Evaluation methodologies in Automatic Question Generation 2013-2018

Jacopo Amidei and **Paul Piwek** and **Alistair Willis**
School of Computing and Communications
The Open University
Milton Keynes, UK
{jacopo.amidei, paul.piwek, alistair.willis}@open.ac.uk

Abstract

In the last few years Automatic Question Generation (AQG) has attracted increasing interest. In this paper we survey the evaluation methodologies used in AQG. Based on a sample of 37 papers, our research shows that the systems' development has not been accompanied by similar developments in the methodologies used for the systems' evaluation. Indeed, in the papers we examine here, we find a wide variety of both intrinsic and extrinsic evaluation methodologies. Such diverse evaluation practices make it difficult to reliably compare the quality of different generation systems. Our study suggests that, given the rapidly increasing level of research in the area, a common framework is urgently needed to compare the performance of AQG systems and NLG systems more generally.

1 Introduction

Evaluation is a critical phase for the development of Natural Language Generation (NLG) systems. It helps to improve performance by highlighting weaknesses, and to identify new tasks to which generation systems can be applied. Given that generation systems and evaluation methodologies should be developed hand in hand, a systematic study of evaluation methodologies for NLG should take a central role in the effort of building machines which are able to reach human-like levels of linguistic communication. Such a study should investigate the current evaluation practices used in various areas of NLG in order to see their weaknesses and suggest directions to improve them.

The aim of this paper is to analyze the evaluation methodologies used in Automatic Question Generation (AQG) as a representative subtask of NLG. To the best of our knowledge, since the introduction of the Question Generation Shared Task Evaluation Challenge (QG-STEC) (Rus et al., 2010), no attempts have been made to introduce a common framework for evaluation in AQG.

To approach this task, we examined the papers in the ACL anthology with a publication date between the years 2013-2018 (more precisely January 2013 to June 2018). Table 1 shows the distribution of the papers involved in the current study across this period. The ACL anthology website represents a resource of inestimable value[1] for this work.

Year of publication	# papers
2018 (Jan-June)	7 (so far)
2017	13
2016	9
2015	5
2014	1
2013	2

Table 1: Number of papers per year describing question generation systems.

We used the single term *question generation* as the search term with the search engine provided in the ACL Anthology website. From the papers that were returned by this query, we focussed only on those papers that were about question generation systems. This gave us 37 papers to analyze, of which 36 were published in conference proceedings and 1 was published in a journal. The number of papers by year is given in Table 1 and illustrated in Figure 1. Figure 1 indicates the rapid increase in publications in this area in recent years.

[1] http://aclweb.org/anthology/

Proceedings of The 11th International Natural Language Generation Conference, pages 307–317,
Tilburg, The Netherlands, November 5-8, 2018.

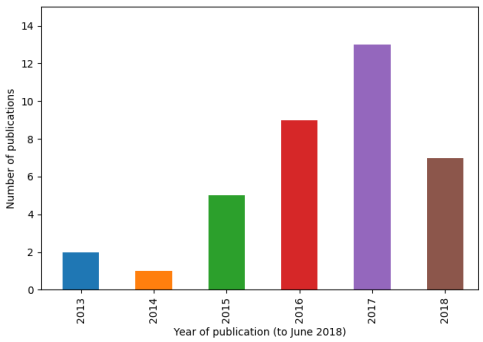

Figure 1: Number of papers on AQG published by year in the ACL anthology.

Note that this study of the literature was carried out in June 2018, and so several major conferences in this area (including ACL, INLG, EMNLP and COLING) had not taken place.[2]

Publication type	Journal or conference name	# of papers
Conference proceeding	INLG	7
	ACL	7
	NAACL-HLT	6
	Workshop on Innovative Use of NLP for Building Educational Application	6
	EMNLP	3
	IJCNLP	1
	EACL	1
	SIGDIAL	1
	COLING	1
	NLPTEA	1
	RANLP	1
	Workshop on Representation Learning for NLP	1
Journal	Computational Linguistics	1

Table 2: Number of papers per conference proceedings or journal.

Before looking more closely at the publications involved, let us introduce the AQG tasks studied in these papers. AQG is the task

> of automatically generating questions from various inputs such as raw text, database, or semantic representation (Rus et al., 2008).

The above definition, adopted by the AQG community, leaves room for researchers to decide what kind of questions and input work with. Following Piwek and Boyer (2012) a particular AQG task can be characterized by three aspects: the input, the output, and finally the relationship between the input and the output. The 37 papers we analyzed can be divided into the following three categories:

1. *Input*: text;
   Output: text;
   Relation: the output question is answered by the input text or the output question asks a clarification question about the input text.

2. *Input*: knowledge base structured data (for example triples ⟨subject, object, subject/object relation⟩);
   Output: text;
   Relation: the output question is answered by the information structure in the input.

3. *Input*: image or image and text or image segmentation annotations;
   Output: text;
   Relation: the output question is answered by the information pictured in the input.

For the sake of simplicity we will denote with *Text2Text* the task expressed by category 1, *Kb2Text* the task expressed by category 2 and finally *Mm2Text* the task expressed by category 3, where *Mm* is short for "Multi-modal". Within each category, we find papers with different aims. We show these in the following list, where the number in brackets shows how many papers fall into that class:

1. *Text2Text* (30)
   - Web searching (1)
   - Chatbot component (1)
   - Creation of comparative questions related to the input topic (1)
   - Clarification questions (1)
   - Question Answering (5)
   - Dataset creation purpose (1)
   - Educational purpose (9)
   - AQG general purposes (11)

2. Kb2Text (4)
   - Question Answering (1)
   - Dataset creation purpose (1)
   - Educational purpose (1)
   - AQG general purposes (1)

3. Mm2Text (3)

[2] A complete list of papers used in this study, as well as useful information to reproduce the results presented in the present paper, can be found at the following link:

https://bit.ly/2IuPJIa

- Data augmentation Visual Question Answering (VQA) purpose (1)
- AQG general purposes (2)

Regarding the papers in the *Text2Text* category, we found some variety in the different types of output. Although in the majority of cases, the system's output was an interrogative sentence, there are 5 papers in which the output is a "fill the gap" question, 3 papers where output is a multiple choice question (with its associated set of distractors) and 3 papers in which the output is a question/answer pair. Also in both the *Kb2Text* and *Mm2Text* categories there is 1 paper each in which the output is a question/answer pair. We also note that one paper in the *Text2Text* category developed a question generator which takes a paragraph of text and an associated answer as input. In this case, the generated question must be answered by the answer given in the input. We conclude this section by specifying that *AQG general purposes* means that the system was not tied to a particular domain or task-dependent setting, whereas *Question Answering* means that the AQG system is developed in order to be used in the Question Answering task.

2 Related Work

Two of the key references for evaluation in NLG are Krahmer and Theune (2010) and Gatt and Krahmer (2018). Both devote an entire section to evaluation. In particular, Section 7 of Gatt and Krahmer's paper gives a helpful description of the methodologies used in NLG for the purpose of evaluation, alongside examples and a discussion of the relevant problems.

Another highly relevant work is that of Gkatzia and Mahamood (2014). Gkatzia and Mahamood studied the use of evaluation methodologies for NLG, performing a study which analyzed a corpus of 79 conference and journal papers published between the years 2005-2014. Their results show the increasing prevalence of automatic evaluation over human evaluation and the prevalence of intrinsic evaluation over extrinsic ones (we discuss intrinsic and extrinsic methods in Section 3). Gkatzia and Mahamood also report that the evaluation approaches are correlated with the publication venue, so that papers published in the same journal or conference tend to use the same evaluation methodologies. Our paper represents a continuation and refinement of the Gkatzia and Mahamood paper, with our specific focus on AQG.

Regarding more specific work on AQG we refer to Rakangor and Ghodasara (2015) and Le et al. (2014), both of which survey AQG, with the latter focussing specifically on educational applications of AQG. For each paper considered, Rakangor and Ghodasara present the methodology used, the generated question type, the language of the generated question, the evaluation methodologies and its results. In contrast, Le et al. report on the educational support type and the evaluation methodologies, in one table and the question type and evaluation results in the other table. Although Le et al. present these results in two tables, the tables in fact have only one paper in common. In comparison, in this paper we focus all our attention on the evaluation methodologies used. For this reason, we report neither the systems' specifications nor the systems' performance.

A final publication of importance to the current work is the report on the Question Generation Shared Task Evaluation Challenge (QG-STEC) (Rus et al., 2010). In the QG-STEC, two tasks, A and B, were defined. Although both tasks shared the same output type, task A took a paragraph and a target question type as input, whereas task B took a single sentence and a target question type as input. Both tasks were evaluated through a human evaluation methodology, based on the 5 criteria: *relevance*, *syntactic correctness and fluency*, *ambiguity*, *question type* and *variety*. However, the Inter Annotator Agreement (IAA) reached in the evaluation phase was low. An attempt to improve the IAA for task B is described in Godwin and Piwek (2016), in which the authors define an interactive process where the annotators discussed their opinions about the criteria used in the evaluation. Although their method improves the IAA, the reproducibility of their results is not guaranteed.

3 Evaluation methodologies for AQG

In this section we present the findings of our analysis. We focus our analysis on two dimensions: *intrinsic evaluation methodology* and *extrinsic evaluation methodology*.

Intrinsic evaluation methods measure the performance of a system by evaluating the system's output "in its own right, either against a reference corpus or by eliciting human judgements of quality" (Gatt and Belz, 2010, p. 264). For example, this could involve measuring the output's grammaticality and fluency. The prevailing intrin-

sic methods are *human evaluation* and *automatic evaluation*. In order to assess the quality of a generated sentence, the former method uses human judgements, while the latter applies an algorithm that automatically calculates a score, for example by checking the similarity between the generated sentence and a set of reference sentences.

Extrinsic methods measure the performance of a system by evaluating the system's output with respect to its ability to accomplish the task for which it was developed. An example of extrinsic evaluation methods is that used to evaluate the STOP system (Reiter et al., 2003). STOP generates "short tailored smoking cessation letters, based on responses to a four-page smoking questionnaire" (p. 41) with the aim of helping people to give up smoking. This system was evaluated "by recruiting 2553 smokers, sending 1/3 of them letters produced by STOP and the other 2/3 control letters, and then measuring how many people in each group managed to stop smoking"[3]. In this case the system was evaluated in the real world to see whether it has the desired effect, of helping people to quit smoking. The results showed that there were no relevant differences between the STOP letters and the control letters.

3.1 A general overview

Table 3 shows the evaluation methodologies used in the papers that we examined. With respect to the frequency of use of intrinsic compared to extrinsic methods, Table 3 confirms the trend identified in Gkatzia and Mahamood (2014). Gkatzia and Mahamood found out that the 74.7% of the papers used the intrinsic evaluation method. In our analysis we found that 83% of the papers used this methodology. However, we note that with respect to Gkatzia and Mahamood's results, we have an inverted trend between the use of an extrinsic method compared to both intrinsic and extrinsic. Indeed, Gkatzia and Mahamood found that 15.2% of the papers used extrinsic methods, against the 6% we get in our analysis, and 10.1% of the papers used both methodologies, where our analysis shows that 11% of the papers use a combination of both.

Evaluation methodologies	# of papers			
	Text2Text	Kb2Text	Mm2Text	Total
Intrinsic human only	13	1	-	14
Intrinsic automatic only	9	-	1	10
Extrinsic (human) only	2	-	-	2
Intrinsic human & Intrinsic automatic	3	2	2	7
Intrinsic human & Extrinsic (human)	2	-	-	2
Intrinsic automatic & Extrinsic (automatic)	1	-	-	1
Intrinsic human & Intrinsic automatic & Extrinsic (automatic)	-	1	-	1

Table 3: Evaluation methodologies used.

Furthermore, our analysis confirms the trend between the use of automatic compared to human intrinsic evaluation methodologies. In Gkatzia and Mahamood (2014) the authors report that in 45.4% of the cases human evaluation is used, whereas in 38.2% of the cases automatic evaluation were adopted. Similarly, our analysis shows that between the papers that prefer intrinsic evaluation methods, 45% used human evaluation, 32% used automatic evaluation and 23% used both human and automatic evaluation.

Table 1 shows that in the period since 2016, there has been a considerable increase in the number of publications in this area. It therefore makes sense to ask whether this increase has been accompanied with a change in the evaluation methodologies used.

Evaluation methodologies	# of papers	
	2013-2015	2016-2018
Intrinsic human only	6	8
Intrinsic automatic only	1	9
Extrinsic (human) only	-	2
Intrinsic human & Intrinsic automatic	1	6
Intrinsic human & Extrinsic (human)	-	2
Intrinsic automatic & Extrinsic (automatic)	-	1
Intrinsic human & Intrinsic automatic & Extrinsic (automatic)	-	1

Table 4: Variation of the evaluation methodologies used between 2013 - 2015 and between 2016 - 2018.

Table 4 shows how the range of evaluation methodologies used has changed. Between the years 2013 - 2015 only intrinsic evaluation methodologies were used – with 75% of papers using human evaluation, 12.5% using automatic evaluation and 12.5% using both methodologies – for the years between 2016 - 2018 extrinsic evaluation methods have also been introduced. Indeed, although the majority of the papers in this period (79%) used intrinsic evaluation methods, 7% of papers used extrinsic evaluation methods and 14%

[3]See Ehud Reiter's blog https://ehudreiter.com/2017/01/19/types-of-nlg-evaluation/

used both the methodologies. We can also see a change in the tendency to use intrinsic methods. Between the years 2016 - 2018, 35% of the papers used human evaluation (a decrease of 40% from the years between 2013 - 2015), 39% of the papers used automatic evaluation (a 26.5% increase on the years between 2013 - 2015) and 26% of the papers used both methodologies (a 13.5% increase on the years between 2013 - 2015).

3.2 Automatic evaluation

Table 5 presents a list of automatic metrics used in the papers studied in the present research. From our analysis it turns out that the most used automatic metric is BLEU followed by METEOR. Note that Table 5 only describes those that use the specified metrics; other papers use metrics that are defined for the specific aims described in the paper that introduces them.

Evaluation methodologies	# of papers			
	Text2Text	Kb2Text	Mm2Text	Total
BLEU (Papineni et al., 2002)	8	3	2	13
METEOR (Banerjee and Laviel, 2005)	4	2	1	7
ROUGE (Lin and Och, 2004)	3	1	-	4
Precision	4	-	-	4
Recall	4	-	-	4
F1	4	-	-	4
Accuracy	2	0	1	3
ΔBLEU (Galley et al., 2015)	-	-	1	1
Embedding Greedy (Rus and Lintean., 2012)	-	1	-	1
Others	5	-	-	5

Table 5: Automatic metrics used.

In our survey we found out that 31% of the papers used just a single metric, whereas the other 69% used more than one. The average is 2 metrics per paper, with a minimum of 1 metric (6 papers) and a maximum of 5 metrics (1 paper). In almost 50% of cases (9 papers), 3 metrics were used. We noticed that only a single paper used an embedding based metric (see Sharma et al. (2017)). In a majority of studies, word-overlap based metrics were used (see Sharma et al. (2017)).

In the last few years, many studies in NLG have shed light on the correlation between human judgement and automatic metrics. The results, which have shown how this correspondence is somewhat weak[4], shed doubt on the feasibility of using these metrics for evaluating the overall quality of a system.

To the best of our knowledge, the area of AQG is currently missing a study which aims to verify the correlation between human judgement and automatic metrics[5]. Such research would have two merits: on one hand, this kind of meta-evaluation study would give a better characterisation of the general problem. On the other hand, the research could provide guidance to researchers about which metric is most appropriate in evaluating a particular model or system.

In conclusion, we believe that research in AQG would benefit from a systematic study that aims to clarify the relation between different evaluation methodologies.

3.3 Human evaluation

Among the various human evaluation methodologies, eliciting *quality judgments* is most common: human annotators are asked to assess the quality of a question based on criteria such as question's grammaticality and fluency. Only two papers used a *preference judgement* methodology, in which the human annotators are asked either to assess pairwise preference between questions or given a couple of questions, one human generated and one automatically generated, assess which one is automatically generated (or which one is the the human generated). One of these papers also used the other methodology of eliciting quality judgements.

Quality judgment methodologies typically ask annotators to use Likert or rating scales to record their judgements. In our analysis, we found that 56% of the papers used some kind of numerical scale. For example, human annotators were often asked to assess the grammaticality of a question on a scale from 1 (worst) to 5 (best). On the other hand, 44% of the papers used a linguistic (or semantic) scale. In these cases, human annotators were typically asked to classify the questions in some category such as coherent, somewhat coherent or incoherent. The number of categories used in the Likert or rating scales by the papers that adopted quality judgment methodologies are shown in Table 6.

Only three papers used more than one scale in the evaluation. One of these uses a free scale in

[4]For an in depth discussion of this point we refer to Reiter and Belz (2009) and to Gatt and Krahmer (2018), especially section 7.4.1 and the references presented there.

[5]Yuan et al. (2017) raise some doubts about the capacity of BLEU to effectively measure the quality of systems used in *Text2Text* tasks.

which the annotators have to choose a positive integer to count the inference steps necessary for answer a question.

Number of categories	# of papers			
	Text2Text	Kb2Text	Mm2Text	Total
2	6	-	-	6
3	6	-	2	8
4	1	1	-	2
5	8	1	-	9
7	-	1	-	1

Table 6: Number of categories used in the Likert or rating scales.

Table 6 shows that the two most common number of categories used in the Likert or rating scales are 3 and 5. In a recent paper, Novikova et al. (2018) suggest that the use of a continuous scale and relative assessments can improve the quality of human judgments. Although in our study, we found 2 papers that used relative assessment, we did not find any papers that use a continuous scale.

Another interesting point is the number of annotators used in the evaluation. This number varies a lot from paper to paper. We found a minimum of 1 annotator (2 papers) to a maximum of 364 annotators (1 paper). Taking the papers which provided information on the number of annotators used (24 papers), and removing five papers that used 53, 63, 67, 81 and 364 annotators – these can be seen as outliers – we found out that the average number of annotators used was almost 4. The most common number was two annotators, used by 29% (7 papers) of the papers. 3 annotators were used by 17% (4 papers) and 4 annotators were used by 13% (3 papers). The others paper used 5, 7, 8 or 10 annotators.

There is a similar breadth in the number of output questions used (that is, the questions generated by the systems), and the criteria (that is, the question features to be checked) used in the evaluation. The number of questions ranged from a minimum of 60 questions (1 paper) to a maximum of 2186 (1 paper). Amongst those papers which actually provide this information (17 papers, or 65%), we found out that the average number of questions used per paper is almost 493. 7 papers (27%) did not report this information, whereas 2 papers (8%) report information about the amount of data from which the questions were generated, without giving the exact number of questions used for the evaluation.

Regarding the criteria used, we noticed that 35% of the papers (8 studies) used an overall quality criterion, that is, a single criterion which was used to evaluate the question's overall quality. On the other hand, 52% of the papers (12 studies) used specific criteria, for example, question grammaticality, question answerability, etc. A full list of these criteria is shown in Table 7. 13% of the papers (3 studies) used both specific criteria and an overall criterion. As Table 7 shows, there is a wide assortment of criteria used across the set of collected papers.

Criterion used	# of papers			
	Text2Text	Kb2Text	Mm2Text	Total
Grammaticality	7	-	-	7
Semantic correctness	4	-	-	4
Answer existence	3	-	-	3
Naturalness	2	1	-	3
Question type	3	-	-	3
Clarity	3	-	-	3
Discriminator quality	3	-	-	3
Relevance	2	-	-	2
Correctness	2	-	-	2
Well-formedness	1	-	-	1
Key selection accuracy	1	-	-	1
Corrected retrieval	1	-	-	1
Fluency	1	-	-	1
Coherence	1	-	-	1
Timing	1	-	-	1
Inference step	1	-	-	1
Question diversity	1	-	-	1
Importance	1	-	-	1
Specificity	1	-	-	1
Predicate identification	-	1	-	1
Difficulty	1	-	-	1
Overall criterion	7	2	2	11

Table 7: Criteria used.

As we can see from Table 7, the specific criteria are mainly used in the *Text2Text* task. Just two criteria are used in the *Kb2Text* task and none in the *Mm2Text*, where an overall quality criterion was preferred. We note that some criteria, for example timing or importance, are specific to one of the aims of the paper in which they are used. Indeed, as shown in the introduction, we can find different aims behind the papers' motivations. We note that among the papers analyzed here, often only little information is provided about the evaluation guidelines[6]. We cannot exclude that, given the evaluation guidelines, some of the criteria presented in Table 7 can be collapsed together. That is, it is possible that different researchers use different names in order to check the same question feature. In order to have a better way to check the quality across systems, we suggest that researchers should publish the evaluation guidelines used in the evaluation, as well as the quantitative results.

[6]Human evaluations are driven by some annotation guideline which is a direct manifestation of some annotation scheme. Whereas the latter characterize the criteria to be evaluated, the first strictly define such criteria and suggest how they should be evaluated.

Table 8 supplies an overview about the IAA reached in the human evaluations. We note that 54% of the papers (14 studies) did not supply this information. Only one of the two papers that used preference judgments reported the agreement between evaluators. In that paper, Fleiss' κ was used to measure the IAA reached between 3 to 5 evaluators. The results, for three batches with different evaluators and questions, were 0.242, 0.234 and 0.182. Table 8 presents the IAA results reported by the papers that used quality judgement methods. Between the papers that reported this information, we found that the IAA was measured in 26 cases and 9 of these were measured with two different coefficients, for a total of 35 IAA values. The agreements were measured for specific criteria or for the overall quality criterion. In one case the agreement over all the criteria was reported. It

Metric used for calculate IAA	# of criteria measured	Average	Min.	Max.
Cohen's κ	14	0.46	0.10	0.80
Krippendorff's α	2	0.143	0.05	0.236
Fleiss's κ	4	0.45	0.33	0.62
Pearson's r	4	0.71	0.47	0.89
Average measure	9	0.80	0.50	0.91
k no better specified	2	0.085	0.08	0.09

Table 8: Measures of Inter-Annotator Agreement.

is notable that the agreement reached in the various evaluations is generally quite low. Indeed, following Artstein and Poesio (2008), only agreement greater than or equal to 0.8 should be considered. Quoting Artstein and Poesio, p. 591:

> Both in our earlier work (Poesio and Vieira 1998; Poesio 2004a) and in the more recent (Poesio and Artstein 2005) efforts we found that only values above 0.8 ensures an annotation of reasonable quality. We therefore felt that if a threshold needs to be set, 0.8 is a good value.

Taking this as an appropriate quality threshold, among the papers that report IAA, very few evaluations should be considered appropriate. More specifically, we found that only 23% (8 over 35 values) of the evaluations reported IAA scores that were greater than or equal to 0.8.[7]

Checking the agreement for number of annotators we found that in the case with 364 annotators the IAA, measured for two criteria and a not better specified κ, was between 0.08 and 0.09. We found only 2 cases for 5 evaluators, which reported a value of 0.05 for Krippendorff's α and an average measure of 0.89. Two papers used 4 annotators altogether: one reported a value of Krippendorff's α of 0.236, with the other reporting a Pearson's r of 0.71. Another paper used 3 evaluators and the Fleiss's κ to measure the IAA for 4 criteria. The results are reported in Table 8.

All other papers reporting an IAA measure were in evaluations that used 2 annotators. The results of these cases are shown in Table 8 highlighting Cohen's κ, the Average measure and Pearson's r.[8]

There are sometimes attempts to design the experimental methodology to improve the level of IAA. In order to improve the agreement, one paper collapsed two score classes into one, whereas two papers allowed a difference of one score between the annotators rating. Two examples of the latter case are the maximum value for Cohen's κ and the maximum value for the average measure reported in Table 8.

We conclude this section by noting that the problem of a low IAA was present also in the Shared Task Evaluation Challenge (QG-STEC) (Rus et al., 2010). In that case, an attempt to improve the IAA for task B was carried out by Godwin and Piwek (2016). Godwin and Piwek define an interactive process in which the annotators can discuss their opinions about the criteria used in the evaluation. At the end of the evaluation process, repeated three times with three annotators on different data each time, they got high IAA with a peak of 0.94 for one of the five criteria used in the evaluation.

Although other papers (see for example Bayerl and Paul (2011), Lommel et al. (2014) and Hwee Tou Ng and Foo (1999)) propose techniques which aim to improve the IAA, in a recent paper (Amidei et al., 2018) we suggest thinking carefully about this practice in the case of NLG tasks. Indeed, if evaluation results have to inform generation sys-

[7] Using the popular Krippendorff's Kappa scales of interpretation (Krippendorff, 1980) – where any data annotation with agreement in the interval [0.8, 1] should be considered good, agreement in the interval [0.67, 0.8) should be considered tentative, and data annotation with agreement below 0.67 should be discarded– we conclude that 43% (15 out of 35) of the evaluations should be considered tentative.

[8] Regarding Pearson's r, we should clarify that in the case of 2 evaluators, IAA was measured for 3 criteria and not 4 as reported in table 8. However, because the Pearson's r measured for the fourth criterion was 0.71, that is the average value, the Pearson's r measure in the case of 2 evaluator is exactly the one shown in Table 8. Furthermore, for the average measure there is a case with 5 annotators. Removing that case, the average for 2 annotators is 0.79.

tems developers of the extent to which they can improve the communicative power of their systems, the aim of attaining a higher IAA runs the danger of biasing system developers towards ignoring important aspects of human language. An unchecked and unquestioned focus on the reduction of disagreement among annotators runs the danger of creating generation goals that reward output that is more distant from, rather than closer to, natural human-like language.

3.4 Extrinsic evaluation

As shown in Table 3, extrinsic evaluation methodologies are rare in the area. As reported by Gkatzia and Mahamood (2014) this is generally true for NLG tasks. Amongst the papers that have chosen to use this kind of evaluation technique, human judges were used in 4 times out of the 6. In the papers where human judges were not used, the Question Generation (QG) system was tested as a component of a Question Answering (QA) system. The performance was evaluated by checking the difference between the QA system without the use of the QG system against the performance of the QA system with the use of the QG system. The aim of those papers was to improve QA systems by creating more accurate question/answer pairs to be used for training purposes.

As a consequence of the different tasks in play, the other papers used humans in different ways. We can find tasks such as: answer the generated questions or use the generated questions in a web page and then answer a survey about the utility of those questions. Or also: engage in a conversation with a chatbot which involves a question-based dialogue, and then rate the conversations.

Also in this case, the number of humans involved in the evaluation varies from paper to paper, ranging from 2 to 81. In contrast to the case of intrinsic human evaluation, in this case the IAA is not reported. We note that human agreement in extrinsic evaluation is not as relevant as in the case of intrinsic evaluation. Indeed, for intrinsic evaluations, agreement is required to have a reliability and validity measure of the evaluation scheme and guidelines (Artstein and Poesio, 2008). The agreement measure should gather evidence that different humans can make similar judgements about the questions evaluated. This fact, following Krippendorff (2011) should allow us to answer the question of: *"how much the resulting data can be trusted to represent something real"?* (page 1). In human intrinsic evaluation, the agreement can be seen as a measure of the replicability of the results. For example, Carletta (1996, p. 1) wrote:

> At one time, it was considered sufficient... to show examples based on the authors' interpretation. Research was judged according to whether or not the reader found the explanation plausible. Now, researchers are beginning to require evidence that people besides the authors themselves can understand and make the judgments underlying the research reliably. This is a reasonable requirement because if researchers can't even show that different people can agree about the judgments on which their research is based, then there is no chance of replicating the research results.

In the case of extrinsic methods, the evaluation aim is to check if the generated sentences fulfil the task for which they were generated. To test this, humans need to use those sentences in real contexts. Now, humans make use of the same tools in different ways, and similarly they answer questions in different ways. For this reason, it is not expected that humans reach similar results in a real context of language use.

4 Discussion

Although systems and tools have been developed in the AQG area over the last few years, Table 1 illustrates that this has not been accompanied by similar improvements in evaluation methodologies. Indeed, with the exception of the Shared Task Evaluation Challenge (QG -STEC) (Rus et al., 2010), no attempts have been undertaken to introduce a common framework for evaluation that allows for comparisons between systems.

We have seen that in human evaluation, different criteria and scales/categories are used. To address this, we recommend that researchers share their evaluation guidelines, and work towards adopting common guidelines that can used to check quality across systems. Furthermore, out of the papers examined here, the problem of evaluation validity emerges. In those studies where it has actually been reported, the IAA is generally low. Also in this case we suggest researchers systematically determine the IAA and share their results, as

well as ideas to attempt to understand and classify any divergences between annotators.

Automatic evaluation can be thought of as a technique to provide a way to standardize the evaluation. Unfortunately, a comparison of human and automatic evaluation is missing in the area. This makes it difficult to understand to what extent the automatic metrics capture the systems' quality. Lacking such comparison, and following Reiter (2018), we suggest considering metrics such as BLEU as tools for systems' diagnostic more than evaluation techniques able to measure the output quality of the systems.

There is scope for more extrinsic evaluation, which can *"provide useful insight of domains' need, and thus they provide better indications of the systems' usefulness and utility"* (Gkatzia and Mahamood, 2014, p. 60). Unfortunately, extrinsic evaluations are not yet widely used.

Though the QG-STEC evaluation scheme has only limited uptake, with the much increased popularity of AQG, it is timely to revisit and address the need for a shared evaluation scheme. The variety of evaluation methodologies, as brought to light by the present work, demonstrates how difficult it currently is to check question quality across generation systems. This prevents us from understanding the actual contributions that are made by new generation systems that are being introduced ever more frequently.

We conclude this section with the following observation. The problem of having a high degree of variation in methodologies is compounded by the use of different datasets in the evaluation phase (see Table 9). The use of a common dataset for evaluation – as suggested by the Shared Task Evaluation Campaign (STEC) (Gatt and Belz, 2010) – could remove bias coming from the training phase. This is particularly true for generation systems that use machine learning techniques. We note that the high variability in the dataset used in the evaluation phase is also due to the variation in the papers' motivations. However, Table 9 suggests that the aim of building a common framework for AQG tasks should involve creation of a dataset to be used only for evaluation purposes. If we want to understand the degree to which a system advances the state of the art, we need to compare different systems on the same dataset, or better, a set of datasets, of course, using the same evaluation methodologies.

Tasks	Dataset or source of test articles
Text2Text	SQuAD; MS-MARCO; WikiQA; TriviaQA; TrecQA; Wikinews; Penn Treebank; QG-STEC datasets; StackExchange; Wikipedia; OMG! website; Project Gutemberg; ReadWorks.org; Engarde corpus; CrunchBase; Newswire (PropBank); textbook from OpenStax and Saylor; not specify TOEFL book; not specify science text books; not specify course Web page; not specify news articles not specify teachers articles; 40 people's personal data.
Kb2Text	Ontology documenting K-12 Biology concepts; SimpleQuestions; Freebase; WikiAnswers.
Mm2Text	COCO-QA; COCO-VQA; IGC_{Crowd}; Bing; COCO; Flickr.

Table 9: Dataset used.

An open evaluation platform in which researchers share their evaluation methodologies and their results can be effective to compare the quality across systems. In such a platform, the shared evaluation methodologies, alongside some datasets used only for evaluation, can be used by researchers to test their systems' performance and the results can be recorded in the open platform. Another benefit of this platform could be to generate an evolutionary process which allows the community to select the evaluation methodologies that are considered more effective.

5 Conclusion

In this paper we have analysed 37 papers which were about AQG. The aim of our work was to study the evaluation methodologies used in the area. Our work confirms the conclusion of Gkatzia and Mahamood (2014) for NLG in general. In AQG we lack a standardised approach for evaluating generation systems. Indeed, our overview shows a quite variegated evaluation landscape which prevents comparison of question quality across generation systems. A careful look at the papers published in the AQG area in the last five years shows how little attention has been given to the evaluation methodology introduced in the QG-STEC. Given the ever-increasing number of publications in the area, a common framework for testing the performance of generation systems is urgently needed.

Acknowledgments

We warmly thanks the anonymous reviewers for their helpful suggestions.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Satanjeev Banerjee and Alon Laviel. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Petra S. Bayerl and Karsten I. Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. δbleu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 26–31.

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. *in E. Krahmer and Mariët Theune (Eds.),* Empirical Methods in Natural Language Generation *Springer-Verlag, Berlin Heidelberg*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65–170.

Dimitra Gkatzia and Saad Mahamood. 2014. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60.

Keith Godwin and Paul Piwek. 2016. Collecting reliable human judgements on machine-generated language: The case of the qg-stec data. In *Proceedings INLG16*, pages 212–216.

Chung Yong Lim Hwee Tou Ng and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop: Standardizing Lexical Resources*, pages 9–13.

Emiel Krahmer and Mariët Theune (Eds.). 2010. *Empirical Methods in Natural Language Generation*. Springer-Verlag, Berlin Heidelberg.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. Retrieved from `http://repository.upenn.edu`.

Nguyen-Thinh Le, Tomoko Kojiri, and Niels Pinkwart. 2014. *Automatic Question Generation for Educational Applications The State of Art. In: van Do T., Thi H., Nguyen N. (eds) Advanced Computational Methods for Knowledge Engineering.* Springer, Cham.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using using longest common subsequence and skip-bigram statistics. In *Proceedings ACL'04*, pages 605–612.

Arle Lommel, Maja Popovic, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In: MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation, Reykjavik, Iceland.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankMe: Reliable human ratings for natural language generation. In *Proceedings of NAACL-HLT*, pages 72–78.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics.*, pages 311–318.

Paul Piwek and Kristy. E. Boyer. 2012. Varieties of question generation: Introduction to this special issue. dialogue and discourse. *Dialogue and Discourse*, 3(2):1–9.

Sheetal Rakangor and Y. R. Ghodasara. 2015. Literature review of automatic question generation systems. *International Journal of Scientific and Research Publications*, 5(1):1–5.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter, Roma Robertson, and Liesl M.Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58.

Vasile Rus, Zhiqiang Cai, and Art Graesser. 2008. Question generation: Example of a multi-year evaluation campaign. In *Rus, V. and A. Graesser (eds.), online Proceedings of 1st Question Generation Workshop*, pages 25–26.

Vasile Rus and Mihai C. Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the Sixth International Natural Language Generation Conference (INLG)*, pages 7–9.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv:1706.09799*.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, and Saizheng Zhang. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25.