Human Evaluation and Statistical Analyses on Machine Reading Comprehension, Question Generation and Open-domain Dialogue

Tianbo Ji

B.Sc., M.Sc.

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

School of Computing

Dublin City University

Supervisors:

Prof. Gareth Jones

Prof. Yvette Graham, Trinity College Dublin

Prof. Qun Liu, HUAWEI Noah's Ark Lab

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Tianbo Ji

ID No.: 17210629

Date: 31 August 2022

Contents

List of Figures

List of Tables

Human Evaluation and Statistical Analyses on Machine Reading Comprehension, Question Generation and Open-domain Dialogue

Tianbo Ji

Abstract

Evaluation is a critical element in the development process of many natural language based systems. In this thesis, we will present critical analyses of standard evaluation methodologies applied in the following Natural Language Processing (NLP) domains: machine reading comprehension (MRC), question generation (QG), and open-domain dialogue. Generally speaking, systems from tasks like MRC are usually evaluated by comparing the similarity between hand-crafted references and system-generated outputs using automatic evaluation metrics, thus these metrics are mainly borrowed from other NLP tasks that have been well-developed, such as machine translation and text summarization. Meanwhile, the evaluation of QG and dialogues is even a known open problem as such tasks do not have the corresponding references for computing the similarity, and human evaluation is indispensable when assessing the performance of the systems from these tasks. However, human evaluation is unfortunately not always valid because: i) human evaluation may cost too much and be hard to deploy when experts are involved; ii) human assessors can lack reliability in the crowd-sourcing environment. To overcome the challenges from both automatic metrics and human evaluation, we first design specific crowd-sourcing human evaluation methods for these three target tasks, respectively. We then show that these human evaluation methods are reproducible, highly reliable, easy to deploy, and cost-effective. Additionally, with the data collected from our experiments, we measure the accuracy of existing automatic metrics and analyse the potential limitations and disadvantages of the direct application of these metrics. Furthermore, in allusion to the specific features of different tasks, we provide detailed statistical analyses on the collected data to discover their underlying trends, and further give suggestions about the directions to improving systems on different aspects.

Acknowledgements

I would firstly like to express my special appreciation and great thanks to my supervisors Prof. Yvette Graham and Prof. Gareth Jones. It is an honour for me to work with outstanding supervisors like them. Many thanks to Yvette for bringing me into the field of NLP evaluation, for helping me to overcome the challenges during my research, for offering experienced suggestions, and for guiding me to the right research direction. I would also like to thank Gareth for providing me valuable and insightful advice on my writing and research skills, which enables me to carry out the publication of papers and accomplish this thesis.

Special thanks to Prof. Qun Liu who was my first year supervisor for letting me join ADAPT Centre. Without his help, I will never have the opportunity to take the first step on the research of NLP. I would also like to thank Qun for inviting me to take an internship in Huawei Noah's Ark Lab, and it is a great experience of working with many talented researchers.

I would like to appreciate the external and internal examiner: Prof. Verena Rieser and Dr. Jennifer Foster, as well as the chairperson of my viva: Dr. Hyowon Lee. Thanks for letting me have an enjoyable experience of defending my PhD, for providing useful comments on this thesis, and for the potential directions of my future work.

I would also like to thank Dr. Longyue Wang, I will never forget the days when we work and cook together. Also, many special thanks to Chenyang Lyu for discussing and conducting the experiments together with me. In addition, I would like to thank Dr. Liting Zhou, Lemin Chan and Quanwei Sun, it is always enjoyable and comfortable to live in the same house with such friendly roommates.

Special thanks to Prof. David Stifter, Dr. Fangzhe Qiu, and Dr. Elliott Lash for inviting me to work as a research assistant in the Chronologicon Hibernicum project from 2017 to 2019. It is very interesting to participate in the research of Early Medieval Irish.

I am very honored to be a member of ADAPT Centre, and I really enjoy working together with so many kind, helpful and experienced colleagues. I thank Science Foundation Ireland for providing funding of my research which allows me to fulfill my dream of being a scientist since childhood. I would also thank the anonymous crowd-sourcing workers to help me complete my work. Thanks to Ireland and lovely Irish people, living in this beautiful country is unforgettable.

Last but certainly not least, I hope to express sincere my gratitude to my family. Thanks to my parents for supporting me in every part of my life.

Chapter 1

Introduction

Natural Language Processing (NLP) enables computers to perform a wide range of language-related tasks. Similar to many fields of engineering, evaluation enables to investigate the strength and weaknesses of NLP technologies and systems (Dušek, Novikova, and Rieser, 2020), and their development is influenced by the evaluation methods. Such methods are applied to evaluating their effectiveness at performing the tasks which they are designed to perform. For example, the improved performance of recent neural network based models that are published at top-tier conferences is generally reported as an increase in their evaluation metric scores (Xu et al., 2021; Zou et al., 2021; Herzig et al., 2021). Machine translation (MT) is the task of automatically translating text from one natural language into another. As one of the most widely-researched and challenging topics in NLP, the evaluation methods used in the assessment of MT have been the focus of much attention. The methods developed for the evaluation of MT have been well-developed and indicate ways in which they highly influence it of other NLP domains. For example, BLEU is a metric proposed for evaluating MT systems (Papineni et al., 2002) that has become the frequently used methods of evaluating many NLP tasks such as image captioning and text summarization. Although BLEU is a relatively mature MT evaluation method, its performance for evaluation of other tasks has been criticised because of its weak correlation with human judgement (Kilickaya et al., 2017; Elliott and Keller, 2013). Despite this criticism of BLEU, its popularity somehow shows no indication of decline (Sai, Mohankumar, and Khapra, 2022). One possible cause is the challenge of designing a new metric for a specific task. For example, ROUGE is a metric proposed for text summarization (C.-Y. Lin, 2004), that is again criticised due to weakly correlating with human judgement (Novikova, Dušek, Cercas Curry, et al., 2017). In addition, challenges also exist when improving current evaluation metrics. For instance, Galley et al. (2015) attempt to improve BLEU's performance by adding a greater number of references. Such human-annotated references however, are generally too costly to be feasible. Therefore, these existing challenges relating to evaluation of NLP tasks give rise to our interest in investigating into and overcoming them.

In this thesis, we examine the evaluation of three NLP tasks: machine reading comprehension (MRC), question generation (QG) and dialogue systems, and investigate the degree to which the standard evaluation methodologies applied in each area are sound and dependable for distinct task. The reason for selecting these NLP tasks is that they either have no established effective evaluation method yet or they still directly use evaluation methods created for distinct tasks.

Take for example the task of machine reading comprehension (MRC), it is a sub-task of question answering (QA) that aims to enable a machine to answer specific questions using documents containing the answers. It requires the use of NLP methods capable of, in some sense, understanding human language. Next, the question generation (QG) task aims at generating meaningful questions according to the corresponding answers given a context (e.g., a set of related documents). QG can be utilized for augmenting a MRC dataset by adding more synthetic questions on which a MRC model can be fine-tuned (Shinoda, Sugawara, and Aizawa, 2021), and to generate related questions using a given passage for educational purpose (Kurdi et al., 2020). Finally, dialogue systems enable human users to communicate with machines and computing applications through natural language. Dialogue can be classified into two main categories: task-oriented and open-domain. A task-oriented dialogue system focuses on helping a user achieve a specific goal, while an open-domain dialogue system aims to enable humans to converse with a machine in natural language in a similar manner to engaging with a real person.

Passage:	SouthPark is a shopping mall named after the affluent SouthPark neighborhood the mall is located in. The mall is located approximately five miles (8 km) south of Uptown Charlotte, North Carolina at the corner of Sharon and Fairview Roads. With 1790000 sqft, SouthPark is the largest mall in Charlotte and the Carolinas, as well as one of the most profitable malls in the country with sales at over $700 per square foot. It is the 10th largest on the East Coast and is the 28th largest in the United States. SouthPark is the most congested shopping area in the United States during Black Friday weekend. Black Friday is the day following Thanksgiving Day in the United States (the fourth Thursday of November).
Question:	What day in November is Southpark most congested?
Answer:	the fourth Thursday

(a) An example of the MRC task where A MRC system takes the **passage** and **question** as the input, and is expected to output the **answer**.

Passage:	Stalybridge Celtic Football Club is an English football club based in Stalybridge, Greater Manchester. They are currently members of the Northern Premier League Premier Division and play at Bower Fold. The team traditionally plays in a blue and white strip. Fred Mace (October quarter 1895 – 5 November 1962) was an English professional footballer who played as a goalkeeper. Born in Hayfield, Derbyshire, he began his playing career in local-league football with Godley Athletic and Copley Celtic. In 1919, he joined Lancashire Combination side Stalybridge Celtic. The club was one of the founder members of the Football League Third Division North two years later, and Mace made one league appearance for them. Stalybridge left the Football League in 1923 to play in the Cheshire County League, where Mace was described as one of the best goalkeepers in the competition.
Answer:	Stalybridge Celtic Football Club
Question:	Fred Mace played for which English football club in 1919?

(b) An example of the QG task where A QG system takes the **passage** and **answer** as the input, and is expected to output the **question**.

History:	A: hello what are doing today? B: I am good, I just got off work and tired, I have two jobs. A: I just got done watching a horror movie. B: I rather read , I've read about 20 books this year.
Utterance (A):	Wow! I do love a good horror movie.
Response (B):	But a good movie is always good.

(c) An example of the open-domain dialogue task where A dialogue system takes the **history** and **utterance** as the input, and is expected to output the **response**.

Figure 1.1: Examples of the three NLP tasks: MRC, QG and dialogue.

Figure 1.1 gives the examples of these NLP tasks. A MRC system is required to provide the **answer** to the given **question** using the **passage**, as shown in Figure 1.1a. Figure 1.1b shows an example of QG, where a QG system is expected to generate the **question** whose corrected answer is the given **answer**, according to the given **passage**. In addition, Figure 1.1c is the example of open-domain dialogue. A dialogue system is expected to provide an appropriate **response** to the **utterance** from a human according to the conversation **history**.

During the development of these NLP technologies, evaluation is a crucial step in understanding how well they perform the task for which they are being developed, and to identify elements which either must or could be improved. Evaluation can also have the effect of influencing system development, as results deemed successful by a given evaluation methodology can naturally steer research in that direction. Despite the importance of evaluation methodologies for individual NLP tasks, automatic evaluation metrics originally developed for other text-based tasks, MT most notably, are commonly employed to evaluate a range of distinct NLP tasks. This is likely due to the fact that devising a suitable evaluation of new tasks is often challenging, time consuming and expensive.

For MRC, a key challenge for evaluation arises from the range of different categories of MRC task. Generally speaking, evaluation of MRC needs to examine whether a system can read and understand unstructured text and then answer questions about it. Based on the types of questions given to the system and the form of answers that a system should provide, the evaluation requirement and methods for MRC can vary. The evaluation of some MRC tasks, such as cloze test and span extraction for instance, is straightforward and carried out by exact match (EM) and F1 score. Evaluation of other MRC tasks, like free answering for example, is far more challenging. EM, for example, simply gives credit for answers that exactly match the gold answer and F1 gives a credit for overlapping parts between the words in system-generated answers and gold references. Since tasks like cloze test expect a fixed system output, metrics are able to accurately reflect the quality of systems. However, evaluating a free answering MRC task, namely a system can freely answer a given question needing no fixed type of outputs, is substantially more challenging, since the task itself does not provide a specific requirement on the form of the output from a corresponding system. Despite its significantly increased complexity, evaluation of the complex task of free answering MRC still employs MT automatic metrics such as BLEU. However, these metrics mostly rely on lexical overlap, which is not effective for proper evaluation of this type. For this task, an appropriate answer to one question could express the same meaning as its ground-truth reference even without any lexical overlap, rendering word overlap based metrics less accurate in this task.

Commonly applied evaluation methods in the QG task, such as METEOR (Banerjee and Alon Lavie, 2005) and NIST (Doddington, 2002), are again word overlap based metrics. These measure the word overlap between a system-generated question and a ground-truth reference. However, the evaluation of QG should take into account the fact that there may exist several appropriate questions. For example, with a passage describing Ireland, the country located in western Europe, two questions $Q1$ and $Q2$, where $Q1$="*What is the capital of Ireland?*" and $Q2$="*Which city in the Leinster province has the largest population?*", can share the same answer "*Dublin*". In other words, it is fairly appropriate for a QG system to generate either $Q1$ or $Q2$ given the same passage and answer, despite few overlap between the meanings of $Q1$ and $Q2$. We deem it the *one-to-many* nature of the QG task, as *one* passage and answer can lead to *many* meaningful questions. A word overlap based metric will however incorrectly assess $Q2$ with a lower score if it takes $Q1$ as the reference because of the lack of word overlap between these two questions. Therefore, it is necessary to accommodate the context and answer for the evaluation of QG. Word overlap based metrics are generally criticized because they ignore the corresponding context and answer when evaluating a question. This makes them incapable of distinguishing appropriate questions in the QG task due to the failure of consideration of the *one-to-many* nature. Additionally, these methods cannot assess QG systems in different dimensions, for example, metrics like BLEU are incapable of telling how appropriate the question words (such as what, who, and so on) are. Furthermore, these metrics thoroughly relies on the ground-truth references, making them impossible to evaluate QG systems if there exist no pre-created reference in practice. For example, a specific application of QG is for educational purpose by automatically generating exam-style questions (Kurdi et al., 2020) using a passage. In this instance, reference questions are generally unavailable.

Our focus next moves to evaluating dialogue systems which, in contrast to aforementioned MRC and QG, is quite removed from MT, but the evaluation challenges facing it are an exacerbated example of those that face MT evaluation. In MT evaluation, it is possible to automatically compare a system output to a human-generated translation to get an approximation of the quality of the system output by a metric like BLEU. Although it does not always produce reliable conclusions as it has been demonstrated that simply improving the BLEU score is unnecessary and insufficient to reflect the real improvement of MT systems (Callison-Burch, Osborne, and Koehn, 2006), in general such a method appears not to correlate too weakly with human judgment and consequently can be viewed as steering MT system development in a suitable direction (Graham, Baldwin, Moffat, et al., 2014). The source of errors in MT evaluation metrics, although not substantial, comes from the fact that many possible ways to adequately translate a sentence into another language exists, when a system happens to produce good output that does not resemble the handcrafted reference, it is unfairly penalised. The source of such errors in these estimates is much worse in the case of dialogue system evaluation. For such a conversational task, there really is no "correct " next statement from the machine and subsequently no way to automatically compare the output with a ground truth of some kind. Evaluation is challenging firstly due to the definition of high quality systems being itself rather challenging to define, and recent research generally lacks a clear and specified definition of the aspect of quality (Howcroft, Belz, et al., 2020). Even given a definition of how a high-quality system should respond to a question appropriately, it is still not clear how to measure "appropriateness". Meanwhile, the popular automatic evaluation methods used in dialogue are again usually borrowed from other domains, including BLEU and METEOR from MT, and ROUGE from automatic summarization, with the latter itself being an adaption of BLEU metric. However, these metrics have been proven to have poor correlation with human judgment when it comes to evaluation of dialogue systems, especially for the task of open-domain dialogue. Human assessment is therefore more prevalent in recent studies of dialogue systems since automatic metrics fail to be close to humans (Finch and Choi, 2020). However, the approach taken by human evaluation is often prohibitively expensive and too time-consuming to be practical. Evaluation of dialogue systems is in dire need of an accurate, affordable and effective human evaluation method.

In summary, all of these tasks incorporate the same over-arching evaluation challenges. In this thesis, we first propose three research questions relating to evaluation of these NLP tasks in Section 1.1, while each question with respect to a certain task will be individually answered in corresponding sections. Also, we overview the structure of this thesis in Section 1.2.

1.1 Research Questions

In general, machine reading comprehension, question generation and dialogue systems share some automatic evaluation metrics originally designed for other tasks, like ROUGE and BLEU. Meanwhile, task-specific metrics are also applied, for example, Answerability (Nema and Khapra, 2018) is proposed for QG while USR is for dialogue systems. The employment of different metrics can potentially result in diverse conclusions for a certain task, where some metrics may even disagree with each other (Peyrard, 2019; Bhandari et al., 2020). Hence, this raises our first research question:

RQ 1: *Within each domain of interest, how accurately do existing automatic metrics measure system performance?*

Moreover, despite the common practice of adopting metrics from distinct tasks in a certain task, the choice of evaluation methods may have a negative influence in the development of systems in each domain. For example, both BLEU and ROUGE are frequently applied in MRC, while they are initially proposed for machine translation and text summarization, respectively. This gives rise to our second research question:

RQ 2 *What are the limitations and disadvantages of the direct application of evaluation metrics from MT and other domains to entirely distinct tasks for system development in each area?*

In addition, since the evaluation of each domain can be challenging because of the potential limitations and disadvantages of existing evaluation methods, a newly proposed evaluation approach for each domain is necessary. Thus, we present our third research question:

RQ 3 *Can more appropriate new methods of evaluation be designed that are feasible given the limited time and resources available in operational settings?*

To address these research questions with regard to involved domains, we propose a new human evaluation method for each domain and conduct relevant experiments. We additionally conduct a self-replication experiment to verify the reliability by investigating the consistency at system-level. Such consistency is assessed by correlation between the results of initial and self-replication experiments. Also, we run statistical significance tests to prevent the system ranking from occurring simply by chance. After the verification, we can then address these research questions. For **RQ 1**, we will measure the performance of metrics by the degree of their correlations with human judgement of which the data is from the experiments of our proposed human evaluation method. In regards to **RQ 2**, we further inquire into effects of employed metrics on the development of systems. For example, we investigate whether the system ranking is consistent when applying distinct automatic evaluation metrics in one specific task. In terms of addressing **RQ 3**, the self-replication experiment is used to investigate whether a newly proposed method is appropriate as such a method should be highly reliable. Results shows that our methods can achieve a nearly perfect correlation, for example, the Pearson correlation coefficient of our proposed evaluation method for MRC reach as high as $r = 0.986$. Furthermore, we report the details of human evaluation experiments including costs and elapsed time, to ensure our proposed methods are feasible within a limited budget of time and resources.

1.2 Thesis Structure

In general, this thesis aims to answer the aforementioned research questions in terms of three distinct NLP tasks, and the structure of the thesis is described as follows:

Chapter 1 briefly introduce the three tasks: machine reading comprehension, question generation and dialogue system, as well as the facing challenges of their evaluation. Section 1.1 describes three relevant research questions, and Section 1.2 introduces the thesis structure.

Chapter 2 firstly provides the information of each task, including the detailed definition and commonly used datasets. Besides, we introduce their prevailing evaluation methods, including both automatic metrics and human evaluation. Section 2.1, Section 2.2 and Section 2.3 are respectively about MRC, QG and dialogue systems, where the performances of evaluation methods and their existing issues in each related task are introduced as well.

In Chapter 3, we introduce a newly proposed human evaluation method for the MRC task which overcomes current problems in MRC evaluation, as well as the corresponding experiment deployed on the crowd-sourcing platform. We provide the details of the experiment, including the involved systems and dataset, the statistics on the data collected from it such as worker pass rates and elapsed time, and the means of control the quality of human workers. We further compute the metric scores on the experiment data and analyze metric performances based on their correlation with the results of our human evaluation methods. In addition, a self-replication experiment, the system consistency on different applied evaluation methods, and the agreement among human raters are also included in this chapter.

Chapter 4 follows the structure similar Chapter 3, where we proposed a new human evaluation method for the QG task to solve existing issues. According to the results from the deployed experiment, we compute metric scores and their correlation with human raters. Additional experiments and analysis such as self-replication experiment, the system consistency and rater agreements are included, while the details of our experiments are available as well. Moreover, we propose a new automatic metric after the deployment of our human evaluation in this chapter.

In terms of the evaluation of dialogue systems, we conduct the related experiments of the proposed crowd-sourcing human evaluation method as introduced in Chapter 5. In general, we provide details of experiments that investigate its capacity of surmounting challenges which current dialogue evaluation is still facing. Meanwhile, the consequent analysis in this chapter can denote whether our evaluation method is highly cost-effective and reliable, and whether it can be deployed on a large scale within an affordable budget.

In regards to MRC, QG and dialogue respectively, the three research questions raised in Section 1.1 are addressed according to the results and analyses of corresponding experiments in Chapter 3, 4 and 5.

Finally, Chapter 6 provides the conclusions of this thesis, as well as the plans for the future researches.

Chapter 2

Current Evaluation Techniques and Existing Challenges

Since this thesis mainly focuses on the evaluation of three NLP tasks: MRC, QG and dialogue systems, in this chapter, we firstly provide their specified definitions, such as what the aim of each task is. We describe what their applications can be as well. In addition, we briefly introduce some commonly applied datasets in each task, including the datasets that systems of each task for our evaluation experiments in this thesis are trained on. Besides, the prevailing methods used in the evaluation of each task are introduced in detail (Section 2.2 for MRC, Section 2.3 for QG and Section 2.3 for dialogue systems). We additionally mention challenges which current evaluations are facing or the known issues that the applied evaluation methods have for these tasks.

2.1 Machine Reading Comprehension

Generally speaking, the aim of machine reading comprehension (MRC) tasks is to develop machines with the ability to automatically provide the correct answer to a question regarding a presented context in the form of natural language. The given context can have diverse forms, ranging from a few short sentences to a set of long documents, and the machine should be able to leverage the entire context in order to

Table 2.1: Commonly applied datasets in each type of the MRC task.

Category	Dataset
Cloze Test	Shmoop (Chaudhury et al., 2019) CliCR (Šuster and Daelemans, 2018) CLOTH (Xie et al., 2017) Who Did What (Onishi et al., 2016) CNN / Daily Mail (Hermann et al., 2015)
Multiple Choice	ReClor (Yu et al., 2020) CosmosQA (L. Huang et al., 2019) SocialIQA (Sap et al., 2019) MCScript (Ostermann, Modi, et al., 2018) SemEval-2018 Task 11 (Ostermann, Roth, et al., 2018) RACE (Lai et al., 2017) MovieQA (Tapaswi et al., 2016) WikiQA (Y. Yang, Yih, and Meek, 2015) MCTest (Richardson, Burges, and Renshaw, 2013)
Span Extraction	SubjQA (Bjerva et al., 2020) DROP (Dua et al., 2019) Quoref (Dasigi et al., 2019) ROPES (K. Lin et al., 2019) DuoRC (Saha et al., 2018) HotpotQA (Z. Yang et al., 2018) SQuAD 2.0 (Rajpurkar, R. Jia, and Liang, 2018) TriviaQA (Joshi et al., 2017) SQuAD (Rajpurkar, J. Zhang, et al., 2016)
Free Answering	TweetQA (Xiong et al., 2019) DuReader (He et al., 2018) NarrativeQA (Kočiský et al., 2018) MS MARCO (Nguyen et al., 2016)

answer questions that are related to the given information. MRC tasks can be classified into the following four categories: Cloze Test, Multiple Choice, Span Extraction and Free Answering (D. Chen, 2018). These categories are principally determined by the form of the answers required by the system to produce, as described as follows:

- **Cloze Tests:** Given the context C with one word (or an entity) $a(a \in C)$ removed, cloze tests ask the machine to fill in the blank with the right word or entity a by maximizing the conditional probability $P(a|C - \{a\})$

- **Multiple Choice:** Given the context C, the question Q and a list of candidate answers $A = \{a_1, a_2, ..., a_n\}$, the multiple choice task is to select the right answers $a_i(a_i \in A)$ from A by maximizing the conditional probability $P(a_i|C, Q, A)$.

- **Span Extraction:** Given the context C, which consists of n tokens, that is $C = \{t_1, t_2, ..., t_n\}$, and the question Q, the span extraction task is to extract the continuous sub-sequence $a = \{t_i, t_{i+1}, ..., t_{i+k}\}(1 \leq i \leq i + k \leq n)$ from context C as the right answer to question Q by maximizing the condition probability $P(a|C, Q)$.

- **Free Answering:** Given the context C and the question Q, the right answer a in a free answering task does not have to be a sub-sequence in the original context C, namely either $a \in C$ or $a \notin C$. The task is to predict the right answer a by maximizing the conditional probability $P(a|C, Q)$.

Table 2.1 shows several popular datasets which are proposed for each type of MRC task. In this thesis, we will further use the results of the systems which are trained on the free-answering dataset NarrativeQA (Kočiský et al., 2018) in the MRC evaluation experiment.

Due to the complexity of MRC, there are a large number of relevant datasets, and Table 2.1 only lists a certain amount of typical datasets of each type. Dzendzik, Foster, and Vogel (2021) provided a comprehensive survey on 60 MRC datasets. According to the survey, Boolean is another type of MRC task where the questions expect a yes/no answer, including BoolQA (C. Clark et al., 2019) and AmazonYesNo (Dzendzik, Vogel, and Foster, 2019). PubMedQA (Jin et al., 2019) is also a Boolean MRC dataset, while it contains questions that can be answered by "maybe". Meanwhile, some MRC datasets in Table 2.1 contain a proportion of yes/no questions (e.g., HotpotQA) or questions that *cannot be answered* (e.g., SQuAD 2.0), and such datasets are sometimes categorized as Mixed (Dzendzik, Foster, and Vogel, 2021).

In addition to classifying by the format of answers, other taxonomies are applied in MRC as well (Rogers, Gardner, and Augenstein, 2021). For example, MRC can be classified into four categories according to the format of questions, including *natural language questions*, *queries*, *cloze*, and *story completion*. Besides, the format of context/passages is another indicator of classification, which can be further characterized in terms of the modality and amount. According to the modality of context, there are seven types of MRC: *unstructured text*, *semi-structured text*, *structured knowledge*, *image*, *audio*, *video* and, *other combinations*. With regard to the amount of context, MRC can be be classified into: *single source*, *multiple sources*, *partial source*, and *no source*.

2.1.1 Evaluation Metrics

Each category of MRC task listed above in Section 2.1 employs a combination of evaluation metric. With the exception of free answering, the answers in these MRC tasks are always predefined, whether a sub-sequence from the given context (span extraction) or a subset of given candidates (cloze tests & multiple choice), meaning that the correctness of such answers is always binary. Because of the simple form of answers, accuracy is effective enough for evaluating models from cloze test and multiple choice, while the performance of a model in a span extraction task can also be simply assessed by accuracy-based metrics such as exact match (EM) and F1 score.

Among these four kinds of MRC tasks, free answering is somehow the most challenging as the machine needs to fully reason over the given context and further generate an answer to the given question in the form of a fluent and natural text. In addition, since the forms of generated answers have no specific restriction in free answering, the composition of a correct answer can range from a single word to a set of sentences in many situations. Due to the difficulty and complexity of the free answering MRC task, its evaluation is also highly challenging. In this case, a common practice is to directly employ existing automatic metrics from other related domains, such as machine translation and text summarization. In the following, we will introduce these aforementioned common evaluation metrics for all four kinds of MRC tasks in detail.

Accuracy

Accuracy of system outputs with respect to the standard references is the most frequently used evaluation method in cloze tests and multiple choice MRC tasks. The accuracy of a system can simply be calculated by following:

$$Accuracy = \frac{n}{m} \tag{2.1}$$

where m is the number of given questions that a model is asked to answer, and n is the number of generated answers that are correct.

As a variety of Accuracy, **Exact match** (EM) is often employed in the span extraction MRC task. Given a question with its corresponding reference answer, EM can measure whether the system-generated answer exactly matches its standard reference answer or not. The value of EM for one single question will be 1 if the predicted answer is 100% the same as the ground-truth answer, and 0 otherwise.

F1 score

F1 score is an evaluation metric used in classification tasks, while it is often utilized together with EM in the span extraction MRC task. Both system-generated candidates and reference answers are treated as a bag of tokens, and the F1 score is computed by:

$$\begin{aligned} P &= \frac{TP}{TP + FP} \\ R &= \frac{TP}{TP + FN} \\ F &= \frac{2 \times P \times R}{P + R} \end{aligned} \tag{2.2}$$

where P is Precision, R is Recall, and F is the F1 score, while the values of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) as described in Table 2.2. In addition, F1 score is in fact a special case of F-β score when $\beta = 1$, of which the general equation is:

$$F_\beta = (1+\beta^2)\frac{P \times R}{\beta^2 \cdot P + R} \tag{2.3}$$

Table 2.2: Definition of TP, FP, TN and FN in MRC evaluation.

	tokens in reference	tokens not in reference
tokens in candidate	TP	FP
tokens not in candidate	FN	TN

BLEU

Bilingual Evaluation Understudy (BLEU) is a method that is originally proposed for evaluating the quality of MT systems (Papineni et al., 2002), and is widely-used in free answering MRC tasks. BLEU computes the level of correspondence between a system-generate answer and the reference answer by calculating the number of n-gram matching segments. These matching segments are thought to be unrelated to their positions in the entire context. The more matching segments there are, the better the quality of the answer is.

In detail, the BLEU score of a predicted answer a and a reference answer r can be computed by Equation 2.4:

$$\begin{aligned}
P_n(a,r) &= \frac{\sum\limits_{k \in a} \min(Count_k, RefCount_k)}{\sum\limits_{k \in a} Count_k} \\
BP &= \begin{cases} 1 & \text{if } l_a > l_r \\ e^{(1-l_r/l_a)} & \text{otherwise} \end{cases} \\
\text{BLEU}{-}N &= BP \cdot \exp\left(\sum_{n=1}^{N} w_N \log P_n(a,r)\right)
\end{aligned} \tag{2.4}$$

where P_n is the modified precision score between a and r, $Count_i$ is the number of times that a unique n-gram k occurs in the system output a, and $RefCount_k$ is the number of time that k occurs in the reference r. In addition, BP is the brevity penalty to penalize outputs that are too short, l_a and l_r are the numbers of words in a and r, N is the number of n-gram, and w_N is the weight for the modified precision of current n-gram whose common value in application is $w_N = N^{-1}$. Note that the corpus BLEU score is based on the modified precision score on the corpus rather than simply averaging the sentence BLEU scores.

In this thesis, the computation of BLEU score depends on the implementation from the Python module "nltk" (see `https://www.nltk.org/_modules/nltk/translate/bleu_score.html`).

GLEU

Since BLEU is initially designed for measuring the corpus performance, GLEU (Google-BLEU) is then proposed to overcome the drawbacks of evaluating a single sentence (Y. Wu et al., 2016). As a variety of BLEU, the GLEU scores is reported to be highly correlated with the BLEU score on a corpus level. GLEU uses the scores of precision and recall instead of the modified precision. The sentence GLEU score for the free answering MRC task can be computed by:

$$\begin{aligned} P &= \frac{\sum_{n=1}^{N} \sum_{k_n \in a} h(k_n, r)}{\sum_{n=1}^{N} \sum_{k_n \in a} h(k_n, a)} \\ R &= \frac{\sum_{n=1}^{N} \sum_{k_n \in a} h(k_n, r)}{\sum_{n=1}^{N} \sum_{k'_n \in r} h(k'_n, r)} \\ \text{GLEU} &= \min(P, R) \end{aligned} \tag{2.5}$$

where P and R are respectively Precision and Recall, function $h(x, y)$ returns the number of times that an n-gram x occurs in a sentence y, a is the predicted answer, r is the reference answer, k_n is the unique n-gram, and 4 is the default value of N

in practice.

Similar to BLEU, we compute GLEU scores using the "nltk" implementation in this thesis (see `https://www.nltk.org/_modules/nltk/translate/gleu_score.html`).

ROUGE

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is an evaluation metric developed for the assessment of Automatic Text Summarization tasks, but originally adapted as a recall-adaptation of the MT metric BLEU (C.-Y. Lin, 2004).

ROUGE-L is the most prevailing variant of ROUGE, where L denotes to longest common subsequence (LCS). The definition of LCS is a sequence of words that appear in the same order in both sentences. In contrast with sub-strings, such as n-gram, the positions of words in a sub-sequence are not required to be consecutive in the original sentence. Equation 2.6 introduces ROUGE-L which uses F-β score by the LCS between the predicted answer and the reference answer when applied in the free answering MRC task:

$$\begin{aligned} R &= \frac{lcs(a,r)}{len(r)} \\ P &= \frac{lcs(a,r)}{len(a)} \\ \text{ROUGE-L} &= (1+\beta^2)\frac{P \times R}{\beta^2 \cdot P + R} \end{aligned} \tag{2.6}$$

where P is Precision and R is Recall, function $lcs(x, y)$ returns the number of words the LCS between sentences x and y contains. The function $len(x)$ returns the number of words in sentence x, and a and r are the predicted and reference answers. β is a hyper-parameter and a practical value is $\beta = P/R$.

Besides ROUGE-L however, ROUGE could have a large number of possible variations according to varying combinations of its parameters, and thus employing ROUGE could result in distinct results for a single task (Graham, 2015). It has eight choices of n-grams/sub-sequences ($\{1, 2, 3, 4\}$-gram;L;W;S;SU), three choices of handling the single score (precision, recall or f-score), two choices of sum up individual scores (mean or median), together with two other binary configurations, including the application of stop-words and word-stemming. Such choices consist of $8 \times 3 \times 2 \times 2 \times 2 = 192$ potential variants of ROUGE in total.

ROUGE scores in this thesis are computed by the python module "rouge" (see `https://pypi.org/project/rouge/`).

METEOR

Metric for Evaluation of Translation with Explicit ORdering (METEOR) was firstly proposed to make up for the disadvantages of BLEU, such as lack of Recall and the inaccuracy of assessing a single sentence (Banerjee and Alon Lavie, 2005). METEOR involves different stages before the computation, including: exact token matching, WordNet synonyms, Porter stemmer and paraphrases.

For the free answering MRC task, METEOR will generate a set of mappings between the predicted answer a and the reference answer r according to a set of given stages, where a mapping is the connection between two unigrams. With the set of mappings, the METEOR score can be computed by Equation 2.7:

$$\begin{aligned} F_{mean} &= \frac{P \times R}{\alpha \cdot P + (1-\alpha) \cdot R} \\ \text{Pen} &= \beta \cdot (\frac{chunk}{m})^{\gamma} \\ \text{METEOR} &= F_{mean} \times (1 - \text{Pen}) \end{aligned} \tag{2.7}$$

where F_{mean} is the weighted harmonic mean of Precision $P = m/u_a$ and Recall $R = m/u_r$ as m is the number of unigrams in mappings, and u_a/u_w is the number of all unigrams in a/r. Pen is the penalty where *chunk* is the number of chunks which consists of a set of abutting unigrams. The default value of hyper-parameters α, β and γ are $\alpha = 0.9$, $\beta = 0.5$ and $\gamma = 3$, respectively (Banerjee and Alon Lavie, 2005), but they can be tuned to maximally correlate with human judgements (Denkowski and A. Lavie, 2011).

Since an official JAVA implement of METEOR is available, we compute the

METEOR score in this thesis by directly invoking corresponding JAVA commands in our Python script as suggested (see instructions in `https://www.cs.cmu.edu/~alavie/METEOR/README.html`).

BERTScore

BERTScore is an automatic metric for evaluating the text generation task (T. Zhang et al., 2020). Instead of using exact match like the overlap-based metrics, BERTScore compute a similarity score between tokens in a candidate sentence and its reference by their contextual embeddings from the pre-trained model BERT (Devlin et al., 2019). Given a candidate answer a that has m tokens and a reference answer r that has n tokens, the BERT model can first generate the representations of a and r as $a = \langle a_1, a_2, \ldots, a_m \rangle$ and $r = \langle r_1, r_2, \ldots, r_n \rangle$, where a_i and r_i respectively mean the contextual embeddings of the i-th token in a and r. Then, the BERT score between the answer and the reference can be computed by Equation 2.8:

$$
\begin{aligned}
P_{\text{BERT}} &= \frac{1}{m} \sum_{a_i \in a} \max_{r_j \in r} a_i^\top r_j \\
R_{\text{BERT}} &= \frac{1}{n} \sum_{r_i \in r} \max_{a_j \in a} a_j^\top r_i \\
F_{\text{BERT}} &= \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}
\end{aligned}
\tag{2.8}
$$

where the final BERT score is the F1 measure F_{BERT} computed by precision P_{BERT} and recall R_{BERT}.

TER

As a derived form of Word Error Rate (WER) (Nießen et al., 2000) which was proposed for the automatic speech recognition task, the MT community proposed the metric Translation Error Rate (TER) by adapting WER for evaluating the quality of MT systems (Snover et al., 2006). Both WER and TER measure the number of edits needed to be taken to make a predicted sequence exactly match its given reference, but the main difference is that, WER uses the classical Levenshtein

Distance (Levenshtein et al., 1966) whose edit operations includes Replacement, Insertion and Deletions, while TER has an extra valid operation Shift to catch phrasal shifts. TER can be applied for the free answering MRC task, as Equation 2.9 represents the formula:

$$\text{TER} = \frac{\#R + \#I + \#D + \#S}{len(r)} \tag{2.9}$$

where # denotes the minimal number of an edit operation (R = Replacement, I = Insertion, D = Deletions and S = Shift) to take to convert the predicted answer a into the reference answer r, and the function len returns the number of words in its input sequence. Each operation of TER is treated as equal cost, and punctuation is isolated and treated in the same way as a word. Unlike other mentioned metrics however, TER score has two notable differences: i) although the TER score is practically within the range of 0-1, its theoretical range is actually 0 to infinity rather than 0 to 1; ii) the lower the TER score of a MRC system is, the higher its quality is.

Human Evaluation

A. Chen et al. (2020) proposed a dataset called MOCHA for evaluating MRC metrics where the $40,000$ human judgement scores on MRC model outputs are collected from a crowd-sourcing human evaluation experiment (A. Chen et al., 2020). First, human workers are asked to read the passage, question, correct answer, and predicted answer. Then, they should evaluate the answer correctness by giving a score that can *best reflects how closely a predicted answer captures the same information as the correct answer*. A 5-point scale is employed which is interpreted as: 1 – *completely wrong answer*, 2 – *mostly wrong*, 3 – *half right*, 4 – *mostly right*, and 5 – *perfect answer*.

A. Chen et al. (2019) trained a multi-hop pointer generator model (Bauer, Y. Wang, and Bansal, 2018) on the training sets of NarrativeQA and SemEval. Then, along with the model outputs, they respectively extracted 500 and 300 data points from the validation sets of NarrativeQA and SemEval for conducting the human evaluation experiment. This is an expert-based human evaluation, as two authors annotated all data points in-house. They are asked to rate *how closely a model output captures the same information as a gold answer* using a 1-5 scale.

2.2 Question Generation

As a way of learning, humans are capable of asking complex and creative questions when exposed to a new environment. One of the concerns emerging recently from the NLP community is whether a machine can also have the ability to ask questions which are appropriate or pertinent given a wide range of input formats. Such a concern then leads to the Question Generation (QG) task. In general, QG aims at generating meaningful questions based on the input information whose format can vary, for example, textual contexts, images, videos or even database. QG has various forms of applications, such as dialogue systems, generating educational content and data augmentation for question answering (Graesser et al., 2005; Shakeri et al., 2020; Lyu et al., 2021).

QG and MRC are deemed two interdependent and highly related tasks. Both tasks utilize textual contexts as input, while a QG system aims to generate questions using given answers and a MRC system can answer the given questions. Interestingly, recent research on QG suggest that the direct employment of MRC datasets on the QG tasks is theoretically feasible (Y. Kim et al., 2019; L. Wang et al., 2020; Cho et al., 2021).

Besides MRC datasets, there are a number of datasets that have been proposed by the QG community as well, with such datasets could have different formats of inputs. For example, LearningQ (G. Chen et al., 2018) is a dataset to encourage generation of questions for educational purpose, and requires a QG system to output a question given a context without the correct answer. Meanwhile, MQR (Multi-

Table 2.3: The information of commonly applied datasets in the QG task, including their initial proposed domains and input formats, where C=*the context*, A=*the answer which is a sub-string of* C, A'=*the answer which may not be a sub-string of* C, and Q=*the question in a wrong format*.

Domain	Dataset	Input format
MRC	HotpotQA (Z. Yang et al., 2018)	$C + A$
	SQuAD (Rajpurkar, J. Zhang, et al., 2016)	$C + A$
	MS MARCO (Nguyen et al., 2016)	$C + A'$
QG	MQR (Chu et al., 2020)	$C + Q$
	INQUISITIVE (Ko et al., 2020)	$C + A$
	LearningQ (G. Chen et al., 2018)	C

domain Question Rewriting) is a dataset that aims at converting an ill-formed question into a well-formed question according to the given context (Chu et al., 2020).

Table 2.3 introduces the commonly used datasets that are applied in the QG tasks. The formats of inputs to a QG system can vary when the applied dataset changes, as described in Table 2.3, where C is the context, A and A' mean the answer to the reference question, and Q is the ill-formed question. A means the answer must be a sub-string in C, while for A' it is not strictly the case. For example, the SQuAD dataset requires a QG system to generate a question according to a context and an answer which is a sub-string of the given context. We subsequently carry out a number of evaluation experiments on the data of QG systems that are trained on the HotpotQA dataset.

2.2.1 Evaluation Metrics

Most metrics for MRC tasks, especially free answering, can act as the evaluation methods for the corresponding QG task, with the exception of accuracy-based metrics such as **Exact Match** and **F1 score**. In the following, we will introduce commonly applied evaluation metrics in the QG task.

Metrics from other NLP domains

Similar to free answering MRC tasks, automatic evaluation metrics in regard to the n-gram overlap between two textual sequences are also suitable for judging the quality of a QG system, including **BLEU**(EQ. 2.4), **ROUGE**(EQ. 2.6), **METEOR**(EQ. 2.7). Meanwhile, BERTScore (EQ. 2.8) can be applied to the QG evaluation as well. These metric scores can be directly computed by the system-generated question, q, and the reference question, r, using their original formulae.

Answerability

Aside from the aforementioned evaluation methods - which are borrowed from other NLP tasks, an automatic metric called Answerability is specifically proposed for the QG task (Nema and Khapra, 2018). Nema and Khapra (2018) suggest combining it with other existing metrics since its aim is to measure how answerable a question is, something not usually targeted by other automatic metrics. For example, given a reference question r: "*What is the address of DCU?*" and two generated questions q_1: "*address of DCU*" and q_2: "*What is the address of*", it is obvious that q_1 is rather answerable since it contains enough information while q_2 is very confusing. However, any similarity-based metric is certainly prone to think that q_2 (ROUGE-L: 90.9; METEOR: 41.4; BLEU-1: 81.9) is closer to r than q_1 (ROUGE-L: 66.7; METEOR: 38.0; BLEU-1: 36.8). Thus, Answerability is proposed to solve such an issue. In detail, for a system-generated question q and a reference question r, the Answerability score can be computed as shown in Equation 2.10:

$$\begin{aligned} P &= \sum_{i \in E} w_i \frac{h_i(q,r)}{k_i(q)} \\ R &= \sum_{i \in E} w_i \frac{h_i(q,r)}{k_i(r)} \\ Answerability &= \frac{2 \times P \times R}{P + R} \end{aligned} \tag{2.10}$$

where i ($i \in E$) represents a certain type of elements in $E = \{R, N, Q, F\}$ (R = Relevant Content Word, N = Named Entity, Q = Question Type, and F = Function

Word). w_i is the weight for type i that $\sum_{i \in E} w_i = 1$. Function $h_i(x, y)$ returns the number of i-type words in question x that have matching i-type words in question y, and $k_i(x)$ returns the number of i-type words occurs in question x. The final Answerability score is the F1 score of Precision P and Recall R.

Along with using Answerability individually, a common practice is to combine it with other metrics as suggested when evaluating QG systems (Y. Chen, L. Wu, and Zaki, 2020; P. Lewis et al., 2020):

$$Metric_{mod} = \beta \cdot Answerability + (1 - \beta) \cdot Metric_{ori} \tag{2.11}$$

where $Metric_{mod}$ is a modified version of an original evaluation metric $Metric_{ori}$ using Answerability, and β is a hyper-parameter. As Nema and Khapra suggested, the values of w_i ($i \in E$) and β can be tuned to obtain a high correlation with human, and a few examples of choosing w_i and β are provided.

In this thesis, we combine it with BLEU to generate the Q-BLEU score using the default value of β, using the official implementation of Answerability which is available in `https://github.com/PrekshaNema25/Answerability-Metric`.

BLEURT

BLEURT is a trained evaluation metric which takes a candidate and its reference as input and gives a score to indicate how the candidate can cover the meaning of the reference (Sellam, Das, and Parikh, 2020). It uses a BERT-based regression model trained on the human rating data from the WMT Metrics Shared Task from 2017 to 2019. Note that BLEURT was proposed for evaluating models on sentence level, and no formal experiments are available for corpus-level evaluation. Therefore, the final BLEURT score of a QG system will be simply computed by the arithmetic mean of all sentence-level BLEURT scores in this QG evaluation experiment, since this is suggested by the authors of BLEURT (see the discussion available on `https://github.com/google-research/bleurt/issues/10`).

Human Evaluation

Although these prevailing automatic metrics mentioned above are widely employed for QG evaluation, criticisms on n-gram overlap-based metrics' ability of evaluating the quality accurately and comprehensively are also raised (Yuan et al., 2017). As a certain answer can potentially have a large number of corresponding plausible questions, simply computing the overlap rate between an output and a reference to reflect the real quality of a QG system seems not convincing. A possible solution is to obtain a larger number of "correct" questions per answer, as n-gram overlap-based metrics would usually benefit from multiple ground-truth references. However, this can bring new issues: i) adding more references over the entire corpora is not inferior to creating a new dataset which could be expensive and time-consuming; ii) it is not easy to identify the "correctness" of a question.

Hence, human evaluation is also involved when evaluating a newly proposed QG systems. A common practice is to randomly sample a few system-generated questions and ask human raters to score these questions on a n-point Likert scale. We will then introduce some examples of human evaluation in recently proposed QG models/systems.

X. Jia et al. proposed EQG-RACE to generate examination-type questions for the educational purpose (X. Jia et al., 2021). 100 outputs of the model are sampled and three expert raters are required to score these outputs in three aspects: Fluency – *whether a question is grammatical and fluent*, Relevancy – *whether the question is semantic relevant to the passage*, and Answerability – *whether the question can be answered by the right answer.* Each aspect uses a 3-point scale, and aspects are reported separately without an overall performance.

Knowledge-Driven Distractor Generation (KD-QG) is a framework with a knowledge base for generating various questions as a means of data augmentation (Ren and Zhu, 2020). For its human evaluation, three proficient experts are individually assigned 50 randomly-sampled items. They first judge whether an assigned item is reliable using a binary scale, and any item with a positive reliability score will be further assessed on its level of plausibility on a 3-point scale that is construed as: 0 – *obviously wrong*, 1 – *somewhat plausible* and 2 – *plausible*. The two aspects, reliability and plausibility, are treated separately without reporting any overall conclusion.

Answer-Clue-Style-aware Question Generation (ACS-QG) aims to generate questions together with the answers from unlabeled textual content (B. Liu et al., 2020). Instead of evaluating the questions alone, a sample is a tuple of (p, q, a) where $p =$ passage, $q =$ question and $a =$ answer. A total of 500 shuffled samples are assigned to 10 volunteers, where each volunteer receives 150 samples to ensure an individual sample is evaluated by 3 different volunteers. A sample is evaluated in three facets, depending on the degree to it is: Well-formed (yes/understandable/no) – *if the question is well-formed*, its Relevancy (yes/no) – *if the question is relevant to the passage*, and Correctness (yes/partially/no) – *if the answer is correct to the question*. The result of each facet is reported as a percentage rather than a summarized score.

Ma et al. proposed a neural QG model consisting of two mechanisms: semantic matching and position inferring (Ma et al., 2020). The model is evaluated by human raters in three aspects: Semantic-Matching, Fluency, and Syntactic-Correctness on a 5-point scale. However, the details about: 1) the number of evaluated samples; 2) the number of involved raters; 3) the type of human raters (crowd-sourcing or experts) are unfortunately not provided.

QURIOUS is a question generation pre-training method, and QURIOUS-based models are expected to outperform other non-QURIOUS models (Narayan et al., 2020). The authors conduct a crowd-sourcing human evaluation experiment to verify this. 30 passages with answers are randomly selected, and human raters will compare the questions from two distinct models. For each single comparison, three individuals are involved for the sake of fairness. In detail, a human rater is presented with a passage, an answer and questions A and B from two models, and is asked which question is better than the other in two aspects: natural - the question is fluent and written in well-formed English, and correct - *the question is correct given the passage and the answer.* Each comparison can have three distinct annotations: $(A = best, B = worst)$, $(A = equal, B = equal)$ or $(A = worst, B = best)$, and the final human score of a system in one aspect is computed by the number of times it is rated as *best* subtracting the number of times it is rated as *worst* following with dividing by the number of times it is evaluated in total.

Although the application of human evaluation somewhat prevalent in the QG evaluation, it still involves some considerable concerns:

1. The examples mentioned above individually use disparate evaluation options with only a few overlaps. Existing human evaluation methods for QG are generally model-specific, because the lack of a standard approach leaves the QG community with no criterion to refer to.

2. The vast majority of QG human evaluations are either expert-based or volunteer-based, while the former are normally expensive and the latter may result in a shortage of raters. In addition, the inconvenience of deploying human evaluation on a large scale can lead to a small sample size that may influence the reliability.

3. The details of human evaluation experiments are ambiguous, sometimes the sample size and number of raters are even unavailable. Although expert-based human evaluation is generally deemed to have a high level of annotator agreement, the corresponding information is never reported, making it hard to guarantee the reliability and validity of the evaluation experiments, especially crowd-sourcing human evaluation.

2.3 Open-Domain Dialogue Systems

Dialogue systems enable a machine to talk to users similar to a real human conversation partner, with the aim of being capable of generating a reasonable response according to the current context of a dialogue system's ongoing conversation in the form of natural language. There are two main categories: task-oriented and open-domain. A task-oriented dialogue system is designed to help a user reach a specific aim, such as road navigation or reserving a restaurant, and it is generally required to be completed within a prescribed number of conversation turns. Meanwhile, an open-domain dialogue system aims at talking to users with no specific goal. The conversation can be about a thing, a topic or complete chitchat. Task-oriented dialogue systems can be simply evaluated by straightforward approaches such as task success rate and F1 score. However, the focus of evaluation of dialogue systems is purely on evaluation of open domain dialogue systems, which is substantially more challenging, comprising one of the most (if not the most) challenging evaluation problems in NLP (Burtsev et al., 2018; Dinan, Logacheva, et al., 2019). We primarily attribute the occurrence of such challenges to the fact that there generally are a vast number of possible appropriate responses in real-world conversations. Additionally, the evaluation of dialogue relies on comparison with handcrafted reference dialogues, while such a means may incur substantial false-negative rates since many appropriate responses are unfairly penalized simply for not corresponding closely with references. Furthermore, evaluation further suffers from challenges with respect to the ability to fully take into account dialogue history.

In regard to automatic evaluation metrics of open-domain dialogue systems, there are two main approaches: referenced-based and reference-free evaluation. The former relies on comparison between a system-generated response and a gold reference and it can be further divided into two categories: word-overlap-based metrics and word-embedding-based metrics, while the latter assesses a dialogue system solely by the conversation history. In addition to automatic metrics, human evaluation is widely applied as well.

2.3.1 Word-overlap-based Evaluation Metrics

Like previous two NLP tasks, the default means of evaluating the quality of a dialogue system appears to be comparing a system output with one or a set of ground-truth references. The details of these metrics are already introduced, including BLEU (EQ. 2.4), GLEU (EQ. 2.5), ROUGE (EQ. 2.6), METEOR (EQ. 2.7). For open-domain dialogue task, the score of a word-overlap-based metric can be computed by a system-generated utterance u and a handcrafted reference r. For example, PersonaChat (S. Zhang et al., 2018) is a dataset consisting of dialogues between participants A and B, which are played by two workers when collecting in the corwd-sourcing platform. A dialogue system trained on the train set can be then evaluated using the test set. Given the history of a dialogue and the latest response from A, the system is expected to generate what B will respond. And the metric score can be computed using the generated response and the real response from B in the dataset.

2.3.2 Word-embedding-based Metrics

A word-embedding-based metric likewise requires an output and a reference, while it measures the quality by the similarity using their meanings instead of the word overlap rate. The metric will first understand each word in a sentence by word embedding - a method proposed by the information retrieval (IR) community that can encode a word into a vector (Salton, Wong, and C. S. Yang, 1975). For example, the aforementioned metric BERTScore (EQ. 2.8) is also a word-embedding-based metrics which is suitable for evaluating dialogue systems. The most popular means of word vector is word2vec (Mikolov et al., 2013). It can approximately represent the meaning of a given word using a vector which is calculated according to its frequency in the corresponding corpus, and concatenating the vectors of all words in a sentence can act as the representation of the sentence. A common approach of a word-embedding-based metric is to calculate the cosine distance between an system output and a reference using their representations (C.-W. Liu et al., 2016).

Embedding Average

Embedding Average (EA) is an algorithm that has been widely used in other NLP tasks such as sentence similarity (Wieting et al., 2015). It first computes a sentence-level embedding by the following:

$$R_s = \frac{\sum\limits_{w \in s} e_w}{|s|} \tag{2.12}$$

where R_s means the representation of a sentence s, e_w is the embedding vector of a word w in s and $|s|$ is the number of words in s. For the dialogue task, the EA score can be subsequently computed using cosine similarity as EA $= cos(R_s, R_r)$ given a system-generated response s and a ground-truth reference r.

Vector Extrema

Vector Extrema (VE) relies on the computation of sentence-level representation as well (Forgues et al., 2014). Given all the word vectors in a sentence, VE yields the i-th element in the sentence vector by Equation 2.13:

$$R_s(i) = \operatorname*{absmax}_{w \in s} e_w^i \tag{2.13}$$

where e_w^i is the i-th element of the word vector e_w and the function absmax returns the number whose absolute value is the largest among all numbers.

Greedy Matching

Instead of sentence-level representations applied in aforementioned metrics, Greedy Matching (GM), initially proposed for assessing intelligent tutoring systems (Rus and Lintean, 2012), uses standalone word embeddings to yield a score. Specifically, given a response s and a reference r for the dialogue task, the GM score can be computed by Equation 2.14:

$$
\begin{aligned}
G(s,r) &= \frac{\sum_{w\in s} \max_{w'\in r} cos(e_w, e_{w'})}{|s|} \\
GM(s,r) &= \frac{1}{2}\Big(G(s,r) + G(r,s)\Big)
\end{aligned}
\tag{2.14}
$$

where e_w is the embedding for the word w, and cos is the function for cosine similarity. As the equation of $G(s,r)$ is asymmetric, the final GM score $GM(s,r)$ should be the average of scores in two directions.

2.3.3 Reference-free Metrics

Evaluating open-domain dialogue systems by referenced-based evaluation is known to have several known issues however. First, in consideration of the fact that there exists a vast number of possible appropriate responses in real-world conversations, such appropriate responses are often unfairly penalized by reference-based metrics for not corresponding closely with references. In addition, these metrics fail to take into account dialogue history. Meanwhile, although the direct application of metrics from other tasks, such as BLEU from MT and ROUGE from text summarization, is the common practice in the evaluation of open-domain dialogue systems, these prevailing metrics have been criticized for their weak correlation with human judgement in other NLP tasks (Reiter, 2018; Graham, 2015; Graham and Qun Liu, 2016).

To address such issues, reference-free metrics are therefore proposed. Distinct from reference-based, reference-free metrics requires no pre-created reference for comparison. Instead, it can score a response of a conversation according to the textual context, while the overall quality of a full conversation history can be assessed by combining all responses in it. Compared to reference-based metrics, reference-free metrics are deemed to perform better according to their correlation with humans. Subsequently, two reference-free metrics, FED and USR, will be introduced in the following.

FED

FED (Fine-grained Evaluation of Dialog) is a reference-free unsupervised metric based on pre-trained models (Mehri and Eskenazi, 2020a). Since a pre-trained model is deemed to have the ability of generating a response according to a given context, FED can assess the quality of a conversation by how a model will respond. Given the content of a conversation, c, FED can score c as follows:

$$\text{FED} = \mathcal{L}_m(r_p|c) - \mathcal{L}_m(r_n|c) \tag{2.15}$$

where m is a pre-trained model, r_p and r_n are predefined positive and negative responses. In addition, $\mathcal{L}_m(r|c)$ computes the likelihood that the model m will generate a response r given c. We employed medium and large DialoGPT (Y. Zhang et al., 2020) as FED scorers, where the full list of predefined positive and negative responses are available in Table A.6 in Appendix A.

USR

The USR (UnSupervised Reference-free) is a supervised evaluation metric that can assess the quality of a conversation (Mehri and Eskenazi, 2020b). USR consists of three sub-metrics for the evaluation of different aspects: USR-MLM is to evaluate the understandability and naturalness, USR-DR(c) and USR-DR(f) are to evaluate the interestingness and consistency. As USR is a supervised evaluation method, it first fine-tunes the pre-trained model RoBERTa (Yinhan Liu et al., 2019) on the training set of the Topical-Chat corpus (Gopalakrishnan et al., 2019) by the open-sourced implementation (Wolf et al., 2019), acting as the USR-MLM sub-metric. Then, the fine-tuned RoBERTa is further fine-tuned the retrieval task using the Ubuntu dialogue dataset (Lowe et al., 2015) that can act as USR-DR(c) and USR-DR(f) sub-metrics. Given a conversation, the scores of the three sub-metrics are first computed, and a trained regression model is employed to combine these sub-metric scores into a final USR score as the overall measurement of the conversation quality.

2.3.4 Human Evaluation

Apart from automatic metrics, human evaluation is additionally widely applied for assessing open-domain dialogue. Human evaluation is commonly adopted in competitions regarding dialogue systems (S. Kim, D'Haro, et al., 2017; S. Kim, D'Haro, et al., 2016; Burtsev et al., 2018), such as Dialog System Technology Challenge (DSTC) and Conversational Intelligence Challenge (ConvAI) . Unfortunately, a common issue occurs that can potentially impact the validity of results when filtering the set of systems to be evaluated via automatic metric scores, since human evaluation is not unfortunately not applied as the initial method of evaluating systems only after filtering according to metric scores. Among a set of dialogue system candidates, only the top N (for example, top 10 or top 20) will move to the human evaluation stage, and others are directly eliminated, depending on the ranking of scores produced by automatic metrics. Nevertheless, this strategy can raise a problem that the best system according to human judgements may be inadvertently filtered out without the chance of participation in human evaluation, since automatic metric scores are known to be a poor substitute for human assessment.

In addition, challenges exists in the live evaluation as such evaluations are reported to be highly challenging. For example, the ConvAI2 competition allows any participant to interact with a dialogue system and to evaluate its performance when the conversation is completed (Dinan, Logacheva, et al., 2019). Unfortunately, the results of live evaluation are ultimately discarded because a large proportion of the collected conversations are deemed to be *senseless*, *offensive*, or *simply not in line with instructions* due to the lack of a means of filtering out invalid data.

On the other hand, competitions that (for one reason or another) do not release data and evaluation techniques into the public domain have reported relative success in terms of human evaluation. However, until such methods can be accessed and independently verified through replication studies, they will unfortunately have little impact. The first Amazon Alexa Socialbot Grand Challenge required human assessors to score how *coherent* and *engaging* conversations were on a 1–5 rating scale by two distinct groups: volunteer Amazon employees (experts), and general Alexa users (crowds) (Ram et al., 2018), are reported to achieve a correlation of overall scores for the two types of human assessors at 0.93. The absolute average rating across all chatbots was reported to be 20% lower for experts compared to general users. In an additional effort to evaluate models, *conversational user experience, coherence, engagement, domain coverage, topical diversity*, and *conversational depth* were assessed (1–5 scale), with combined scores reported to correlate with those of general users at $r = 0.66$. In addition to methods and data not being publicly available, correlations are difficult to interpret since no detail is provided about the number of judgments on which the correlation is calculated for example.

With the exception of dialogue-related competitions that generally aim to include human evaluation of systems, the evaluation of newly proposed automatic evaluation metrics for open-domain dialogue usually involves human evaluation as well. In other words, these metrics, such as aforementioned USR and FED, require a human evaluation data set on which to evaluate the results of themselves as the proving of superiority and validity. However, this often raises new issues because of the application of inappropriate statistics, and these misuses are introduced in the following.

When proposing USR, Mehri and Eskenazi (2020b) conduct the human evaluation experiment that can assess the quality of dialog for a range of criteria using various rating scales: *understandable* (0–1 rating scale), *natural* (1–3), *maintains context* (1–3), *interesting* (1–3), *uses knowledge* (0–1); *overall quality* (1–5). Despite human evaluation being carried out by experts inter-annotator agreement levels varied depending on criteria being measured, ranging from as low as 0.298. Additionally, although correlations between human assessments are reported as significant at $p < 0.01$, despite such statistics often being reported for correlations, they are unfortunately not very meaningful in terms of their impact on correlation interpretation and can be somewhat misleading. Contrary to common expectations, even small effect sizes (low r) can produce very low p-values (strong significance) in such tests. Aiming to achieve a *significant correlation* is an extremely low bar to reach in terms of consistency, since a low p-value in this case simply rejects the null hypothesis that *the correlation is zero*.

Furthermore, Pang et al. (2020) deploy a crowd-sourcing human evaluation experiment on Amazon Mechanical Turk using a 1–5 rating scale when proposing the GPT-2 based holistic metric for the automatic evaluation of dialogues on four distinct aspects. The inter-rater agreement, reported as $r = 0.61$, is computed by averaging the Pearson correlations between pairs of human assessors. However, mean correlations are unfortunately difficult to interpret, since correlation coefficients are not additive, averages calculated in the usual way cannot be assumed to reflect central tendency, and unfortunately, the distribution of correlations is not reported (Alexander, 1990).

In addition to the aforementioned issues, human evaluation of open-domain dialogue systems rarely take into account the fact that differences in performance can occur simply by chance, significance tests are therefore necessary.

In general, the following challenges still remain in the evaluation of open-domain dialogue systems, with respect to both automatic metrics and human evaluation:

- Reference-based metrics: 1) may unfairly penalize an appropriate response because of its lack of conformity with the given reference; 2) cannot consider the dialogue history; 3) are reported as weakly correlated with humans;

- Human evaluation experiments associated with evaluation metrics in cases have applied inappropriate statistics, and the necessary significance test are not included;

- Using automatic metrics prior to human evaluation in conversational competitions can potentially filter out the best system according to human judgement.

- Lacking a quality control method potentially makes live human evaluations inapplicable;

- The results of some human evaluations are not reproducible as the accompanying data and detailed evaluation techniques are unavailable to the public.

Chapter 3

Evaluations on Machine Reading Comprehension

The evaluation of free-answering MRC is challenging as introduced in Chapter 1, since it only employs automatic metrics from other NLP tasks and still lacks a specific evaluation method. The applied metrics generally assess a system output according to its lexical overlap with the reference. In this chapter, to solve existing issues of MRC evaluation, we propose a new human evaluation method for the free-answering MRC task. This method can evaluate the adequacy of a system-generated according to its reference, and its fluency without the reference. We will introduce its methodology in detail, including the employed quality control method to prevent from collecting useless data from unreliable human workers. We will conduct corresponding human evaluation experiments based on the output of several MRC systems. A self-replication experiment based on the assessment of adequacy is deployed as well to ensure the reliability of our proposed method. With the results of the experiments, we can additionally examine the three research questions (**RQ**s) which we proposed in Section 1.1, with respect to the MRC task.

To begin with, Section 3.1 provides a brief review of the experiment design, including the human evaluation method we use and the interface shown to workers. Following that, Section 3.2 introduces details of the quality control method for crowd-sourcing human evaluations of MRC as well as the structure of tasks that are assigned to human workers. Section 3.3 then introduces used in this experiment, NarrativeQA. We also examine core statistics on the dataset, including distribution of word lengths of references and question categories. Finally, in this Section, the systems to be evaluated in the experiment are introduced. The final section of the chapter, Section 3.4, then provides the results of the experiment. We first report statistics of the collected data, including the pass rates of workers and assignments, the average duration and final expenditure. Additionally, we report the human scores and corresponding automatic metric scores to examine **RQ1** and **RQ2** according to the comparison between them. Additional statistical analyses on the experiment data show the validity of this method and provide suggestions for the improvement of MRC systems, with respect to **RQ3**.

3.1 Experiment Design

In this section, we provide information about the methodology of the proposed evaluation method, the details of the corresponding experiment design. We additionally introduce the means of choosing the sample size of human ratings.

3.1.1 Methodology

To overcome the above previous challenges of evaluating MRC task, we proposed a new crowd-sourcing human evaluation method by adapting Direct Assessment (DA) which has been successfully applied in other related fields, such as MT (Graham, Baldwin, Moffat, et al., 2016), multilingual surface realisation (Mille et al., 2020) and video captioning (Graham, Awad, and Smeaton, 2018). DA was first employed in evaluation on large-scale machine translation shared tasks at the Conference on Machine Translation (WMT) in 2016 and is subsequently the official human evaluation approach for ranking systems (Ondřej Bojar, Chatterjee, Federmann, Graham, Haddow, Huck, et al., 2016; Ondřej Bojar, Chatterjee, Federmann, Graham, Haddow, S. Huang, et al., 2017; Ondrej Bojar, Federmann, et al., 2018; Loic Barrault et al., 2019). DA is further adapted to evaluation of other tasks, such as automatic video captioning and multimodal machine translation (Awad, Butt, Fiscus, et al., 2017; Awad, Butt, Curtis, et al., 2018).

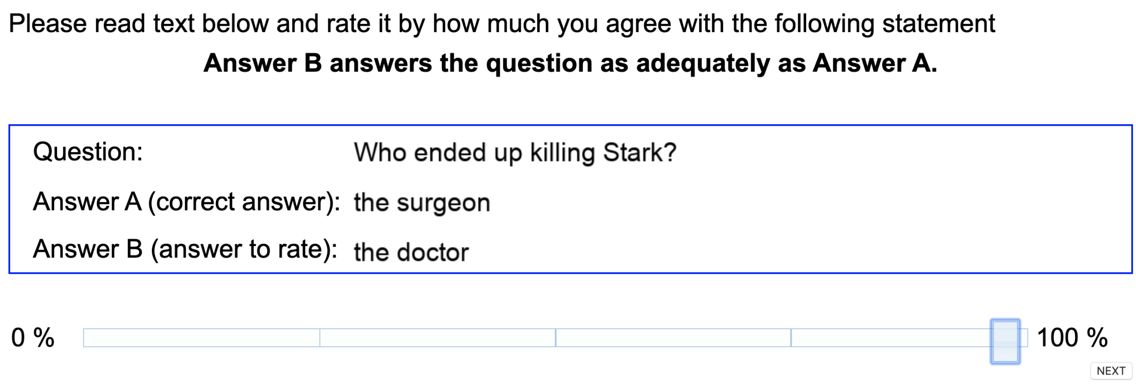
The following describes the advantages of DA compared to other existing human evaluation methods:

- Rather than commonly applied n-point scales, DA utilizes a continuous rating scale which facilitates the further fine-grained analyze of collected human rating scores.

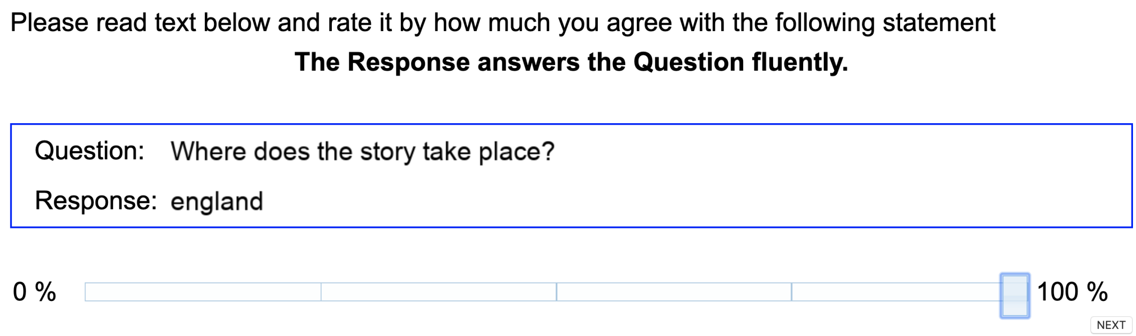
- Using statistical means, DA can employ quality-controlling mechanisms to ensure that the data we used for further analysis is valid and reliable.

- DA enables human evaluation to be deployed on a substantially larger scale at a feasible cost.

Accordingly, we adapt DA to the free-answering MRC task to evaluate MRC systems in two separate aspects: Adequacy and Fluency, where the former assesses the adequacy of system-generated answers and the latter is the measurement of how fluent a human will think an answer is. Amazon's Mechanical Turk (AMT) is the a crowd-sourcing platform where we deployed this human evaluation experiment (see `http://www.mturk.com`). In AMT, each worker will be assigned a "human intelligence task" (HIT). Figure 3.1 provides two examples of the interface shown to human assessors when evaluating adequacy and fluency. For the adequacy assessment, human assessors are asked to rate the degree to the agreement on how *Answer B answers the question as adequately as Answer* as Figure 3.1a shows, where *Answer A* is the human-generated reference answer and *Answer B* is the answer generated by the system to be evaluated. Separately, only the question and system-generated answer are shown to workers when evaluating the textual fluency of answers, with the instructional Likert statement replaced by *The response answers the question fluently* in Figure 3.1b.



(a) adequacy assessment



(b) fluency assessment

Figure 3.1: The interface of adequacy and fluency assessment as shown to AMT assessors separately

Although we collect ratings of adequacy and fluency, our further analyses are mainly based on the results from adequacy, while the stand-alone fluency will not be used for judging and ranking system. However, fluency can help to examine a system's ability of generating high quality texts from a different angle other than adequacy, we can thus apply fluency assessment as a secondary evaluation mechanism for our human evaluation, especially for re-ranking systems when systems have very similar levels of adequacy.

During the evaluation, each assessor is shown a question, a reference answer (Answer A) and a system-generated answer (Answer B) on a single screen at a time , meaning that one rater can evaluate only one answer at a time As it can be seen from Figure 3.1a. This is because multiple outputs on one screen can influence each other's scores resulting in a biased conclusion (Ondřej Bojar, Ercegovčević, et al., 2011). In addition, human assessors are not permitted to review previous answers, change previous judgments, or skip to the next answer. Particularly, the questions, system-generated answers and reference answers are actually rendered as images rather than real texts in order to prevent submissions from robots.

In this experiment, the passage will not be shown to workers during the evaluation of both adequacy and fluency. For adequacy, the passage is not included because we think the reference has contains enough information for workers to judge the adequacy of a generated answers, since adequacy measures the degree to which the information in the reference answer is preserved in the system-generated answer. Meanwhile, fluency determines whether the generated answer is well-formed in natural language, and it can be evaluated independently without the passage. However, passages are necessary when other evaluation criteria are examined. For example, a worker must refer to the passage if we would like to measure criteria such as *correctness* - "how correct is the answer to the question according to the given passage", or *relevancy* - "how relevant is the answer is to the given content". Nevertheless, we suggest to provide passages to workers when deploying our MRC human experiments in the future, since the information in the passage can also be useful for evaluating the adequacy, especially when a candidate answer and its reference share the same meaning but are expressed in different way. For example, we can allow workers to read the passage via an external link or a popup window.

3.1.2 Crowd-sourcing versus Experts

Human evaluation of NLP tasks is generally implemented by experts, such as linguists, professionals and relevant researchers. Despite the fact that expert-based human evaluation is deemed to properly estimate the performance of a NLP model/system, such evaluation is nonetheless confronted with challenges (Celikyilmaz, E. Clark, and Gao, 2020). First, the deployment of expert-based human evaluation usually relies on a traditional laboratory environment which is costly and time-consuming (Iskender, Polzehl, and Möller, 2020). In addition, experts are prone to injecting subjective opinions and biases during evaluation (Amidei, Piwek, and Willis, 2018).

Meanwhile, crowd-sourcing is proposed as an alternative to the expert which can be deployed on a large scale within an affordable budget and limited time. Although the quality of crowd-sourcing evaluation is always the concern since it may suffer from inaccurate assessment according to the results of experts (Gillick and Yang Liu, 2010; Lloret, Plaza, and Aker, 2013; Fabbri et al., 2020), evidence indicates that crowd-sourcing has the ability of generating high quality evaluation in NLP tasks (Snow et al., 2008; Nowak and Rüger, 2010; Graham, Baldwin, Moffat, et al., 2016; Graham, Awad, and Smeaton, 2018). Therefore, we conduct our experiment in a crowd-sourcing environment. And we believe this proposed crowd-sourcing human evaluation method can help to overcome the challenge that current human evaluation lacks a standard procedure which may result in high degree of variation (Lee et al., 2019).

3.1.3 Selecting the Sample Size

Before deploying the experiment, choosing a suitable sample size is also an indispensable part to ensure the conclusivity of rankings in human evaluation. Experiments with insufficient samples usually have low statistical power, and the conclusion drawn from such experiments may simply result from false negatives caused by low powered tests, such as Chinese to English news translation (Hassan et al., 2018).

Therefore, we followed a practical guide regarding sample sizes in machine translation (Graham, Haddow, and Koehn, 2020), and since we will employ an adaption of DA and the same significance test to identify differences between systems. Applying a similar sample size in this experiment should be appropriate for our purpose. It is however impossible to ascertain the actual sample size required prior to this experiment, since the variance of rating distributions for systems is not yet known. In future research, we suggest employing appropriate methods to estimate the sample size of human evaluation experiments via a power simulation (Howcroft and Rieser, 2021).

3.2 Quality Control In Our Experiment

In this section, we will introduce and motivate the application of quality control to crowd-sourcing data, the structure of a single HIT that is assigned to a rater as well as various quality controlling approaches.

3.2.1 Quality Control

Due to the anonymous nature of crowd-sourcing human assessment, evaluation quality can vary. For example, some workers just randomly rate everything because the anonymity of crowd-sourcing platform, rendering their ratings as useless. Quality control mechanisms are thus necessary for filtering out unreliable data generated by such human assessors.

A common-applied quality control strategy in crowd-sourcing is to include a gold standard set of items and only accept the data from workers who rate high scores to such items (J. Le et al., 2010; S.-W. Huang and Fu, 2013). Such a method is not effective however since human workers can easily "game" this strategy by simply assigning high scores to every item they are rating.

Instead of only applying a gold-standard quality-controlling method, we will operate three quality control methods in our research to improve the robustness to the "gaming turkers", including bad reference, repeat and reference answers:

- **Bad Reference**: a set of system-generated answers are randomly selected, then the degraded versions of them are automatically generated and paired with the original answers. The scores of such degraded answers are expected to be significantly lower than those of original answers if the human evaluators are credible. Such a strategy proposes a means of verifying how reliable a single individual is without comparing with any another human assessor, and the agreement with experts is no longer needed as well.

- **Repeat**: a set of system-generated answers are randomly selected, then they are directly copied to make a new set of repeated answers. The scores of repeated answers should be highly close to the original ones. They will be subsequently used for analysing the degree to raters' agreement.

- **Reference Answer**: a set of system-generated answers are randomly selected, they are then replaced by their references of according questions. Such answers are expected to received a extremely high score (≈ 100). Note that this approach is actually an application of gold-standard quality control as mentioned above, and we only employ it as an auxiliary method since its stand-alone application is at the risk of being gamed by crowd-sourcing workers.

3.2.2 Structure of HIT

With the three above methods, and answers generated by different systems to be rated, we can now create the HITs for our DA experiment. Figure 3.2 provides the graphical depiction of the items contained in a single 100-answer HIT, where each colored circle indicates a different kind of item. The composition of a HIT is described as follows:

- 10 system-generated answers and 10 "bad reference" versions (comprising a total of 20 answers)
- 10 system-generated answers and 10 "repeat" versions (comprising another total of 20 answers)
- 10 system-generated answers and 10 "reference answer" versions (comprising another 20 answers)
- 40 additional system-generated answers

In other words, a HIT containing 100 items is made up of: (a) 70 ordinary system-generated answers; (b) 10 bad-reference answers corresponding to 10 of above 70; (c) 10 exact repeats corresponding to 10 of above 70; (d) 10 reference answers corresponding to 10 of above 70. Although Figure 3.2 seems a hierarchical structure,

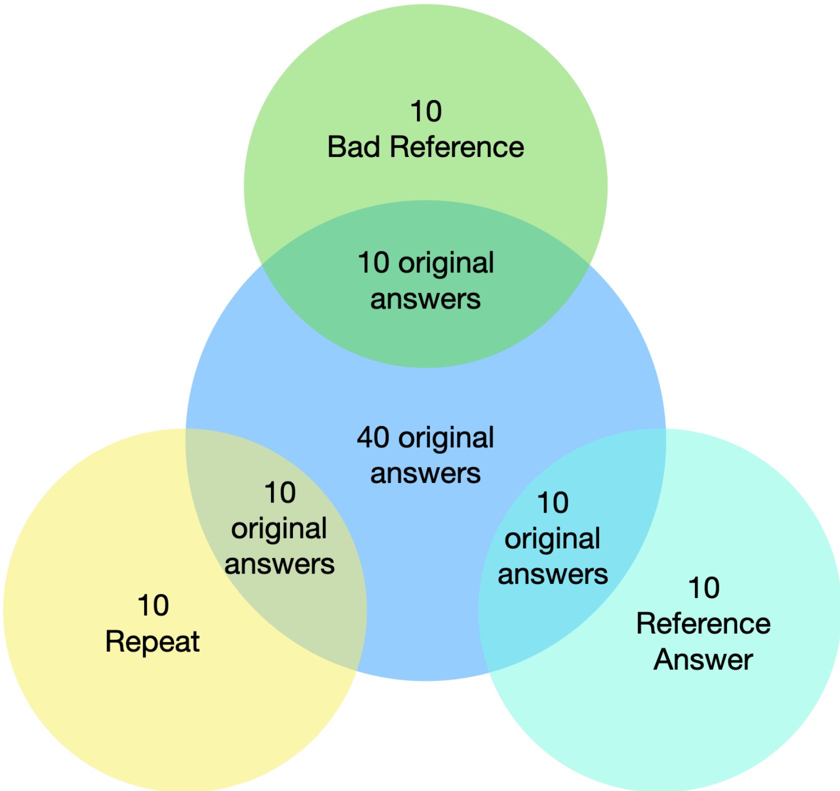


Figure 3.2: The structure of a 100-answer HIT

a HIT will be fully shuffled before assigned to a worker to guarantee an unbiased assessment.

3.2.3 Detailed Mechanisms of Quality Control

Since we have introduced the structure of HITs, we will now provide the details of generating quality-controlled versions of answers using the above mechanisms.

Bad Reference: To create a *bad reference* answer, we took the original system-generated answer and degraded its level of adequacy and fluency by replacing a random short sub-sequence s from the original answer with another string s', where s' is a randomly extracted sub-string from the human-generated answer belonging to another question, and s should have the same number of words as s'.

Given the original answer that consists of n words, the number of words that string s (and s') should have is subsequently decided on the following rules:

- for $1 \leq n \leq 3$, s comprises 1 word.

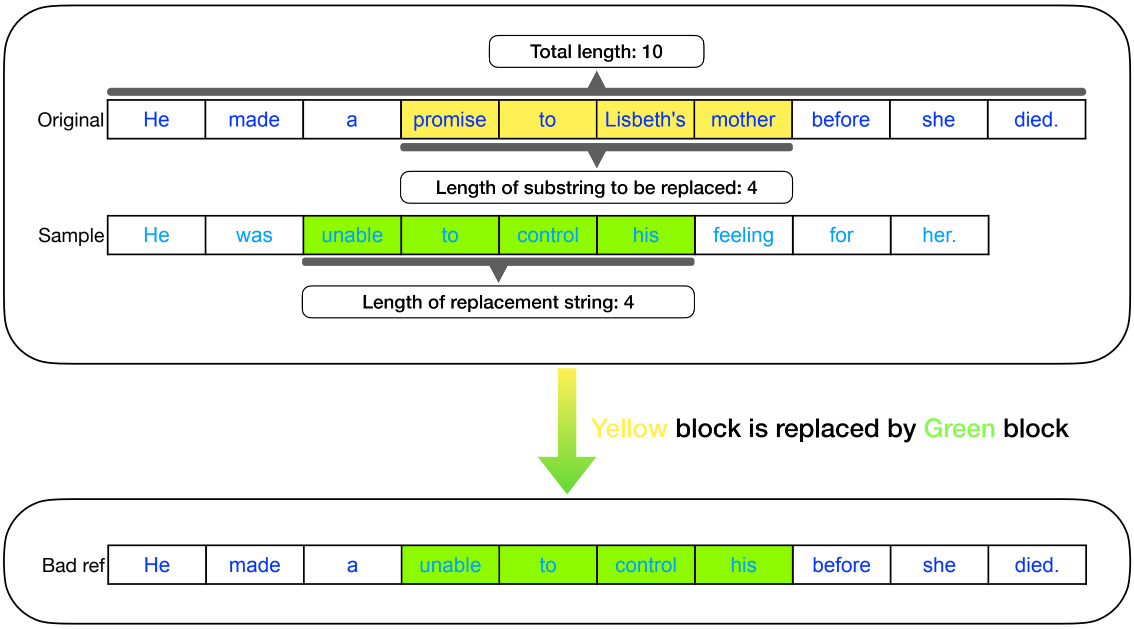


Figure 3.3: The process of generating the degraded version of an answer.

- for $4 \leq n \leq 4$, s comprises 2 words.
- for $6 \leq n \leq 8$, s comprises 3 words.
- for $9 \leq n \leq 15$, s comprises 4 words.
- for $16 \leq n \leq 20$, s comprises 5 words.
- for $n \geq 21$, s comprises $\lfloor n/5 \rfloor$ words.

Figure 3.3 provides the graphical process of degrading an ordinary answer to a "bad reference", where the positions of s (yellow block) and s' (green block) in their corresponding sentences are also randomly selected. Specially, the first and final words will never be included when selecting s and s', unless the original or candidate sentence has only no more than two words. Note that the rule of deciding the number of words is retained as the same as DA was initially developed for MT, but we also adapt the rule of deleting s to replacing s by s'. Compared to the initial rule, the inserted string s' itself is in the form of a fluent text from a human-generated sentence, and human assessors cannot easily identify it without reading the entire answer.

Repeat: A *repeat* answer is generated by simply copying one system-generated answer with its original question and reference, there will be two exact same answers for one single question shown to workers during assessment. Although we expect workers to give similar scores to these two answer, we will not use it as the means to discard results from workers. Instead, we will take the mean score of repeat and original as the final rating for an answer to reduce the bias of judgement.

Reference Answer: When creating a *reference answer*, the system-generated answer is replaced with the reference answer to its according question, meaning that the answer to a question is exactly the same as its reference.

3.3 Dataset and Systems

To test our human evaluation method on free-answering MRC task, we employ NarrativeQA dataset which requires machines to reason over the entire books or movie scripts before generating answers in order to encourage deeper language comprehension (Kočiský et al., 2018). The whole dataset, including training, validation and test, contains approximately 46K questions where 10,557 of them belongs to the test set. Our experiment is thus based on the answers generated by a range of systems with respect to the NarrativeQA testset.

3.3.1 Core Statistics on NarrativeQA Test Set

Our research mainly relies on the results of test set, however Kočiský et al. (2018) only reported the statistical information about the training and validation set when NarrativeQA was proposed. Therefore, we provide analysis the test set as well, including the distributions of word lengths of the ground-truth references, and the frequency of question categories. Figure 3.4 provides a bar chart describing numbers of words in references in the test set and their distribution. We can find that more than half of the references consist of no more than three tokens (sum up to $\approx 52\%$), which implies an unbalanced distribution in the test set. The influence of word

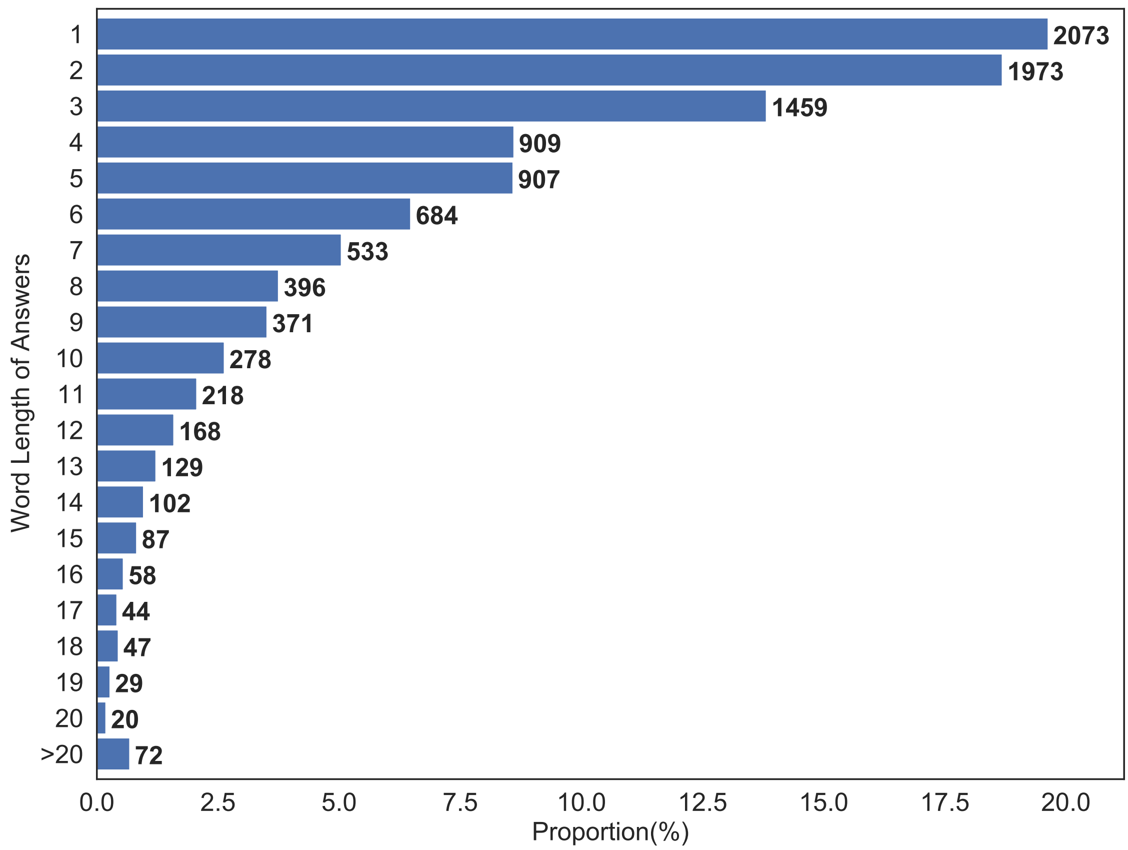


Figure 3.4: The distribution of ground-truth answers with different word lengths, where each bar and the number on its right side indicate the proportion and the real number of that length in the NarrativeQA test set.

lengths of references will be investigated in Section 3.4.6. In addition, we follow the same step with NarrativeQA to reported the distribution of question categories in the test set. Table 3.1 shows the results of question categories, where 300 questions are randomly sampled from the test set and they are manually labelled with their question categories. As an aside, compared with the values on validation set provided by NarrativeQA, the diversity of question categories on test set appears to be more balanced than the validation set.

3.3.2 Systems for Evaluation

Since the NarrativeQA test set provides two separate human-generated answers to each single question, we extract one of those two in order to establish a set of outputs from a system called "Human" which we employ to estimate the performance of

Table 3.1: Question categories on a set of 300 randomly-sampled questions from the test set, compared to the values counted on validation set (Kočiský et al., 2018).

Category	Frequency	Frequency (validation)
Person	26.67%	30.54%
Description	19.33%	24.50%
Entity	13.33%	4.03%
Location	10.33%	9.73%
Why	10.00%	9.40%
How	8.00%	8.05%
Object	5.00%	3.36%
Numeric	3.33%	3.02%
Event	2.00%	4.36%
Duration	1.67%	1.68%
Relation	0.33%	1.34%

humans in the free-answering MRC task according to Mturk workers. Meanwhile, the other human-generated reference answer for each question will act as the ground-truth reference.

Besides the *human* system, we also employ seven MRC systems that are trained on the NarrativeQA train set in order to provide a realistic evaluation of metrics. We firstly include two variants of the Commonsense system (Bauer, Y. Wang, and Bansal, 2018), in which grounded multi-hop relational commonsense information is selected and used to fill in gaps of reasoning between context hops, initially with different hyperparameters. Additionally, we run two baseline systems proposed in (Bauer, Y. Wang, and Bansal, 2018), each of which comprised the same multi-hop pointer-generator model but with different hyperparameters. Further to this, we ran a simple heuristic system for MRC (Sugawara et al., 2018) with a gated-attention reader (Dhingra et al., 2017) and a attention-guided answer distillation system, which transfers knowledge from an ensemble model to a single model by knowledge distillation (Hu et al., 2018). Finally, we include an example system from the recent Neural Network for NLP course (available on `https://www.phontron.com/class/nn4nlp2017/`).

Each system above will automatically generate answers on the test data, meaning that every system produces a set of 10,557 outputs. In total, we now have 84,456

Table 3.2: Core statistics on the HITs and workers in our experiment.

(a) Numbers and pass rates of workers and collected ratings before and after quality control for fluency and two runs of adequacy.

Evaluation modality		Worker Involved			Ratings		
		Total	Passed	Pass Rate	Total	Passed	Pass Rate
Adequacy	Run 1	72	37	51.39%	9,226	6,785	73.54%
	Run 2	77	41	53.25%	6,720	3,520	52.38%
Fluency		103	49	47.57%	9,200	4,560	49.57%

(b) Average durations of HITs to complete in minutes, and the average number of HITs a worker completes.

Evaluation modality		Avg HIT Duration (min)			Avg Assigned HITs		
		Passed	Failed	Overall	Passed	Failed	Overall
Adequacy	Run 1	41.5	30.9	38.7	2.57	1.00	1.82
	Run 2	22.8	20.3	21.6	1.07	1.14	1.11
Fluency		33.7	33.2	33.5	1.16	1.07	1.12

(10557×8) answers from eight various MRC systems for deploying our human evaluation experiment, and they are subsequently placed in a pool of system-generated answers from which the 70 ordinary answers are randomly sampled to create the basis for generating a 100-answer HIT. We employ sampling without replacement in this case.

3.4 Experiment Results

3.4.1 HITs and Workers

Table 3.2 provides the core statistics, with 3.2a showing pass rates of workers and ratings, and 3.2b containing information on the time and number of HITs completed.

Table 3.2a indicates the numbers of individual human assessors who completed HITs on AMT platform, those who passed quality control, as well as the overall pass rate. The numbers of ratings before and after applying quality control are also reported together with the corresponding pass rate. In order to tentatively verify the quality control mechanism, we ran a preliminary adequacy experiment prior to final deployment of this entire experiment. This preliminary experiment consists of HITs of a smaller size, where a HIT can randomly contains 30 to 80 items. The data collected from it is also included when reporting the final results of the adequacy experiment.

Table 3.2b shows the average completion time of passed, failed and all HITs in minutes for Adequacy, including the first and second runs, and Fluency, along with the number of HITs that a worker completed on average, showing that as expected a passed HIT costs a worker more time than a failed one in all three experiments. Meanwhile, passed workers usually take more than two HITs in the first run of our Adequacy experiment despite our suggestion that each worker should complete only a single HIT when experiments are deployed on the AMT platform.

Implement of Quality Control

Since we expect the "bad references" ratings to be lower than their ordinary counterpart answer, a statistical test is applied for comparison between these two paired groups of ratings. As we cannot directly assume that the ratings are normally distributed, the non-parametric Wilcoxon signed-rank test with lower-tailed alternative hypothesis is then applied for the significant test between the ratings of "bad references" and those of their according ordinary answers. Additionally, the significance test will be tested on individuals rather than HITs because our propose is to discard unqualified workers. Therefore, we extract a set, B, "bad references" ratings and a set, O, of corresponding ordinary answer ratings (system produced ratings) from all HITs a worker completed and subsequently apply the significance test to compare B and O to test the reliability of a given worker. The test with $p \geq \alpha$ indicates the failure to reject the null hypothesis H_0 that B and O has no significant difference. Data from workers whose $p \geq \alpha$ is thus filtered out as these workers do not conform to our initial expectation. In this case, we use the conventional threshold of $\alpha = 0.05$ for p-values.

Initially, we likewise plan to employ "reference answers" (gold standard) as our secondary approach, where the workers who have passed "bad reference" quality control with average "reference answer" scores below the threshold will be further filtered out. However, tentative results show that such a method has no effect on rankings of systems since their scores barely change applying it. We accordingly suggest that employing "bad reference" as the stand-alone quality control approach is sufficient enough for this human evaluation, and this is also the approach in MT (Graham, Baldwin, Moffat, et al., 2016; Ondřej Bojar, Chatterjee, Federmann, Graham, Haddow, S. Huang, et al., 2017) and video captioning (Graham, Awad, and Smeaton, 2018) evaluations.

Cost of the Experiment

The payment strategy in our experiment follows the original DA experiment which is proposed for MT evaluation (Graham, Baldwin, Moffat, et al., 2016), where each 100-item HIT costs $0.99. In this MRC human evaluation experiment, a worker who passed the quality control is paid $0.99 per completed HIT, and this entire experiment costs no more than $200. It is notable that the expenditure of experiments for further relevant research can decrease because we will investigate the validity of our method via the results of self-replication experiment in Section 3.4.3. The second run of the adequacy experiment is thus not essential in future applications. Additionally, since the fluency experiment acts as the secondary ranking method, it is also not necessary for research which only focus on evaluating systems' level of adequacy. Assessing one system with about 850 valid ratings costs less than $12 according to our first run. Therefore, the new human evaluation is highly cost-efficient.

We provide the following to failed workers when rejecting their work: *Due to the anonymous nature of crowd-sourcing, it is unfortunately necessary to include quality checks within HITs. The data you submitted did not meet the minimum required quality level for approval. Note that the rejection is automatically carried out by the program script, and you can contact us if you think you work is improperly rejected.* For failed workers who contacted us, they may still get paid after manually

Table 3.3: Human evaluation results for MRC systems in terms of average fluency and adequacy scores, where n is the number of collected ratings, z is the standardized mean scores and raw means averaged scores. Systems includes: Human = human performance estimate, nn4nlp = Neural networks for NLP, Attention-guided AD = attention-guided answer distillation. Rows above horizontal lines indicate the systems significantly outperforming all systems in a lower ranking.

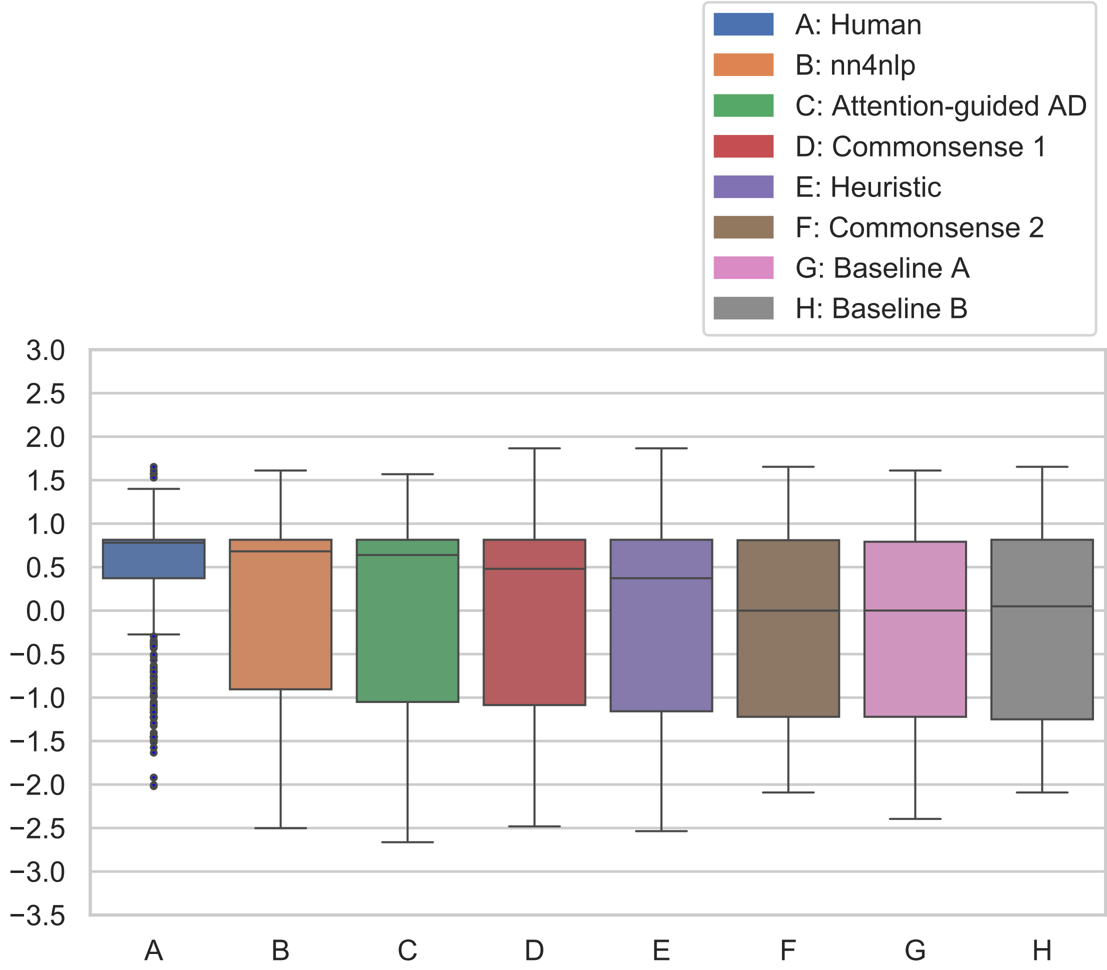
System	Adequacy			Fluency		
	n	z	raw	n	z	raw
Human	831	0.446	81.9	584	0.464	82.5
nn4nlp	842	0.092	66.8	634	0.063	67.6
Attention-guided AD	789	0.055	64.3	557	0.015	66.1
Commonsense 1	849	−0.015	62.1	520	0.057	68.0
Heuristic	892	−0.074	59.6	561	−0.002	65.1
Commonsense 2	851	−0.191	54.0	578	−0.050	63.6
Baseline A	885	−0.194	54.4	538	0.054	68.5
Baseline B	846	−0.199	54.7	588	−0.043	63.0

checking their work as long as no obvious attempt to game the work was found. However, the data collected from such workers will not be used to compute the human score even they got paid after contacting us.

3.4.2 Human Scores

After quality-controlling the workers, we can now compute the human evaluation scores of each system in both Adequacy and Fluency. Since the second run of Adequacy is special for testing reliability of the proposed method, we henceforth use the result of the first run when reporting stand-alone Adequacy.

Table 3.3 shows results of the human evaluation in terms of fluency and adequacy, where n denotes the number of collected ratings. If an ordinary answer has a "repeat" version, the mean value of their ratings will be used as the final score of that answer. Certainly, "bad references" and "reference answers" are excluded for computing the human scores. The raw score of each evaluated system Table 3.3 is calculated by the arithmetic mean of all ratings of its ordinary answers, and the systems in it are ranked by their standardized mean score z in Adequacy experiment where the z score of a system can be computed by the following method:



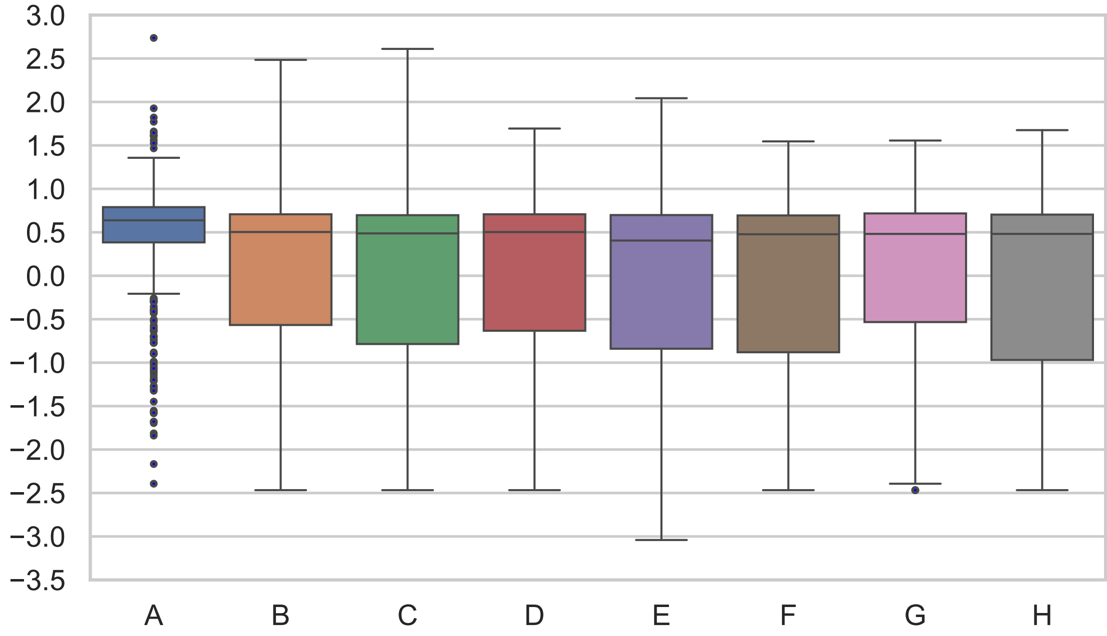
(a) Adequacy

Figure 3.5: The distribution of standardized human scores (z) of MRC systems in the adequacy (first run) and fluency experiments.

1. For a worker w, we compute the mean μ_w and the standard deviation σ_w over all ratings that were completed by w, including the ordinary answers and three various quality-control versions;

2. For an ordinary answer a, its standardized score z_a is computed by $z_a = (r_a - \mu_w)/\sigma_w$ where r_a is the rating of a in the continuous scale and w is the worker who rated the answer a;

3. The final z score of a system can be then computed by $z = \frac{1}{n}\sum_{i=1}^{n} z_{a_i}$, where n is the number of all rated ordinary answers belonging to that system and z_{a_i} is the standardized score of the i-th ordinary answer a_i.



(b) Fluency

Figure 3.5: The distribution of standardized human scores (z) of MRC systems in the adequacy (first run) and fluency experiments.

We also apply Wilcoxon Rank-sum test on system z scores to indicate whether one evaluated system can significantly outperform all systems in a lower-ranking cluster, and such systems are depicted via a horizontal line below their row.

Besides the raw and z scores, we also provide the distribution of z scores of the systems in both Adequacy and Fluency experiments, as shown in Figure 3.5. In general, systems have very similar third quartiles, and the z scores in Adequacy is more concentrated than Fluency. Additionally, although the distributions of system z scores differs from each other, they are generally negatively skewed. For both experiments, system Human has the most concentrated distributed z zscores, meanwhile it has the largest amount of outliers.

3.4.3 System Consistency

To assess the reliability of this newly proposed approach, we carry out a self-replication experiment where we deploy the adequacy experiment for two separate runs consisting of distinct HITs. The HITs in both runs are randomly generated from the same pool of candidate items, and the workers in both runs are automatically allocated by the AMT platform. Both runs share the same instructions and user interfaces, while the second run is conducted after the first one is completed.

As indicated in Table 3.2, the two runs of adequacy have different numbers and pass rates of collected ratings and involved workers, and we can think that run 1&2 as two independent runs. The system rankings from run1&2 show that the second ranking correlates almost perfectly with the first run, with Pearson correlation coefficient $r = 0.986$, revealing that this human evaluation method is capable of providing reproducible results.

The rankings between each pair of systems, however, can be possibly a chance occurrence for such an empirical evaluation, significance tests are therefore essential to tell avoid drawing conclusions simply to chance. We therefore apply a one-sided Wilcoxon rank-sum test on individual standardized ratings between each pair of systems for the two separate runs of adequacy. Figure 3.6 shows the result of the significance test, where the left and right heatmaps respectively indicate the data collected from run 1 and run 2 and the order of systems is the same as Table 3.3. Results draw the exact conclusions from the same distributions of colored cells with only very minor differences in p-values between two runs at $p < 0.1$. We additionally observe that the result with $p < 0.05$ will remain the same conclusion. Hence, we can conclude that the result of this human evaluation is valid and reliable.

3.4.4 Automatic Metrics

Taking the human z scores from our experiment as the gold standard, we can now investigate how accurate the commonly applied automatic metrics are by analysing the level of correlation between metric scores and human z scores in our human evaluation.

(a) Run 1

(b) Run 2

Figure 3.6: Significance test results for systems in two distinct runs where a colored cell means that the system in that row significantly outperformed the system in that column ($p < 0.1$).

System Scores

First, we compute the scores of commonly applied automatic metrics as described in Table 3.4. From left to right, these metrics are ordered by their correlation with human z scores, where the ranking indicates that ROUGE-L performs best while the performance of GLEU is the worst with respect to human opinions. For a more straightforward comparison, we represent the rankings of systems by our human evaluation scores and automatic metrics in Figure 3.7. It can be observed that, Human remains the top one, and the rankings of all other systems can fluctuate with applied metrics when the applied metric changes. Some systems are relatively stable, for example, the ranking of Attention-guided AD will keep the third for most metrics, and it only drops from the third to the forth when ROUGE-l is applied. However, when it comes to other systems, their rankings can significantly fluctuate according to the changes of metrics. For example, when the current evaluation method changes from the human evaluation to BLEU-1, the ranking of nn4nlp system will drop from the second to the seventh, while Baseline B can significantly

Table 3.4: The system scores of different automatic metrics along with the z scores,where the automatic metrics are ordered by their values of the correlation with z scores from left to right.

System	z	ROUGE-L	METEOR	BLEU-4	BLEU-1	GLEU
Human	0.446	57.29	25.92	17.69	45.80	22.58
nn4nlp	0.092	45.90	17.29	5.13	17.08	7.05
Attention-guided AD	0.055	41.13	18.90	13.37	31.28	15.96
Commonsense 1	−0.015	45.75	19.45	13.79	36.57	17.19
Heuristic	−0.074	34.56	12.23	1.75	6.76	2.67
Commonsense 2	−0.191	37.04	14.38	8.01	24.83	12.75
Baseline A	−0.194	37.04	14.91	8.93	26.22	12.95
Baseline B	−0.199	36.23	14.58	8.06	27.32	12.46

increase from the last one to the fourth.

Correlation Coefficients and Williams Test

To explore these metrics' accuracy of measurement in terms of system performance, we compute three correlation coefficients by comparing metrics scores for our systems with our human evaluation results. Table 3.5 shows the Pearson, Spearman and Kendall's Tau correlations between the system z scores and according metric scores. It shows that ROUGE-L correlates best with the human evaluation at $p = 0.929$ while GLEU performs poorly only reaching $p = 0.514$. Since all these metrics are not initially designed for MRC, it is interesting that the performances of these metrics can have such difference. Additionally, although GLEU performs worse than BLEU-4&1 in terms of Pearson correlation, it correlates higher to human than BLEU-4&1 according to the Spearman and Kendall's Tau correlations. Since the correlations are computed between scores of evaluation methods in the same data set, we cannot simply assume these correlations are independent. The degree to the correlation with each pair of automatic metrics should also be calculated, and Williams test is thus utilized for the assessment of difference in such correlations (Williams, 1959). The values marked with superscript ** in Table 3.5 depicts a metric that can outperform metrics in the lower cluster at $p < 0.01$ according to Williams test of differences in dependent correlations Graham and Baldwin, 2014. We also compute Williams test between ROUGE-L and METEOR, and the result

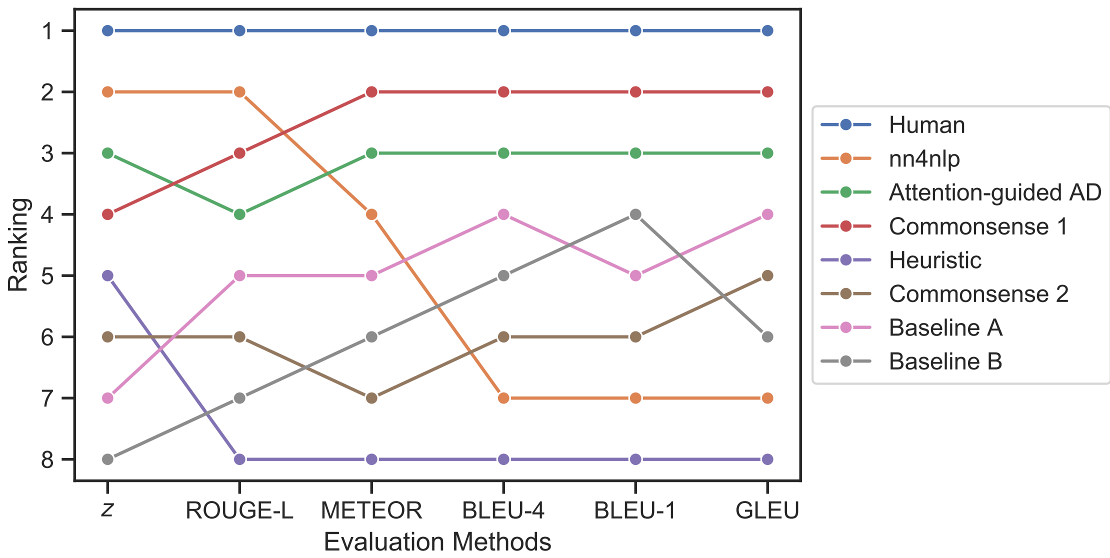


Figure 3.7: The ranking of MRC systems according to different evaluation methods on the collected data from the first run of the adequacy experiment, where these methods are ordered by their correlation with humans according to the results of our experiment.

Table 3.5: Correlation of commonly applied automatic metrics with human evaluation of the adequacy of answers; r = Pearson correlation; ρ denotes Spearman correlation; τ denotes Kendall's Tau correlation; metrics with Pearson correlation that significantly outperforms BLEU-4&1 and GLEU at $p < 0.01$ according to Williams test denoted by **

	ROUGE-L	METEOR	BLEU-4	BLEU-1	GLEU
r	0.929**	0.896**	0.599	0.534	0.514
ρ	0.810	0.690	0.333	0.310	0.381
τ	0.643	0.429	0.214	0.143	0.286

$p = 0.2$ indicates that we cannot conclude ROUGE-L correlates significantly better than METEOR with human, although ROUGE-L has a higher Pearson correlation coefficient than METEOR.

The aforementioned analysis provides insight for answering the **RQ 1** which we proposed in Chapter 1. In terms of this task, **RQ 1** can be described as follows:

- **RQ 1**: *How accurately do existing automatic metrics measure the performance of free-answering MRC systems?*

Subsequently, **RQ 1** can be answered as: according to the results of our human evaluation experiments, prevailing automatic metrics generally have different degrees

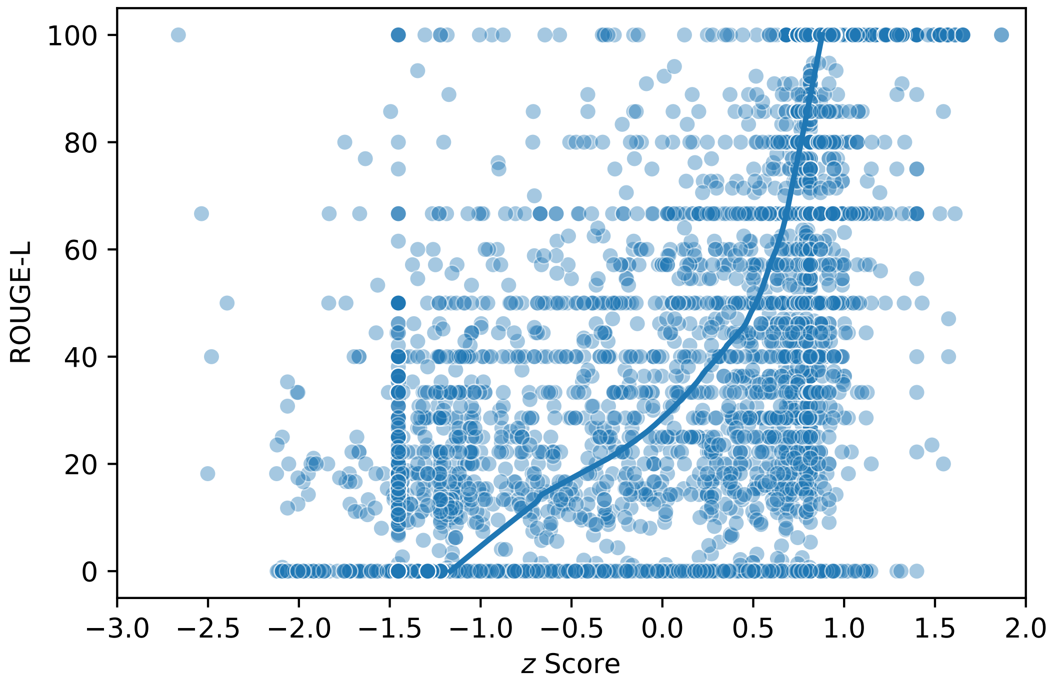


Figure 3.8: The relations between ROUGE-L scores and z scores at the sentence-level.

of accuracy when evaluating the performance of free-answering MRC systems, since ROUGE-L and METEOR can correlate well with human judgement, where other metrics fail to achieve high correlation.

Sentence Scores

System scores for automatic metrics are reported by calculating the mean of their sentence scores using different approaches to averaging depending on the metric. For example, ROUGE-L, METEOR and BLEU score use the arithmetic mean, harmonic mean and geometric mean, respectively. Our previous analysis shows that some metrics like ROUGE-L and METEOR can have a considerably high correlation with human evaluation according to the system score, and we want to further investigate the relation between the sentence metric scores and sentences z scores.

Figure 3.8 and Figure 3.9 show two scatter plots with the trend lines that describe the joint distribution of the z scores and the corresponding metric score at the sentence-level, where each point represents the z score (y-axis) and the metric

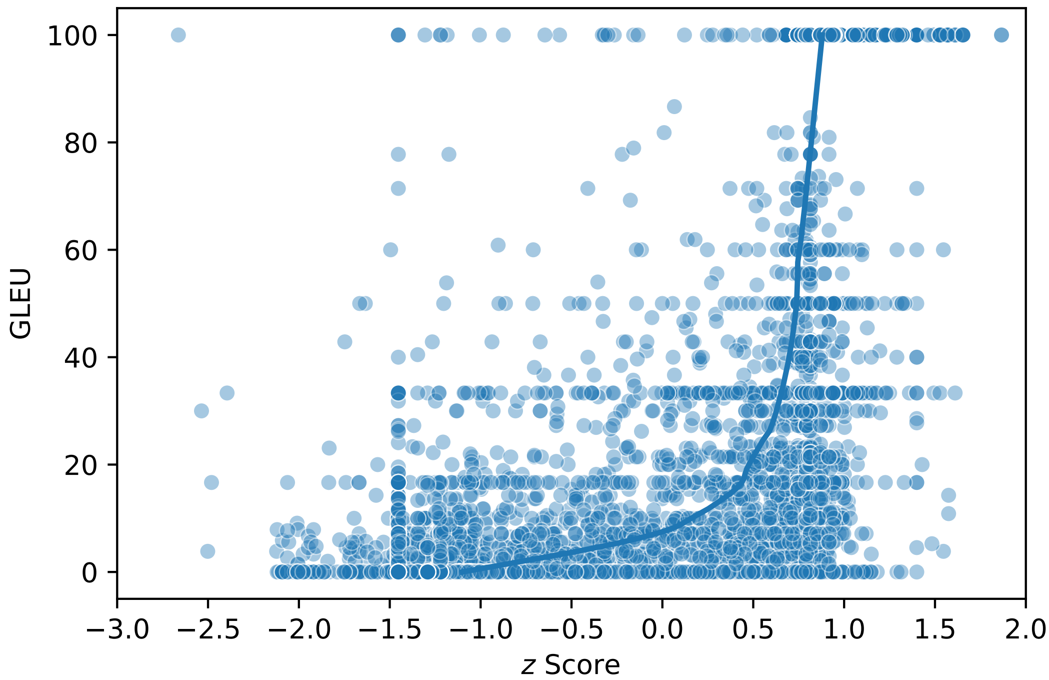


Figure 3.9: The relations between GLEU scores and z scores at the sentence-level.

score (y-axis) of a sentence. Trend lines are drawn according to locally weighted scatterplot smoothing (LOWESS) method to provide a clear depiction of the relationship between metric scores and z scores (Cleveland, 1979). Figure 3.8 is based on ROUGE-L which correlates best with human evaluation, while Figure 3.9 is the GLEU metric which has the lowest correlation with human evaluation.

We conjecture that the reason ROUGE-L correlates better than GLEU with human opinion is that the trend line of ROUGE-L shows a closer relation to a *linear* relation, indicating that ROUGE-L can better fit the human ratings compared with GLEU. In addition, the overall distribution of both ROUGE-L and GLEU is not concentrated linearly and there are still a large number of sentences scores diverging from trend lines. However, ROUGE-L scores are relatively centrally distributed while GLEU scores are mainly on the right bottom. Although the sentence scores of ROUGE-L can contain a high level of random error, its relatively balanced joint distribution likely helps to counteract the effects of positive errors (lower z scores with higher metric scores) and negative errors (higher z scores with lower metric scores) when the system scores are computed by the averaging of sentence scores.

Question:	To be object for subject is the same as what?
Reference:	to be our representation or mental picture is the same thing
Answer:	to be a mental picture or our representation
Scores	raw: 100 (z: 0.81) ROUGE-L: 42.1; METEOR: 59.3

(a) An example where automatic metrics underestimate the generated answer according to human judgement.

Question:	How is Wesley killed?
Reference:	he gets shot by sheriff ballard when distracted while watching an episode of a reason to love
Answer:	wesley is distracted by watching an ankle holster and commits a reason to love
Scores	raw: 0 (z: -1.21) ROUGE-L: 45.2; METEOR: 44.6

(b) An example where automatic metrics overestimate the generated answer according to human judgement.

Figure 3.10: Examples of mismatched human scores and automatic metric scores from the first run of our adequacy human evaluation experiment at the sentence-level.

This is a potential cause that ROUGE-L can correlate highly with human judgement at the system-level.

Figure 3.10 provides two examples where human judgements disagree with automatic metrics when evaluating the performance of MRC systems at the sentence-level. Figure 3.10a shows an example where human workers think the system output is perfect (raw score: 100) while ROUGE-L only rate it as 42.1. We find that human workers can successfully figure out that the system output shares the same meaning with the reference. However, metrics which rely on the word overlap fail to give a high score. Figure 3.10b is an example where automatic metrics overestimate the system output (ROUGE-L: 45.2) according to human judgements (raw score: 0). The question asks how *Wesley is killed*, and the correct answer is expected to include information such as *get shot*. We find that the system-generated answer contains no such information, therefore human workers only rate it as low as 0. However, the answer has a high rate of word overlap with the reference, which misleads the automatic metrics to give a high score.

We can thus answer **RQ 2** regarding this task, which is described as follows:

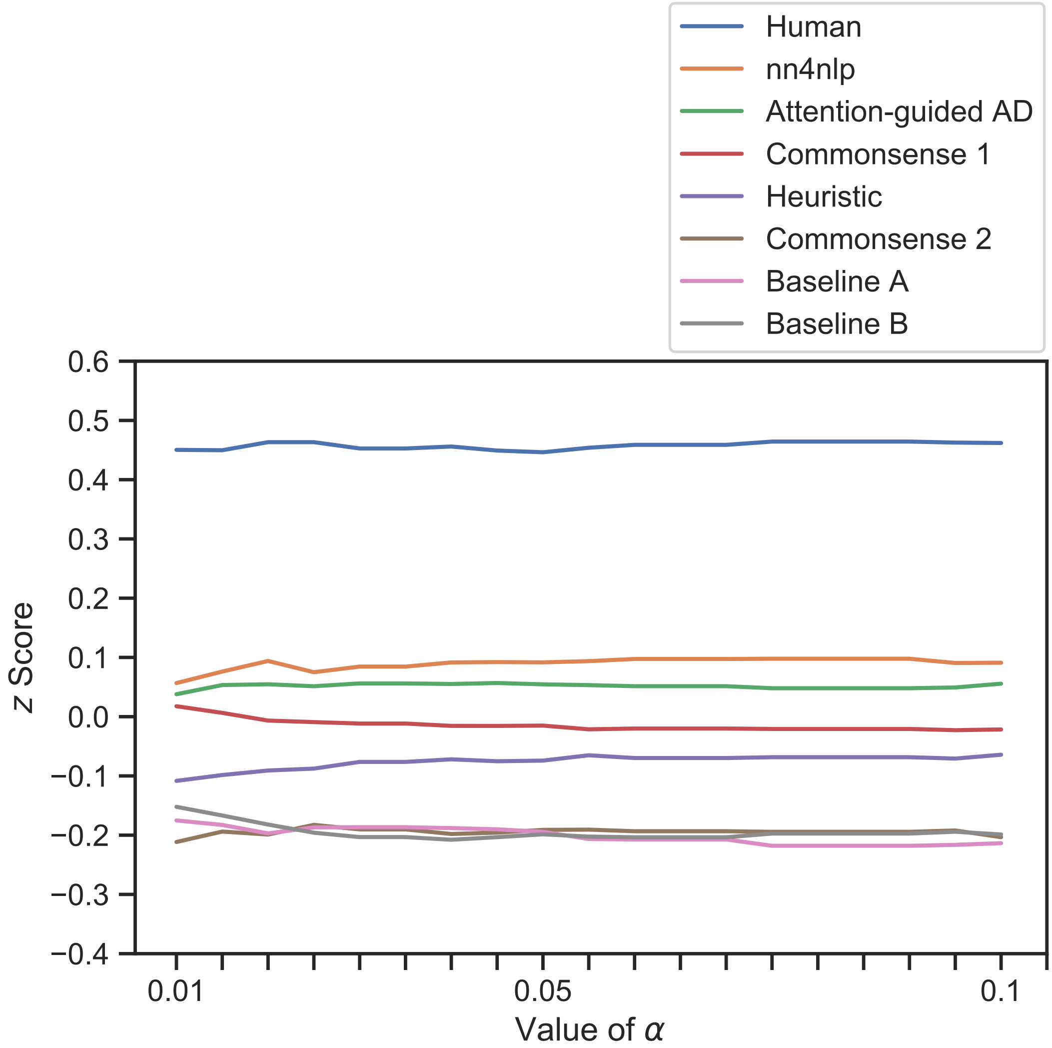
- **RQ 2** *What are the limitations and disadvantages of the direct application of evaluation metrics from MT and other domains to entirely distinct tasks for system development in free-answering MRC?*

Firstly, according to Figure 3.7, applying different metrics may inconsistently rank the systems, resulting in a confusion of system ranking when applied metric changes. In addition, although automatic metrics, such as ROUGE-L, can have a high correlation with human judgement at the system-level, the scores produced by these metrics are not always capable of reflecting the real quality when evaluating a single sentence. A potential disadvantage of the application of these metrics in free-answering MRC is that, a system with a high metric score at the system-level may perform weakly in sentence-level evaluation. In other words, metrics, such as ROUGE-L and METEOR in this case, are acceptable to employ as the approach of ranking systems, but it is meanwhile non-negligible that, such metrics lack the ability of accurately assessing individual system outputs.

3.4.5 Influence of α on Quality Control

As we described in Section 3.4.1, the ratings from workers whose p-value of Wilcoxon signed-rank test equals or exceeds the threshold α are filtered out, where we empirically choose $\alpha = 0.05$ as it is a frequently-used value. Nevertheless, 0.01 and 0.1 are also typical values, so we apply various values of α to check how it can alter system z scores.

Figure 3.11 represents z scores of systems when incremental values of α are applied in the range of 0.01 and 0.1 for the two runs of Adequacy. Figure 3.11a show that z scores only slightly fluctuate with the changes of α, while the ranking of systems remains unchanged except the last three systems as it has been described in Figure 3.6 that these three systems show no significant difference. We can draw a similar conclusion from Figure 3.11b, but the ranking of Attention-guided AD

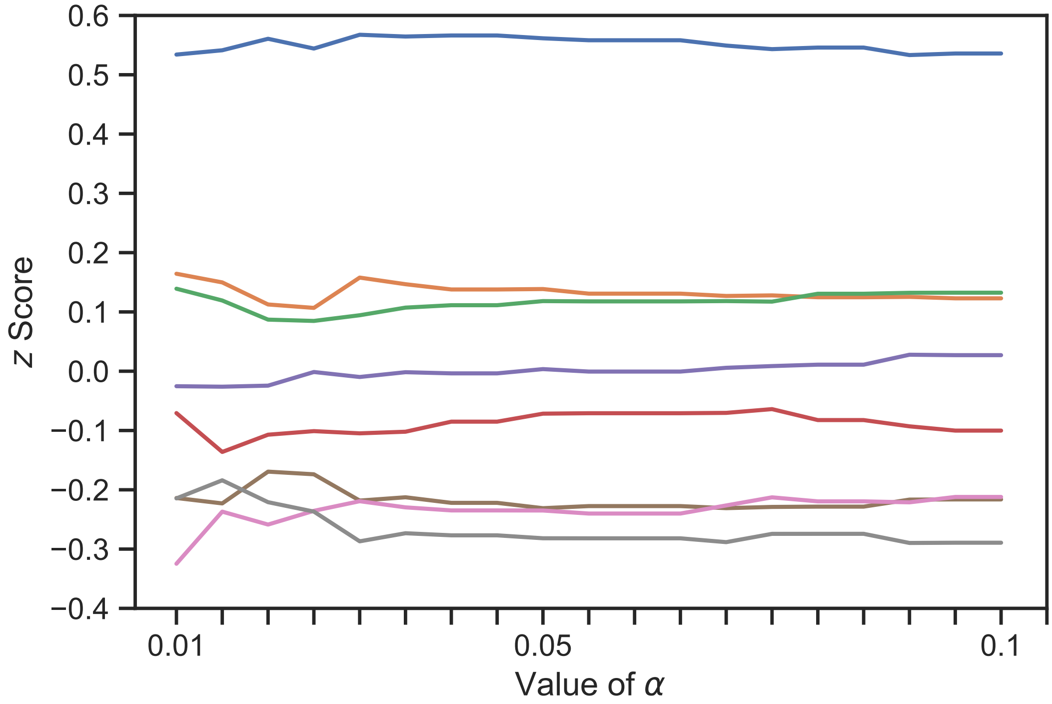


(a) Run 1

Figure 3.11: The changes of system z scores when the threshold α applied for the quality-control method ranges from 0.01 to 0.1.

and nn4nlp will change when α reaches 0.075 as no significant difference is found between them.

Since the scores barely change with α, we can conclude that the quality control mechanism in our human evaluation method is robust to the values of threshold α as long as it is within a reasonable range. Compared with the other two typical values, the correlation between run1&2 on $\alpha = 0.05$ can reach $r = 0.986$, while r will slightly drop to 0.950 and 0.976 if $\alpha = 0.01$ and $\alpha = 0.1$ are applied, respectively. Hence, we think the empirical employment of $\alpha = 0.05$ for quality controlling the workers in this experiment is valid and acceptable.



(b) Run 2

Figure 3.11: The changes of system z scores when the threshold α applied for the quality-control method ranges from 0.01 to 0.1.

3.4.6 Word Lengths of References

As it has been observed in Section 3.3 that the references in the data set have an imbalanced distribution of word lengths, we want to investigate whether the number of words in references can have an influence on the final system z scores. In detail, the following mechanism is utilized: the ratings to system-generated answer whose corresponding references have *more than* n words will be filtered out and a new z score is then computed on the remaining ratings. The applied mechanism can ensure that the sample size is still sufficient. With a range of n ($1 \leq n \leq 20$ in this case), we are able to examine the impact of word lengths on our human evaluation.

Figure 3.12 compares z scores of the eight systems when answering questions with references having no more than n words. Results show that, in general, most systems performs best on $n = 2$, while perhaps unsurprisingly the z score of systems will encounter a decline with the increase of n. In general, it seems that, the increase of word lengths of references will cause decreasing system z scores. We think this

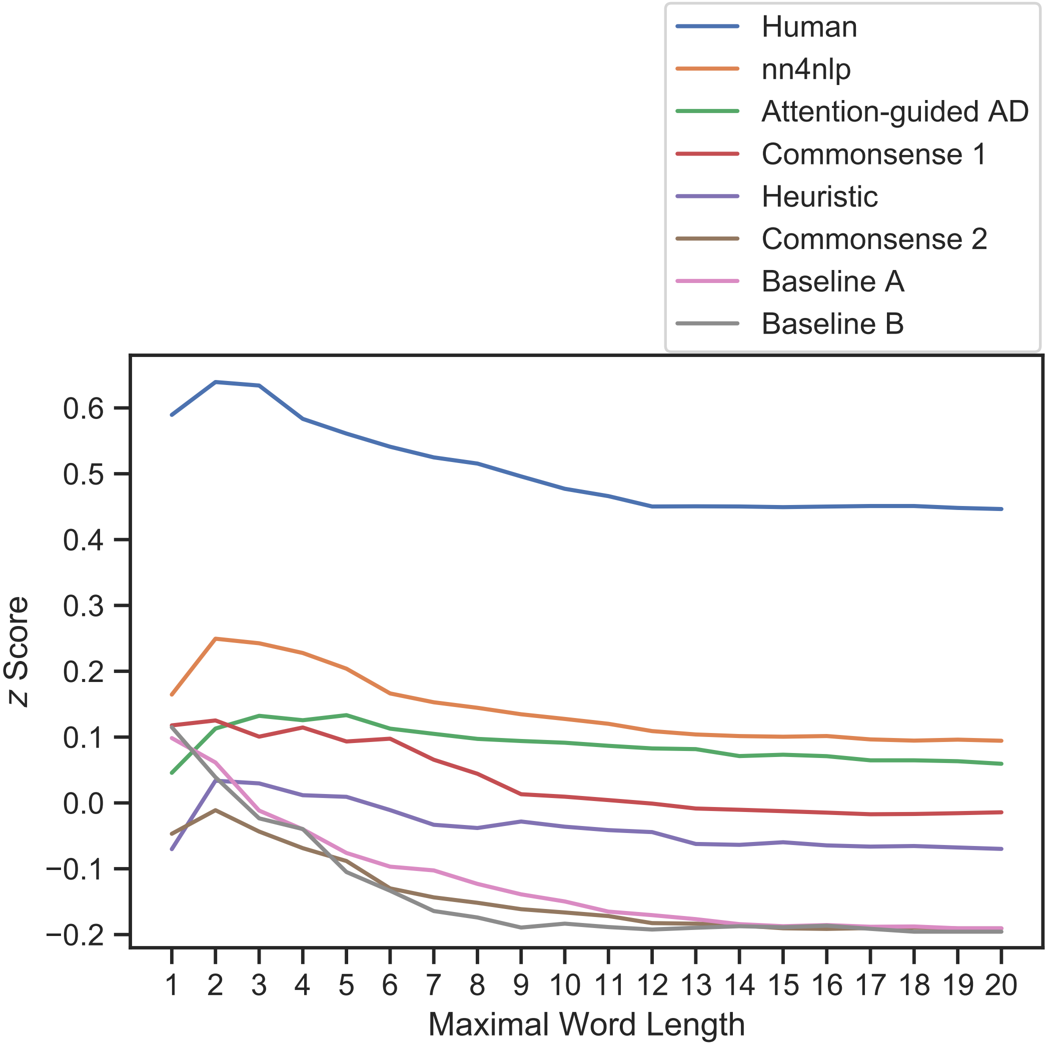


Figure 3.12: The system z scores change with the maximal number (n) of words that ground-truth references have.

is likely to relate to the search space of the decoder, which are employed in neural generative MRC systems in our experiment for the generation of answers in the form of natural language. Decoding is nromally identified as a search problem (Zarrieß, Voigt, and Schüz, 2021), where the length of references indicate the search space for the decoder to generate the answer. Generally speaking, the more words an answer has, the larger the search space, resulting in the model being more prone to producing an answer containing errors.

3.4.7 Human Assessor Consistency

A human evaluation method is commonly deemed to be empirical and such a method generally requires to report its rater agreement as a measure of the reliability of the ranking (Ondrej Bojar, Buck, et al., 2014). We therefore examine the consistency of individual assessors for the newly proposed human evaluation method. In doing this, we expect assessor consistency from passed workers to surpass that from failed workers to demonstrate that our strict quality control mechanism has a positive impact on our evaluation method. We will calculate the individual assessor consistency in two aspects: intra-annotator agreement and inter-annotator agreement, where the former is reported as Cohen's kappa coefficient (Cohen, 1960) and the later is reported as correlation coefficients.

Intra-annotator Agreement

Cohen's kappa coefficient (κ) is the conventional means of measuring the agreement among annotators in previous WMT shared tasks (Callison-Burch, Koehn, Monz, Post, et al., 2012; Ondřej Bojar, Chatterjee, Federmann, Graham, Haddow, Huck, et al., 2016). Since κ is only appropriate for the interval-level scale, we first need to convert a raw score into several categories (called n-cat) by the formula $cat(r, n) = min(\lfloor \frac{r \times n}{100} + 1 \rfloor, n)$, where r is the original value of a raw score which ranges from 0-100 and n is the target number of categories. Cohen's kappa κ is initially proposed for the intra-annotator agreement of one rater, the direct application of κ is however unsuitable for this experiment since a large number of assessors are involved and each assessor rates different items. Instead, we compute the degree of agreement between the repeat answers and the paired ordinaries regardless of their actual raters, and the computed κ can act as the intra-annotator agreement.

For the computation of κ, we first need to generate a confusion matrix using the converted n-cat scores from repeats and ordinaries, as shown in Table 3.6. Each cell a_{ij} in the table indicates the number of pairs of the repeat answer and its ordinary answer whose n-cat scores are respectively i and j. In addition, s_k and

Table 3.6: The n-cat confusion matrix for paired repeats and ordinaries.

		Repeat (i)				
		1	2	...	n	
Ordinary (j)	1	a_{11}	a_{21}	...	a_{n1}	s_1
	2	a_{12}	a_{22}	...	a_{n2}	s_2
	...	...	...	...	...	...
	n	a_{1n}	a_{2n}	...	a_{nn}	s_n
		t_1	t_2	...	t_n	

Table 3.7: Cohen's kappa which indicates the agreement of passed and failed raters on n categories ($n \in \{2, 4, 5, 10\}$).

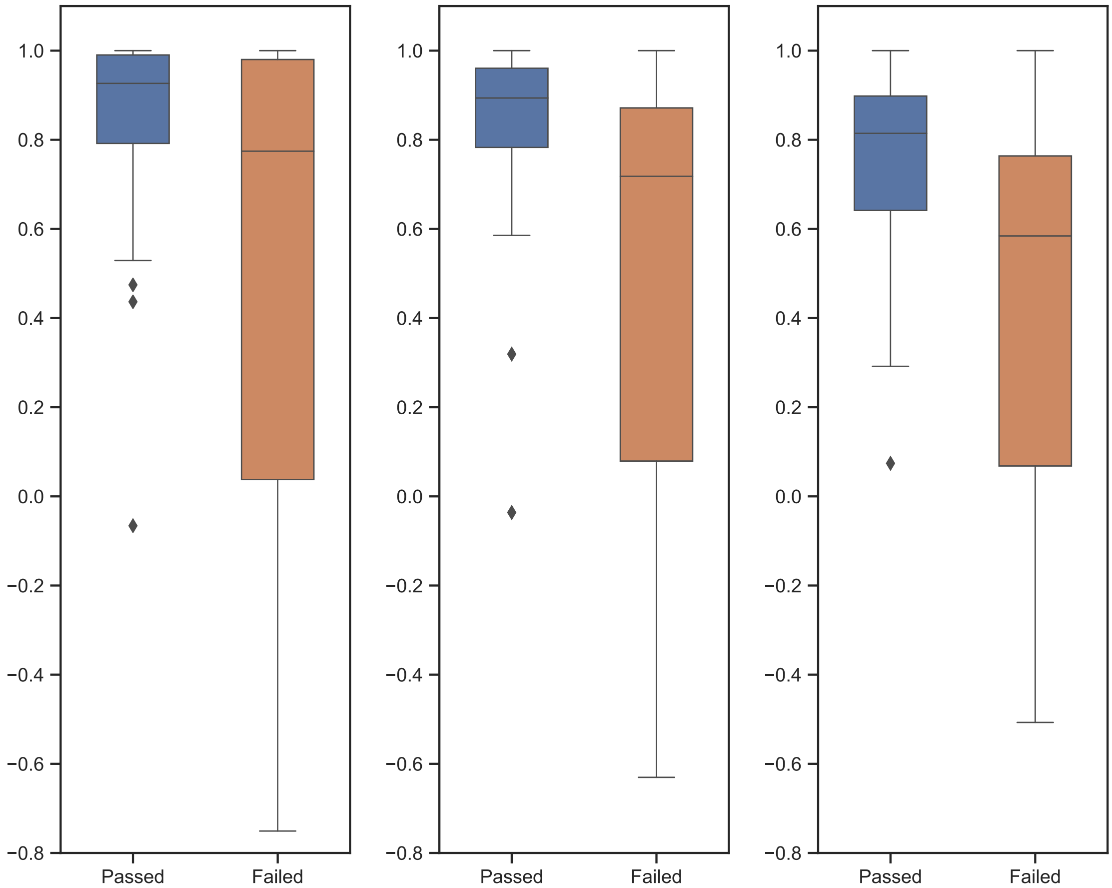
n	Passed			Failed		
category	p_o	p_e	κ	p_o	p_e	κ
10	0.686	0.278	0.565	0.559	0.188	0.456
5	0.775	0.349	0.654	0.665	0.275	0.539
4	0.832	0.393	0.723	0.673	0.316	0.521
2	0.897	0.520	0.786	0.822	0.545	0.609

t_k represent the sum of the according row and column, where $s_k = \sum_{i=1}^{n} a_{ik}$ and $t_k = \sum_{j=1}^{n} a_{kj}$ ($k \in 1, 2, \ldots, n$). According to the result of the confusion matrix, we can use Formula 3.1 to compute the value of κ as follows:

$$\begin{aligned} \kappa &= \frac{p_o - p_e}{1 - p_e} \\ p_o &= \frac{1}{N} \sum_{k=1}^{n} a_{kk} \\ p_e &= \frac{1}{N^2} \sum_{k=1}^{n} \left(s_k \times t_k \right) \end{aligned} \tag{3.1}$$

where N is the sum of all cells from Table 3.6 ($N = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$), p_o is the probability of observed agreement and p_e is the probability of random agreement.

Table 3.7 shows results for p_o, p_e and κ in terms of rater agreement of workers who passed or failed quality control, where n is the number of categories, p_o denotes the proportion of ratings that the repeats agree with its ordinaries, p_e denotes the probability of scoring the repeat and ordinary answers in a random manner, and κ is the intra-annotator agreement. We observe that κ of passed raters can get a rather



(a) Pearson (b) Spearman (c) Kendall

Figure 3.13: The distribution of the Pearson, Spearman and Kendall's τ correlations between the ratings of repeat answers and those of paired ordinaries on individual raters who passed (blue) and failed (orange) the quality control method.

high increase (at least 0.109) compared to failed workers for every kind of n-cat, especially a significant promotion (0.202) of κ is available on 4-cat. In accordance with the typical interpretation (Landis and Koch, 1977), we find that, for both 4-cat and 5-cat, the level of intra-annotator agreement increases from moderate (0.4-0.6) to substantial (0.6-0.8) when raters changes from the failed to the passed.

As a result, we believe the applied quality control method is capable of improving the reliability according to the intra-annotator agreement.

Inter-annotator Agreement

Although we can use Cohen's kappa κ to represent the intra-annotator agreement, it is inappropriate for assessing the degree of agreement between raters as we mentioned previously. Besides, κ has other considerable limitations. First, the continuous ratings have to be converted into the form of interval scales before the computation. Meanwhile, the values of converted n-cat ratings are treated independently and their ordinal nature is somehow neglected. For example, 4 should be expected to correlate higher with 5 than 1. Hence, we decide to use correlation coefficients rather than κ. However, we cannot simply use one single numerical value like κ to estimate the overall level of agreement when using correlation coefficients, because they are not additive (Alexander, 1990). In other words, it is inconsequential to represent the rater agreement by computing the average over a set of correlations. Instead, we use the distribution of correlations to represent the trend and compare between the results from both passed and failed workers.

As shown in Figure 3.13, we calculate three types of correlations between repeats and ordinaries for each worker, including Pearson, Spearman and Kendall's τ, where the blue and orange boxplots represent the results of workers who passed and failed the quality control, respectively. We observe that the correlations of passed workers are significantly higher than failed workers with a more concentrated distribution for all types of correlations. For the Pearson (Fig 3.13a) and Spearman correlation (Fig 3.13b), the correlations of passed workers are usually between 0.8 and 1.0 and those of passed workers can range from 0.0 to 1.0, showing that passed workers are highly self-consistent while failed workers are more likely to be randomly guessing. For the Kendall's τ correlation (Fig 3.13c), passed workers mildly drops to the range of 0.6-0.9 while failed workers become slightly more concentrated. No matter what type of correlation is applied however, we can draw the same conclusion: the inter-annotator agreement of workers who passed the quality control is incontrovertibly higher than workers who failed.

3.5 Summary

In this chapter, we present a new human evaluation method adapted from DA to evaluate free-answering MRC task. A MRC can be evaluated on two aspects: adequacy and fluency, and the method is robust as two distinct runs of adequacy experiments shows a extremely high correlation.

We introduced the detailed approaches for quality controlling workers, the dataset and systems participating in this search. The structure of a HIT and the sample size are also introduced to provide a clear means for other researchers to reproduce or modify our approach.

After the experiments are completed on AMT platform, we then compute the number and pass rate of individuals and ratings and report the raw and z scores of different systems on our evaluation method. In addition, we compared the results of commonly applied automatic metrics with z scores in the level of systems and sentences. The correlation between system scores shows that ROUGE-L and METEOR have a higher degree of accuracy than BLEU and GLEU in the matter of ranking systems. Meanwhile the relations between sentence metric scores and z scores indicate that these automatic metrics lack the ability of authentically evaluating a single sentence although a metric may correlate highly with human judgements when ranking systems. Besides, we provide statistical analyses on the collected data in various aspects, including the influence of α on the quality control and the word lengths of references to investigate the rationality of our experiment.

In conclusion, we propose a crowd-sourcing human evaluation method which is cost-effective and highly efficient, and show its reliability and validity. In consideration of the **RQ 3** in terms of MRC which is decribed as follows:

- **RQ 3**: *Can more appropriate new methods of evaluation be designed that are feasible given the limited time and resources available in operational settings?*

it can likewise be answered that a more appropriate new human evaluation method is available for the free-answering MRC task.

Chapter 4

Evaluations on Question Generation

In this chapter, we will focus on solving the existing issues involving in the question generation task, since its evaluation, including automatic metrics and human evaluation, suffer from several known issues yet. As described in Chapter 1 and Chapter 2, applied automatic metrics generally fail to take into account the *one-to-many* nature of QG. Meanwhile, the human evaluation of QG still lack a standard approach. In this chaper therefore, we propose a new crowd-sourcing human evaluation method for QG, and deploy the corresponding experiments. We additionally report the results of our experiments and provide related analysis. Furthermore, we propose an unsupervised reference-free evaluation metric that can automatically evaluate a QG system using a pretrained language model. According to our proposed human evaluation method and automatic metric, we subsequently answer **RQ**s which we proposed in Section 1.1, in terms of the QG task.

In detail, the essential methodology of the crowd-sourcing human evaluation method we designed for the QG task will be introduced In Section 4.1. The experiment for the verification of the method is also introduced with its detailed settings, including the design of interface shown to human raters, mechanisms for guaranteeing the quality of evaluation. Section 4.2 introduces the constitution of the assignment for each human rater in our human evaluation experiment, as well as the detailed implement of the quality control technique. Section 4.3 presents the HotpotQA dataset which is employed in this research, as well as the 11 involved QG systems. Section 4.1 is generally about the experimental results and analysis, such as statistic data of the experiments, human evaluation scores based on collected data, and corresponding significance test. In addition, Section 4.5 introduces the details of a newly proposed automatic evaluation metric that can evaluate system-generated outputs without any reference. Its performance is reported as well, by computing the correlation between its scores at the system-level and the results of our human evaluation experiment. We further answer **RQ**s regarding the QG task in Section 4.1 and 4.5. Finally, Section 4.6 concludes this chapter.

4.1 Experiment Design

In this section, the essential methodology of the crowd-sourcing human evaluation method we proposed for the QG task will be introduced. The experiment for the verification of the method is also introduced with its detailed settings, including the design of interface shown to human raters, mechanisms for guaranteeing the quality of evaluation, and the evaluation criteria.

4.1.1 Methodology

QG receives a context with a sentence as the input and generates a textual sequence as the output, with automatic metrics reporting the computation of word/n-gram overlap between the generated sequence and the reference question. However, human evaluation can vary. When evaluating MRC systems via crowd-sourced human evaluation, raters are asked to judge system-generated answers with reference to gold standard answers because a correct answer to the given question should be, to some degree, similar to the reference answer (Ji, Graham, and Jones, 2020).

Whereas, simply applying the same evaluation is not ideal since evaluating a QG

1. Each time you will see a question together with a passage, and your task is to rate the questions after reading the given passage.
2. Each HIT contains one certain passage with 20 various questions to rate.
3. The highlighted content in the passage is expected to be the answer to the presented question.
   A passage with no highlighted content means the question should be a "yes-or-no" question.
4. Chrome is preferred, other browsers may cause some errors.
5. There is a feedback box at the end of the HIT. If you encounter any problems, please enter them in this box or email our MTurk account.

Figure 4.1: The full instruction shown to a human worker that the worker should read and then click the "I understand" button before starting the current HIT.

system is more challenging due to its *one-to-many nature* as described in Section 1, namely a QG system is capable of producing a question that is appropriate but distinct from the reference. Such evaluation may unfairly underestimate the quality of a generated question because of its inconformity with the reference. To avoid this situation in our experiment, we ask a human rater to directly judge the quality of a system-generated question only according to the passage and answer present, instead of providing a reference question.

Experiment Interface

Since our crowd-sourced evaluation method can involve workers who have no specific knowledge of the related field, a minimal level of guidance is necessary to concisely introduce the evaluation task. Prior to each HIT, a list of instructions followed by button labelled *I understand* is provided, with the human rater beginning a HIT by clicking the button. The full list of instructions is described in Figure 4.1. In terms of the fourth instruction, we present the HTML element "range control" embedded with hash marks, but not all browsers can fully support this feature. For example, this feature is completely unsupported by Firefox. Hence, Chrome is recommended for ensuring the stability of our experiment.

Within each HIT, a human assessor is required to read a passage and a system-generated question with the input (correct) answer, then rate the quality of the

Passage: The Battle of Saint-Mihiel was a major World War I battle fought from 12–15 September 1918 , involving the American Expeditionary Force (AEF) and 110,000 French troops under the command of General John J. Pershing of the United States against German positions . The U.S. Army Air Service (which later became the U.S. Air Force) played a significant role in this action . General of the Armies John Joseph "Black Jack" Pershing (September 13 , 1860 – July 15 , 1948) was a senior United States Army officer . His most famous post was when he served as the **commander of the American Expeditionary Force (AEF) on the Western Front** in World War I , 1917–18.

Question: What was the most famous post of the man who commanded American and French troops against German positions during the Battle of Saint-Mihiel ?

Figure 4.2: The interface shown to human workers, including a passage with highlighted contents and a system-generated question. The worker is then asked to rate the question.

question according to the given passage and the answer. Since the answer is a subsequence of the passage, we directly emphasize the answer within the passage. Figure 4.2 provides an example of the interface employed in experiments, where a worker is shown a passage whose highlighted contents are expected to be the answer to the generated question. Meanwhile, workers may see a passage without any highlighted content since a fraction of the answers are simply "yes-or-no".

Evaluation Criteria

Human raters assess the system output question in regards to a range of different aspects, as opposed to directly providing a single overall score. Figure 4.3 provides an example rating criterion, where a human rater is shown a Likert statement and asked to indicate the level of agreement with it through a range slider from *strongly disagree* (left) to *strongly agree* (right).

The full list of evaluation criteria we employed in this experiment is available in Table 4.1, where the labels are in reality not shown to the workers during the evaluation. As an empirical evaluation method, these criteria are those most commonly employed in current research (see Section 2.2.1) but can be substituted for distinct

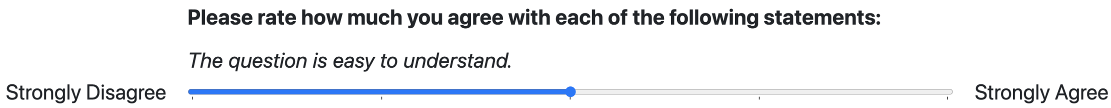


Figure 4.3: The example of a Likert statement of an evaluation criterion shown to a human worker.

Table 4.1: The rating criteria of assessing the quality of a system-generated question, where only the Likert statements are available for human workers and the labels are not shown in the experiment.

Label	Likert statement
Understandability	*The question is easy to understand.*
Relevancy	*The question is highly relevant to the content of the passage.*
Answerability	*The question can be fully answered by the passage*
Appropriateness	*The question word (where, when, how, etc.) is fully appropriate.*

criteria if necessary. Since our contribution focuses on proposing a human evaluation approach that can act as a standard or a framework for judging QG systems, rather than proposing a fixed combination of evaluation criteria, the criteria we employed are neither immutable nor hard-coded. And we encourage adjusting, extending and pruning them if necessary. Additionally, the rating criterion "answerability" in Table 4.1 should not be confused with the automatic metric Answerability, while the former in our experiment will be called Q-BLEU (see Section 2.2.1).

4.2 Quality Control

Similar to human evaluation experiments in other tasks, such as MT (Graham, Baldwin, Moffat, et al., 2016) and MRC (Chapter 3), quality-controlling the crowd-sourced workers is likewise necessary for the QG evaluation. Since no ground-truth reference will be provided for the comparison with system-generated questions, the quality control methods involve no "reference question". Two methods - *bad reference* and *repeat* - are employed the means of quality-controlling the crowd to filter out incompetent results.

The methods of quality controlling the crowd are:

- **Bad Reference**: a set of system-generated questions are randomly selected, and their degraded versions are automatically generated to make a set of "bad references". The degradation mechanism is broadly similar to MRC evaluation (see Section 3.2), while the replacement samples are extracted from the entire set passages rather than all the references. Note that the initial and final words are not included for questions with more than two words, and the passage regarding the current question is also excluded.

- **Repeat**: a set of system-generated questions are randomly selected, and they are copied to make a set of "repeats".

In order to implement quality control, we will apply a significance test between the paired bad references and their associate ordinary questions on all rating types. In this case, a non-parametric paired significance test, Wilcoxon signed-rank test, is utilized as we cannot assume the scores are normally distributed. We use two set $Q = \{q_1, q_2, \dots\}$ and $B = \{b_1, b_2, \dots\}$ to represent the ratings of ordinary questions and bad references, where q_i and b_i respectively represent the scores of n rating criteria for an ordinary question and its related bad reference. For this experiment, we have 4 rating criteria as described in Table 4.1. We then compare the p-value produced by the significant test between Q and B with a selected threshold α to test whether the scores of ordinary questions are significantly higher than those of bad references. We apply the significance test on each worker, and the HITs from a worker with resulting $p < \alpha$ are kept. We choose $\alpha = 0.05$ as our threshold as it is demonstrated to be appropriate in Section 3.4.5.

Structure of HIT

The questions to be evaluated as well as their passages and answers are generated on the HotpotQA test set by 11 various systems, including one system called "Human" that can simulate the human performance, and 10 neural-network-based QG systems, the details of which will be introduced in Section 4.3.

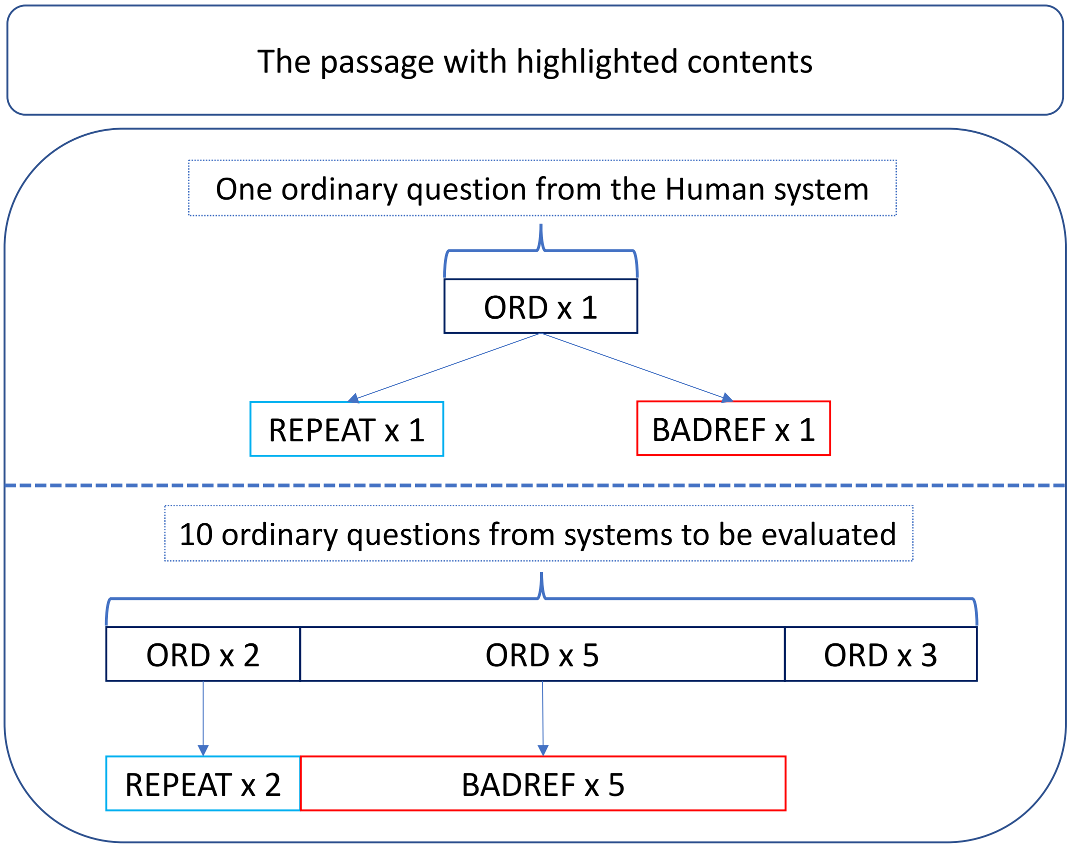


Figure 4.4: The structure of a single HIT in the QG evaluation experiment, where a HIT contains a certain passage, 11 system-generated questions and 9 variant questions for the purpose of controlling the quality. Meanwhile, ORD, REPEAT and BADREF respectively represent ordinary, repeat and bad reference questions.

For other tasks involving crowd-sourced human evaluation, a single HIT is made up of 100 items to rate (see Section 3.2.2). However, HITs with similar size are inappropriate in this case as a passage containing several sentences should be provided for workers, and a 100-item HIT means a highly oversized workload for an individual. The reading quantity in a single HIT is one of concern as our preliminary experiment shows that a HIT with too many contents to read can significantly decrease the workers' efficiency. Instead, we organize the structure of HITs in the QG evaluation experiment as follows:

- 1 original question, 1 repeat and 1 bad reference from the Human system (comprising a total of 3 questions);
- 2 original questions and their repeats from 2 of the 10 neural QG systems

(comprising a total of 4 questions);

- 5 original questions and their bad references from other 5 of the 10 normal systems (comprising a total of 10 questions);
- 3 original questions from the rest 3 of the 10 normal systems (comprising a total of 3 questions).

where all these questions in one HIT share the identical passage and the correct answer. In other words, each HIT in the QG evaluation experiment consists of 20 items to rate, including: (i) 11 ordinary system-generated questions; (ii) 6 bad reference question corresponding to 6 of these 11; (iii) 3 exact repeats corresponding to 3 of these 11. Figure 4.4 provides the detailed structure of a HIT, where ORD=ordinary question, REPEAT=repeat question and BADREF=bad reference question. Although the hierarchical structure in Figure 4.4 organizes the 20 items in a certain order, they will be fully shuffled before the deployment.

4.3 Dataset and Systems

We conduct the experiment on the HotpotQA dataset (Z. Yang et al., 2018), initially proposed for the multi-hop question answering task (see `https://hotpotqa.github.io/`). The term, multi-hop, means that a machine should have the ability to answer given questions by extracting useful information from several related passages. The documents in the dataset are extracted from Wikipedia articles, and the questions and answers are created by crowd workers. A worker is asked to provide the questions whose answers requires reasoning over all given documents. Each question in the dataset is associated with one correct answer and multiple passages, where the answer is either a sub-sequence from the passage or simply yes-or-no. These multiple passages are treated as a simple passage to show to human raters during the experiment. Note that the original HotpotQA test set provides no answer for each question, and such a set is inappropriate for the QG task as an answer is necessary for a QG system to generate a question. Instead, a common practice is to randomly sample a fraction from the training set as the validation set, and the original validation set can act as the test set when training or evaluating a QG system based on a QA dataset. The test set we used to grab system-generated outputs for the QG evaluation is in fact the validation set.

Besides, HotpotQA dataset provides two forms of passages: full passages and supporting facts. For each question, its full passages, on the average, consist of 41 sentences while the average number of sentences in its supporting facts is eight. Since the reading quantity is one of our concerns, we use the sentences from supporting facts to constitute the passage to prevent workers from reading too many sentences per assignment. And this is also the reason we choose HotpotQA, as such supporting facts are not always available in other datasets.

4.3.1 QG Systems for Evaluation

To analyze the performance of our proposed human evaluation method, 11 systems will be evaluated, including 10 systems that are trained on the HotpotQA dataset and the Human system that can represent the performance of humans on generating questions. The Human system is directly made up of the questions extracted from the HotpotQA testset. The 10 trained systems are from the following neural network models:

- **T5 (small & base)**: a model using a text-to-text transfer transformer (T5) that is pre-trained on a large text corpus (Raffel et al., 2020).
- **BART (base & large)**: a denoising autoencoder using a standard sequence-to-sequence Transformer architecture (M. Lewis et al., 2020).
- **Att-GGNN**: an attention-based gated graph neural network (Pan et al., 2020).
- **Att-GGNN (plus)**: the Att-GGNN model combined with the context switch mechanism (Ji, Lyu, et al., 2021).

- **H-Seq2seq**: a hierarchical encoding-decoding QG model (Ji, Lyu, et al., 2021).
- **H-Seq2seq***: the H-Seq2seq model using a larger dictionary for the avoidance of generating the unknown token ⟨UNK⟩.
- **GPT-2**: A large transformer-based language model whose parameter size reaches 1.5B (Radford et al., 2019).
- **RNN**: a vanilla RNN-based seq2seq model.

These systems then generate questions on the HotpotQA testset, and each system is guaranteed to have at least one question to be evaluated within each HIT.

4.4 Experiment Results

In this section, we design an experiment to investigate of our proposed human evaluation method on the AMT platform, and we report the details of experiments. We also report the human score of QG systems at the system-level based on the collected data, and apply significance tests on each pair of QG systems. We also compute the scores of alternative automatic metrics and investigate their performances via correlation with human judgements according to the results of human evaluation experiment.

4.4.1 HITs and Workers

Two runs of experiments are deployed on the AMT platform, where the second run is designed to serve as a self-replication experiment to ensure the reliability of experimental findings. We then compute the correlation between the human scores of two runs at the system-level to examine the consistency of our method, which will be introduced in Section 4.4.3. The HITs in the two experimental runs are randomly sampled from a HIT pool, which is generated as the outputs from the

Table 4.2: Statistical information of the collected experiment data.

(a) The numbers of both workers and HITs before and after the quality-controlling mechanism as well as their pass rates for two runs.

Experiment	Worker			HIT		
	Passed	Total	Pass rate	Passed	Total	Pass rate
Run1	123	356	34.55%	334	786	42.49%
Run2	105	283	37.10%	282	598	47.16%

(b) The average elapsed time per HIT needed to be completed in minutes, and the average number of HITs that a worker is assigned.

Experiment	Elapsed time (per HIT in minutes)			Assigned HIT (per worker)		
	Passed	Failed	Total	Passed	Failed	Total
Run1	33.24	26.93	29.61	2.72	1.94	2.21
Run2	38.68	25.79	31.87	2.69	1.78	2.11

aforementioned QG systems. Table 4.2 provides statistical information with regard to the data of workers and HITs collected from our human evaluation experiments.

Table 4.2a shows the numbers of human raters who participate in the QG evaluation experiment on the AMT platform, who passed the quality control and their pass rate for two distinct runs. The quality control method is as described in Section 4.2. The number of HITs before and after quality control, as well as the pass rate are also reported. For the first run, we collected 334 passed HITs resulting in a total of $18,704$ valid ratings. Specifically, a non-human system on average received $1,603$ ratings and the human system received $2,672$ ratings, which is a sufficient sample size since it exceeds the minimum acceptable number (approximately 385) according to the related research of statistical power in MT (Graham, Haddow, and Koehn, 2020).

Table 4.2b shows the average duration of a HIT and how many HITs a worker takes on the average according to the influence of the quality control method for both runs. Human raters whose HITs pass the quality control threshold usually spend a longer time completing a HIT than raters of failed HITs.

Cost of the Experiment

Similar to previous crowd-sourcing human experiments on the AMT platform, such as MT (Graham, Baldwin, Moffat, et al., 2016) and our MRC evaluation in Chapter 3, a worker who passed the quality control was paid 0.99 USD per completed HIT. This entire experiment cost less than 700 USD in total. For research using our proposed evaluation method in the future, the total cost should be approximately half of this since we ran the experiment an additional time to investigate reliability, which generally is not required. The resulting correlation between the system scores of the two separate data collection runs was $r = 0.955$, sufficient to ensure reliability of results. Failed workers were often still paid for their time, where they could claim to have made an honest attempt at the HIT. Only obvious attempts to game the HITs are rejected. In general, according to the cost of our first data collection run, assessing a QG system with nearly $1,600$ valid ratings in fact costed about 30 USD (total cost 334 USD $\div$ 11 models $\approx$ 30.4 USD). However, the experimental cost in future research may vary, depending on the sample size of collected data.

4.4.2 Human Scores

Human raters may have different scoring strategies, for example, some strict raters tend to give a lower score to the same question compared with other raters. Therefore, we use the average standardized (z) scores instead of the original score, in order to iron out differences resulting from different strategies. Equation 4.1 is the computation of the average standardized scores for each evaluation criterion and the overall score of a QG system:

$$\begin{aligned} z_q^c &= \frac{r_q^w - \mu_w}{\sigma_w} \\ z^c &= \frac{1}{|Q|} \sum_{q \in Q} z_q^c \\ z &= \frac{1}{|C|} \sum_{c \in C} z^c \end{aligned} \tag{4.1}$$

Table 4.3: Human evaluation standardized z scores of overall and all rating criteria in the first run, where a bold value indicates the system receives the highest score among systems except the Human system, and N indicates the number of evaluated questions of a system; systems (described in Section 4.3) are sorted by the overall score.

System	N	Overall	Understandability	Relevancy	Answerability	Appropriateness
Human	668	0.322	0.164	0.262	0.435	0.429
BART$_{large}$	400	**0.308**	**0.155**	**0.255**	0.420	**0.403**
BART$_{base}$	401	0.290	0.135	0.234	**0.430**	0.360
T5$_{base}$	395	0.226	0.051	0.241	0.395	0.217
RNN	395	0.147	−0.050	0.128	0.222	0.289
Seq2Seq	404	0.120	−0.030	0.022	0.180	0.309
T5$_{small}$	405	0.117	−0.108	0.106	0.260	0.210
Baseline$_{plus}$	408	0.076	−0.133	0.076	0.196	0.165
Seq2Seq*	396	0.053	−0.055	−0.039	0.088	0.217
Baseline	396	−0.008	−0.186	−0.032	0.155	0.032
GPT-2	408	−0.052	−0.202	−0.126	0.050	0.068

where the standardized score z_q^c on the criterion c of a system-generated question q is computed by its raw score r_q^c and the mean μ_w and the standard deviation σ_w of its rater w, z^c is the system-level standardized score on the criterion c of a QG system, Q is the set consisting of all rated questions (q) belonging to the QG system, and the overall average standardized scores z is computed by averaging the z^c of all criteria (C).

Table 4.3 shows the standardized human scores of all systems based on the ratings from all passed workers in the first run as well as the sample size N, where overall is the arithmetic mean of the scores of understandability, relevancy, answerability and appropriateness. A highlighted value indicates the system in the row outperforms every other system excluding the human Human question for that rating criterion. For the calculation of standardized z scores, the scores of bad references are not included, and for repeat questions the mean score of both evaluations for that question are combined into the final score.

As described in Table 4.3, the Human system receives the best z scores among all evaluation aspects, which is as expected since it consists of human-generated questions. For all QG systems excluding Human, BART$_{large}$ outperforms all other systems overall, and individually for understandability, relevancy and appropriateness. We also find that BART$_{base}$ somehow performs better than BART$_{large}$ at the answerability criterion. This is interesting as the performance of a model should generally increase if it is trained on a larger corpus. We think this implies that training models on a larger scale may potentially reduce the ability to generate high quality questions in terms of some aspects, namely answerability in this case. This is probably because a larger corpus may contain more noise which can negatively influence some aspects of a model, and it is worth investigating in future work.

4.4.3 System Consistency

To assess the reliability of the proposed human evaluation method, two distinct runs of the experiment are deployed with different human raters and HITs on the AMT platform. We think a robust evaluation method should be able to have a high correlation between the results of two independent experiment runs. Following a similar setting in the MRC self-replication experiment, the HITs in the two runs of the QG experiment are randomly generated from the same pool of candidate items as well, and both runs share the same instructions and user interfaces. We also conduct the second run after the first is completed, and workers in both runs are automatically allocated.

Table 4.4 shows the human evaluation results on the second run of our experiment, where the systems follows the order in the first run. We additionally compute the correlation coefficients between the standardized z scores of both runs as shown in Table 4.5, where r, ρ and τ represent Pearson, Spearman and Kendall's tau correlation, respectively. We observe that the overall scores of two distinct experimental runs can reach $r = 0.955$, while Person correlation of other evaluation criteria ranges from 0.865 (Relevancy) to 0.957 (Answerability). We believe such correlation values

Table 4.4: Human evaluation standardized z scores of overall and all rating criteria in the second run, where these systems follows the order in Table 4.3, and N indicates the number of evaluated questions of a system.

System	N	Overall	Understandability	Relevancy	Answerability	Appropriateness
Human	564	0.316	0.188	0.279	0.386	0.410
BART_{large}	342	0.299	0.180	0.277	0.380	0.359
BART_{base}	338	0.306	0.181	0.299	0.397	0.347
T5_{base}	329	0.294	0.158	0.298	0.396	0.326
RNN	342	0.060	−0.040	−0.008	0.072	0.217
Seq2Seq	332	0.086	−0.053	0.064	0.115	0.217
T5_{small}	340	0.157	−0.012	0.166	0.248	0.224
Baseline_{plus}	341	0.069	−0.094	0.081	0.134	0.157
Seq2Seq*	348	0.083	−0.014	0.077	0.104	0.163
Baseline	329	−0.025	−0.200	−0.023	0.042	0.083
GPT-2	343	−0.047	−0.122	0.000	−0.036	−0.031

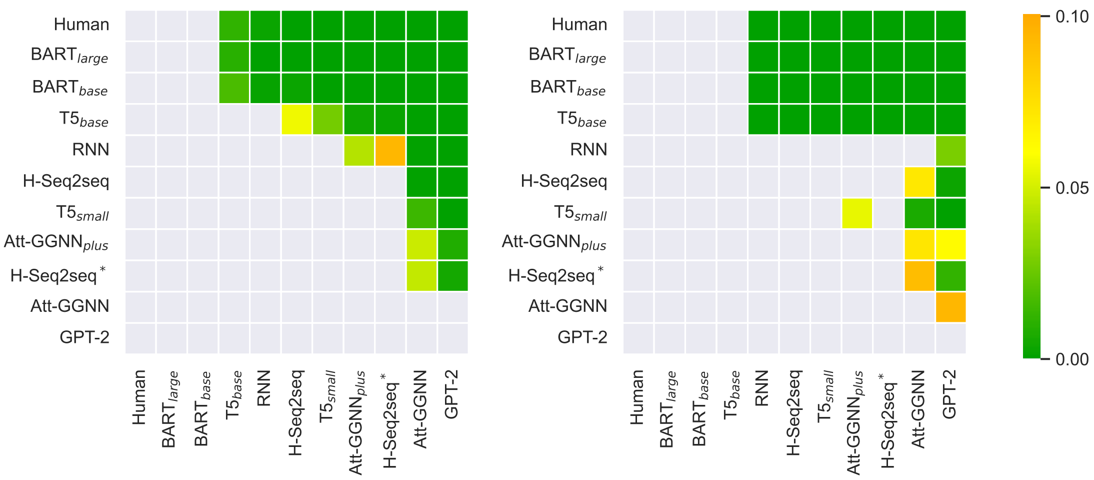
Table 4.5: The Pearson (r), Spearman (ρ) and Kendall's tau (τ) correlations between the standardized z scores of two runs of the experiment, including overall and four evaluation criteria.

	Overall	Understandability	Relevancy	Answerability	Appropriateness
r	0.955	0.953	0.865	0.957	0.884
ρ	0.882	0.891	0.718	0.882	0.845
τ	0.745	0.709	0.527	0.745	0.709

are high enough to ensure the robustness of our proposed human evaluation method.

4.4.4 Significance Tests

We apply the Wilcoxon rank-sum test to each pair of systems based on their human evaluation z scores of overall. The pairwise results between systems for first and second runs are shown in Figures 4.5a and Figures 4.5a respectively , where systems are sorted by the overall z scores and a coloured cell indicates the system in the row can significantly outperform the system in the column at $p < 0.1$. We observe that the heatmaps for two runs overlap at a very high proportion ($\approx 84\%$), which further indicates the reliability of the proposed method. Furthermore, overlap can



(a) Run 1 (b) Run 2

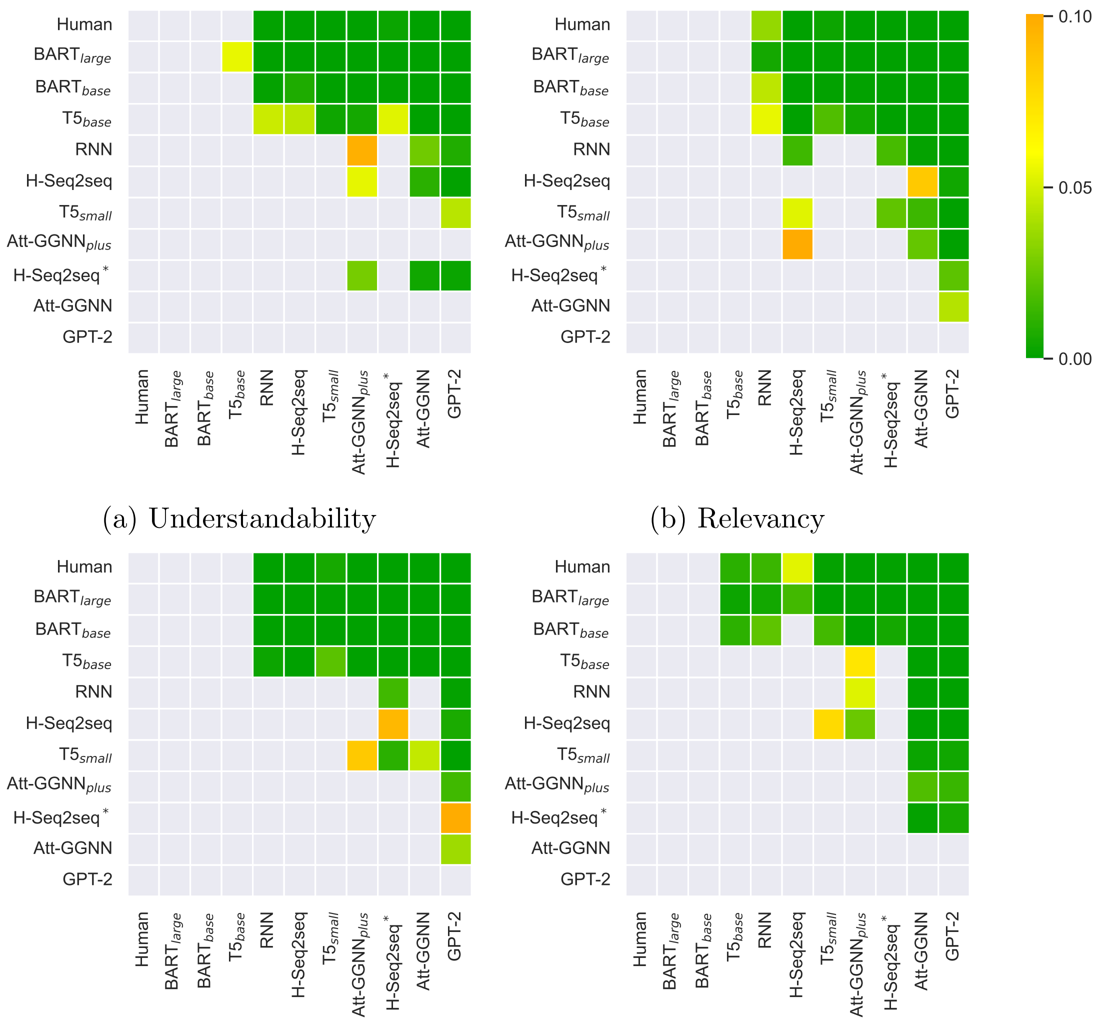
Figure 4.5: The results of significance test on the Overall z scores for QG systems in the first and second run, where the systems follows the orders in Table 4.3.

still reach 80% at $p < 0.05$. In addition, an interesting observation is that both runs indicates that the Human system, which represents the human performances on the QG task, cannot significantly outperform BART$_{large}$ and BART$_{base}$. This could be interpreted as a statistical tie with human performance for the task for BART within the context of this evaluation setting.

Figure 4.6 provides the results of significance tests on the four individual evaluation criteria: Understandability (Fig. 4.6a), Relevancy (Fig. 4.6b), Answerability (Fig. 4.6c) and Appropriateness (Fig. 4.6d).

4.4.5 Human Assessor Consistency

We report the distribution of three types of correlation coefficients as the inter-annotator agreement in the first run of our experiment from both passed and failed workers. Figure 4.7 is the distributions of correlations, including Pearson (r), Spearman (ρ) and Kendall (τ), computed on the ratings of each pair of an ordinary question and its repeat, where the results of workers who passed and failed the quality control methods are respectively marked as blue and orange. As shown in Figure 4.7a, the Pearson correlations of passed workers tend to lie in 0.7–1.0 while those of failed workers are mostly located around 0.4. Figure 4.7b and Figure 4.7c re-



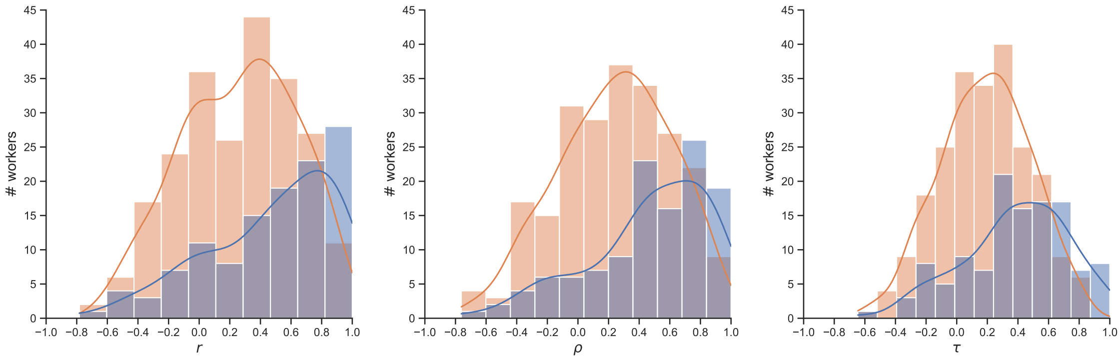
(a) Understandability (b) Relevancy (c) Answerability (d) Appropriateness

Figure 4.6: Significant test on scores in four rating criteria, where the QG systems follows the orders in Table 4.3.

port the results of the Spearman and Kendall correlations on both types of workers which show a similar tendency with Pearson. Generally speaking, passed workers, according to the distributions of all three types of correlation, have a higher level of rater agreement than failed workers, and our quality control method is capable of filtering out low-quality workers.

4.4.6 Automatic Metrics

In Section 2.2, we introduced automatic metrics which are commonly applied for the QG evaluation. In this section, we investigate the accuracy of these metrics when



(a) Pearson (b) Spearman (c) Kendall

Figure 4.7: The distribution of the Pearson (r), Spearman (ρ) and Kendall's tau (τ) correlations between the ratings of pairs of repeat and ordinary questions on individual raters who passed (blue) and failed (orange) the quality control method.

evaluating QG systems in our experiment by calculating the correlation between these automatic metric scores and our human evaluation z scores.

System Scores

Table 4.6 shows the automatic metric scores, where the metrics are ranked by their correlations with the Overall z score. Since automatic metrics focus on the similarity between a candidate and a reference, we cannot compute a metric score for the Human system as its candidates are actually the references in the dataset. We observe that METEOR performs best among these metrics, while perhaps surprisingly Q-BLEU1, which is proposed for the QG task, has the lowest correlation with humans.

Figure 4.8 provides further analysis of the ranking of systems by each evaluation metric, where the automatic metrics are ordered by highest correlation with human assessment. It can be seen that the rankings of the first systems barely change across metrics, as BART$_{large}$ and BART$_{base}$ only switch when Q-BLEU4 and Q-BLEU1 are applied. Since Q-BLEU is proposed to assess how likely a question is able to be answered, we believe that such a change of ranking is reasonable since our human evaluation also shows that BART$_{base}$ has a higher answerability score than BART$_{large}$ (see Table 4.3). An interesting observation is that all automatic metrics

Table 4.6: Automatic metric scores of systems as well as the standardized human evaluation z score of Overall. These metrics are sorted by the values of their Pearson Correlation (r) with z. The Human system is excluded since its automatic metric scores are unavailable.

System	METEOR	ROUGE-L	BERTScore	BLEURT	Q-BLEU4	Q-BLEU1
BART_{large}	30.18	47.58	90.85	-0.363	43.77	51.47
BART_{base}	29.66	47.13	90.74	-0.381	44.14	51.65
T5_{base}	27.99	41.60	88.44	-0.682	37.78	44.84
RNN	15.46	26.77	84.59	-1.019	9.68	15.92
H-Seq2seq	17.50	29.86	85.49	-0.953	10.51	17.74
T5_{small}	23.62	32.37	86.34	-0.860	26.73	32.92
Att-GGNN_{plus}	21.77	36.31	86.27	-0.784	12.63	19.86
H-Seq2seq*	18.23	31.69	85.83	-0.866	11.12	18.36
Att-GGNN	20.02	33.60	86.00	-0.802	11.13	18.67
GPT-2	16.40	29.98	86.44	-0.899	24.83	31.85

consider the RNN system worst, while humans rank it as highly as fourth position among systems. We think a potential cause is that, the questions produced by RNN may be appropriate according to humans, but they do not have much overlap with references, resulting in low metric scores. For other systems, their rankings fluctuate when the evaluation method changes. We also find that Q-BLEU4 and Q-BLEU1 consistently produce the same ranking.

Correlation Coefficients and Williams Test

Table 4.7 shows correlation of automatic metrics with human evaluation according to Pearson (r), Spearman (ρ) and Kendall's tau (τ) correlation coefficients. METEOR is the only metric with a Pearson correlation reaching 0.8, while other metrics are only above 0.7. We cannot conclude from the increased correlation of METEOR that it is statistically significantly better than the other metrics, and a significance test is necessary (Graham and Baldwin, 2014). We apply the Williams test (Williams, 1959) for the assessment of differences in correlations in this experiment. Since the correlations are computed between scores of evaluation methods in the same data set, we cannot assume they are independent, while Williams test is suitable

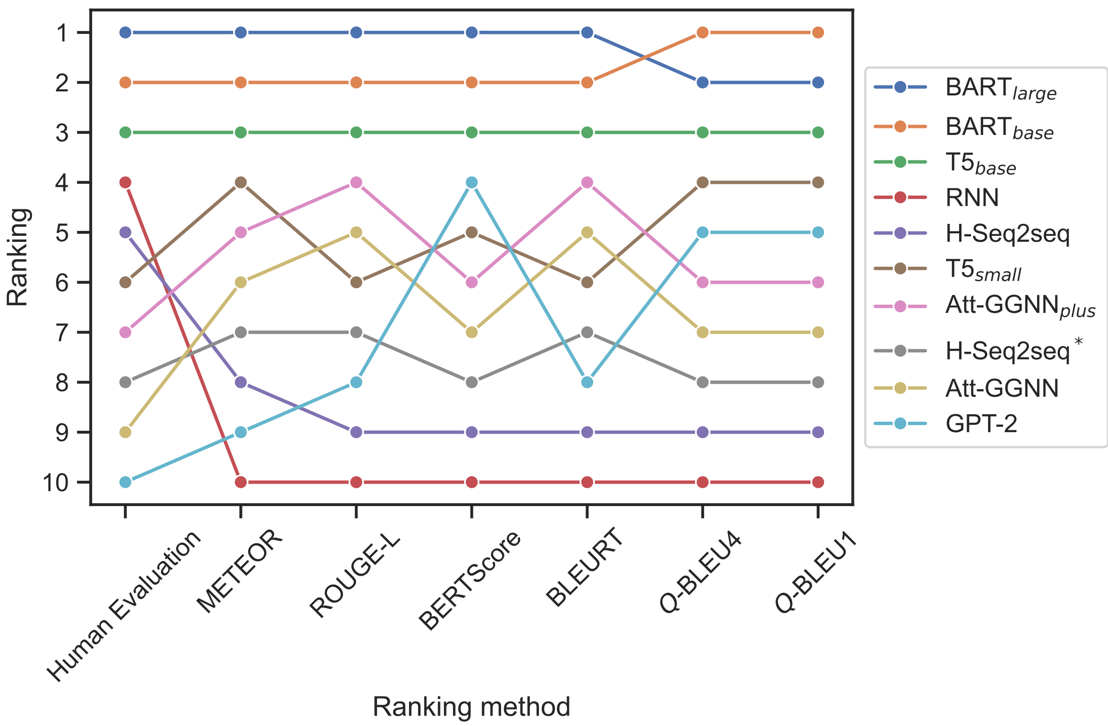


Figure 4.8: Ranking of QG models according to the Overall z scores and automatic metrics for the first run.

for dependent correlations (Steiger, 1980). For each pair of automatic metrics, we apply Williams test to their correlations with human judgement, $p < 0.01$ indicates that the difference in their correlations is statistically significant. However, we unfortunately find that no automatic metric can outperform metrics in the lower cluster at $p < 0.01$, as even the Williams test between the best metric (METEOR) and the worst metric (Q-BLEU1) is $p = 0.248$. Therefore, we cannot conclude that METEOR is significantly better than other metrics in this case, and its increase of Pearson correlation is likely to occur by chance.

With the aforementioned results, we can therefore answer **RQ 1** with regard to the QG task, which is described as follows:

- **RQ 1**: *How accurately do existing automatic metrics measure QG system performance?*

The answer to **RQ 1** is that, automatic metrics generally fail to accurately measure QG systems since they failed to achieve high correlation with human judgement. In addition, even METEOR which has the highest correlation with human judgement

Table 4.7: The Pearson (r), Spearman (ρ) and Kendall's tau (τ) correlation between automatic metric scores and the Overall scores in the first run, where the metrics are sorted by r.

	QAScore	METEOR	ROUGE-L	BERTScore	BLEURT	Q-BLEU4	Q-BLEU1
r	0.864	0.801	0.770	0.761	0.739	0.725	0.724
ρ	0.827	0.612	0.503	0.430	0.503	0.467	0.467
τ	0.709	0.511	0.378	0.289	0.378	0.289	0.289

cannot significantly outperform other metrics in this case.

4.5 QAScore - Evaluating QG Systems using Pretrained Language Model

In this section, we describe our newly proposed pretrained-model-based unsupervised QG metric. From the results described in Section 4.4, we find that current automatic metrics fail to achieve a high correlation with human evaluators, whether overlap-based metrics or trained models. To solve this issue, we propose a new automatic evaluation for the QG task. It should be taken into account that there are many possible correct questions for the same answer and passage, meaning that multiple distinct questions can legitimately share the same answer due to the *one-to-many* nature of QG as described in Section 4.1. Additionally, one disadvantage is that all existing automatic evaluation metrics entirely rely on the comparison between a candidate and a ground-truth reference, which can additionally answer **RQ 2** regarding QG:

- **RQ 2**: *What are the limitations and disadvantages of the direct application of evaluation metrics from MT and other domains to entirely distinct tasks for system development in QG?*

And we can answer **RQ 2** as that, current applied metrics are inappropriate for the evaluation of QG since a system is able to generate an eligible question with few overlap with the reference. Hence, we believe a reference-free metric is more appropriate since there can be several correct questions for a given pair of an answer and a passage.

Another drawback is that metrics like BERTScore and BLEURT perform even worse than classical evaluation metrics since they are initially proposed and trained for distinct tasks. For example, BLEURT is a BERT-based model and it is further fine-tuned on the task-specific human ratings to achieve a high correlation with human in the MT task, while our experiment shows its performance on the QG task is unsatisfactory. Although a possible way to improve its performance is to fine-tuning them to achieve a higher correlation with human, it requires extra resources which are usually expensive and time-consuming. Thus, our proposed metric has the advantage of being unsupervised. Pretrained language models are demonstrated to contain plenty of useful knowledge since they are trained on large scale corpus (Shin et al., 2020a). Therefore, we plan to directly employ a pretrained language model to act as a evaluation metric without using other training data or supervision, as introduced in Section 4.5.1.

Hence, we propose a new automatic metric which can directly use a pretrained model as its scorer in a plug-and-play manner. Compared with previous evaluation metrics, our approach has three main advantages: i) it can evaluate a standalone candidate question with no need to compute the similarity with any human-generate reference; ii) it is easy to deploy as it directly takes a pretrained language model as the scorer and requires no extra data for further fine-tuning; iii) it correlates better with human judgement according to the results of our human evaluation experiment.

4.5.1 Methodology

Since QG and QA are two complementary tasks, we can naturally conjecture that a QG-system-generated question can be evaluated according to the quality of the answer generated by a QA system. We take the passage and the answer A, "commander of the American Expeditionary Force (AEF) on the Western Front", in

Figure 4.2 as an example. We show two distinct question $Q1$ and $Q2$, where $Q1$ is "*What was the most famous post of the man who commanded American and French troops against German positions during the Battle of Saint-Mihiel?*" and $Q2$ is "*What was the Battle of Saint-Mihiel?*". It can be found that, A is the correct answer to $Q1$ rather than $Q2$. In this case therefore, a QA model is *more* likely to generate A when given $Q1$, and it is expected *not* to generate A when given $Q2$. In other words, the likelihood that a QA model can produce a given $q1$ is more than that given $Q2$, meaning that the proposed metric will score $Q1$ higher than $Q2$. The detailed scoring mechanism will be introduced in Section 4.5.1.

RoBERTa

We chose to employ the masked language model RoBERTa (Yinhan Liu et al., 2019) in a MRC manner to examine the likelihood of an answer, and its value can act as the quality of the target question to be evaluated. RoBERTa (**R**obustly **o**ptimized **BERT a**pproach) is a BERT-based approach for pretraining a masked language model. Compared with the original BERT, RoBERTa is trained on a larger dataset with a larger batch size and longer elapsed time. It also removes the next sentence prediction (NSP) step and leverages full-sentences (sentences that reach the maximal length). For text encoding, RoBERTa employs a smaller BPE (Byte-Pair Encoding) vocabulary from GPT2 instead of the character-level BPE vocabulary employed in the original BERT. We believe these approaches enable RoBERTa to determine improper answers when receiving the passage and the question.

Process of Scoring

Given the passage, the correct answer, and the QG-system-generated question, we first encode and concatenate the passage and the answer. Figure 4.9 provides a visualization of the process of scoring a generated question using its passage and answer using the masked language model RoBERTa. First, the passage and the question are concatenated by the end-of-sequence token $\langle \text{eos} \rangle$, which represents the

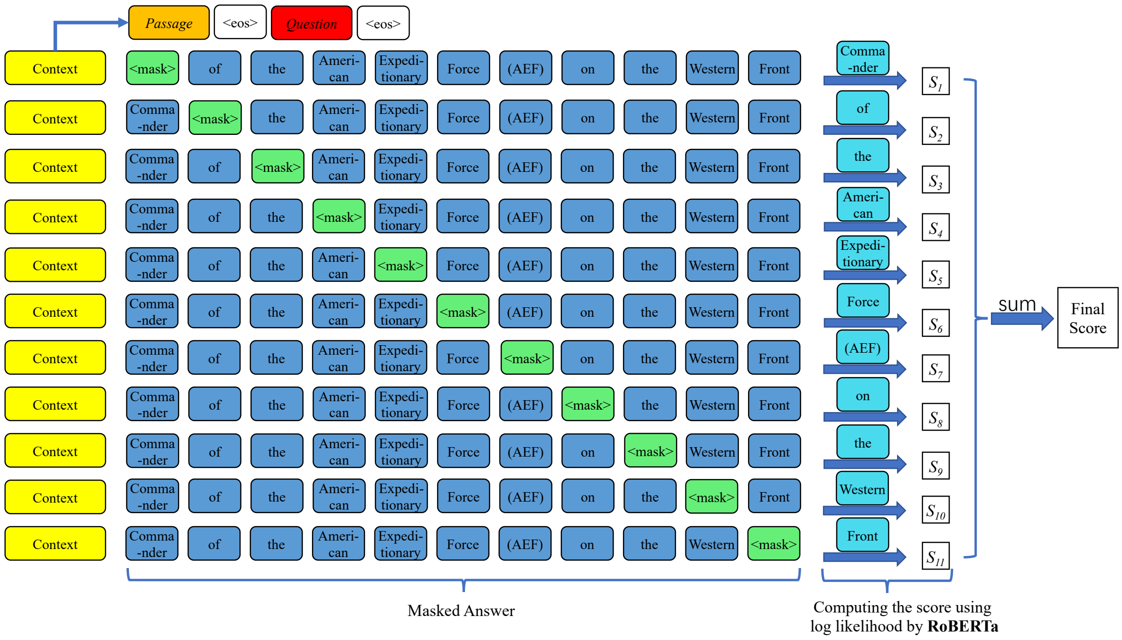


Figure 4.9: The process of scoring a question by RoBERTa, where the context (yellow) contains the passage and the question (to be evaluated), the score of a single word is the likelihood that RoBERTa can predict the real word (cyan) which is replaced by the mask token ⟨mask⟩ (green) in the original answer, and the final metric score is the sum of scores of all words in the answer.

context for the masked language model. Next, the masked answer containing one masked word is concatenated by the context together with the ⟨eos⟩ token as the input for the model. The model is then asked to predict the real value of the masked word using the context and the masked answer. The log likelihood that RoBERTa can generate the true word can act as the score for that masked word. For the evaluation of a single question, all words in the given answer will be masked in a one-at-a-time manner. The final metric score of the question q can be computed by Equation 4.2:

$$\text{QAScore}(p) = \sum_{w \in a} LL(w|p, q, a_{\widetilde{w}}) \tag{4.2}$$

where p is the passage, a is the answer, w is a word in the answer, $a_{\widetilde{w}}$ is the answer of which the word w is replaced by the mask token, and LL is the function which returns the log likelihood that RoBERTa can produce the real word w given the passage p, question q and the answer $a_{\widetilde{w}}$ with one word masked.

4.5.2 Results

Since this proposed metric leverages a means of QA to assess QG-system-generated questions, we call it QAScore. Table 4.8 shows the metric scores of QG systems evaluated using QAScore, and Table 4.7 describes how QAScore correlates with human judgement according to our human evaluation experiment. Since our metric does not rely on a ground-truth reference, we can include the result of the Human system unlike other automatic metrics. It can be seen that our metric correlates with human judgement at 0.864 according to the Pearson correlation coefficient, where even the best automatic metric METEOR can only reach 0.801 (see Table 4.7). Also, compared with the other two pretrained-model-based metrics BERTScore and BLEURT, our metric can outperform them at > 0.1. In terms of Spearman, our metric achieves $\rho \approx 0.8$ where other metrics can only reach at most $\rho \approx 0.6$. In addition, our metric also outperforms other metrics according to Kendall's tau since it reach at $\tau \approx 0.7$ and other metrics merely achieve at most $\tau \approx 0.5$. We can conclude that our metric correlates better with human judgements with respect to all three categories of correlation coefficients. Nevertheless, we did not carry out the Williams test between QAScore and other metrics because QAScore evaluates 11 systems while other metrics only evaluate 10 systems, while Williams test requires two metrics has the same number of evaluated systems.

Together with the results of the proposed human evaluation method, we can therefore provide the answer to **RQ 3** in terms of QG:

- **RQ 3**: *Can more appropriate new methods of evaluation be designed that are feasible given the limited time and resources available in operational settings?*

This can be answered from two aspects: i) in consideration of the cost and efficiency, we successfully propose a new crowding-sourcing human evaluation method which is appropriate for the QG task within limited time and resources; ii) we also propose a new reference-free automatic evaluation method which has the best performance according to the correlation with human and needs no extra resources.

Table 4.8: The scores of all QG systems based on this proposed evaluation metric QAScore as well as the overall z score from the first run of our human evaluation experiment, where systems follow the order in Table 4.3.

System	QAScore	Overall (z)
Human	−0.985	0.322
BART$_{large}$	−1.020	0.308
BART$_{base}$	−1.030	0.290
T5$_{base}$	−1.037	0.226
RNN	−1.064	0.147
H-Seq2seq	−1.076	0.120
T5$_{small}$	−1.049	0.117
Att-GGNN$_{plus}$	−1.065	0.076
H-Seq2seq*	−1.045	0.053
Att-GGNN	−1.068	−0.008
GPT-2	−1.108	−0.052

4.6 Summary

In this chapter, we propose a new crowd-sourcing human evaluation method for the question generation task. Each candidate question is evaluated on four various aspects: Understandability, Relevancy, Answerability and Appropriateness. To investigate the robustness of our method, we deployed a self-replication experiment that the correlation between the results from two independent runs can reach as high as $r = 0.955$. We also provide the means of filtering out unreliable data from unqualified workers. We introduce the structure of a HIT, the dataset we used and the involved QG models to encourage the community to repeat our experiment.

With the data we collected from the completed experiment, we first analyse information of human raters and assigned HITs, including the pass rates and elapsed time. We report the standardized scores, including the overall score and four individual evaluation criteria. Using the standardized scores, we examine significance test of both runs. We also compute the scores of prevailing automatic metrics on collected data, and results show that METEOR correlates best with human at $p = 0.801$. However, William test shows no automatic metric can significantly outperform other metrics.

To overcome the disadvantages of current automatic metrics, we therefore propose a new unreferenced pretrained-model-based metric. Compared to existing prevailing evaluation metrics, this proposed metric can achieve the best performance on the collected data set, according to correlations with the results of our human evaluation experiment.

In conclusion, we propose a new crowd-sourcing evaluation method for the question generation task with high robustness and efficiency. This method can be deployed on a large scale within a limited budge of time and resources. We additionally propose a reference-free automatic metric which achieves the highest correlation with human among other automatic metrics.

Chapter 5

Evaluations on Open-domain Dialogue Systems

We presented existing challenges in evaluation of open-domain dialogue systems in Chapter 1 and Chapter 2. For example, there lacks a clear definition of a high-quality dialogue, and automatic metrics are usually criticized for weakly correlating with human judgements. In addition, the human evaluation of open-domain dialogue is generally too expensive and time-consuming to be practical. Therefore, open-domain dialogue indeed requires an evaluation method which is accurate, affordable and effective.

To overcome the challenges of open-domain dialogue evaluation, we introduce a new evaluation method based on human assessment of live conversations with systems in this chapter. We also conduct corresponding human evaluation experiments. According to the results of our experiments, we can therefore investigate the proposed **RQ**s in terms of open-domain dialogue.

Section 5.1 focuses on the design of a human evaluation experiment, consisting of the detailed methodology and according user interfaces in the platform. Section 5.2 introduces a method for quality controlling of the human workers, including a preliminary method that failed to operate as expected. Section 5.3 describes the PersonaChat dataset which is used in our investigation, and details of alternative dialogue systems. Section 5.4 provides details and analysis of the data we collected from deployed experiments. Section 5.4.5 investigates influence of preassigned profiles, called persona, on system performances. Finally, Section 5.5 concludes this chapter.

5.1 New Method for Evaluation of Open-domain Dialogue Systems

Evaluation of open-domain dialogue systems is highly challenging. Importantly existing automatic metrics do not provide reliable indication of what may or may not be a high-quality conversation. In this chapter, we propose a new human evaluation method which seeks to overcome the challenges of evaluation of open-domain dialogue systems introduced in Section 2.3. Our methodology is introduced in detail in this section, including choice of an appropriate evaluation procedure, the mechanism for testing reliability, selection of rating scale for human raters, the design of a user interface for use when deploying evaluation experiments on crowd-sourcing platforms, and the evaluation criteria for use in assessment of different aspects of a conversation.

5.1.1 Methodology

Evaluation Procedure: Static or Live

According to the underlying procedure, human evaluation can be generally divided into two categories: static and live evaluation. The former is in an offline process where human assessors are asked to evaluate according to the given dialogue history which is randomly from a corpus of conversations (Z. Lin et al., 2019; Du and Black, 2019). In live evaluation, human assessors are responsible for participating in both chatting with a system and evaluating it according to their interaction experience. Since static evaluation dispenses with interaction between a human worker and the dialogue system, the worker only needs to focus on reviewing the content of a conversation. Accordingly, static evaluation is deemed to have a lower cost with respect to time and required human labour.

However, static evaluation is regarded as having less validity than live evaluation (Finch and Choi, 2020). A human rater in static evaluation is absent from the interaction with the system, and this will influence the accuracy of their evaluation. Ghandeharioun et al. (2019) think an external assessor, namely an assessor in static evaluation, can only subjectively estimate the degree of a user's satisfaction with the system according to the provided conversation, making it even less accurate. The open-domain dialogue task itself generally has a less specific objective, resulting in a lack of an objective indicator, hence the success of a conversation can only be judged by its participant. Therefore, our proposed method of evaluating open-domain dialogue systems, seeks to address the weaknesses of static assessment by using human assessment of live conversations.

Test of Reliability

Although live human evaluation of dialogue systems has the advantage of having high validity, its reliability unfortunately lacks an efficient means of assessment. Additionally, developing methods of evaluation for language tasks that achieve high rater consistency has been challenging, often resulting in low levels of agreement between annotators (Finch and Choi, 2020; Callison-Burch, Koehn, Monz, and Zaidan, 2011; Callison-Burch, Koehn, Monz, Post, et al., 2012; Ondřej Bojar, Buck, et al., 2013; Ondrej Bojar, Buck, et al., 2014). Being aware of these challenges, we conduct a self-replication experiment for our proposed method to examine its reliability. In detail, we will deploy the same human evaluation experiment for two distinct runs. The order of dialogue systems in HITs of both runs are random, and the participant human workers are randomly allocated by the crowd-sourcing platform. When both runs are completed, the reliability of our evaluation method is assessed by testing the correlation between the resulting human scores of the systems from those two runs.

Human Ratings of Dialogue Quality

Similar to the aforementioned MRC and QG human evaluation methods, a continuous (0–100) rating scale is employed for dialogue systems, with three main motivation points (Graham, Baldwin, Moffat, et al., 2013; Novikova, Dušek, and Rieser, 2018; Li, Weston, and Roller, 2019; Santhanam and Shaikh, 2019; Santhanam, Karduni, and Shaikh, 2020; Mille et al., 2020; Loïc Barrault et al., 2020; Howcroft, Belz, et al., 2020), as follows:

1. When comparing the performance of competing dialogue systems, the continuous scale can reduce the potential bias by enabling score standardization. The score distribution of each human assessor is standardized according to the overall mean and standard deviation of all ratings provided by that assessor, thus removing any adverse effects of those employing overly harsh (or indeed lenient) scoring strategies.

2. The continuous scale allows us to leverage statistical approaches to help determine which systems can significantly outperform the others. Namely, we can apply standard significance tests on the distributions of human scores of dialogue systems.

3. The continuous rating scale enables crowd-sourcing human evaluation to facilitate highly accurate quality control of anonymous workers so that the evaluation can be deployed on a large scale while still maintaining its validity at a low cost. This is possibly most important because the results of a live evaluation can be rendered meaningless due to the lack of an approach to discard invalid data.

Table 5.1: The evaluation criteria employed to assess models in our human evaluation in the form of Likert statements; corresponding evaluation labels *not* shown to human assessors.

Label	Likert statement
Robotic	*It was obvious that I was talking to a chatbot as opposed to another human user.*
Interesting	*The conversation with the chatbot was interesting.*
Fun	*The conversation with the chatbot was fun/enjoyable.*
Consistent	*The chatbot was consistent throughout the conversation.*
Fluent	*The chatbot's English was fluent and natural throughout the conversation.*
Repetitive	*I felt that the chatbot kept being repetitive during the conversation.*
Topic	*The chatbot stays on topic.*

5.1.2 Process of Evaluation and User Interface

Similar to previous works, such as MRC in Chapter 3 and QG in Chapter 4, our evaluation method is deployed on the crowd-sourcing platform AMT as well, where the assignment for each worker is called a HIT. Before starting a HIT, we present instructions to each crowd-sourcing worker to introduce the current task and provide clear guidance. Figure A.1 in Appendix A shows the full list of instructions shown to workers in detail. In particular, we require the workers to interact with a dialogue system, which is in fact a "chatbot" from the perspective of workers, in a realistic and non-repetitive manner. Two examples are provided for the workers to help avoid generating unnecessary conversations, as such live evaluations can result in meaningless data (Dinan, Logacheva, et al., 2019). In addition, the minimal required number of inputs and the total number of systems to be evaluated are shown to workers.

The static evaluation procedure was employed for human evaluation of MRC and QG (See Chapter 3 and 4), where all evaluation data can be included in a comma separated values (CSV) file. However, a single CSV file is not sufficient in live evaluation of dialogue systems since this needs the ability to respond to requests from workers. Hence, we deploy all the dialogue systems to be evaluated on a server that can handle the requests, and a CSV file is used to indicate which system the

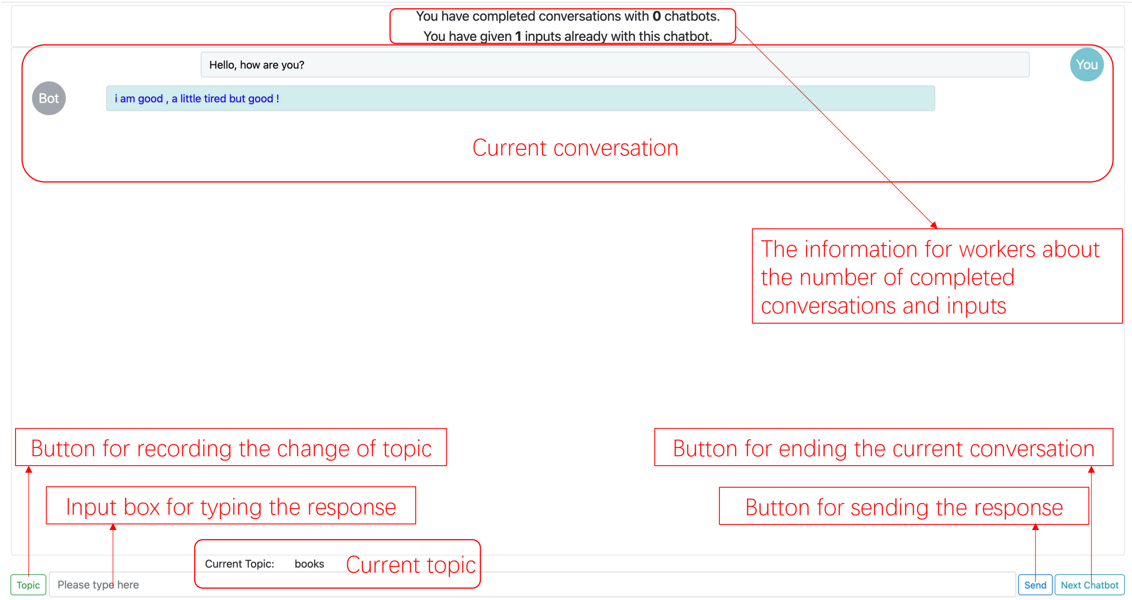


Figure 5.1: The user interface for workers when interacting with a system.

worker should chat with. This relates to instruction 11: *The chatbot may take a few seconds to respond, please be patient*, because the communication may takes a few seconds until the worker receives the response from the server. In addition, we found that special symbols (for example, &, #) in the input sentence may cause errors when sending a HTTP request during debugging. Meanwhile, some features are not fully supported by other browsers. Thus we provide instruction 12: *Please use Chrome and avoid special symbols if possible* to prevent our experiment from potential errors.

Crowd-sourcing workers can click the “I understand” button to start the evaluation, or quit the current assignment by closing the page if needed. If the current assignment is accepted, each human assessor is asked to carry out a live conversation with a randomly selected system before rating the quality. When the minimal requirement is satisfied, namely 10 conversational input sentences as introduced in Figure A.1, the assessor should manually end the interaction with dialogue system and start evaluate the quality of the conversation under a number of criteria. We set this minimal number to prevent the collected dialogues from being too short, and we believe 10 is an appropriate value. However, this can be adjusted in future research.

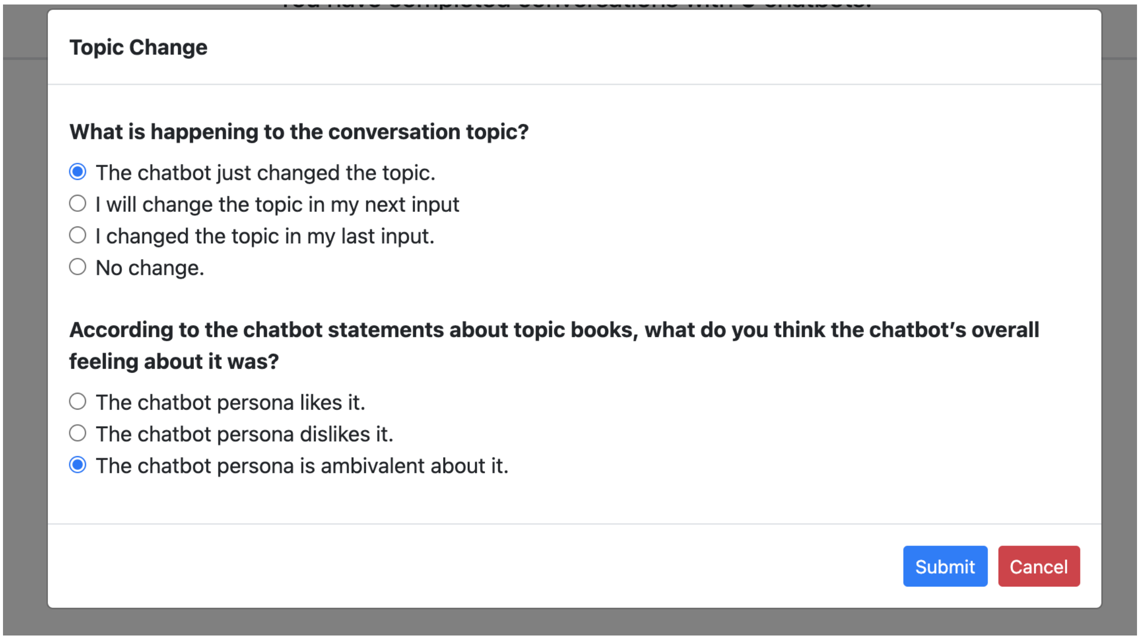


Figure 5.2: The popup window when recording the change of topic by clicking the Topic button as shown at the bottom left in Figure 5.1.

Table 5.1 shows the evaluation criteria employed in our experiment. These criteria are not completely immutable, and we encourage to extend, remove and adjust them for future researches as necessary.

Figure 5.1 shows our user interface for chatting with a system. At the top, the assessor can see how many systems have been completed, as well as the number of conversational inputs to the current system. The middle of the interface shows the conversation history between the assessor (right) and the system (left). At the bottom, workers can type input sentences and send them by clicking the send button. In addition, workers can see the current topic, information regarding topics is given in Section 5.4. Also, workers are required to record it if the topic is changed by either themselves or the current system, by clicking the Topic button at the bottom left. A popup window for recording the change then appears, as shown in Figure 5.2. The changes of topic includes: i) the worker find the chatbot changed the topic, ii) the worker changed the topic in the last input, and iii) the worker want to change the topic in next input. Workers are additionally record the chatbot's feeling of the topic, including *like*, *dislike* and *ambivalent*. Although the topic of a conversation is not directly related to our proposed research questions, we think this is useful for

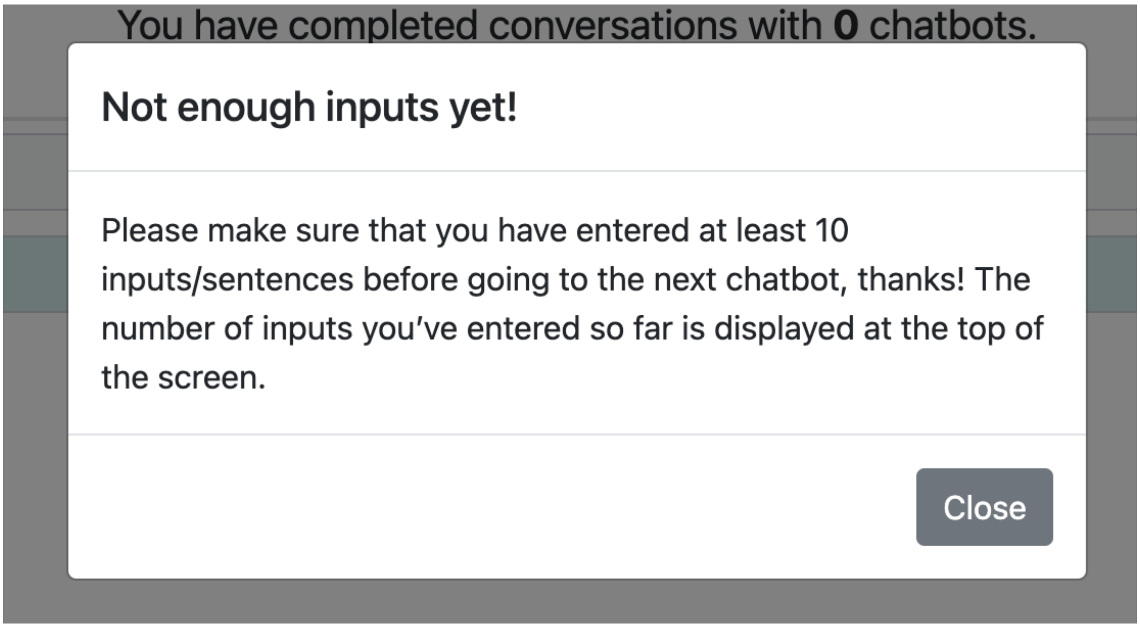


Figure 5.3: Popup warning when a worker clicks the Next Chatbot button without sufficient conversation turns.

future work, such as investigating the influence of changing topics on the quality of a conversation. The worker can manually complete the current conversation by clicking the Next Chatbot button. However, workers are not allowed to end the interaction with the current chatbot if the minimal requirement of conversation turns (10 in this case) has not yet been satisfied. A warning window appear as shown in Figure 5.3.

Subsequently, we ask assessors to rate each completed conversation under the seven aforementioned measurement criteria (see Table 5.1) according to their experience of the interaction. Figure 5.4 shows the interface for rating the conversation.

Although a continuous rating scale is advantageous for several reasons as previously introduced, the employment of such a scale raises the question of how it should be labeled. In evaluation of language tasks, adjectival scale labels, such as *poor, low, medium, high, perfect/ okay, good, excellent*, and so on, are often employed despite their likely contribution to annotator inconsistency (Loukina et al., 2020; Sorodoc et al., 2017). This is despite evidence of adjectival scale labels being problematic in terms of bias resulting from positively and negatively worded items not being true opposites of one another, and items intended to have neutral intensity in fact proving to have specific conceptual meanings. Alexandrov (2010) provides a summary of issues associated with adjectival labels.

To avoid any such causes of inconsistency, we structure each rating as a simple Likert declarative statement and ask human assessors to rate the degree to which they agree with each of these statement, making it possible to keep the rating scale constant while only changing the statement for each measurement criteria. Each criterion, along with its corresponding Liker statement, is labeled only at each extreme with *strongly disagree* (left) and *strongly agree* (right). Workers can evaluate by dragging the slider, but its real value in the continuous scale is invisible to them.

By clicking the NEXT button, a worker completes the evaluation of the current conversation and moves to the next system to be evaluated. It is not permitted to skip the current system, review previous conversations or modify completed ratings. When all systems have been completed, workers are encouraged to leave their feedback for future analyses and improving experiences of interaction and experiment design before ending their participation in the HIT.

5.2 Quality Control

Human evaluation in dialogue systems has been found to suffer from low-quality or even meaningless collected data (Dinan, Logacheva, et al., 2019), and low degree of rater agreement, regardless of either experts or crowd-sourcing workers are employed (Mehri and Eskenazi, 2020b; Finch and Choi, 2020). In addition, crowd-sourcing workers may seek to game the work, making no proper attempt at the task due to the anonymous nature during data collection, sometimes results may even be created by some automated robotic process (Loic Barrault et al., 2019). Therefore, an appropriate approach to identify and discard unusable data is necessary.

Many existing approaches to controlling the quality of crowd-sourcing workers depend on employing pre-created gold-standard items as quality checks (J. Le et al.,

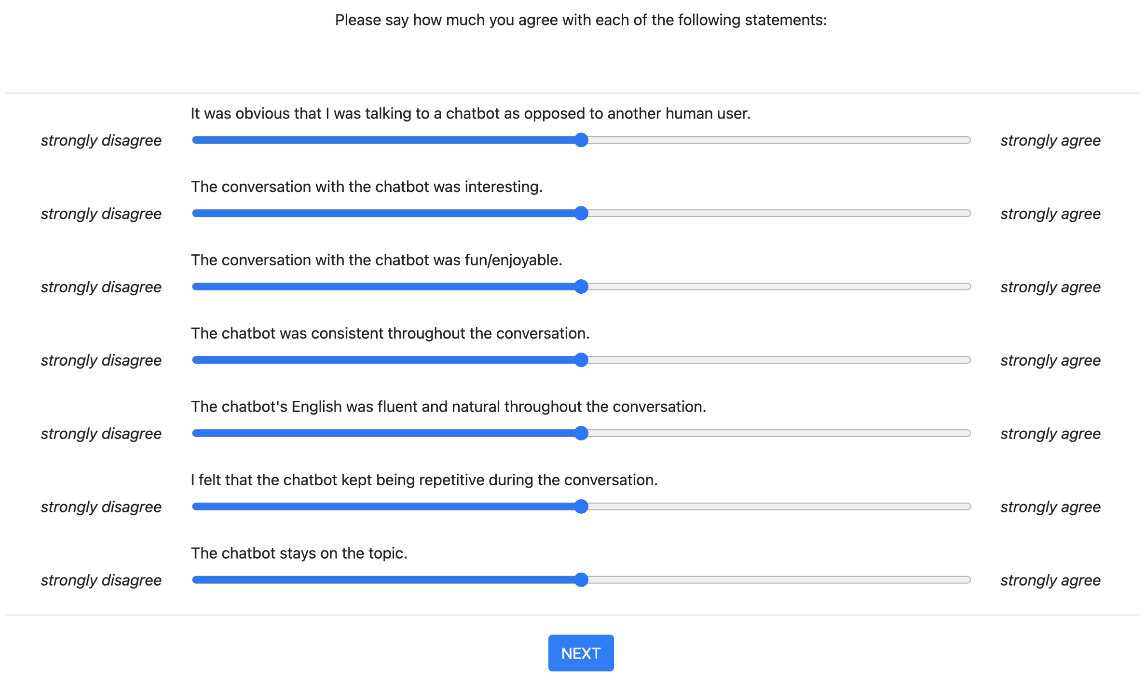


Figure 5.4: The interface shown to a worker to evaluate the conversation with a system after clicking the Next Chatbot button in Figure 5.1. Once evaluation of the current conversation is done, worker clicks the NEXT button to move to the next system. If all conversations are completed, the worker is redirected to a feedback page to leave the feedback and finish the HIT, as shown in Figure 5.5.

2010; S.-W. Huang and Fu, 2013; Qiang Liu, Ihler, and Steyvers, 2013; Lasecki, Teevan, and Kamar, 2014). This approach filters out the workers who fail to give high scores to gold-standard items. However, it is highly likely to allow low quality data to pollute the resulting evaluation, since any worker willing to assign high scores to *all items* will undeservedly pass this check. Although we did not employ this approach, we anecdotally found that many human workers on the crowd-sourcing platform attempted this strategy, namely they assigned high scores to every conversation in our experiments. The approach also runs in contrast to our aim of the same individual who took part in a live conversation to also assess its quality, as applying this approach to our experiments relies on the use of pre-created gold-standard conversations.

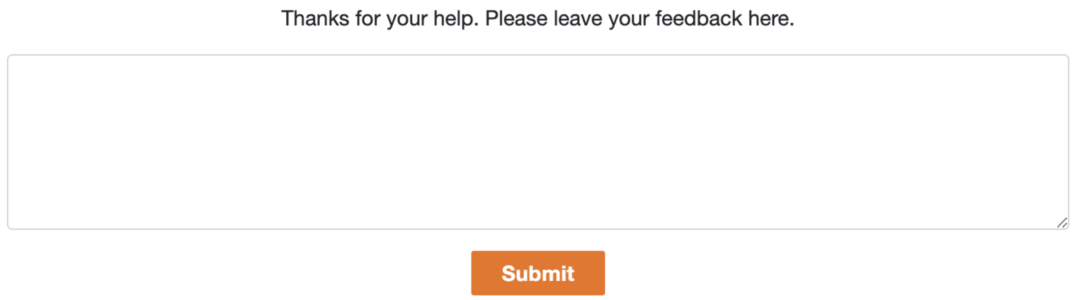


Figure 5.5: The interface shown to workers when all systems in a single HIT are completed, where they are welcome to leave their feedback in this page.

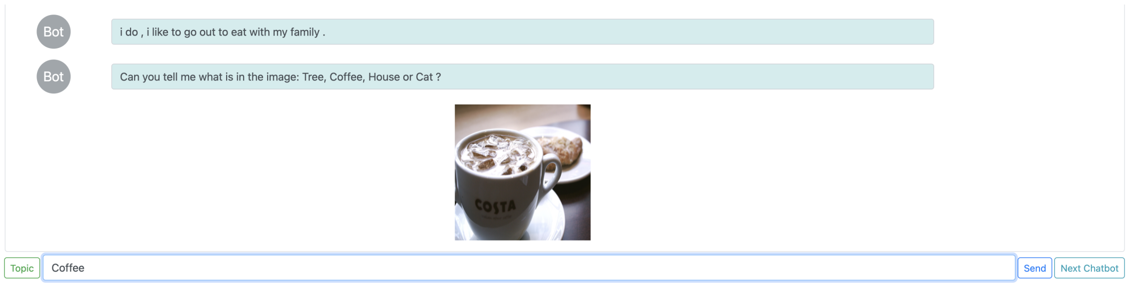


Figure 5.6: Workers are asked to take a sub-task of image recognition during the interaction.

5.2.1 A Failed Preliminary Attempt to Quality Control

Instead of evaluating pre-created conversations that are usually infeasible to achieve in practice, we initially decided to provide a compulsory sub-task during the interaction with a dialogue system. We employed image recognition as there are many available public datasets. By checking whether the image recognition sub-task can be correctly completed, we can approve or reject the work of current worker. We employed the Caption-Quality dataset (available on `https://github.com/google-research-datasets/Image-Caption-Quality-Dataset`), a dataset consisting of images paired with their captions (Levinboim et al., 2019). We extracted images whose captions belong to a set of certain topics which is expected to be clear and easy to distinguish, such as coffee, dogs, cats and so on.

Figure 5.6 shows an example of conducting such a quality control method, where an image describing a cup of coffee is presented and the worker is expected to select the right answer, coffee, from four different things. First, the worker converse with a chatbot as normal: worker types a sentence and receives the response. During the interaction, the sub-task will be inserted during a random turn, and the worker is asked to complete it. When the worker-provided answer is recorded, the conversation will continue by repeating the response before the sub-task, as shown in Figure 5.7.

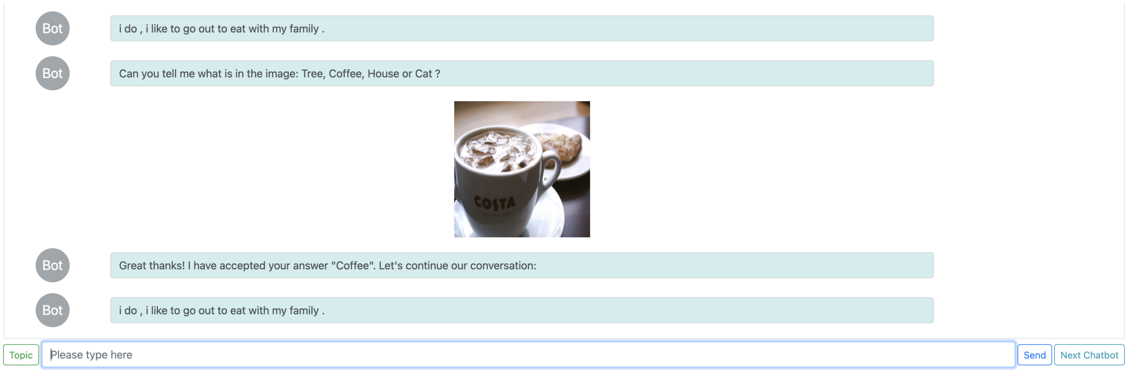


Figure 5.7: The chatbot will continue the conversation when the worker complete the sub-task.

However, such a quality control method fails to perform as expected. We find that the results of this preliminary experiment using this method showed the pass rate of workers reach 100%. We then manually checked the ratings of conversations, and found that workers may incorrectly score a conversation but still passed this method (for example, a worker may assign high scores to all conversations regardless of their real quality). It means that this method failed to filter out any unqualified worker. We think this is probably because the intention of this inserted sub-task is too obvious that a worker can easily pass it without taking it seriously. This sub-task based quality method was therefore abandoned because of its ineffectiveness.

5.2.2 Quality Controlling the Crowd

As alternative to the aforementioned sub-task based method, we decided to explore use of a statistic-based approach to quality-controlling human workers. In detail, each worker was asked to hold six conversations in a HIT, using a shuffled arrangement of five dialogue systems and a single quality control (qc) model. This was intended to collect a sufficiently rich score distribution from each individual worker who participated, where each HIT can collect 6 models $\times$ 7 rating criteria $=$ 42 ratings. This quality control approach is designed to achieve reliable quality control by deploying dialogue systems in live conversations that have *known distinct performance levels* instead of asking workers to assess the quality of pre-created known high quality conversations (gold-standard) or carrying out a sub-task (our previous failed method).

Within a HIT, the five systems, m, produce *some quality level of conversation* and the qc model, l, produces known *lower* quality dialogues (lower than the five systems). Meanwhile, the details of HITs, such as the information of the systems employed, are invisible to workers, meaning that the five dialogue systems and the qc model should be equally treated as "chatbots" by the workers according to provided instructions in Figure A.1 in Appendix A.

For a single worker who takes part in conversations with m and l, we then check how consistently the worker rated the conversations of l lower than m. This results in a quality control mechanism that does not ask workers to be consistent with other workers or to correctly rate gold standard dialogues, but only assesses worker consistency by how *consistently they distinguish between known distinct performance systems and only with respect to their own conversation ratings.*

From a practical standpoint, creating a low performance model, namely the qc model l in this case, is additionally far less challenging and costly than pre-creating a known set of dialogues that are known to have *high quality.* Generating low conversations can operate fully automatically by the qc model, where it produces outputs via generating *random responses* which are further degraded by the *meaning distortion* operation, which will be introduced in the next paragraph.

In detail, for *random response* degradation, low quality responses are generated by randomly sampling responses from conversations in the PersonaChat training set, which will be introduced in Section 5.3. Since the responses are randomly selected regardless of the worker inputs, the qc model is likely to be perceived as low quality since its responses have low relevance. To reduce the quality of conversations further, we apply *meaning distortion*: each response, r, is altered to distort its meaning by randomly selecting a sequence of words within that response and replacing it with a sequence of words sampled from another dialogue in the training set, with the length of the replaced word sequence being determined by the number of words in r. The following provides specific details for generating a random response:

- for $1 \leq n \leq 3$: r is 1 word;
- for $4 \leq n \leq 5$, r is 2 words;
- for $6 \leq n \leq 8$, r is 3 words;
- for $9 \leq n \leq 15$, r is 4 words;
- for $16 \leq n \leq 29$, r is 5 words;
- for $n \geq 30$, r is $\lfloor n/5 \rfloor$ words.

where the position of the word sequence to be replaced is additionally random with the exception of response of length $n \geq 3$, the sequence of replaced words does not include the response initial or final words.

Crowd-sourced workers converse with each chatbot - either a genuine dialogue system or the qc model - before rating conversation quality, and statistical significance tests are then applied to score distributions of workers for the ratings they attributed to ordinary dialogue systems, m, relative to the qc model having a low quality, l. The resulting p-value is then employed as a means of rating worker consistency. Any worker with $p >= 0.05$ shows no significant difference between low and ordinary model quality and is filtered out. Figure 5.8 provides two examples of conversations from a human rater's interaction with (1) a genuine dialogue system (left), and (2) the qc model (right) which only returns with a degraded random response, where we expect a qualified worker should be able to score *normal* conversation significantly higher than a *degraded* one.

Normal		Degraded	
User:	Hello, how are you?	User:	How are you?
Bot:	i am good , a little tired but good !	Bot:	I do not have a job. I have broken have a job.
User:	I am tired as well, but good too	User:	Sorry about that?
Bot:	i am tired too , but from being out all day .	Bot:	Thanks. I the love watching old.
User:	What do you like doing?	User:	Tell me about it?
Bot:	i enjoy movies , reading books , and even just doing nothing	Bot:	that is cool, do you i supposed to I you travel?
User:	I love reading books	User:	I do not like travelling
Bot:	i enjoy reading too ! what is your favorite book ?	Bot:	I teach meet you, sorry elementary school.
User:	I love the novel "voice of the heart"	User:	Okay, do you like to swim?
Bot:	i have been wanting to read that , i love reading too	Bot:	do you live sports are u skin am in a house.
User:	You should read too	User:	I love sports, especially soccer
Bot:	i do love to read when i have time	Bot:	I like but i should go, books
User:	What book was the last book you read?	User:	How often do you read?
Bot:	the notebook , a reread lol	Bot:	i are your think u r i am a pilot , so i do not have much time .
User:	I haven't read that	User:	I envy you, I would love to be a pilot
Bot:	i read it yesterday . what do you do for fun ?	Bot:	I i am 4 felicia!
User:	I love going hiking	User:	I am Ben
Bot:	i have not gone hiking a while	Bot:	Oh, how nice. do you do for you have?
User:	I think you should	User:	I do not understand
Bot:	i should	Bot:	I is your favorite don't I am sure you will find out soon.

Figure 5.8: A typical human-system conversation (left) and a conversation between a human and the qc model (right), where *random response* and *meaning distortion* techniques have been applied to degenerate model responses.

5.3 Dataset and Systems

The dialogue systems involved in our experiments are trained on the open-domain conversational dataset, PersonaChat (S. Zhang et al., 2018), which was also utilized in the ConvAI2 competition (Dinan, Logacheva, et al., 2019). PersonaChat was proposed for solving known problems of chit-chat models: lacking specificity, consistent personality and captivation. It is a crowd-sourced dataset of which each conversation is paired with a given profile, called a "persona", each comprised of no less than five sentences. The data collection follows three stages: (1) a set of persona were firstly collected whose sentences are generated from human workers, and (2) another group of workers were asked to rewrite these personas to increase the difficulty of the task, and finally (3) two paired workers were asked to chat with each other according to the revised persona assigned with them, resulting in a dataset consisting of 1,155 valid personae, and 10,907 dialogues containing 162,064 utterances.

The test dialogue systems in our experiments are from models as follows:

- Poly-Encoder Transformer: a model having an improved architecture of transformer (Vaswani et al., 2017) which learns self-attention features in the global, instead of token, level. (Humeau et al., 2019).

- Bi-Encoder Transformer: a model containing two transformer-based encoders, one is to independently encode knowledge and dialogue context while the other is to re-encode the combined information of knowledge and dialogue context after knowledge selection (Dinan, Roller, et al., 2018).

- Key-Value Memory Networks: a model which enables to encode prior knowledge, such as dialogue history in this experiment, and store in key-value memories (Miller, Fisch, et al., 2016).

- Sequence to Sequence: a model that follows the vanilla seq2seq architecture (Sutskever, Vinyals, and Q. V. Le, 2014).

- LSTM-based model: a language model using the LSTM neural network (Hochreiter and Schmidhuber, 1997).

Within the evaluation setting of ConvAI2, each dialogue model is given a persona consisting of approximately five textual statements to emulate a personality. For each of the five models, we additionally include a version that is assigned no persona, to increase the diversity of our experiment and to provide an interesting comparison, resulting in 10 competing dialogue systems. These systems are from the dialogue research software platform ParlAI (Miller, Feng, et al., 2017), and have been fine-tuned according to the setting of ConvAI2 (available on `https://parl.ai/docs/zoo.html#convai2-models`).

5.4 Experiment Results

In section, we report the results of our experiments, and provide analysis based on the collected data, including information about HITs and workers, the human scores at the system-level, the consistency of our method, significance tests, human rater consistency and so on.

5.4.1 Meta-Evaluation

HITs were posted on the crowd-sourcing AMT platform, similar to our previous experiments, such as evaluation of MRC and QG in Chapter 3 and 4. Firstly, and in order to evaluate the *open-domain* models in as realistic a setting as possible, we allow workers to *choose the topic* of conversation and input their chosen topic in a text field. The open nature of conversations should be noted however as something that influences the difficulty of producing consistent results in our self-replication experiment. The fact that we allow human assessors to freely choose the topic of conversation means that differences in ratings could result from legitimate differences in performance when different topics are chosen by human assessors. We nonetheless test our evaluation allowing the user to choose the topic as this is part of our core aim for developing evaluation of dialogue truly in the open domain.

Besides choosing a topic, we additionally asked workers to input their opinion of the topic they chose to discuss with the systems, categorizing the topic as either *liked, ambivalent about it, or disliked.* For example, if the topic they chose to discuss was *dogs*, we were curious to know if this was motivated by the fact that the worker liked or disliked dogs or indeed that they had chosen to discuss something they had no particular feeling about. Figure 5.9 provides the interface when choosing a topic.

In contrast to free topic choice, we further investigate the performance of systems in a slightly easier setting where the topic under discussion is known to the dialogue system, by selecting a sentence from its persona statement, which we refer to as an *ice-breaker topic statement.* An ice-breaker topic statement is then provided to

Figure 5.9: The popup window where a worker can freely type a topic and record the opinion of this topic, before starting the conversation.

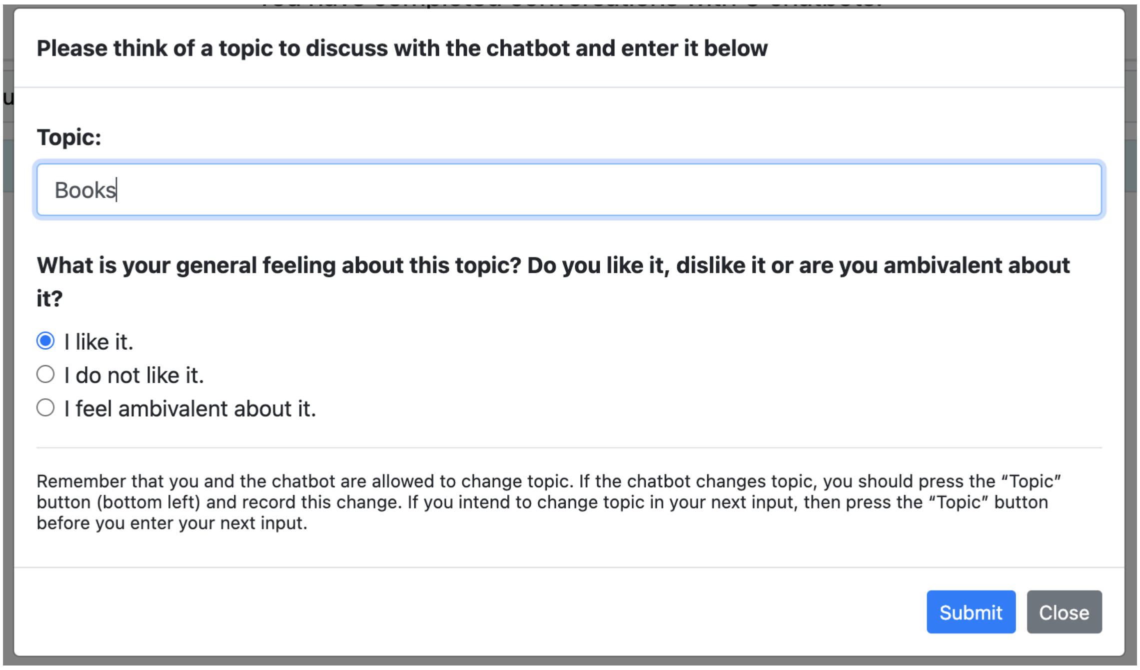


Table 5.2: Proportions of freely-chosen (free topic run 1&2) and preassigned (Ice-breaker) topics that are reported by workers as liking (Like), being ambivalent towards (Ambivalent) or disliking (Dislike), including passed and failed workers.

Opinion	Free Topic (run 1)		Free Topic (run 2)		Ice-breaker	
	Passed	Failed	Passed	Failed	Passed	Failed
Like	83.88%	88.58%	86.35%	93.75%	61.52%	71.43%
Ambivalent	7.44%	3.75%	6.22%	2.34%	18.87%	13.29%
Dislike	8.68%	7.68%	7.43%	3.91%	19.61%	15.28%

human assessors at the beginning of each conversation, and the assessor is instructed to chat with the system about this topic. The process and user interfaces of ice-breaker evaluation resemble those of free topic, except the stage when determining a topic before conversation. Figure A.2 in Appendix A provides relevant user interface in ice-breaker evaluation statement.

Table 5.2 shows subsequent proportions of workers' opinions on topics, including worker-typed topics and given topics. As results of both free runs show that, perhaps unsurprisingly, the vast majority of workers chose to discuss something they *liked*. For instance, in the first run, nearly 84% workers who passed our quality control

Figure 5.10: Word cloud of topics chosen by human workers in the first run of free topic evaluation.

method liked the topic they chose. Nonetheless, 7% of them were *ambivalent* about the topic they chose and 9% chose a topic they reported as *disliking*. In addition, compared with Free Topic, less workers reported they liked the topics in Ice-breaker together with the increased proportions of ambivalent and disliking.

Figure 5.10 shows the topics chosen by workers in the first data collection run in the free topic evaluation statement. We find that the majority of topics are unsurprisingly common and familiar, such as food and pets. We think these collected topics together with human opinions are useful for the development of open-domain dialogue systems. For example, a dialogue system can start with a topic which is generally liked by human. Although the results of Table 5.2 and Figure 5.10 are not directly related to our research questions, we believe they can be used for future research.

5.4.2 HITs and Workers

Table 5.3 provides the information of human workers, assessed dialogues and deployed HITs in this experiments. There are different topic-related statements: free topic where workers are permitted to freely choose a topic before a conversation including the initial run (run 1) and a repeated experiment (run 2), and a set-up where the topic of a conversation is predetermined by extracting a sentence from the persona of the current system.

Table 5.3: Statistical information about workers, dialogues and HITs in our experiments, where workers freely chose the topic (free run 1); precisely the same experiment set-up was repeated (free run 2); and where the topic was prescribed via selecting directly from the persona of the system (ice-breaker).

(a) Numbers of workers who took part in human evaluation of systems and the total number of assessed dialogues before (Total) and after (Passed) quality control, together with according pass rates.

Topic	Worker			Dialogue		
Statement	Total	Passed	Pass rate	Total	Passed	Pass rate
Free Run 1	249	173	69.48%	1525	1075	70.49%
Free Run 2	248	139	56.05%	1480	835	56.42%
Ice-breaker	248	171	68.95%	1450	1030	71.03%

(b) Average time taken per dialogue in minutes (min) and average number of HITs per worker took, before and after quality control.

Topic	Ave. Duration (min)			Ave. Taken HIT		
Statement	Passed	Failed	Total	Passed	Failed	Total
Free Run 1	6.53	7.04	6.68	1.24	1.18	1.22
Free Run 2	6.87	7.58	7.18	1.21	1.18	1.19
Ice-breaker	6.60	6.70	6.63	1.20	1.09	1.17

Table 5.3a shows the number of workers who participated in three statements, numbers of dialogues assessed in total before and after quality controlling, and the proportions of workers and dialogues that passed quality checks. These assessed dialogues only count in conversations between humans and the 10 genuine systems, excluding the qc model. The results of free run 1 amounts to 1,525 dialogues $\times$ 7 criteria = 10,675 human ratings, while Table 5.3a indicates equivalent statistics with respect to free run 2 in which a total of 1,480 dialogues $\times$ 7 ratings = 10,360 human ratings were collected in total. In addition, the ice-breaker statement results in collecting 10,150 human ratings in total.

Table 5.3b reports how long the evaluation of a dialogue averagely takes in minutes and how many HITs a worker takes on average, where Passed and Total is the same as Table 5.3a and Failed represents those which are filtered out by the quality control method. In general, evaluating a dialogue after the interaction normally costs 6-7 minutes, in all three experiments, while a failed worker averagely spends more time than a worker who passed the quality control. Meanwhile, a worker averagely takes approximately 1.2 HIT, while the number of passed workers is slightly higher than that of failed workers.

Implementation of Quality Control

In general, we use statistical significance tests to control the quality of crowd-sourcing workers, as described in Section 5.2.2. Given two sets M and L respectively representing the rating distributions of ordinary dialogue systems m and the qc model l from all HITs taken by a human worker, we apply the non-parametric one-sided Wilcoxon rank sum test between L and M, with the alternative hypothesis that values in L are more likely to be less than those in M. The consequent p-value then acts as the indication of whether a worker can consistently score l lower than m during evaluation. Any worker with $p \geq \alpha$ is therefore rejected since no significant difference occurs, and α is 0.05 in this case.

However, M and L do not contain the scores under the negative evaluation criteria: *robotic* and *repetitive*, because of two main reasons. First, we deployed a small experiment with a few HITs to preliminarily investigate our quality control method. We found that workers normally tend to give a terrible score when evaluating how *robotic* a dialogue is, regardless of the system. This possibly results from the gap in existence between performances of existing dialogue systems and humans which makes workers feel *all systems are like a robot*. Thus, we think robotic is impractical for distinguishing qc model from genuine systems. Furthermore, as the qc model only respond *randomly*, it is deservedly non-repetitive since it never returns a duplicated sentence. Hence, *repetitive* scores are eliminated from quality control.

Cost of the Experiment

Each worker was paid 0.99 USD for a HIT which consists of 5 valid human-system dialogues with 35 ratings, excluding the qc model. For example, free topic run 1 costs approximately 220 USD resulting in 1,075 dialogues and 7,525 ratings from 10 genuine systems, meaning that each system with collected ≈107 dialogues (≈750 ratings) costs no more than 22 USD. It is notable that the quality control method we applied for removing unreliable data is not the only criterion for deciding whether a human worker can get paid, namely a worker whose data is discarded can still receive the payment. The entire experiments consisting of three topic-related statements cost no more than 600 USD. Also, free topic run 2, namely the self-replication experiment, is primarily utilized to test the repeatability of this newly proposed method, which can be omitted to reduce the cost in future researches.

5.4.3 Human Scores of Dialogue Systems

This section shows the calculation of average standardized scores of dialogue systems in this experiment, and results of *free topic* and *ice-breaker* (see Section 5.4.1) experimental runs.

Calculating System-Level Scores

Scores are collected from workers who rate models on a continuous rating scale ranging from 0 to 100, and we refer to these initial scores as *raw* scores. In particular, scores for negative attributes, namely *robotic* and *repetitive* in this experiment, are firstly reversed by 100 subtracting the original rating. This results in revised scores possessing the same tendency according with positive attributes for ease of further computation and comparison, and these revised scores will act as the *raw* scores of negative attribute. A distribution of scores is extracted for each worker, including rating scores of qc model, and a *raw* score is standardized into a z score according to each worker's mean (μ) and standard deviation (σ) by $z = (raw - \mu)/\sigma$ in order to eliminate any differences in worker scoring strategy Average standardized z scores for each criteria are firstly calculated, and an overall z score is then calculated as the average of all measurement criteria.

Evaluating with Freely Chosen Topics

Table 5.4: Average standardized scores for dialogue systems in initial data collection run; workers were free to choose the topic of conversation (Free run 1); where A=Bi-Encoder Transformer, B=Poly-Encoder Transformer, C=Key-Value Memory Network, D=Sequence to Sequence, and E=LSTM-based Model; a system with p means it holds a persona; scores for *robotic* and *repetitive* have been reversed; n is number of ratings; systems are ordered by overall score; a underlined score means the highest score of that evaluation criterion.

	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
A	798	$\underline{0.534}$	$\underline{0.564}$	$\underline{0.602}$	$\underline{0.711}$	0.863	0.964	$\underline{-0.038}$	0.069
B	798	0.419	0.474	0.481	0.614	$\underline{0.875}$	$\underline{0.994}$	−0.431	−0.075
A_p	707	0.318	0.399	0.372	0.443	0.821	0.404	−0.330	0.116
C	791	0.262	0.491	0.379	0.028	0.636	−0.066	−0.316	$\underline{0.680}$
C_p	714	0.189	0.409	0.373	0.159	0.672	−0.114	−0.521	0.349
B_p	707	0.173	0.230	0.197	0.369	0.673	0.320	−0.395	−0.187
D	707	−0.087	−0.190	−0.208	0.166	0.311	0.401	−0.637	−0.449
D_p	798	−0.201	−0.308	−0.234	0.092	0.312	0.025	−0.625	−0.669
E_p	763	−0.217	−0.181	−0.201	−0.196	0.380	−0.455	−0.605	−0.264
E	742	−0.243	−0.165	−0.160	−0.142	0.329	−0.407	−0.745	−0.411

We firstly compute the scores of competing systems based on collected data to investigate system performances when topics of conversations are provided by humans. Table 5.4 shows the average standardized (z) scores of different rating criteria at the system level, resulting from the initial data collection run of our human evaluation experiments on AMT (free topic run 1), where systems are ordered by highest overall score. We can find that system A, the Bi-Encoder Transformer based system without persona, receives the highest overall score, as well as *interesting*, *fun*, *consistent* and *robotic* criteria. In addition, system B outperforms others on *fluent* and *topic*, while system C has the best performance on *repetitive*.

Evaluating with Prescribed Topics

In addition to the free topic statement, we subsequently deploy the ice-breaker statement experiment where a topic which is extracted from the persona is given to the human assessor before the interaction with a chatbot.

The numbers of workers who participated in the ice-breaker experiment run are provided in Table 5.3a, while a breakdown of results for each system and overall average scores are shown in Table 5.5 as well as the correlation between scores for systems when a topic is freely chosen. System A again, is the best performing system according to the overall score, together with the highest *consistent*, *fluent* and *topic* scores. System A_p has the best performances of *fun* and *robotic*, while System C outperforms others on *interesting* and *repetitive*. Compared the initial results of freely chosen topics (Table 5.4), the best chatbot, system A, achieves highest *fluent* and *topic* scores, nonetheless is deemed worse in *interesting*, *fun* and *robotic* when evaluated in ice-breaker conversations.

Except standardized system scores, raw average scores for systems in the ice-breaker run are provided in Table A.4 in Appendix A. Interestingly, in terms of absolute differences in raw scores, system A in the ice-breaker evaluation achieves higher fluency, consistency but is deemed more robotic, compared with those in free topic run 1 which are available in Table A.2 in Appendix A.

Relatively speaking, when it comes to system rankings, no meaningful difference in relative performance is observed when systems are tested in a scenario where the worker chooses a topic and when one is prescribed with an ice-breaker statement, as can be seen from the strong correlation (r) between scores at system-level in free topic run 1 and ice-breaker evaluation as shown in Table 5.5.

5.4.4 System-level Consistency

In order to test whether this proposed method is reliable, we additionally compute the standardized system scores in the self-replication experiment (introduced in Section 5.1.1), namely the free topic run 2, where the results are reported in Table

Table 5.5: Average standardized scores for dialogue systems in the experiment in which workers were given the topic of conversation (ice-breaker); where the detailed system names are the same as those in Table 5.4; a system with p means it holds a persona, and the ice-breaker statement is subsequently unknown to systems without p; scores for *robotic* and *repetitive* have been reversed; n is number of ratings; systems follow the order in Table 5.4; a underlined score means the highest score of that evaluation criterion; r is the correlation between current assessment criterion and that in the first run of free topic (Table 5.4).

	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
A	721	0.552	0.565	0.527	0.873	1.018	1.011	−0.287	0.156
B	721	0.376	0.379	0.340	0.634	0.769	0.820	−0.221	−0.087
A_p	742	0.422	0.589	0.560	0.518	0.718	0.527	0.009	0.034
C	784	0.322	0.615	0.537	0.190	0.631	0.061	−0.344	0.565
C_p	700	0.222	0.402	0.337	0.089	0.654	−0.068	−0.376	0.514
B_p	658	0.273	0.406	0.340	0.414	0.633	0.423	−0.369	0.063
D	728	−0.139	−0.277	−0.204	0.123	0.349	0.295	−0.638	−0.620
D_p	721	−0.267	−0.426	−0.402	−0.011	0.234	0.000	−0.628	−0.636
E_p	714	−0.198	−0.172	−0.203	−0.054	0.316	−0.343	−0.533	−0.396
E	721	−0.240	−0.125	−0.161	−0.196	0.318	−0.393	−0.631	−0.489
r	–	0.984	0.967	0.944	0.958	0.951	0.981	0.715	0.950

A.1 in Appendix A. Meanwhile, Table 5.6 shows the consistency of the evaluation between each experimental run via the three categories of correlation of scores for each measurement criteria as well as consistency overall, according to results from Table 5.4 and Table A.1. Across the board, consistency of Pearson correlation (r) for example, is very high, exceeding a correlation of $r \geq 0.9$ in almost all cases with the exception of *robotic* which nonetheless achieved a correlation of around $r = 0.7$. With regards to ρ and τ, similar results to r appear. As can be observed from Table 5.6, the correlation reached in terms of overall scores for systems is $r = 0.969$, which is very close to a perfect correlation, showing extremely high levels of reliability for the evaluation, evidence that the approach overcomes substantial challenges with respect to evaluating open-domain dialogue systems, where assessors are legitimately free to choose distinct topics of conversation.

Moreover, the raw scores at system-level in the two runs of free topic experiments

Table 5.6: Correlations between the system scores from the initial and second runs of free-topic experiments (free topic run 1&2), including Pearson (r), Spearman (ρ) and Kendall's tau (τ) correlation coefficients.

	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
r	0.969	0.952	0.927	0.899	0.960	0.951	0.646	0.936
ρ	0.903	0.802	0.855	0.806	0.939	0.915	0.673	0.939
τ	0.733	0.674	0.733	0.600	0.822	0.778	0.467	0.822

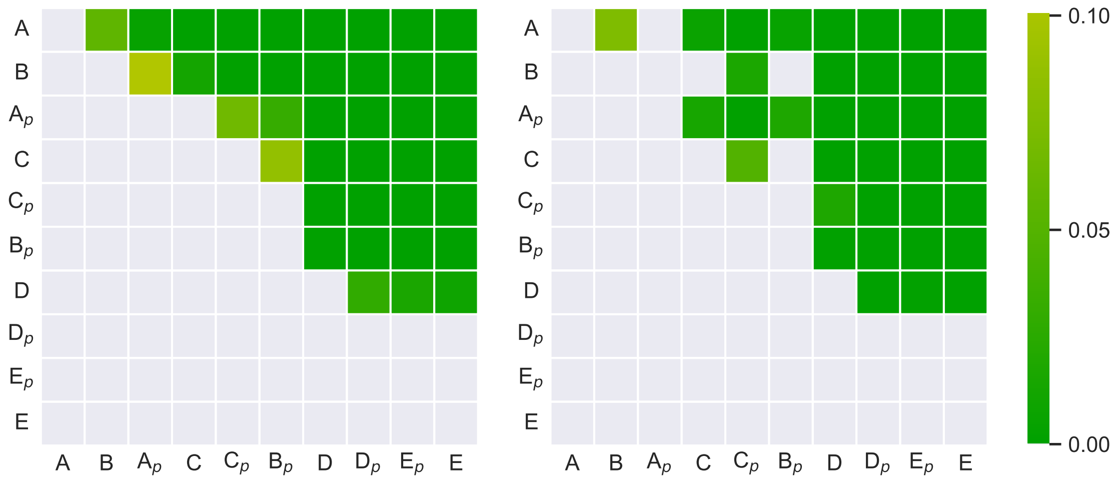
Table 5.7: Correlation of assessed criteria with others in free topic run 1; correlations in the upper right half correspond to Pearson (r) while lower left are Spearman correlations (ρ).

	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
Overall	-	0.959	0.976	0.861	0.966	0.796	0.916	0.674
Interesting	0.927	-	0.992	0.691	0.949	0.599	0.875	0.840
Fun	0.903	0.988	-	0.753	0.961	0.660	0.889	0.783
Consistent	0.842	0.673	0.636	-	0.811	0.969	0.770	0.210
Fluent	0.879	0.939	0.915	0.648	-	0.724	0.857	0.667
Topic	0.745	0.552	0.503	0.915	0.503	-	0.676	0.122
Robotic	0.867	0.830	0.782	0.648	0.867	0.491	-	0.642
Repetitive	0.673	0.770	0.782	0.261	0.770	0.055	0.758	-

are available in Table A.2 and Table A.3 in Appendix A, where the latter also reports the correlation of raw scores between two runs, again indicating a high reliability according to $r = 0.595$ on overall *raw* score.

Consistency of Evaluation Criteria

We examine how these evaluation criteria correlate with each other. Results are reported in Table 5.7 based on the data collected from the first run of free topic. Perhaps expectably, *fun* and *interesting* receive extremely high correlation at $r = 0.992$ and $\rho = 0.988$ since they somewhat have overlap of meaning. Meanwhile, we can observe *repetitive* correlates weakly with both *consistent* and *topic*, implying that system performances can vary when emphasizing different evaluation criterion.



(a) Run 1 (b) Run 2

Figure 5.11: Results of significance tests on overall system scores for two runs of free topic, where a colored cell means that the system in the row outperforms that in the column due to the test; A=Bi-Encoder Transformer, B=Poly-Encoder Transformer, C=Key-Value Memory Network, D=Sequence to Sequence, and E=LSTM-based Model; systems without p contain no persona; system order follows Table 5.4.

5.4.5 Significance Test

In any empirical evaluation, statistical significance tests should be applied to take into account the fact that small differences in scores between systems can occur simply by chance. We employ the standard statistical approach, Wilcoxon rank sum test, on scores of each pair of competing systems. Such a non-parametric test is applied because there is no guarantee of the normally distributed scores.

Figure 5.11 shows the results of significance tests on the rating distributions of pairwise dialogue systems based on the standardized overall scores from two distinct data collection runs of free topic experiment, where the definitions and order of systems accords with Table 5.4. A coloured cell in this figure indicates that the system in the row can significantly outperform it in the column at $p < 0.1$ resulting from the significance test. Results shows a very high proportion of identical conclusions, 84%, drawn from the results of both runs of free topic as respectively shown Figure 5.11a and Figure 5.11b at $p < 0.1$. The proportion can remain 84% even when the condition changes to $p < 0.05$. The results of significance tests in the ice-breaker run are additionally provided in Figure A.3 in Appendix A.

Persona Contribution to System Performance

The personas in the PersonaChat dataset are somewhat arbitrary statements about a hypothetical person. This is still a lack of relevant investigation into whether such personas in fact enhance the quality of dialogues to any meaningful degree, the influence of persona on system performance is therefore worthy of examining. Since we have verified the reliability of the human evaluation at system-level, we take a closer look at the results of both free topic and ice-breaker experiments, and further examine whether the employment of a persona influences the system performance.

Results in Table 5.4 show that in general a system in the free topic experiment is rated more favorably by human assessors when they carry out dialogues *without* a persona, excluding system E which achieves a lower score than E_p. Nonetheless, results of significance test in Figure 5.11a show there exists no significant difference between system E with and without persona, implying the higher score of E_p may occur by chance. Similar observation can be made for ice-breaker experiment results shown in Table 5.5 and Figure A.3 in Appendix A. We can conclude that, although it seems counterintuitive, a system assigned no persona generally has a better performance when conversing with a human. We think this may be because the persona assigned to a system somehow becomes the noise when chatting with a real person, while a system without persona can focus more on the current conversation, resulting in a conversation having a higher quality.

5.4.6 Human Assessor Consistency

Although the overall aim of our evaluation is to produce reliable results at the system level, which we previously tested in Section 5.4.4 by comparison of results in self-replication experiment, we additionally examine ratings of workers at the level of individual dialogue ratings. A Pearson correlation coefficient is applied since standard agreement measures such as the Kappa coefficient are not applicable according to aforementioned limitations in Section 3.4.7.

The distribution of Pearson correlation (r) for pairs of workers who assessed the

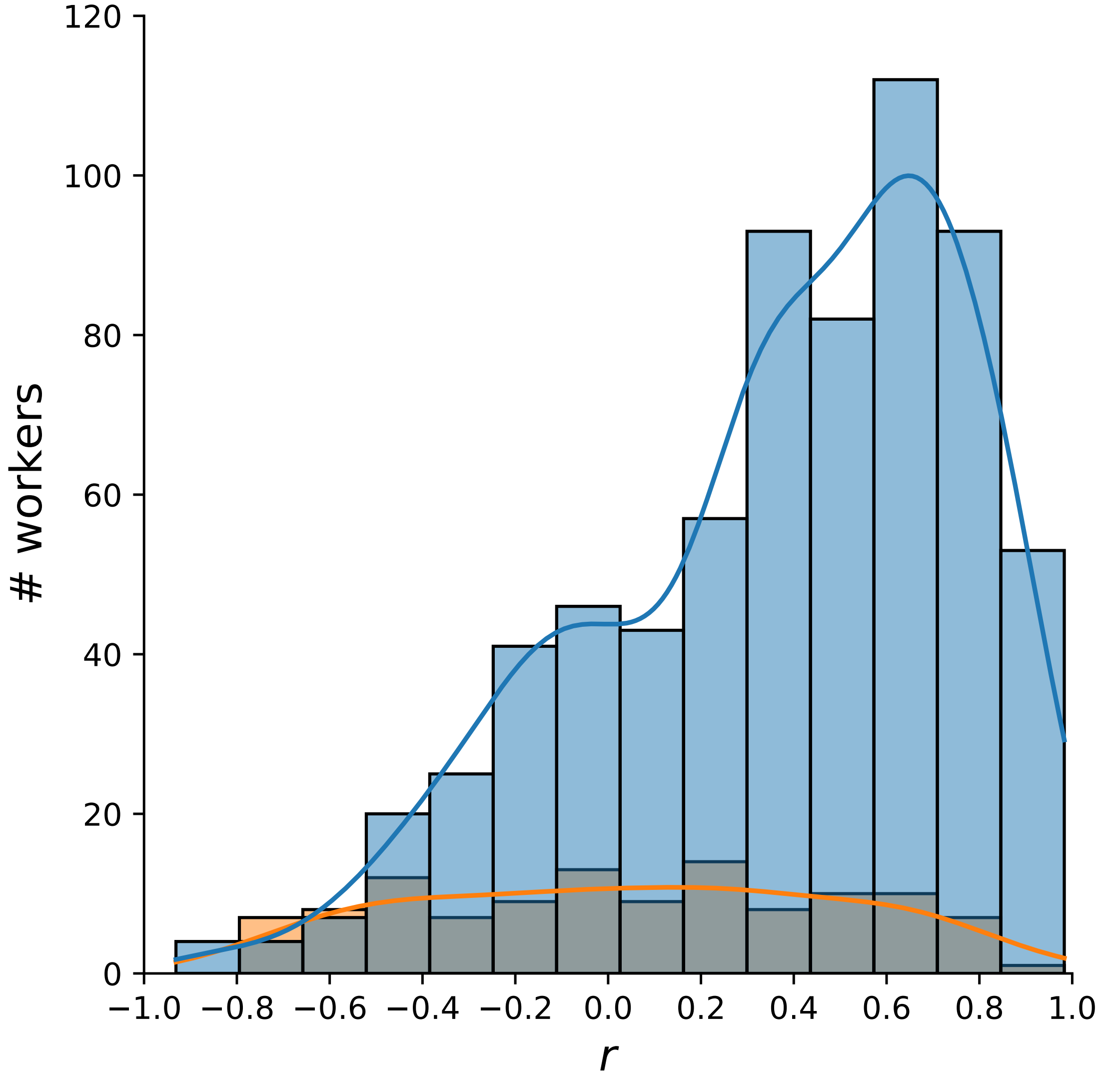


Figure 5.12: Distribution of agreement between pairs of human assessors as measured by the Pearson correlation (r) of ratings provided by workers who passed (blue) and failed (orange) quality control.

same system is depicted in Figure 5.12, including results of passed (blue) and failed (orange) workers. The total number of pairs of workers who completed the same system and passed quality control is 680, with a total of 115 pairs who failed. We can observe that the likelihood of agreement between pairs who failed quality control is close to random as the distribution is approaching uniformity across almost the range of possible coefficients. In contrast, for pairs of passed workers, the peak of agreement r is in the range of 0.6 and 0.7, indicating a high degree of agreement in general between such annotator pairs.

Some of the observed disagreements are likely to be the result of legitimate differences between scores of two workers who chose distinct topics to discuss with the same model however, an unavoidable source of inconsistency when testing models with respect to the open domain. Furthermore, remaining disagreements at the level of individual ratings are probably not problematic at the level of overall scores in relation to aggregation of ratings collected on a continuous rating scale. Technically speaking, the most meaningful reliability measures for continuous ratings scales test consistency of *aggregate* (system-level) results because although a high level of random error is expected in individual continuous rating scale scores, when aggregates are calculated for large samples of ratings, positive and negative errors that are truly random effectively cancels themselves out, and does not negatively impact consistency. In other words, the employed rating scale does not rely on consistency at the level of individual ratings. The consistency of individual raters is nevertheless examined since it is the standard approach in practice (Qiang Liu, Ihler, and Steyvers, 2013). However, it is notable that results provided in this part are not crucial when testing reliability for an evaluation carried out via a continuous rating scale, and consistency in overall system-level results is more important.

5.4.7 Comparison with Automatic Evaluation Metrics

We examine the performance of the main open-domain dialogue evaluation metrics in terms of their correlation with human judgements from results of our human evaluation experiments. Despite the recurring criticism against automatic metrics due to their poor correlation with human judgements (Sai, Mohankumar, and Khapra, 2022), it is nevertheless worth reporting corresponding performance, since applying these automatic metrics is still a common approach in the development of dialogue systems. Two categories of automatic evaluation metrics are considered in this part of our investigation: word-overlap-based and reference-free.

Table 5.8: Correlation between metric scores and the average standardized overall scores in free topic run 1 at system-level, including Pearson (r), Spearman (ρ) and Kendall's tau (τ), where metrics are ordered by r.

	METEOR	BLEU-1	ROGUE-L	GLEU	BLEU-4
r	-0.321	-0.707	-0.799	-0.816	-0.883
ρ	-0.328	-0.705	-0.705	-0.681	-0.766
τ	-0.225	-0.494	-0.494	-0.494	-0.584

Word-overlap-based Metrics

In this experiment, we report the scores of systems using four prevailing word-overlap-based metrics: METOER, BLEU which includes BLEU-4 and BLEU-1, ROGUE-L and GLEU, resulting in five metric scores for each system. Such metrics, as previously described in Section 2.3, rely on the overlap between a system output and a precreated reference to assess the system performance. However, the conversations from our experiments have no *ground-truth reference*, meaning that these metrics are infeasible to evaluate the conversations on our collected data. Instead, we compute scores on the testset of PersonaChat, which is also used in ConVAI2 competition.

Table 5.8 reports the results of different correlation coefficients between system scores from commonly used word-overlap-based evaluation metrics and the overall average z scores in the first run of free topic experiment, where the metric scores are additionally provided in Table A.5 in Appendix A. Unfortunately, instead of achieving a strong correlation, the results indicate that all these metrics even failed to correlate positively with human judgement, as shown in Table 5.8. This negative correlation implies that such evaluation approaches may rank system in reverse to human assessment, confirming that the common practice of applying human evaluation on systems after ranking by automatic metrics in current conversational competitions like ConvAI2 will produce invalid system rankings.

Return to **RQ 2** regarding open-domain dialogue:

- **RQ 2**: *What are the limitations and disadvantages of the direct application of evaluation metrics from MT and other domains to entirely distinct tasks for* *system development in open-domain dialogue?*

we can address it as: applying metrics from other domains to dialogue evaluation suffers from the negative correlation with human judgement, which can result in improperly filtering out high quality systems according to human judgement since such metrics are likely to produce reversed system rankings.

Reference-free Metrics

In addition to traditional word-overlap-based metrics which are borrowed from other NLP tasks, two reference-free automatic metrics, USR and FED, are employed. They are specifically proposed for dialgoue evaluation, of which the details are available in Section 2.3. Both metrics utilize pretrained language models and require no reference, since they have the ability of assessing a dialogue using the conversation content only. Therefore, we compute the score of unreferenced evaluation metrics on the data we collect from the initial run of free topic.

In this part, we use pretrained medium and large DialoGPT (Y. Zhang et al., 2020) as the scorers of FED. Since FED requires a set of predefined positive and negative responses to compute relevant likelihood of each evaluation attribute, we also provide the full list of those responses in Table A.6 in Appendix A. Utterances of some rating attributes, such as *interesting*, *consistent*, *fluent*, *topic*, and *repetitive* in this case, are available in the official implement of FED (see `https://github.com/Shikib/fed`) and we use them off-the-shelf. For other criteria including *fun* and *robotic*, we adapt their positive and negative utterances correspondingly.

As shown in Table 5.9, results of reference-free metrics correspond better than word-overlap-based metrics and are more encouraging in terms of reference-free metrics. FED has the ability of distinguishing *repetitive* models, but for other criteria, it correlates weakly or even negatively with human. Meanwhile, despite USR only correlating marginally with human in terms of *consistency* and *topic loyalty*, USR-DR(f) correlates closest to human among the three sub-metrics, while it performs best on evaluating *consistency* and *topic loyalty*. The system scores of FED and USR

Table 5.9: Pearson correlation (r) between reference-free metric scores and human evaluation (free topic run 1), where FED_m and FED_l respectively use medium and large DialoGPT, USR is the overall USR score computed according to the three sub-metrics; USR_m=USR-MLM, USR_c=USR-DR(c) and USR_f=USR-DR(f).

Criterion	FED_m	FED_l	USR	USR_m	USR_c	USR_f
Overall	0.590	0.530	−0.230	−0.419	0.046	0.205
Interesting	0.028	−0.042	−0.451	−0.235	−0.238	−0.081
Fun	−0.339	0.115	−0.378	−0.319	−0.131	0.032
Consistent	0.236	0.227	0.214	−0.620	0.518	0.652
Fluent	−0.138	−0.054	−0.227	−0.374	0.028	0.151
Robotic	0.528	0.461	−0.070	−0.290	0.106	0.191
Repetitive	0.841	0.752	−0.713	0.182	−0.690	−0.568
Topic	0.046	0.004	0.222	−0.754	0.606	0.746

computed on data from first run of free topic are respectively provided in Table A.7 and Table A.8 in Appendix A.

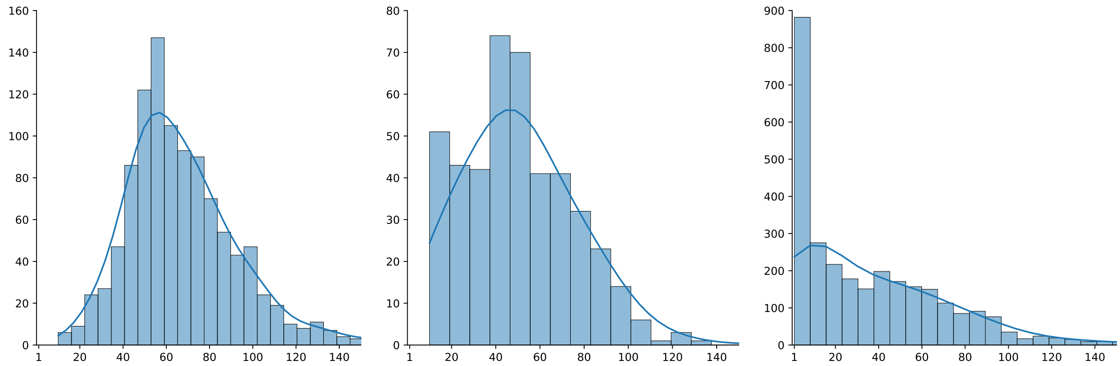
According to the performance of both word-overlap-based and reference-free metrics, we can therefore address how accurately existing automatic metrics can measure dialogue system performance, namely **RQ 1** regarding dialogue systems:

- **RQ 1**: *How accurately do existing automatic metrics measure open-domain dialogue system performance?*

We find that, current metrics in dialogue evaluation generally fail to achieve a high degree of evaluation accuracy due to their poor or even stark low correlation with human judgement.

5.4.8 Comparison with ConvAI2 Live Evaluation

Since data collected from other live evaluation was previously deemed as useless due to its negative attributes such as senselessness and offensiveness (Dinan, Logacheva, et al., 2019), we provide comparisons between the data collected from our first run of free topic experiment and that from the ConvAI2 live evaluation. Figure 5.13 represents the distribution of words in conversations from our experiment and ConvAI2, where the abscissa means the number of words and ordinate is the number of conversations. We observe that the distribution in ConvAI2 live evaluation is



(a) Passed (b) Failed (c) ConvAI2

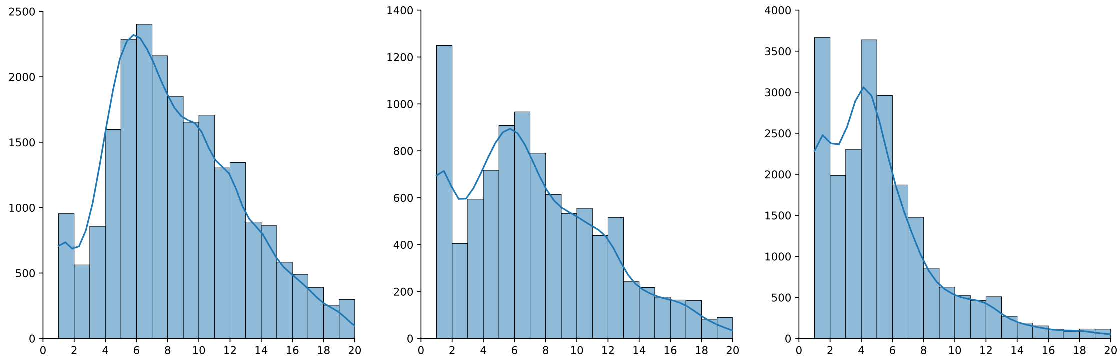
Figure 5.13: Words per conversation from workers who passed (5.13a) and failed the quality control (5.13b) in our human evaluation (free run 1); as well as workers from ConvAI2 live evaluation (5.13c).

rather unbalanced in that vast majority of the conversations in it consist of less than 10 words. We believe this occurs because no minimal number of inputs is required, which makes raters prone to end a dialogue in advance of generating a full conversation. In addition, Figure 5.14 shows the distribution of words in utterances that are input by workers in free topic run 1 and those in ConvAI2. It can be found that workers who failed our quality control and who participate in ConvAI2 live evaluation tend to respond with only one word. However, such responses should generally only occur as greetings at the beginning or the end of a conversation, and a conversation consists of too many are likely to be less meaningful. In contrast, passed workers will reply longer utterance when conversing with a system. Moreover, Figure A.4 and Figure A.5 in Appendix A provide the relative information at character level about conversations and user input utterances, as well as median number of words and characters for conversations and inputs in Table A.9 in Appendix A.

Together with the experiment results in aforementioned sections, we can finally address the **RQ 3** in terms of dialogue systems:

- **RQ 3**: *Can more appropriate new methods of evaluation be designed that are feasible given the limited time and resources available in operational settings?*

We successfully design a new dialogue evaluation method which is appropriate and feasible within limited time and resources. This proposed human evaluation method



(a) Passed (b) Failed (c) ConvAI2

Figure 5.14: Words per input utterance from workers who passed (5.14a) and failed the quality control (5.14b) in our human evaluation (free run 1); as well as workers from ConvAI2 live evaluation (5.14c).

for dialogue is more appropriate as it is highly reliable with self-replication experiment at $r = 0.969$, as well as high degree of rater agreement. Also, this method can be deployed using reasonable resources according to the reported cost in our experiment.

5.5 Summary

In this chapter, we proposed a new crowd-sourcing human evaluation method for the open-domain dialogue system task, as the development of reliable evaluation of this has been highlighted as a known open-problem. For a dialogue, seven various evaluation criteria are assessed: *interesting*, *fun*, *consistent*, *fluent*, *topic*, *robotic*, and *repetitive*, which is adjustable for further research. We also find an appropriate approach to controlling the quality of crowd. We overcome previous challenges and provide a new human evaluation methodology shown as highly consistent, with results for models correlating at $r = 0.969$ in two separate data collection runs. In addition, our evaluation has the advantages of differences in scoring strategies to be ironed out via score standardization, applicability of standard significance testing, and increasing the reliability of results.

With the data we collected from the completed experiment, we first analyse information of human raters and assigned HITs, including the pass rates and elapsed time. We then report the average standardized scores at system-level, as well as significance test. Agreement of human annotators are also examined, showing that our method has a high rater consistency. We also compute the scores of prevailing automatic metrics and compare them with our human evaluation results, concluding that word-overlap-based metrics have perishing performances, while reference-free metrics perform relatively better but still fail to achieve a very high correlation with humans.

In conclusion, we propose a new crowd-sourcing approach for the evaluation of open-domain dialogue systems, which is easy to deploy in a large scale within appropriate costs and resources. In the future, we hope to use this method for longitudinal evaluation of dialogue systems to measure improvements over time.

Chapter 6

Conclusion and Future Work

In this thesis, we have investigated existing problems and challenges in the evaluation of three distinct NLP tasks: machine reading comprehension, question generation (QG), and open-domain dialogue. To address these problems and challenges for each task, we respectively developed a new human evaluation method that is demonstrated to be appropriate and highly reliable, together with the capacity of be deployed in a large scale within a reasonable budget.

We started with introducing how evaluation effects the development of NLP technologies and the evaluation challenges encountered in current research, where we took the BLEU metric as an example in Chapter 1. We then briefly introduced three distinct tasks which we examined in our investigations, as well as the reasons of selecting these tasks. We provided a review of a number of commonly applied evaluation metrics in each task and the issues and problems which need to be address in their evaluation. Based on this analysis, we identified three research questions relating to the evaluation of NLP tasks, the identified research questions are as follows:

- **RQ 1**: *Within each domain of interest, how accurately do existing automatic metrics measure system performance?*
- **RQ 2** *What are the limitations and disadvantages of the direct application of evaluation metrics from MT and other domains to entirely distinct tasks for*

system development in each area?

- **RQ 3** *Can more appropriate new methods of evaluation be designed that are feasible given the limited time and resources available in operational settings?*

After giving a comprehensive review of the evaluation methods and corresponding issues for our three NLP tasks in Chapter 2, we then began to address the research questions raised task by task in Chapter 3, 4 and 5. For the MRC task, we designed a new human evaluation method based on Direct Assessment, and showed that this method has high reliability using a self-replication experiment. We then used the results of our human evaluation method to compare with automatic evaluation metric scores in order to address **RQ 1**. This showed that not all automatic metrics can accurately evaluate MRC systems at system-level. We also investigated the difference between the performances of human evaluation and automatic metrics when evaluating a single MRC to answer **RQ 2**. The results show that all current metrics perform weakly in MRC evaluation at the sentence-level. Finally, we examined the human rater consistency, and the consequent high rater agreement together with the reported practical costs in experiments thus addressed **RQ 3**.

We subsequently moved to the QG task in Chapter 4 where we proposed a human evaluation method and conducted corresponding experiments. Similarly, we answered **RQ 1** about QG by comparing performance of automatic metrics to results of our proposed method, showing that such metrics cannot achieve high accuracy in evaluating QG. Furthermore, in terms of **RQ 2**, we found that these metrics entirely rely on human-generated references which is unsuitable for evaluating QG due to the one-to-many nature which is described in Chapter 1. In addition, we demonstrated that the proposed QG evaluation method is highly reliable which is feasible for large scale deployment given limited resources, in regard to **RQ 3**.

Chapter 5 inquired into the three **RQ**s regarding dialogue evaluation. We again firstly proposed a new human evaluation method for dialogue, and a self-replication experiment was included to show how consistent the resulting system scores of this method can be. We then tested the correlation between our human scores and metric scores at the system-level. Results showed that all the automatic metrics achieve a low correlation with human judgements, of which some traditional metrics even correlate negatively with human judgements. For **RQ 1** therefore, current metrics are generally incapable of accurately assess dialogue systems. On the meantime, we addressed **RQ 2** regarding the application of metrics from other domains to dialogue evaluation, showing that this can result in inappropriately discarding systems which are deemed high quality according to human judgement. Like the previous two tasks, we also reported high consistency and cost-effectiveness of our proposed human evaluation method to answer **RQ 3** with respect to dialogue.

6.1 Contributions

In this thesis, we mainly focus on solving problems and overcoming challenges in the evaluation of three distinct NLP tasks by proposing a new human evaluation method for each. In summary, the contributions to these tasks in this thesis can be described as follows:

- New human evaluation methods were proposed for each of the NLP tasks that we examined.
- We provided comprehensive analysis of performances of automatic metrics for each task compared with human judgements via our proposed methods.
- The details of corresponding crowd-sourcing experiments were introduced and self-replication experiments were conducted to ensure their reproducibility.
- We additionally proposed a new automatic reference-free evaluation metric for the QG task.
- Collected data and code has been made available as open-source to encourage future research.

Generally speaking, we designed a new crowd-sourcing evaluation method for each task to address their existing evaluation limitations described in Chapter 1 and 2. Instead of a task-specific evaluation, each proposed method rather acts as an evaluation standard since the evaluation criteria and settings in it can easily be adjusted. In the meantime, a quality control approach is necessarily involved in each proposed method to prevent collection of unusable data, by filtering out such data from workers who failed the quality control method. Besides, each method included a self-replication experiment. The nearly perfect correlation between results of this and initial experiment in each case indicated that our proposed method is highly reliable and reproducible for all these NLP tasks, where the Pearson correlation (r) in MRC (Chapter 3), QG (Chapter 4) and dialogue (Chapter 5) achieve 0.986, 0.955 and 0.969, respectively. In addition, since human evaluation methods generally require to report rater agreement (Ondrej Bojar, Buck, et al., 2014), the consistency between human raters was examined. Results showed that our proposed method in each case has a high annotator agreement. The details of these methods and corresponding experiments are provided, including the designed user interface, total cost and statistic data such as numbers and pass rate of human workers. We open-sourced the collected data and the processing scripts to enable application in practice and encourage future research.

We also examined the performance of the automatic evaluation metrics in each domain by computing their correlation with human judgements according to the results of the proposed human evaluation methods. For the MRC task, we found that prevailing metrics are incapable of evaluating performance at the sentence-level. We also found such metrics may produce different system rankings, resulting in potential confusion of ranking MRC systems when distinct metrics are applied. In terms of QG, similar to MRC, we reported that the applied metrics do not correlate highly with human judgements, and additionally that they failed to produce consistent system rankings. In the case of dialogue, we found the performance of the automatic metrics to be worse since most metrics even failed to achieve a positive correlation with human evaluation. This revealed that a high quality dialogue system may be incorrectly filtered out before implementing human evaluation when applying such metrics as the common practice in dialogue competitions.

In Chapter 3, we comprehensively analyzed how the choice of α can influence the statistical quality control method, which can be used to instruct the implementation of quality control in future research. In addition, we investigated the fluctuation of system performance when the word lengths of references were changed.

In Chapter 4, we found current metrics do not take into account the one-to-many nature of QG, we therefore proposed a new unsupervised and reference-free metric. This metric receives a question together with the passage and answer, and assesses the input question in a QA manner based on a pretrained language model. Our results indicate that it can achieve higher correlation with human judgements than other metrics.

In terms of dialogue evaluation in Chapter 5, it had reported in ConvAI2 that live evaluation suffer in ineffective because the conversations collected was deemed senseless or offensive (Dinan, Logacheva, et al., 2019). Hence, we additionally compared the data collected from our human evaluation experiment to the data in ConvAI2. Our results showed that live dialogue evaluation is able to avoid such issues via following our human evaluation method.

6.2 Future Work

Since we have successfully proposed evaluation methods for three NLP tasks that used to suffer from known issues, we are interested in overcoming the challenges that are encountered in evaluation of other NLP tasks, such as multimodal machine translation, text summarization, automatic code generation. For example, text summarization requires a source document to abstractively shortened into a condensed summary, which has a specific purpose of providing readers with contents of the given document in a concise and precise manner (Saggion and Lapalme, 2002). However, evaluation metrics used for the summarization task, such as ROUGE (Gu et al., 2016; Nallapati, Zhai, and Zhou, 2017) and BLEU (X. Zhang and Lapata, 2017; Pasunuru, Guo, and Bansal, 2017), merely produce a overall score, resulting in a lack of interpretability in terms of the performance of task. In detail, a low metric score does not give enough information to indicate what aspects a condensed summary may be weak in. For instance, the summary may be too verbose to be concise, or it may be far away from the original document causing it to be poorly related to the original document. In future work therefore, we plan to design a method for evaluation of text summarization which can take into account its task purpose, of which the criteria should contain at least *concision* and *precision*.In the meantime, proposing new human evaluation methods for different NLP tasks is useful to enable measurement of improvements in corresponding NLP technologies over time, while such longitudinal evaluation of NLP tasks is our concern in our future work, following that of machine translation (Graham, Baldwin, Moffat, et al., 2014).

Despite the successful deployment of human evaluation experiments in this thesis, these methods still have with limitations. The methods are generally task-specific and researchers should carefully choose the correct method when applying them to a certain task. Also, the user interfaces of the evaluation experiments in different NLP tasks can vary when deploying them on a crowd-sourcing platform, making these methods less convenient. Hence, in the future work we plan to integrate these methods into a more general framework, of which the functions includes generating the HITs for the deployment on AMT platform, processing raw data files, and providing the analysis of collected data. This integration can continue the preliminary Python program using the PySimpleGUI module (see `https://pysimplegui.readthedocs.io/en/latest/`), where a demonstration that we applied in dialogue evaluation is shown in Figure 6.1. We firstly plan to integrate MRC, QG and dialogue together, so that researchers on these three domain can directly generate corresponding files and receive results and analyses in future research. In the future, more NLP tasks could be included, such as the aforementioned text summarization.

In addition, we notice that some systems we included in this thesis, such as

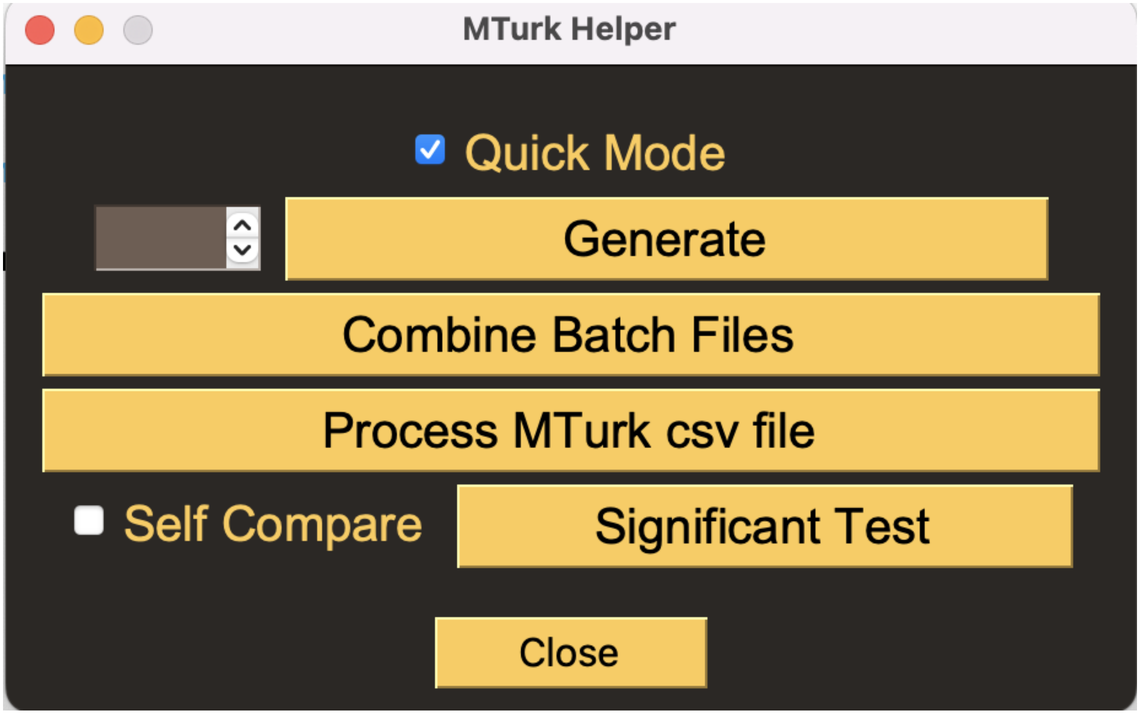


Figure 6.1: Demonstration of our Python program which is only currently available for dialogue evaluation.

the answer distillation MRC system in Chapter 3 and the LSTM-based language model in Chapter 5, are no longer state-of-the-art (SOTA) to some degree, with the development of NLP technologies. We plan to conduct experiments on the outputs of more SOTA systems or models in order to test the performance of our proposed methods. For instance, GPT-3 is a trained language model consisting of 175 billion parameters and researchers reported that it is somewhat hard for human annotators to distinguish between GPT-3 generated news articles and human-written articles (Brown et al., 2020). We think it is well-motivated to investigate whether our human evaluation method can accurately evaluate the performance of GPT-3 when it is applied for open-domain dialogue.

Furthermore, we would like to provide fine-grained analyses on QG and dialogue systems, following those in MRC in Chapter 3. For example, we can test the performance of a dialogue system on a certain topic, such as pets or food as shown in Figure 5.10 in Chapter 5, to investigate how to improve system performances using the collected topic data. Also, our collected dialogue data recorded the change of topics during the interaction with a dialogue system, it would be interesting to know whether changing the current topic can affect system performance.

As reported in Chapter 3, current MRC evaluation metrics failed to correlate well with human judgment at the sentence-level, a metric that can overcome such an issue is also one of our concerns in the future. Recent research indicates that pre-trained language models are capable of correlating more highly at the sentence-level by training on human assessment data. For example, BERTHA is a BERT-based evaluation metric trained on data of video captioning human assessment via transfer learning, which has been found it can achieve higher correlation with human judgement at the caption-level (Lebron et al., 2022). We think the transfer learning is likewise appropriate for the MRC task since we successfully collected the data of MRC human evaluation data. A potential scheme would be to utilize a language model to predict a score of a MRC output, with the goal of training to minimize the distance, or maximize the correlation, between predicted scores and human scores. Moreover, our analyses in Chapter 5 revealed that dialogue evaluation metrics generally perform poorly at the system-level, we think such a transfer learning based scheme would enable to develop a more appropriate metric for evaluating dialogue at system-level the as well.

Although the proposed QAScore metric has been demonstrated to outperform other QG evaluation metrics at the system-level, it is again faced with the aforementioned issue that it only produces an overall score. We are interested in an approach of improving QAScore to allow assessing QG systems in different aspects. At this point, a preliminary idea is to utilize the recently prevailing *prompting*, a paradigm that converts various downstream tasks into a language model format (Han et al., 2021). Recent pretrained models are generally trained on large-scale corpus, such as RoBERTa that is used in our QAScore metric. Prompting can directly elicit knowledge from these models, Shin et al. (2020b) demonstrated that prompting RoBERTa can improve its performance on tasks such as sentiment analysis and relation extraction. Therefore, we think applying prompting to RoBERTa-based QAScore is feasible as different evaluation criteria can be regarded as different downstream tasks. A next step would be to seek prompting templates, for example, the Likert statements provided in Table 4.1 in Chapter 4 are likely to be suitable for each criterion. Together with the collected data in our QG human evaluation experiments, we can improve QAScore to achieve a high correlation with human judgement on all criteria.

Due to the failure of the sub-task in dialogue evaluation, we would like to know whether any other sub-task is available. This sub-task should be more related to the main task, and the intention should not be too obvious. Since the personae of a dialogue system are hidden from workers, we can ask workers about persona-related questions. For example, if a system has the persona "I love dogs", we can let workers to answer "What pet do you think is the favorite of the chatbot: dog, cat or rabbit? " after the conversation is completed.

Bibliography

Alexander, Ralph A (1990). "A note on averaging correlations". In: *Bulletin of the Psychonomic Society* 28.4, pp. 335–336.

Alexandrov, Aliosha (2010). "Characteristics of single-item measures in Likert scale format". In: *The Electronic Journal of Business Research Methods* 8.1, pp. 1–12.

Amidei, Jacopo, Paul Piwek, and Alistair Willis (Aug. 2018). "Rethinking the Agreement in Human Evaluation Tasks". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 3318–3329. URL: https://aclanthology.org/C18-1281.

Awad, George, Asad A. Butt, Keith Curtis, et al. (2018). "TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search". In: *Proceedings of TRECVID 2018*. Gaithersburg, MD.

Awad, George, Asad A. Butt, Jonathan Fiscus, et al. (2017). "Trecvid 2017: Evaluating Ad-hoc and Instance Video Search, Event Detection, Video Captioning and Hyperlinking". In: *Proceedings of TRECVID 2017*. Gaithersburg, MD.

Banerjee, Satanjeev and Alon Lavie (June 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*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: https://aclanthology.org/W05-0909.

Barrault, Loic et al. (Aug. 2019). “Findings of the 2019 Conference on Machine Translation (WMT19)”. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pp. 1–61. URL: http://www.aclweb.org/anthology/W19-5301.

Barrault, Loïc et al. (Nov. 2020). “Findings of the 2020 Conference on Machine Translation (WMT20)”. In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 1–54. URL: https://www.aclweb.org/anthology/2020.wmt-1.1.

Bauer, Lisa, Yicheng Wang, and Mohit Bansal (Oct. 2018). “Commonsense for Generative Multi-Hop Question Answering Tasks”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4220–4230. DOI: 10.18653/v1/D18-1454. URL: https://www.aclweb.org/anthology/D18-1454.

Bhandari, Manik, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu (Dec. 2020). “Metrics also Disagree in the Low Scoring Range: Revisiting Summarization Evaluation Metrics”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5702–5711. DOI: 10.18653/v1/2020.coling-main.501. URL: https://aclanthology.org/2020.coling-main.501.

Bjerva, Johannes, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein (Nov. 2020). “SubjQA: A Dataset for Subjectivity and Review Comprehension”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5480–5494. DOI: 10.18653/v1/2020.emnlp-main.442. URL: https://aclanthology.org/2020.emnlp-main.442.

Bojar, Ondrej, Christian Buck, et al. (June 2014). “Findings of the 2014 Workshop on Statistical Machine Translation”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 12–58. URL: http://www.aclweb.org/anthology/W/W14/W14-3302.

Bojar, Ondrej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz (Oct. 2018). "Findings of the 2018 Conference on Machine Translation (WMT18)". In: *Proceedings of the Third Conference on Machine Translation*. Belgium, Brussels: Association for Computational Linguistics, pp. 272–307. URL: http://www.aclweb.org/anthology/W18-64028.

Bojar, Ondřej, Christian Buck, et al. (Aug. 2013). "Findings of the 2013 Workshop on Statistical Machine Translation". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1–44. URL: http://www.aclweb.org/anthology/W13-2201.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, et al. (Sept. 2017). "Findings of the 2017 Conference on Machine Translation (WMT17)". In: *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 169–214. URL: http://www.aclweb.org/anthology/W17-4717.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, et al. (Aug. 2016). "Findings of the 2016 Conference on Machine Translation". In: *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, pp. 131–198. URL: http://www.aclweb.org/anthology/W/W16/W16-2301.

Bojar, Ondřej, Miloš Ercegovčević, Martin Popel, and Omar Zaidan (July 2011). "A Grain of Salt for the WMT Manual Evaluation". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 1–11. URL: https://www.aclweb.org/anthology/W11-2101.

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners". In: *CoRR* abs/2005.14165. arXiv: 2005.14165. URL: https://arxiv.org/abs/2005.14165.

Burtsev, Mikhail et al. (2018). "The First Conversational Intelligence Challenge". In: *The NIPS '17 Competition: Building Intelligent Systems*. Ed. by Sergio Escalera and Markus Weimer. Cham: Springer International Publishing, pp. 25–46. ISBN: 978-3-319-94042-7.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (June 2012). "Findings of the 2012 Workshop on Statistical Machine Translation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, pp. 10–51. URL: https://aclanthology.org/W12-3102.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan (July 2011). "Findings of the 2011 Workshop on Statistical Machine Translation". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 22–64. URL: http://www.aclweb.org/anthology/W11-2103.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (Apr. 2006). "Re-evaluating the Role of Bleu in Machine Translation Research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics, pp. 249–256. URL: https://aclanthology.org/E06-1032.

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao (2020). "Evaluation of Text Generation: A Survey". In: *CoRR* abs/2006.14799. arXiv: 2006.14799. URL: https://arxiv.org/abs/2006.14799.

Chaudhury, Atef, Makarand Tapaswi, Seung Wook Kim, and Sanja Fidler (2019). "The Shmoop Corpus: A Dataset of Stories with Loosely Aligned Summaries". In: *CoRR* abs/1912.13082. arXiv: 1912.13082. URL: http://arxiv.org/abs/1912.13082.

Chen, Anthony, Gabriel Stanovsky, Sameer Singh, and Matt Gardner (Nov. 2019). "Evaluating Question Answering Evaluation". In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, pp. 119–124. DOI: `10.18653/v1/D19-5817`. URL: `https://aclanthology.org/D19-5817`.

— (Nov. 2020). "MOCHA: A Dataset for Training and Evaluating Generative Reading Comprehension Metrics". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6521–6532. DOI: `10.18653/v1/2020.emnlp-main.528`. URL: `https://aclanthology.org/2020.emnlp-main.528`.

Chen, Danqi (2018). "Neural Reading Comprehension and Beyond". PhD thesis. Stanford University.

Chen, Guanliang, Jie Yang, C. Hauff, and G. Houben (2018). "LearningQ: A Large-Scale Dataset for Educational Question Generation". In: *International AAAI Conference on Web and Social Media*.

Chen, Yu, Lingfei Wu, and Mohammed J. Zaki (Apr. 2020). "Reinforcement Learning Based Graph-to-Sequence Model for Natural Question Generation". In: *Proceedings of the 8th International Conference on Learning Representations*.

Cho, Woon Sang, Yizhe Zhang, Sudha Rao, Asli Celikyilmaz, Chenyan Xiong, Jianfeng Gao, Mengdi Wang, and Bill Dolan (Apr. 2021). "Contrastive Multi-document Question Generation". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 12–30. URL: `https://aclanthology.org/2021.eacl-main.2`.

Chu, Zewei, Mingda Chen, Jing Chen, Miaosen Wang, Kevin Gimpel, Manaal Faruqui, and Xiance Si (2020). "How to Ask Better Questions? A Large-Scale Multi-Domain Dataset for Rewriting Ill-Formed Questions". In: *Proc. of AAAI*.

Clark, Christopher, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova (June 2019). "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2924–2936. DOI: 10.18653/v1/N19-1300. URL: https://aclanthology.org/N19-1300.

Cleveland, William S (1979). "Robust locally weighted regression and smoothing scatterplots". In: *Journal of the American statistical association* 74.368, pp. 829–836.

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46. DOI: 10.1177/001316446002000104.

Dasigi, Pradeep, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner (Nov. 2019). "Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5925–5932. DOI: 10.18653/v1/D19-1606. URL: https://aclanthology.org/D19-1606.

Denkowski, M. and A. Lavie (2011). "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pp. 85–91.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dhingra, Bhuwan, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov (July 2017). "Gated-Attention Readers for Text Comprehension". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1832–1846. DOI: `10.18653/v1/P17-1168`. URL: `https://www.aclweb.org/anthology/P17-1168`.

Dinan, Emily, Varvara Logacheva, et al. (2019). "The Second Conversational Intelligence Challenge (ConvAI2)". In: *CoRR* abs/1902.00098. arXiv: `1902.00098`. URL: `http://arxiv.org/abs/1902.00098`.

Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston (2018). "Wizard of Wikipedia: Knowledge-Powered Conversational agents". In: *CoRR* abs/1811.01241. arXiv: `1811.01241`. URL: `http://arxiv.org/abs/1811.01241`.

Doddington, George (2002). "Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics". In: *Proceedings of the Second International Conference on Human Language Technology Research*. HLT '02. San Diego, California: Morgan Kaufmann Publishers Inc., pp. 138–145.

Du, Wenchao and Alan W Black (July 2019). "Boosting Dialog Response Generation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 38–43. DOI: `10.18653/v1/P19-1005`. URL: `https://aclanthology.org/P19-1005`.

Dua, Dheeru, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner (June 2019). "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2368–

2378. DOI: 10.18653/v1/N19-1246. URL: https://aclanthology.org/N19-1246.

Dušek, Ondřej, Jekaterina Novikova, and Verena Rieser (2020). "Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge". In: *Computer Speech & Language* 59, pp. 123–156. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2019.06.009. URL: https://www.sciencedirect.com/science/article/pii/S0885230819300919.

Dzendzik, Daria, Jennifer Foster, and Carl Vogel (Nov. 2021). "English Machine Reading Comprehension Datasets: A Survey". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 8784–8804. DOI: 10.18653/v1/2021.emnlp-main.693. URL: https://aclanthology.org/2021.emnlp-main.693.

Dzendzik, Daria, Carl Vogel, and Jennifer Foster (June 2019). "Is It Dish Washer Safe? Automatically Answering "Yes/No" Questions Using Customer Reviews". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1–6. DOI: 10.18653/v1/N19-3001. URL: https://aclanthology.org/N19-3001.

Elliott, Desmond and Frank Keller (Oct. 2013). "Image Description using Visual Dependency Representations". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1292–1302. URL: https://aclanthology.org/D13-1128.

Fabbri, Alexander R, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev (2020). "SummEval: Re-evaluating Summarization Evaluation". In: *arXiv preprint arXiv:2007.12626*.

Finch, Sarah E. and Jinho D. Choi (July 2020). "Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols".

In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, pp. 236–245. URL: https://aclanthology.org/2020.sigdial-1.29.

Forgues, Gabriel, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay (2014). "Bootstrapping dialog systems with word embeddings". In: *Nips, modern machine learning and natural language processing workshop*. Vol. 2.

Galley, Michel et al. (July 2015). "deltaBLEU: 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 (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, pp. 445–450. DOI: 10.3115/v1/P15-2073. URL: https://aclanthology.org/P15-2073.

Ghandeharioun, Asma, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard (2019). "Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Gillick, Dan and Yang Liu (June 2010). "Non-Expert Evaluation of Summarization Systems is Risky". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Los Angeles: Association for Computational Linguistics, pp. 148–151. URL: https://aclanthology.org/W10-0722.

Gopalakrishnan, Karthik, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür (2019). "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations". In: *Proc. Interspeech 2019*, pp. 1891–1895. DOI: 10.21437/Interspeech.2019-3079.

Graesser, A. C., P. Chipman, B. C. Haynes, and A. Olney (2005). "AutoTutor: an intelligent tutoring system with mixed-initiative dialogue". In: *IEEE Transactions on Education* 48.4, pp. 612–618. DOI: 10.1109/TE.2005.856149.

Graham, Yvette (Sept. 2015). "Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 128–137. DOI: 10.18653/v1/D15-1013. URL: https://aclanthology.org/D15-1013.

Graham, Yvette, George Awad, and Alan Smeaton (Sept. 2018). "Evaluation of automatic video captioning using direct assessment". In: *PLOS ONE* 13.9, pp. 1–20. DOI: 10.1371/journal.pone.0202789. URL: https://doi.org/10.1371/journal.pone.0202789.

Graham, Yvette and Timothy Baldwin (Oct. 2014). "Testing for Significance of Increased Correlation with Human Judgment". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 172–176. DOI: 10.3115/v1/D14-1020. URL: https://aclanthology.org/D14-1020.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel (Dec. 2013). "Crowd-Sourcing of Human Judgments of Machine Translation Fluency". In: *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*. Brisbane, Australia, pp. 16–24. URL: https://www.aclweb.org/anthology/U13-1004.

— (Apr. 2014). "Is Machine Translation Getting Better over Time?" In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 443–451. DOI: 10.3115/v1/E14-1047. URL: https://aclanthology.org/E14-1047.

— (Jan. 2016). "Can machine translation systems be evaluated by the crowd alone". In: *Natural Language Engineering* FirstView, pp. 1–28. ISSN: 1469-8110. DOI: 10.

1017/S1351324915000339. URL: http://journals.cambridge.org/article_S1351324915000339.

Graham, Yvette, Barry Haddow, and Philipp Koehn (Nov. 2020). "Statistical Power and Translationese in Machine Translation Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 72–81. DOI: 10.18653/v1/2020.emnlp-main.6. URL: https://www.aclweb.org/anthology/2020.emnlp-main.6.

Graham, Yvette and Qun Liu (2016). "Achieving accurate conclusions in evaluation of automatic machine translation metrics". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1–10.

Gu, Jiatao, Zhengdong Lu, Hang Li, and Victor O.K. Li (Aug. 2016). "Incorporating Copying Mechanism in Sequence-to-Sequence Learning". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1631–1640. DOI: 10.18653/v1/P16-1154. URL: https://aclanthology.org/P16-1154.

Han, Xu et al. (2021). "Pre-Trained Models: Past, Present and Future". In: *CoRR* abs/2106.07139. arXiv: 2106.07139. URL: https://arxiv.org/abs/2106.07139.

Hassan, Hany et al. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". In: *CoRR* abs/1803.05567. arXiv: 1803.05567. URL: http://arxiv.org/abs/1803.05567.

He, Wei et al. (July 2018). "DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications". In: *Proceedings of the Workshop on Machine Reading for Question Answering*. Melbourne, Australia: Association for Computational Linguistics, pp. 37–46. DOI: 10.18653/v1/W18-2605. URL: https://www.aclweb.org/anthology/W18-2605.

Hermann, Karl Moritz, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). "Teaching Machines to Read and Comprehend". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 1693–1701.

Herzig, Jonathan, Thomas Müller, Syrine Krichene, and Julian Eisenschlos (June 2021). "Open Domain Question Answering over Tables via Dense Retrieval". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 512–519. DOI: 10.18653/v1/2021.naacl-main.43. URL: https://aclanthology.org/2021.naacl-main.43.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Howcroft, David M., Anya Belz, et al. (Dec. 2020). "Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions". In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, pp. 169–182. URL: https://aclanthology.org/2020.inlg-1.23.

Howcroft, David M. and Verena Rieser (Nov. 2021). "What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 8932–8939. DOI: 10.18653/v1/2021.emnlp-main.703. URL: https://aclanthology.org/2021.emnlp-main.703.

Hu, Minghao, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou (Oct. 2018). "Attention-Guided Answer Distillation for Machine

Reading Comprehension". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2077–2086. DOI: 10.18653/v1/D18-1232. URL: https://www.aclweb.org/anthology/D18-1232.

Huang, Lifu, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (Nov. 2019). "Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2391–2401. DOI: 10.18653/v1/D19-1243. URL: https://www.aclweb.org/anthology/D19-1243.

Huang, Shih-Wen and Wai-Tat Fu (2013). "Enhancing Reliability Using Peer Consistency Evaluation in Human Computation". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. CSCW '13. San Antonio, Texas, USA: Association for Computing Machinery, pp. 639–648. ISBN: 9781450313315. DOI: 10.1145/2441776.2441847. URL: https://doi.org/10.1145/2441776.2441847.

Humeau, Samuel, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston (2019). "Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring". In: *CoRR* abs/1905.01969. URL: http://arxiv.org/abs/1905.01969.

Iskender, Neslihan, Tim Polzehl, and Sebastian Möller (Nov. 2020). "Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation". In: *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Online: Association for Computational Linguistics, pp. 164–175. DOI: 10.18653/v1/2020.eval4nlp-1.16. URL: https://aclanthology.org/2020.eval4nlp-1.16.

Ji, Tianbo, Yvette Graham, and Gareth J.F. Jones (2020). "Contrasting Human Opinion of Non-Factoid Question Answering with Automatic Evaluation". In:

New York, NY, USA: Association for Computing Machinery, pp. 348–352. ISBN: 9781450368926. URL: https://doi.org/10.1145/3343413.3377996.

Ji, Tianbo, Chenyang Lyu, Zhichao Cao, and Peng Cheng (2021). "Multi-Hop Question Generation Using Hierarchical Encoding-Decoding and Context Switch Mechanism". In: *Entropy* 23.11. ISSN: 1099-4300. DOI: 10.3390/e23111449. URL: https://www.mdpi.com/1099-4300/23/11/1449.

Jia, Xin, Wenjie Zhou, Xu Sun, and Yunfang Wu (2021). "EQG-RACE: Examination-Type Question Generation". In: *AAAI*.

Jin, Qiao, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu (Nov. 2019). "PubMedQA: A Dataset for Biomedical Research Question Answering". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 2567–2577. DOI: 10.18653/v1/D19-1259. URL: https://aclanthology.org/D19-1259.

Joshi, Mandar, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer (July 2017). "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1601–1611. DOI: 10.18653/v1/P17-1147. URL: https://www.aclweb.org/anthology/P17-1147.

Kilickaya, Mert, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem (Apr. 2017). "Re-evaluating Automatic Metrics for Image Captioning". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 199–209. URL: https://aclanthology.org/E17-1019.

Kim, Seokhwan, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino (2016). "The fifth dialog state tracking challenge". In: *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 511–517.

Kim, Seokhwan, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, and Matthew Henderson (2017). "The fourth dialog state tracking challenge". In: *Dialogues with Social Robots*. Springer, pp. 435–449.

Kim, Yanghoon, Hwanhee Lee, Joongbo Shin, and Kyomin Jung (July 2019). "Improving Neural Question Generation Using Answer Separation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 6602–6609. DOI: 10.1609/aaai.v33i01.33016602. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4629.

Ko, Wei-Jen, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li (Nov. 2020). "Inquisitive Question Generation for High Level Text Comprehension". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6544–6555. DOI: 10.18653/v1/2020.emnlp-main.530. URL: https://aclanthology.org/2020.emnlp-main.530.

Kočiský, Tomáš, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette (2018). "The NarrativeQA Reading Comprehension Challenge". In: *Transactions of the Association for Computational Linguistics* 6, pp. 317–328. DOI: 10.1162/tacl_a_00023. URL: https://www.aclweb.org/anthology/Q18-1023.

Kurdi, Ghader, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari (2020). "A systematic review of automatic question generation for educational purposes". In: *International Journal of Artificial Intelligence in Education* 30.1, pp. 121–204.

Lai, Guokun, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy (Sept. 2017). "RACE: Large-scale ReAding Comprehension Dataset From Examinations". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 785–794. DOI: 10.18653/v1/D17-1082. URL: https://www.aclweb.org/anthology/D17-1082.

Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, pp. 159–174. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2529310.

Lasecki, Walter S., Jaime Teevan, and Ece Kamar (2014). "Information Extraction and Manipulation Threats in Crowd-Powered Systems". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '14. Baltimore, Maryland, USA: Association for Computing Machinery, pp. 248–256. ISBN: 9781450325400. DOI: 10.1145/2531602.2531733. URL: https://doi.org/10.1145/2531602.2531733.

Le, John, Andy Edmonds, Vaughn Hester, and Lukas Biewald (2010). "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution". In: *In SIGIR 2010 workshop*, pp. 21–26.

Lebron, Luis, Yvette Graham, Kevin McGuinness, Konstantinos Kouramas, and Noel E. O'Connor (2022). "BERTHA: Video Captioning Evaluation Via Transfer-Learned Human Assessment". In: *CoRR* abs/2201.10243. arXiv: 2201.10243. URL: https://arxiv.org/abs/2201.10243.

Lee, Chris van der, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer (Oct. 2019). "Best practices for the human evaluation of automatically generated text". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 355–368. DOI: 10.18653/v1/W19-8643. URL: https://aclanthology.org/W19-8643.

Levenshtein, Vladimir I et al. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union, pp. 707–710.

Levinboim, T., A. Thaplfziyal, P. Sharma, and R. Soricut (2019). “Quality Estimation for Image Captions Based on Large-scale Human Evaluations”. In: *arXiv preprint arXiv:1909.03396*.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020). “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: `10.18653/v1/2020.acl-main.703`. URL: `https://aclanthology.org/2020.acl-main.703`.

Lewis, Patrick et al. (2020). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 9459–9474. URL: `https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf`.

Li, Margaret, Jason Weston, and Stephen Roller (2019). “Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons”. In: *arXiv preprint arXiv:1909.03087*.

Lin, Chin-Yew (July 2004). “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: `https://www.aclweb.org/anthology/W04-1013`.

Lin, Kevin, Oyvind Tafjord, Peter Clark, and Matt Gardner (Nov. 2019). “Reasoning Over Paragraph Effects in Situations”. In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, pp. 58–62. DOI: `10.18653/v1/D19-5808`. URL: `https://aclanthology.org/D19-5808`.

Lin, Zhaojiang, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung (Nov. 2019). “MoEL: Mixture of Empathetic Listeners”. In: *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 121–132. DOI: 10.18653/v1/D19-1012. URL: https://aclanthology.org/D19-1012.

Liu, Bang, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He (2020). "Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus". In: *Proceedings of The Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, pp. 2032–2043. ISBN: 9781450370233. DOI: 10.1145/3366423.3380270. URL: https://doi.org/10.1145/3366423.3380270.

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau (Nov. 2016). "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2122–2132. DOI: 10.18653/v1/D16-1230. URL: https://aclanthology.org/D16-1230.

Liu, Qiang, Alexander T Ihler, and Mark Steyvers (2013). "Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?" In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2013/file/cc1aa436277138f61cda7039 Paper.pdf.

Liu, Yinhan et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692. arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

Lloret, Elena, Laura Plaza, and Ahmet Aker (June 2013). "Analyzing the capabilities of crowdsourcing services for text summarization". In: *Language Resources and*

Evaluation 47.2, pp. 337–369. ISSN: 1574-0218. DOI: 10.1007/s10579-012-9198-8. URL: https://doi.org/10.1007/s10579-012-9198-8.

Loukina, Anastassia, Nitin Madnani, Aoife Cahill, Lili Yao, Matthew S. Johnson, Brian Riordan, and Daniel F. McCaffrey (July 2020). "Using PRMSE to evaluate automated scoring systems in the presence of label noise". In: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA → Online: Association for Computational Linguistics, pp. 18–29. DOI: 10.18653/v1/2020.bea-1.2. URL: https://www.aclweb.org/anthology/2020.bea-1.2.

Lowe, Ryan, Nissan Pow, Iulian Serban, and Joelle Pineau (Sept. 2015). "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: Association for Computational Linguistics, pp. 285–294. DOI: 10.18653/v1/W15-4640. URL: https://aclanthology.org/W15-4640.

Lyu, Chenyang, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang, and Qun Liu (2021). "Improving Unsupervised Question Answering via Summarization-Informed Question Generation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4134–4148.

Ma, Xiyao, Qile Zhu, Yanlin Zhou, and Xiaolin Li (Apr. 2020). "Improving Question Generation with Sentence-Level Semantic Matching and Answer Position Inferring". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 8464–8471. DOI: 10.1609/aaai.v34i05.6366. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6366.

Mehri, Shikib and Maxine Eskenazi (July 2020a). "Unsupervised Evaluation of Interactive Dialog with DialoGPT". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, pp. 225–235. URL: https://www.aclweb.org/anthology/2020.sigdial-1.28.

Mehri, Shikib and Maxine Eskenazi (July 2020b). “USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 681–707. DOI: 10.18653/v1/2020.acl-main.64. URL: https://www.aclweb.org/anthology/2020.acl-main.64.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1301.3781.

Mille, Simon, Anya Belz, Bernd Bohnet, Thiago Castro Ferreira, Yvette Graham, and Leo Wanner (Dec. 2020). “The Third Multilingual Surface Realisation Shared Task (SR’20): Overview and Evaluation Results”. In: *Proceedings of the Third Workshop on Multilingual Surface Realisation*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 1–20. URL: https://www.aclweb.org/anthology/2020.msr-1.1.

Miller, Alexander H., Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston (2017). “ParlAI: A Dialog Research Software Platform”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 - System Demonstrations*. Ed. by Lucia Specia, Matt Post, and Michael Paul. Association for Computational Linguistics, pp. 79–84. DOI: 10.18653/v1/d17-2014. URL: https://doi.org/10.18653/v1/d17-2014.

Miller, Alexander H., Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston (2016). “Key-Value Memory Networks for Directly Reading Documents”. In: *CoRR* abs/1606.03126. arXiv: 1606.03126. URL: http://arxiv.org/abs/1606.03126.

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou (2017). "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, pp. 3075–3081.

Narayan, Shashi, Gonçalo Simões, Ji Ma, Hannah Craighead, and Ryan T. McDonald (2020). "QURIOUS: Question Generation Pretraining for Text Generation". In: *CoRR* abs/2004.11026. arXiv: 2004.11026. URL: https://arxiv.org/abs/2004.11026.

Nema, Preksha and Mitesh M. Khapra (Oct. 2018). "Towards a Better Metric for Evaluating Question Generation Systems". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3950–3959. DOI: 10.18653/v1/D18-1429. URL: https://www.aclweb.org/anthology/D18-1429.

Nguyen, Tri, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng (2016). "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset". In: *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*. Ed. by Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne. Vol. 1773. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-1773/CoCoNIPS%5C_2016%5C_paper9.pdf.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney (May 2000). "An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research". In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/278.pdf.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser (Sept. 2017). "Why We Need New Evaluation Metrics for NLG". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, pp. 2241–2252. DOI: 10.18653/v1/D17-1238. URL: https://aclanthology.org/D17-1238.

Novikova, Jekaterina, Ondřej Dušek, and Verena Rieser (June 2018). "RankME: Reliable Human Ratings for Natural Language Generation". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* New Orleans, Louisiana: Association for Computational Linguistics, pp. 72–78. DOI: 10.18653/v1/N18-2012. URL: https://aclanthology.org/N18-2012.

Nowak, Stefanie and Stefan Rüger (2010). "How Reliable Are Annotations via Crowdsourcing: A Study about Inter-Annotator Agreement for Multi-Label Image Annotation". In: *Proceedings of the International Conference on Multimedia Information Retrieval.* MIR '10. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, pp. 557–566. ISBN: 9781605588155. DOI: 10.1145/1743384.1743478. URL: https://doi.org/10.1145/1743384.1743478.

Onishi, Takeshi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester (Nov. 2016). "Who did What: A Large-Scale Person-Centered Cloze Dataset". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas: Association for Computational Linguistics, pp. 2230–2235. DOI: 10.18653/v1/D16-1241. URL: https://aclanthology.org/D16-1241.

Ostermann, Simon, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal (May 2018). "MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1564.

Ostermann, Simon, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal (June 2018). "SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge". In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 747–757. DOI: 10.18653/v1/S18-1119. URL: https://aclanthology.org/S18-1119.

Pan, Liangming, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan (July 2020). "Semantic Graphs for Generating Deep Questions". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1463–1475. DOI: 10.18653/v1/2020.acl-main.135. URL: https://aclanthology.org/2020.acl-main.135.

Pang, Bo, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu (July 2020). "Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 3619–3629. DOI: 10.18653/v1/2020.acl-main.333. URL: https://www.aclweb.org/anthology/2020.acl-main.333.

Papineni, Kishore, 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 on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

Pasunuru, Ramakanth, Han Guo, and Mohit Bansal (Sept. 2017). "Towards Improving Abstractive Summarization via Entailment Generation". In: *Proceedings of the Workshop on New Frontiers in Summarization*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 27–32. DOI: 10.18653/v1/W17-4504. URL: https://aclanthology.org/W17-4504.

Peyrard, Maxime (July 2019). "Studying Summarization Evaluation Metrics in the Appropriate Scoring Range". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5093–5100. DOI: 10.18653/v1/P19-1502. URL: https://aclanthology.org/P19-1502.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language Models are Unsupervised Multitask Learners". In.

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

Rajpurkar, Pranav, Robin Jia, and Percy Liang (July 2018). "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 784–789. DOI: 10.18653/v1/P18-2124. URL: https://aclanthology.org/P18-2124.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (Nov. 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://www.aclweb.org/anthology/D16-1264.

Ram, Ashwin et al. (2018). "Conversational AI: The Science Behind the Alexa Prize". In: *CoRR* abs/1801.03604. arXiv: 1801.03604. URL: http://arxiv.org/abs/1801.03604.

Reiter, Ehud (Sept. 2018). "A Structured Review of the Validity of BLEU". In: *Computational Linguistics* 44.3, pp. 393–401. ISSN: 0891-2017. DOI: 10.1162/

coli_a_00322. eprint: https://direct.mit.edu/coli/article-pdf/44/3/393/1809172/coli_a_00322.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00322.

Ren, Siyu and Kenny Q. Zhu (2020). "Knowledge-Driven Distractor Generation for Cloze-style Multiple Choice Questions". In: *CoRR* abs/2004.09853. arXiv: 2004.09853. URL: https://arxiv.org/abs/2004.09853.

Richardson, Matthew, Christopher J.C. Burges, and Erin Renshaw (Oct. 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*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 193–203. URL: https://www.aclweb.org/anthology/D13-1020.

Rogers, Anna, Matt Gardner, and Isabelle Augenstein (2021). "QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension". In: *CoRR* abs/2107.12708. arXiv: 2107.12708. URL: https://arxiv.org/abs/2107.12708.

Rus, Vasile and Mihai Lintean (June 2012). "A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics". In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montréal, Canada: Association for Computational Linguistics, pp. 157–162. URL: https://aclanthology.org/W12-2018.

Saggion, Horacio and Guy Lapalme (2002). "Generating Indicative-Informative Summaries with SumUM". In: *Computational Linguistics* 28.4, pp. 497–526. DOI: 10.1162/089120102762671963. URL: https://aclanthology.org/J02-4005.

Saha, Amrita, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan (July 2018). "DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1683–1693.

DOI: 10.18653/v1/P18-1156. URL: https://www.aclweb.org/anthology/P18-1156.

Sai, Ananya B., Akash Kumar Mohankumar, and Mitesh M. Khapra (Jan. 2022). "A Survey of Evaluation Metrics Used for NLG Systems". In: 55.2. ISSN: 0360-0300. DOI: 10.1145/3485766. URL: https://doi.org/10.1145/3485766.

Salton, G., A. Wong, and C. S. Yang (Nov. 1975). "A Vector Space Model for Automatic Indexing". In: *Commun. ACM* 18.11, pp. 613–620. ISSN: 0001-0782. DOI: 10.1145/361219.361220. URL: https://doi.org/10.1145/361219.361220.

Santhanam, Sashank, Alireza Karduni, and Samira Shaikh (2020). "Studying the effects of cognitive biases in evaluation of conversational agents". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Santhanam, Sashank and Samira Shaikh (Oct. 2019). "Towards Best Experiment Design for Evaluating Dialogue System Output". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 88–94. DOI: 10.18653/v1/W19-8610. URL: https://aclanthology.org/W19-8610.

Sap, Maarten, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi (Nov. 2019). "Social IQa: Commonsense Reasoning about Social Interactions". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 4463–4473. DOI: 10.18653/v1/D19-1454. URL: https://aclanthology.org/D19-1454.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh (July 2020). "BLEURT: Learning Robust Metrics for Text Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. URL: https://aclanthology.org/2020.acl-main.704.

Shakeri, Siamak, Cicero dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang (2020). "End-to-End Synthetic Data Generation for Domain Adaptation of Question Answering Systems". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5445–5460.

Shin, Taylor, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh (Nov. 2020a). "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4222–4235. DOI: 10.18653/v1/2020.emnlp-main.346. URL: https://aclanthology.org/2020.emnlp-main.346.

— (Nov. 2020b). "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4222–4235. DOI: 10.18653/v1/2020.emnlp-main.346. URL: https://aclanthology.org/2020.emnlp-main.346.

Shinoda, Kazutoshi, Saku Sugawara, and Akiko Aizawa (Aug. 2021). "Improving the Robustness of QA Models to Challenge Sets with Variational Question-Answer Pair Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. Online: Association for Computational Linguistics, pp. 197–214. DOI: 10.18653/v1/2021.acl-srw.21. URL: https://aclanthology.org/2021.acl-srw.21.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul (Aug. 2006). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng (Oct. 2008). "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 254–263. URL: https://aclanthology.org/D08-1027.

Sorodoc, Ionut, Jey Han Lau, Nikolaos Aletras, and Timothy Baldwin (Apr. 2017). "Multimodal Topic Labelling". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 701–706. URL: https://www.aclweb.org/anthology/E17-2111.

Steiger, James H. (1980). "Tests for comparing elements of a correlation matrix." In: *Psychological Bulletin* 87, pp. 245–251.

Sugawara, Saku, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa (2018). "What Makes Reading Comprehension Questions Easier?" In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4208–4219. DOI: 10.18653/v1/D18-1453. URL: https://www.aclweb.org/anthology/D18-1453.

Šuster, Simon and Walter Daelemans (June 2018). "CliCR: a Dataset of Clinical Case Reports for Machine Reading Comprehension". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1551–1563. DOI: 10.18653/v1/N18-1140. URL: https://www.aclweb.org/anthology/N18-1140.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, pp. 3104–3112.

Tapaswi, Makarand, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler (2016). "MovieQA: Understanding Stories in Movies through Question-Answering". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4631–4640. DOI: 10.1109/CVPR.2016.501.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, Liuyin, Zihan Xu, Zibo Lin, Haitao Zheng, and Ying Shen (Dec. 2020). "Answer-driven Deep Question Generation based on Reinforcement Learning". In: *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5159–5170. DOI: 10.18653/v1/2020.coling-main.452. URL: https://aclanthology.org/2020.coling-main.452.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu (2015). "Towards Universal Paraphrastic Sentence Embeddings". In: *CoRR* abs/1511.08198.

Williams, Evan J (1959). "Regression Analysis". In: *Wiley, New York, USA* 14.

Wolf, Thomas et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *CoRR* abs/1910.03771. arXiv: 1910.03771. URL: http://arxiv.org/abs/1910.03771.

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *CoRR* abs/1609.08144. arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

Xie, Qizhe, Guokun Lai, Zihang Dai, and Eduard H. Hovy (2017). "Large-scale Cloze Test Dataset Designed by Teachers". In: *CoRR* abs/1711.03225. arXiv: 1711.03225. URL: http://arxiv.org/abs/1711.03225.

Xiong, Wenhan, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang (July 2019). "TWEETQA: A Social Media Focused Question Answering Dataset". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5020–5031. DOI: 10.18653/v1/P19-1496. URL: https://aclanthology.org/P19-1496.

Xu, Hongfei, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Meng Zhang (Aug. 2021). "Multi-Head Highly Parallelized LSTM Decoder for Neural Machine Translation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 273–282. DOI: 10.18653/v1/2021.acl-long.23. URL: https://aclanthology.org/2021.acl-long.23.

Yang, Yi, Wen-tau Yih, and Christopher Meek (Sept. 2015). "WikiQA: A Challenge Dataset for Open-Domain Question Answering". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2013–2018. DOI: 10.18653/v1/D15-1237. URL: https://aclanthology.org/D15-1237.

Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning (2018). "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yu, Weihao, Zihang Jiang, Yanfei Dong, and Jiashi Feng (2020). "ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HJgJtT4tvB.

Yuan, Xingdi, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler (Aug. 2017). "Machine Comprehension by Text-to-Text Neural Question Generation". In: *Pro-*

ceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, Canada: Association for Computational Linguistics, pp. 15–25. DOI: 10.18653/v1/W17-2603. URL: https://aclanthology.org/W17-2603.

Zarrieß, Sina, Henrik Voigt, and Simeon Schüz (2021). "Decoding Methods in Neural Language Generation: A Survey". In: *Information* 12.9. ISSN: 2078-2489. DOI: 10.3390/info12090355. URL: https://www.mdpi.com/2078-2489/12/9/355.

Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston (July 2018). "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2204–2213. DOI: 10.18653/v1/P18-1205. URL: https://www.aclweb.org/anthology/P18-1205.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020). "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, Xingxing and Mirella Lapata (Sept. 2017). "Sentence Simplification with Deep Reinforcement Learning". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 584–594. DOI: 10.18653/v1/D17-1062. URL: https://aclanthology.org/D17-1062.

Zhang, Yizhe et al. (July 2020). "DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 270–278. DOI: 10.18653/v1/2020.acl-demos.30. URL: https://www.aclweb.org/anthology/2020.acl-demos.30.

Zou, Yicheng, Zhihua Liu, Xingwu Hu, and Qi Zhang (Nov. 2021). "Thinking Clearly, Talking Fast: Concept-Guided Non-Autoregressive Generation for Open-

Domain Dialogue Systems". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2215–2226. DOI: 10.18653/v1/2021.emnlp-main.169. URL: https://aclanthology.org/2021.emnlp-main.169.

Appendix A

Experiment Design and Results

Table A.1: Average standardized scores for dialogue systems in the second data collection run; workers were free to choose the topic of conversation (Free run 2); where the detailed system names are the same as those in Table 5.4; a system with $_p$ means it holds a persona; score for *robotic* and *repetitive* have been reversed; n is number of ratings; systems follow the order in Table 5.4.

	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
A	623	0.455	0.635	0.629	0.728	0.924	0.922	−0.443	−0.212
B	553	0.344	0.464	0.407	0.554	0.763	0.822	−0.338	−0.266
A_p	539	0.423	0.747	0.763	0.555	0.728	0.474	−0.348	0.040
C	539	0.245	0.576	0.492	0.229	0.585	0.043	−0.545	0.337
C_p	609	0.154	0.453	0.390	0.027	0.544	−0.200	−0.515	0.382
B_p	630	0.260	0.464	0.372	0.560	0.581	0.496	−0.412	−0.238
D	595	0.002	0.009	−0.064	0.389	0.282	0.656	−0.720	−0.541
D_p	679	−0.258	−0.285	−0.304	0.033	0.209	−0.226	−0.550	−0.683
E	567	−0.202	−0.063	−0.044	−0.075	0.300	−0.346	−0.646	−0.539
E_p	511	−0.218	−0.152	−0.143	0.043	0.426	−0.352	−0.702	−0.646

Table A.2: Average raw scores for dialogue systems in the initial data collection run (free topic run 1); where the detailed system names are the same as those in Table 5.4; a system with p means it holds a persona; score for *robotic* and *repetitive* have been reversed; n is number of ratings; systems follow the order in Table 5.4; a underlined score means the highest score of that evaluation criterion.

	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
A	798	<u>52.49</u>	<u>53.03</u>	<u>54.07</u>	<u>58.12</u>	61.78	65.24	<u>35.73</u>	39.47
B	798	50.41	51.39	51.68	56.37	<u>64.50</u>	<u>67.84</u>	25.63	35.45
A_p	707	45.53	47.38	46.23	48.52	60.17	47.50	28.30	40.62
C	791	43.96	50.50	47.53	35.85	55.73	33.98	27.35	<u>56.76</u>
C_p	714	41.21	47.13	46.26	39.25	55.05	32.07	21.85	46.84
B_p	707	39.93	41.35	40.06	44.93	53.74	43.72	25.25	30.49
D	707	33.71	30.28	29.95	41.72	45.92	49.07	17.30	21.72
D_p	798	29.38	26.19	27.97	37.53	44.19	35.26	17.46	17.06
E	742	28.99	30.75	30.65	31.27	46.42	23.60	15.10	25.13
E_p	763	28.65	29.34	28.50	29.13	47.07	21.30	17.82	27.41

Table A.3: Average raw scores for dialogue systems in the second data collection run (free topic run 2); where the detailed system names are the same as those in Table 5.4; a system with p means it holds a persona; score for *robotic* and *repetitive* have been reversed; n is number of ratings; systems follow the order in Table 5.4; r is the correlation between current assessment criterion and that in the first run of free topic.

	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
A	623	51.67	56.62	56.27	59.21	64.69	64.04	27.11	33.74
B	539	49.07	52.42	50.66	54.88	60.86	63.73	29.57	31.38
A_p	553	50.56	59.95	60.23	54.28	60.61	52.06	27.59	39.22
C	630	45.87	55.60	53.02	45.16	54.70	38.72	24.40	49.50
C_p	539	42.27	51.19	49.61	37.90	54.17	30.42	22.74	49.84
B_p	609	46.71	51.92	49.95	54.62	56.01	52.85	28.48	33.10
D	595	38.17	38.31	35.39	50.99	46.38	57.94	16.09	22.08
D_p	567	30.89	31.07	30.37	38.37	44.64	31.47	21.85	18.48
E	679	31.70	35.67	36.32	35.26	46.91	26.79	18.98	21.99
E_p	511	31.66	33.63	33.26	38.77	51.53	26.99	17.63	19.79
r	–	0.959	0.947	0.919	0.880	0.951	0.951	0.783	0.945

1. Your task is to have 6 conversations with a chatbot, and a different chatbot will talk to you in each conversation.
2. Before each conversation, you should think of a topic to talk about with the chatbot (your choice of topic). You will be asked to enter this topic before the conversation starts.
3. The current topic will be displayed to you throughout the conversation.
4. If the chatbot changes the topic to a new one, you should record this by updating the conversation topic using the Topic button (bottom left).
5. You are also allowed to change the topic, you should use the same button to do this (bottom left).
6. At the end of each conversation, you should tell us what you think about the chatbot.
7. In each conversation, you should type in a minimum of 10 inputs/sentences.
8. The purpose of these HITs where you will generate conversations with chatbots is to test how realistic their conversations are with users. In order for your data to be useful to us we require that your half of the conversation is also realistic. For example, your data will not be useful to us if you do the following:
   User: Hi
   Bot: Hi
   User: Hi
   Bot: Hi
   .. and so on.
9. Another example, if you are too repetitive or your responses are not appropriate given what the chatbot has just said, this will not be a useful test for them. For example, the following conversation is not ok:
   User: Hi
   Bot: Hi
   User: wow (not appropriate response)
   Bot: I saw a good movie last night
   User: wow (repetitive)
   Bot: Do you like football?
   User: I have two children and one dog. (not appropriate response)
   .. and so on.
10. We need realistic conversations, so please do your best to talk to the bot as if the bot was another person you actually want to talk to. Obvious attempts to game the process and ones that don't make a real effort will unfortunately be rejected.
11. The chatbot may take a few seconds to respond, please be patient.
12. Please use Chrome and avoid special symbols if possible.
13. There is a feedback box at the end of the HIT. If you encounter any problems, please enter them in this box or email our MTurk account.

Figure A.1: Instructions shown to crowd-sourcing workers before starting the open-domain dialogue human evaluation.

Figure A.2: The popup window where a worker is given a topic and record the opinion of this topic, before starting the conversation.

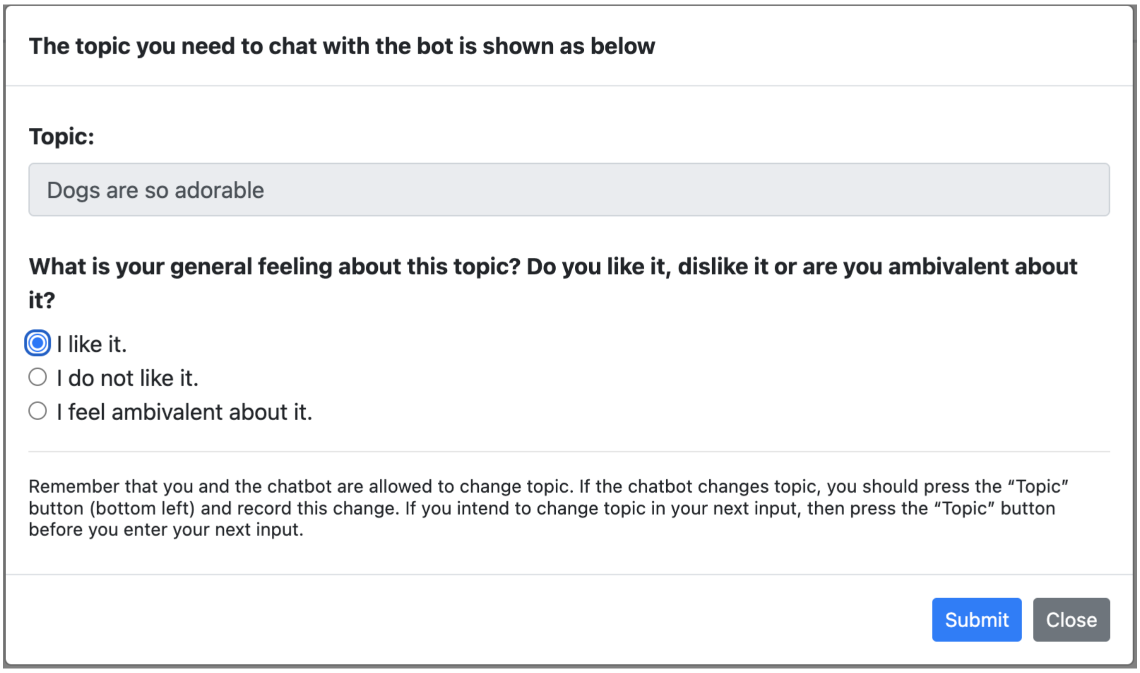


Table A.4: Average raw scores for dialogue systems in the ice-breaker experiment; where the detailed system names are the same as those in Table 5.4; a system with p means it holds a persona and the ice-breaker statement is subsequently unknown to systems without p; score for *robotic* and *repetitive* have been reversed; n is number of ratings; systems follow the order in Table 5.4; r is the correlation between current assessment criterion and that in the first run of free topic; a underlined score means the highest score of that evaluation criterion.

	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
A	721	53.43	53.65	52.35	63.24	67.28	66.97	28.17	42.32
A_p	721	50.21	54.53	53.50	52.84	58.83	53.18	38.87	39.70
B	742	49.55	49.23	47.76	57.79	60.64	62.22	32.56	36.65
C	784	47.93	56.18	53.69	43.15	56.88	40.46	29.61	55.54
B_p	700	44.94	48.83	46.70	49.58	55.86	49.21	25.82	38.61
C_p	658	42.41	47.98	45.48	37.66	54.51	32.50	26.00	52.72
D	728	35.14	30.32	33.13	42.90	49.92	48.51	20.11	21.09
E_p	721	31.58	31.73	30.82	35.44	47.12	27.06	21.90	26.97
E	721	30.09	33.17	31.95	31.14	47.12	24.90	19.10	23.23
D_p	714	27.22	22.56	22.53	35.22	41.70	34.98	17.44	16.09
r	–	0.970	0.955	0.918	0.949	0.928	0.972	0.738	0.968

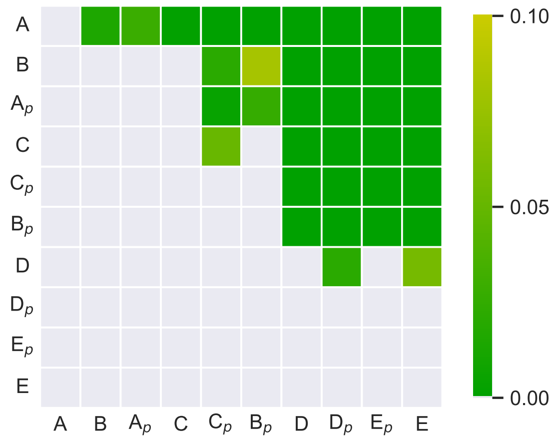
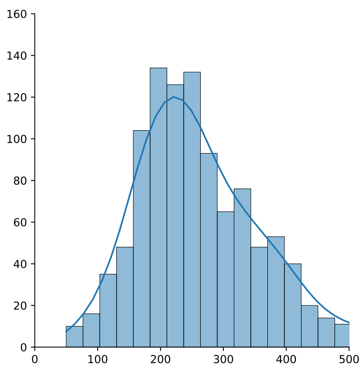


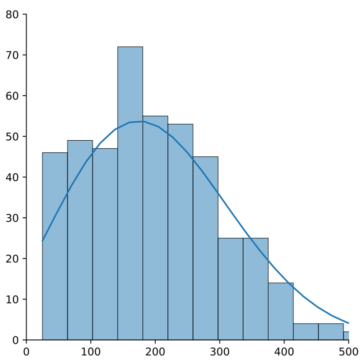
Figure A.3: Results of significance tests on overall system scores in ice-breaker experiment, where a colored cell means that the system in the row outperforms that in the column due to the test; systems and their order follows those in Table 5.4.

Table A.5: System scores of word-overlap-based automatic evaluation metric, where metrics follows the order in Table 5.8.

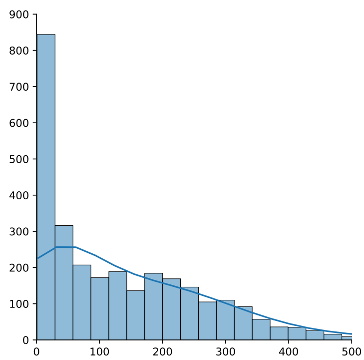
System	METEOR	BLEU-4	ROGUE-L	GLEU	BLEU-4
A	6.50	15.96	12.73	4.30	1.01
B	6.12	14.73	11.96	4.02	0.91
A_p	7.29	16.63	13.72	4.75	1.33
C	5.95	13.32	11.26	3.77	0.80
C_p	6.16	14.31	11.87	3.94	0.90
B_p	7.09	16.75	13.73	4.69	1.34
D	6.50	16.51	15.00	4.71	1.63
D_p	7.27	18.56	16.28	5.89	2.43
E_p	6.68	18.19	14.96	5.63	2.13
E	6.68	18.19	14.96	5.63	2.13



(a) Passed



(b) Failed



(c) ConvAI2

Figure A.4: Characters per conversation from workers who passed quality control (A.4a); failed quality control (A.4b) in our human evaluation; ConvAI2 live evaluation (A.4c).

Table A.6: Positive and negative utterances employed for the FED metric. For criteria that are available in original FED (interesting, consistent, fluent, topic and repetitive), we use their utterances off-the-shelf. In addition, we adapt the utterances for criteria run and robotic.

Criterion	Positive	Negative
Interesting	Wow that is really interesting. That's really interesting! Cool! That sounds super interesting.	That's not very interesting. That's really boring. That was a really boring response.
Fun	Wow that is very fun. Chat with you is enjoyable. You are fun.	That's not very fun. I am not having fun.
Consistent	-	That's not what you said earlier! Stop contradicting yourself!
Fluent	That makes sense! You have a good point.	Is that real English? I'm so confused right now! That makes no sense!
Topic	-	Stop changing the topic so much. Don't change the topic!
Robotic	-	You are robot. You do not sound like a person.
Repetitive	-	Stop saying the same thing repeatedly. Why are you repeating yourself? Stop repeating yourself!

Table A.7: FED scores of different evaluation criteria at system-level.

Table A.8: USR and its sub-metric scores at system-level.

Table A.9: Median numbers of words and characters in conversations and inputs provided by workers who passed quality control; failed quality control in our human evaluation; ConvAI2 live evaluation.

		Passed	Failed	ConvAI2
Characters	Median in an Input	27	22	16
	Median in a Conversation	249	188	105
Words	Median in an Input	8	6	4
	Median in a Conversation	63	48	28

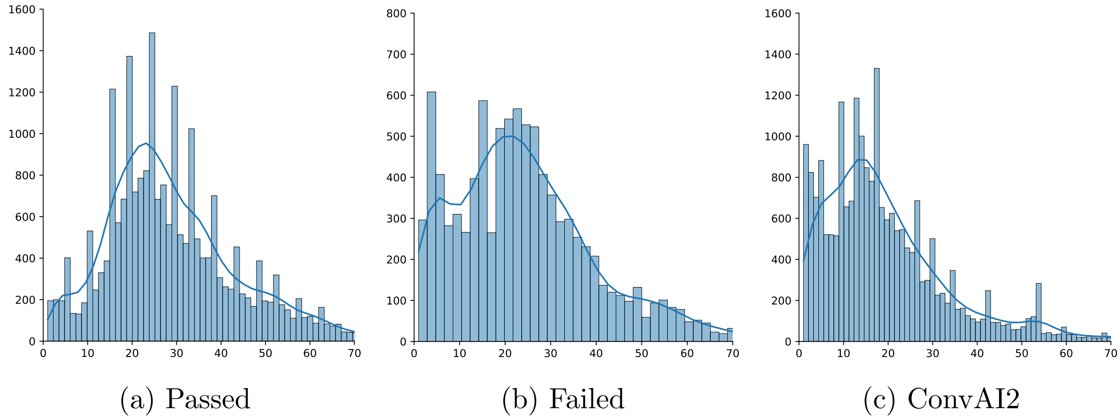


Figure A.5: Characters per input utterance from workers who passed quality control (A.5a); failed quality control (A.5b) in our human evaluation; ConvAI2 live evaluation (A.5c).