with manually generated questions and then add automatically generated ones. As our focus is

[2]https://allenai.org/data/data-all-2.html
[3]https://trec.nist.gov/data/qamain.html

reading comprehension questions in an educational domain and most existing data sets for such questions focus on sciences, this was also our focus for text selection. Many data sets (see Section2.) have questions manually generated via crowd-sourcing. There are only a few data sets providing data from realistic educational contexts, most prominently the LearningQ data, where questions are either produced by teachers (the TED-Ed subset) or by students (the Khan Academy subset). We thus looked for topics which occurred both in SQuAD (Rajpurkar et al., 2016), as the most frequently used data set for automatic question generation, as well as the Khan Academy and TED-Ed data in LearningQ (Chen et al., 2018). Furthermore, we restricted ourselves to topics which would also be suitable for data collection in our additional languages of interest, Basque and German. Thus, we checked the availability of these topics in the German and Basque Wikipedia as well as a Wikipedia-like source tailored specifically towards children, KLexikon in German[4] and Txikipedia in Basque[5] to get suitable articles. Applying these criteria, we arrived at two topics from the biology domain, *brain* and *gene*.

We selected one text per topic for Khan and TED. In SQuAD, the amount of questions per text was four to five, which is why we randomly picked four texts per topic to have a sufficient amount of questions. As the Khan and TED texts are quite long, especially compared to the rather concise SQuAD paragraphs, we shortened them.

German Wikipedia and KLexikon articles were split into their paragraphs, from which we then excluded those that mostly introduced technical terms or contained lists or image descriptions. In a few cases, two subsequent shorter paragraphs were combined to form one longer paragraph. This resulted in 7 and 8 Wikipedia paragraphs for gene and brain, respectively, as well as 3 KLexikon paragraphs for each of the topics. In the case of Basque Wikipedia and Txikipedia articles we applied the same criteria and tried to select similar texts regarding the topic.This resulted in 7 and 7 Wikipedia paragraphs, as well as 1 and 3 Txikipedia paragraphs for gene and brain, respectively.

4.2. Question Selection and Generation

Table 2 gives an overview of the total amount of reading texts as well as manually and automatically generated questions we use for each language.

For the German texts, we collected manually generated reading comprehension questions from biology students training to become biology teachers in the German school system. They were given one text per person and asked to write down questions which they could imagine to ask high-school students about the text. They spent about 15 minutes reading the text and writing questions.

For Basque texts, we collected manually generated reading comprehension questions from 3 biology teachers teaching at the teaching school in the University of the Basque Country (UPV/EHU). They were given 18, 8 and 6 texts, respectively; that means that some teachers wrote questions for the same text. They were asked to write down open-ended questions which they could imagine to ask high-school students about the Wikipedia texts and secondary school students about the Txikipedia texts. As shown in table 2, the number of questions for the Basque data set is higher than for the German data set, even though the teachers also spent about 15 minutes or less reading each text and writing the questions.

For the English texts we collected all questions available in the respective data sets, amounting to a total of 64 manually generated English questions. For English, we also automatically generated questions based on a rule-based system and one state-of-the-art deep neural network system. Concretely, we used Heilman (2011)'s system, a rule-based system that overgenerates and ranks questions from text. And, an implementation of the deep neural network system proposed by Du et al. (2017) which incorporates the global attention mechanism (Luong et al., 2015) into an encoder-decoder sequence learning framework during the decoding process. The attention mechanism focuses on relevant information in the source text to generate the questions in order to mimic the human's problem solving process. The system was originally trained on SQuAD.

For SQuAD we obtained 5 questions per source text and method and 20 questions per source text for TED and Khan. In the case of the rule-based system we randomly selected 5 questions out of the top 10 ranked questions for SQuAD and the top 20 for TED and Khan. The questions generated by the NQG system were accordingly selected based on the predicted log-likelihood estimation.

4.3. Annotation Setup

For English, we collected annotations from two groups of annotators, in order to compare the level of expertise and familiarity with linguistics and educational topics needed to handle the task. Our *expert* annotators are colleagues from our respective departments with some background in NLP as well as teaching. The *crowd-worker* annotators were recruited via Amazon Mechanical Turk. In the case of Basque and German, we only used expert annotators.

For each annotator group, we requested two annotations per question. Note that the goal of this annotation is not to arrive at a final gold-standard used for training or testing an automatic classifier but rather to check the feasibility of our annotation scheme and its results for different kinds of questions. Therefore we deemed two annotations per condition sufficient and have no need to perform, e.g., a majority vote among workers.

Expert annotators were not paid. Based on the time needed by the experts we estimated the average time requirement per HIT and paid AMT workers 0.30 $ per HIT in order to make sure that annotators were paid an average amount of 10 Euros per hour. Workers were first introduced to the task by means of some examples. They were presented with a reading text and a question about it and worked then through the questionnaire as described in Section 3. Expert annotations originated from 14 individual annotators while 211 crowd-workers contributed to our evaluation.

[4]https://klexikon.zum.de

[5]https://eu.wikipedia.org/wiki/Txikipedia:Azala

	Experts					Crowd-workers				
	all		applicable-only			all		applicable-only		
evaluation category	**%agree**	κ	**%agree**	κ	**#pairs**	**%agree**	κ	**%agree**	κ	**#pairs**
understandable	0.8	0.56	0.8	0.56	299	0.73	0.31	0.73	0.31	299
domainRelated	0.78	0.54	0.95	0.4	170	0.65	0.24	0.87	−0.07	180
grammatical	0.65	0.44	0.73	0.32	170	0.58	0.25	0.74	0.13	180
clear	0.61	0.4	0.65	0.21	170	0.52	0.2	0.64	−0.01	180
rephrase	0.71	0.22	0.86	0.25	28	0.61	−0.02	0.64	0.18	14
answerable	0.73	0.52	0.91	0.53	140	0.57	0.26	0.71	0.1	173
informationNeeded	0.72	0.5	0.83	0.19	118	0.53	0.23	0.69	0.14	112
central	0.71	0.5	0.81	0.16	118	0.59	0.25	0.86	0.12	112
wouldYouUseIt	0.58	0.34	0.47	0.03	118	0.48	0.2	0.57	0.01	112

Table 3: Inter-annotator agreement for expert annotators and crowd-workers.

	Basque					German				
	all		applicable-only			all		applicable-only		
evaluation category	**%agree**	κ	**%agree**	κ	**#pairs**	**%agree**	κ	**%agree**	κ	**#pairs**
understandable	0.95	−0.02	0.95	−0.02	152	1.0	1.0	1.0	1.0	95
domainRelated	0.95	−0.02	0.99	0.0	145	0.97	0.39	0.97	−0.02	94
grammatical	0.84	0.53	0.88	0.6	145	0.89	0.32	0.89	0.23	94
clear	0.87	0.04	0.91	−0.04	145	0.86	0.08	0.86	−0.06	94
rephrase	0.81	0.48	0.95	−0.02	21	0.79	0.07	1.0	0.0	2
answerable	0.86	0.07	0.92	−0.04	142	0.77	0.37	0.78	0.34	92
informationNeeded	0.68	0.11	0.76	0.06	131	0.53	0.33	0.6	0.33	63
central	0.8	0.13	0.91	−0.05	131	0.63	0.27	0.76	0.07	63
wouldYouUseIt	0.61	0.03	0.68	−0.08	131	0.41	0.07	0.43	−0.21	63

Table 4: Inter-annotator agreement for Basque and German data (only expert annotators).

5. Annotation Evaluation

In this section, we examine the collected annotations closer in the light of the research questions from section 1..

5.1. RQ 1 and 2: Reliability of Evaluations

Inter-annotator-agreement is a first indicator when assessing new annotation guidelines as to whether they can be reliably applied. We evaluate agreement pairwise and separately for each evaluation category as well as for each annotator group.

Because of our hierarchical evaluation setup, after the first question about understandability, a certain annotation can always be non-applicable for an individual annotator. In order to be more comparable to previous non-hierarchical evaluation setups we present one agreement evaluation (referred to as *all*) where we introduce a non-applicable label whenever an annotator did not see that particular question. To get a more realistic estimate of those cases that have actually been annotated by two humans, we also provide the *applicable-only* evaluation where we only included evaluations where the question was applicable to both annotators. We evaluate both Cohen's kappa as well as percentage agreement. For *applicable-only*, we additionally report the number of annotation pairs it has been computed on.

Table 3 shows the evaluation results for the English data, Table 4 for Basque and German. We can see that there is often a relative high percentage agreement for both expert and crowd-workers. For expert annotators, kappa values are in a modest range for the *all* category, while they are considerably lower for crowd-workers, especially when only applicable cases are considered. Annotation label distributions that are often quite skewed lead to relatively low kappa values. We conclude from the low IAA values of crowd-workers that evaluation is a quite subjective task that needs some sort of expertise and that crowd-sourcing in the way we did it in our studies is not a reasonable way of evaluating reading comprehension questions.

If we compare evaluations for individual categories, we see for experts that questions about linguistic quality have a higher agreement in the *applicable-only* condition than those annotations targeting more educational appropriateness. The last evaluation category *Would you use it?* is clearly by design a very subjective one and unsurprisingly we observe here the lowest agreement between annotators.

On the English data, we also compared inter-annotator-agreement individually for each question generation method (neural, rule-based and manual) and found them to be in the same range.

5.2. RQ3: Quality of Generated and Hand-Crafted Questions

As the evaluation of inter-annotator-agreement showed that crowd-workers did not annotate questions reliably, our subsequent evaluations are based on the expert annotations only. We aggregate in this evaluation over all individual annotations by counting how often each annotation label

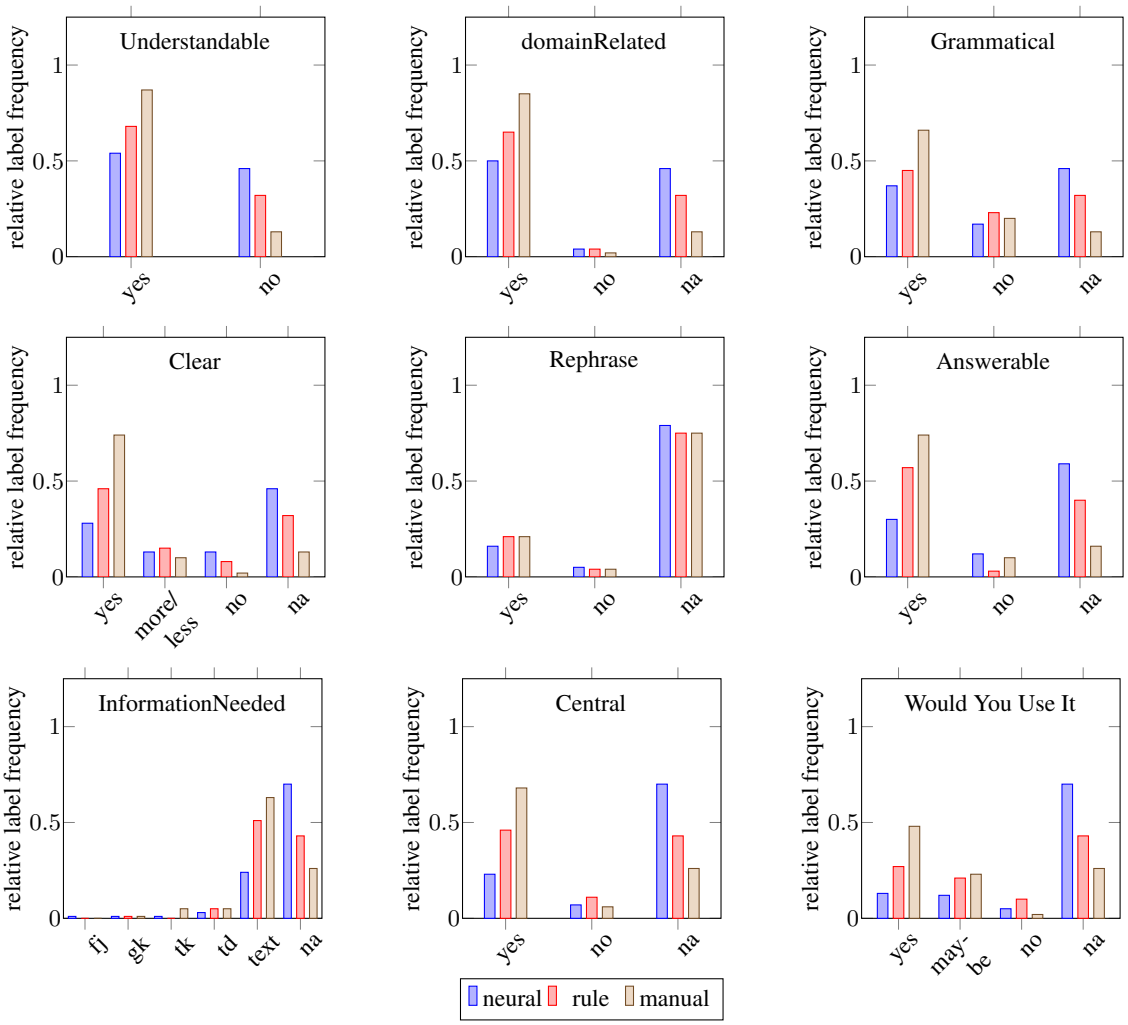

Figure 2: Comparison of the three question generation methods. Only expert annotations were taken into account. The information needed to answer a question corresponds to whether it can be found in one (text) or a combination of different (td) places in the text, needs text and additional knowledge (tk) or just general knowledge (gk) to answer, or asks for feelings or judgements (fj).

was assigned per category.

We first compare evaluations between the three different data sources: *manual* questions, *neural* and *rule-based* questions. Figure 2 shows the results. We can clearly see that manually created questions are generally evaluated higher (i.e. more understandable, grammatical etc) in almost all categories than both kinds of automatically generated questions, while the rule-based questions are scored higher than those produced by the neural system.

That means that, in line with previous findings, neural and rule-based system do not yet reach human performance. However, our evaluation also shows, that differences between manually and automatically generated questions are somewhat larger for the more advanced categories such as whether the question is answerable or clear instead of being domain related or understandable. Also for the, as discussed previously, very subjective category of whether the evaluator would use a question, differences between systems are huge with almost three times as many questions that the annotator would use for the manual than neurally-generated system. This underlines that it is not enough to generate well-formed questions, but that they have to make sense pedagogically as well. The rule-based system, which produced questions of higher linguistic quality than the neural system, might produce acceptable questions, only if we do not care too much about the higher evaluation categories.

Note that the findings on the neural system are certainly in part due to the neural data being trained on the SQuAD dataset, i.e. data from crowd-workers. This again highlights the need for pedagogically motivated training data that could be used as training material for automatic question generation.

5.3. RQ4: Evaluation of hand-crafted questions by origin

As a next step, we zoom in on the manually created questions, distinguishing in our analysis between the three data sources SQuAD, Ted and Khan for English, as well as the German and Basque data. We see from Figure 3 that the Khan questions are in many categories scored lower than questions from the other sources. This is an

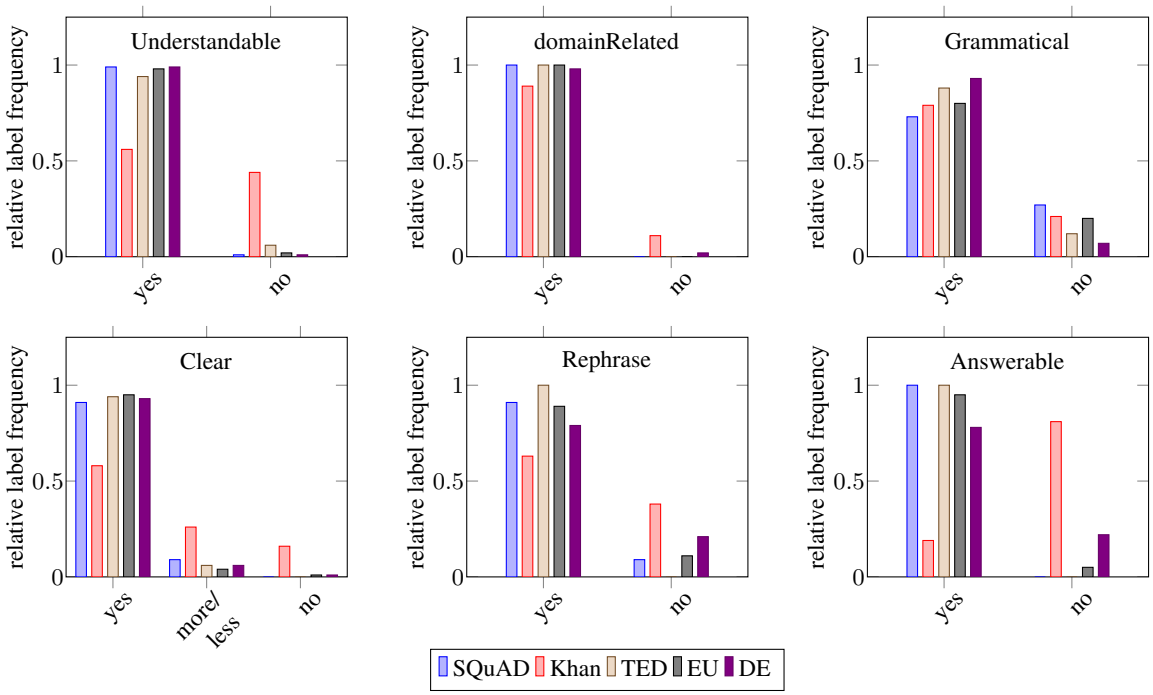

Figure 3: Comparison of the manually created questions from SQuad, Khan, Ted, the new Basque (EU) and German (DE) data sets. Only expert annotations were considered and only cases in which the respective labels were found to be applicable were taken into account.

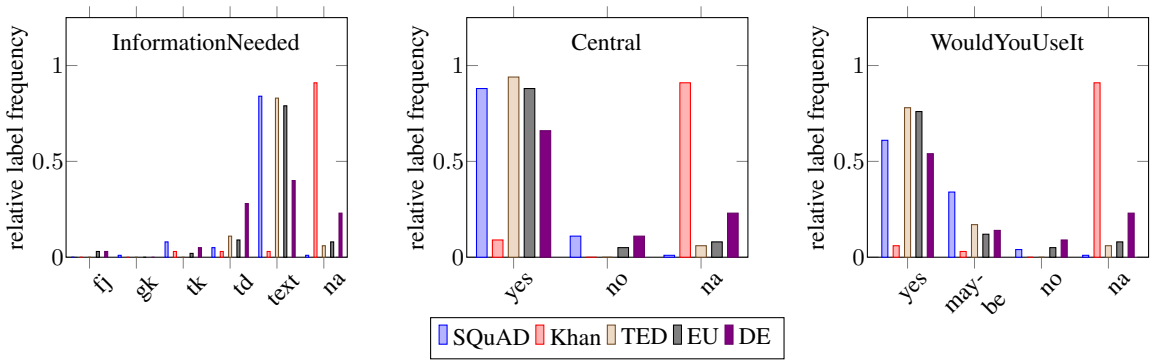

Figure 4: Comparison of the manually created questions. Again, only expert annotations were taken into account. The information needed to answer a question corresponds to whether it can be found in one (text) or a combination of different (td) places in the text, needs text and additional knowledge (tk) or just general knowledge (gk) to answer, or asks for feelings or judgements (fj).

at first glance surprising finding given that SQuAD was produced by crowd-workers and TED and Khan by teachers and students respectively and our initial hypothesis was that crowd-workers might be more unreliable in producing questions. However, we could argue, that crowd sourced questions probably stick closer to the text and are therefore considered more domain-related and understandable while the student-created questions come from a message board context where students exchange about learning material and are therefore more informal and are often of a lower linguistic quality.

Figure 4 shows the label distribution for the 3 last categories. In contrast to Figure 3, we also included the *not applicable* label, as we have a considerable amount of answers with such a low linguistic quality that they could not be annotated with the more advanced evaluation categories. In this evaluation, we can see that the two data sets created by teachers, the Ted and Basque data set contain the highest amount of actually usable and central questions.

6. Conclusion

In this paper, we proposed a new annotation scheme for the evaluation of automatically and manually crafted reading comprehension questions. An annotation study showed that the scheme apparently needs some sort of expertise to apply it. In our setup, crowd-workers were not able to apply the scheme with acceptable agreement and it was also not a trivial task for annotators with a teaching and linguistics background. Manually generated questions were in most aspects evaluated better than automatically generated ones,

especially neurally generated. A literature review revealed that most manually created datasets were created by crowdworkers. To evaluate potential differences between expert content creators and crowdworkers, we collected two new evaluation data sets for Basque and German. Our evaluations showed that surprisingly, these datasets still contain a high number of literal questions similar to crowd-sourced data. However, we found that the teacher crafted questions had higher percentages of central and actually usable questions. This highlights the need for larger expert-generated datasets to support the generation of high-quality data.

7. Acknowledgements

This work has been partly supported by the Spanish Ministry of Economy and Competitivenes under the deepReading Project RTI2018-096846-B-C21 (MCIU/AEI/FEDER,UE) and by the Stifterverband fÃ¼r die Deutsche Wissenschaft e.V. It has also been funded by the Horizon 2020 Framework Programme of the European Union under the enetCollect CA16105 COST action.

Finally, we would like to thank Igone Palacios, Arantza Rico and Aritz Ruiz, the expert teachers who manually created the Basque dataset, as well as the student teachers of Professor Philipp Schmiemann, who created the German dataset.

8. Bibliographical References

Amidei, J., Piwek, P., and Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018.

Ananthakrishnan, R., Bhattacharyya, P., Sasikumar, M., and Shah, R. M. (2007). Some issues in automatic evaluation of english-hindi mt: more blues for bleu. *ICON*.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Chen, G., Yang, J., Hauff, C., and Houben, G.-J. (2018). Learningq: a large-scale dataset for educational question generation. In *Twelfth International AAAI Conference on Web and Social Media*.

Day, R. and Park, J.-s. (2005). Developing reading comprehension questions. *Reading in a Foreign Language*, 17:60–73, 01.

Du, X., Shao, J., and Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352.

Dunn, M., Sagun, L., Higgins, M., Guney, V. U., Cirik, V., and Cho, K. (2017). Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

FitzGerald, N., Michael, J., He, L., and Zettlemoyer, L. (2018). Large-scale qa-srl parsing. *arXiv preprint arXiv:1805.05377*.

Heilman, M. (2011). Automatic factual question generation from text. *Language Technologies Institute School of Computer Science Carnegie Mellon University*, 195.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., and Hajishirzi, H. (2017). Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., and Roth, D. (2018). Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Le, N.-T., Kojiri, T., and Pinkwart, N. (2014). Automatic question generation for educational applications â the state of art. *Advances in Intelligent Systems and Computing*, 282:325–338, 01.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Mazidi, K. and Nielsen, R. D. (2014). Linguistic considerations in automatic question generation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 321–326.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Nema, P. and Khapra, M. M. (2018). Towards a better metric for evaluating question generation systems. *arXiv preprint arXiv:1808.10192*.

Ostermann, S., Modi, A., Roth, M., Thater, S., and Pinkal, M. (2018). Mcscript: A novel dataset for assessing machine comprehension using script knowledge. *arXiv preprint arXiv:1803.05223*.

Ozuru, Y., Briner, S., Kurby, C. A., and McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(3):215.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Smith, N. A., Heilman, M., and Hwa, R. (2008). Question generation as a competitive undergraduate course project. In *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, pages 4–6.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2016). Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Welbl, J., Lui, N. F., and Gardner, M. (2017). Crowdsourcing multiple choice science questions. In *Workshop on Noisy User-generated Text*.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Yang, Y., Yih, W.-t., and Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

9. Language Resource References

Chen, Guanliang and Yang, Jie and Hauff, Claudia and Houben, Geert-Jan. (2018). *LearningQ: a large-scale dataset for educational question generation*.

Pranav Rajpurkar and Jian Zhang and Konstantin Lopyrev and Percy Liang. (2016). *SQuAD: 100, 000+ Questions for Machine Comprehension of Text*.