It's Common Sense, isn't it? Demystifying Human Evaluations in Commonsense-enhanced NLG systems

Miruna Clinciu[1*], **Dimitra Gkatzia**[2*✉], and **Saad Mahamood**[3*]

[1]Heriot-Watt University, Edinburgh, Scotland, UK
[2]Edinburgh Napier University, Edinburgh, Scotland, UK
[3]trivago N.V., Düsseldorf, Germany
✉ Corresponding author: d.gkatzia@napier.ac.uk

Abstract

Common sense is an integral part of human cognition which allows us to make sound decisions, communicate effectively with others and interpret situations and utterances. Endowing AI systems with commonsense knowledge capabilities will help us get closer to creating systems that exhibit human intelligence. Recent efforts in Natural Language Generation (NLG) have focused on incorporating commonsense knowledge through large-scale pretrained language models or by incorporating external knowledge bases. Such systems exhibit reasoning capabilities without common sense being explicitly encoded in the training set. These systems require careful evaluation, as they incorporate additional resources during training which adds additional sources of errors. Additionally, human evaluation of such systems can have significant variation, making it impossible to compare different systems and define baselines. This paper aims to demystify human evaluations of commonsense-enhanced NLG systems by proposing the *Commonsense Evaluation Card (CEC)*, a set of recommendations for evaluation reporting of commonsense-enhanced NLG systems, underpinned by an extensive analysis of human evaluations reported in the recent literature.

1 Introduction

Commonsense knowledge is vital for human communication, as it helps us make inferences without explicitly mentioning the context. Recently, there has been an interest in developing Natural Language Generation (NLG) systems that exhibit commonsense abilities (e.g. (Lin et al., 2020)). Although everyone understands what common sense is, defining it remains a challenge as it is highly context-dependent. Common sense can be defined as "simple wisdom" (Oxford English Dictionary online), "the ability to use good judgment in making decisions and to live in a reasonable and safe way" (Cambridge dictionary), or as a "sound and prudent judgment based on a simple perception of the situation or facts" (Mirriam Webster). Common sense involves language understanding and reasoning abilities, representing a key factor for establishing effective interactions between humans and machines (Minsky, 1991). In his pioneering work, McCarthy (1959) proposes that "a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows".

Traditionally, commonsense knowledge has been injected in NLG systems either implicitly in the form of rules and/or explicitly with semantic representations in the form of external knowledge bases or ontologies. For instance, expert domain NLG systems (such as the BabyTalk system (Portet et al., 2008)) have incorporated external knowledge in the form of a clinical ontology. In these expert domain NLG systems, knowledge (which might include procedural knowledge) is represented in rules that are built into the system and have been acquired through experts via interviews, observations or other approaches (Reiter et al., 2003). Most recent challenges have focused on injecting commonsense knowledge into neural NLG models in two ways: through pre-trained models and through utilising commonsense graphs or knowledge bases. The former assumes that pre-trained models already contain commonsense knowledge (Petroni et al., 2019). The latter incorporate entity relationships derived from semantic graphs (e.g. ConceptNet (Speer et al., 2016)) or knowledge bases (e.g. (Sydorova et al., 2019)).

It is clear that the incorporation of external knowledge of some form has always been at the heart of NLG system development. In this paper,

[*] Equal Contribution

we are interested in examining how commonsense-enhanced NLG systems are evaluated and whether the accuracy of the underlying commonsense knowledge is assessed by the system creators. To our knowledge, there are no automatic metrics available for commonsense evaluation, and therefore we focus only on human evaluations.

Human evaluation is an area that has received an increasing amount of scrutiny within the wider NLG research community. Previous work has highlighted issues with regards to missing details in evaluations, lack of proper analysis of results obtained, variability in the use of names and definitions of evaluated aspects of output quality (van der Lee et al., 2019; Amidei et al., 2018) and a mismatch on evaluation methods chosen which is correlated with the publication venue rather than the NLG task (Gkatzia and Mahamood, 2015). After examining the last twenty years of human evaluations in NLG, recent survey work has found systemic issues with high levels of diversity of evaluation approaches, inconsistencies and variability in quality criterion names, missing definitions, and fundamental reporting gaps (Howcroft et al., 2020). These issues mean there is a pressing need to better understand the state of human evaluations in other niche areas of NLG such as those systems enhanced with commonsense knowledge.

The contributions of this paper are three-fold: (1) we firstly present an annotated dataset of papers reporting commonsense-enhanced NLG systems published between 2018–2020 in ACL conferences; (2) we present a detailed analysis on human evaluation including reporting on what criteria researchers have most commonly used and whether they have evaluated the underlying commonsense knowledge on its own right and through the generated text; and (3) finally we present the *Commonsense Evaluation Card*, a set of recommendations for human evaluation reporting of commonsense-enhanced NLG systems with the aim to improve not only reproducibility but also improve understanding of such systems.

2 Background

2.1 Commonsense Knowledge in NLG

NLG systems have typically been built with the aim of integrating some form of expertise in their application domain (Jacobs, 1986; Reiter and Dale, 1997). However, as NLG systems find greater general use cases there is a need to incorporate a form of knowledge that is much broader to make up for the differences between human and machine language understanding in decision making, known as common sense (Davis and Marcus, 2015; Lin et al., 2020; Zhang et al., 2020).

The incorporation of commonsense knowledge is considered a challenging task within AI. This challenge is due to the fact that commonsense reasoning or knowledge is considered a black box, as there is uncertainty on how to represent knowledge in order to solve commonsense reasoning problems (Zhang et al., 2020). The reliance on existing knowledge bases to incorporate this type of broad-based knowledge might not be sufficient as it may, in many cases, fail to incorporate explicit fundamental knowledge (Tandon et al., 2018; Ji et al., 2020).

Pre-trained models, on the other hand, have capabilities of learning relational patterns and can achieve commonsense reasoning without explicit knowledge representation, as conveyed in the traditional pipelines (Ji et al., 2020; Vinyals and Le, 2015). However, it remains unclear how the reasoning is performed and how prior knowledge is learned in the training phase (Rajani et al., 2020).

2.2 External Knowledge

In the last few years, several attempts have been made to incorporate commonsense knowledge in NLG systems, using external knowledge bases, such as ConceptNet or Atomic (Bauer et al., 2018; Ji et al., 2020). ConceptNet consists of nearly 120K triples obtained from the Open Mind Commonsense knowledge entries in ConceptNet 5 (Speer and Havasi, 2012) that contains world facts and informal relationships between common concepts that convey some prior knowledge (Zhou et al., 2018). ATOMIC is an atlas of everyday commonsense knowledge and contains 880k triples about causes and effects of human activities and annotated by crowd-sourced workers. ATOMIC is organized as if-then relations and can be categorised based on causal relations (Sap et al., 2019; Guan et al., 2020). COMET is a framework for automatic construction of commonsense knowledge bases, known also as COMmonsense Transformers. This model generates commonsense knowledge based on pre-trained language models (Bosselut et al., 2019). Recent research has also focused on injecting triples into sentences in order to create domain-specific knowledge (Liu et al., 2020; Wang et al., 2020b) or incorporating commonsense knowledge directly in the training data (Huang et al., 2019).

2.3 Pre-trained language models (PTLMs)

An alternative to using explicit external models for commonsense knowledge is the use of PTLMs. Training deep learning models requires extensive amounts of data to prevent over-fitting. This can be problematic for NLG tasks, where collecting and annotating data represents a time-consuming and costly process (Qiu et al., 2020). PTLMs, on the other hand, have the potential to solve the problem of data scarcity, as they do not rely on many resources for training models' parameters.

In the field of NLG, PTLMs have been applied to open-ended non-expert domains, such as question answering, where commonsense knowledge should serve as a link between the performance of these models and human evaluation (Lin et al., 2019). However, transferring commonsense knowledge using PTLMs comes with certain limitations corresponding to each pre-trained model.

PTLMs using domain-specific information from knowledge graphs or unstructured information are highly dependent on the training data quality. For instance, the knowledge extracted from the triples is unable to capture semantic relationships between entities (Zhou et al., 2018; Ji et al., 2020) and solving this can instil commonsense knowledge in NLG systems.

An ongoing discussion about the inherent biases of the training data exposed different types of bias that significantly influence natural language generation systems, such as gender bias, geographical and political bias among others (Papakyriakopoulos et al., 2020). Also, the frequency of the words that influence training data might not correspond to the real-life scenarios and can lead to false facts (Shah et al., 2019). This is also known as "the black sheep problem": when querying a system using GPT−3 to tell the colour of sheep, it will suggest "black" as often as "white", being impossible to distinguish between the linguistic meaning and the visual recognition of "a black sheep" (Gordon and Van Durme, 2013). Solving these issues can represent a first step in building NLG systems that integrate commonsense knowledge.

2.4 Commonsense knowledge evaluation

Understanding commonsense knowledge of natural language text is still a limited task. For humans, it is easy to understand both implicit and explicit meanings of a given sentence, whereas for machines this still remains a challenging task.

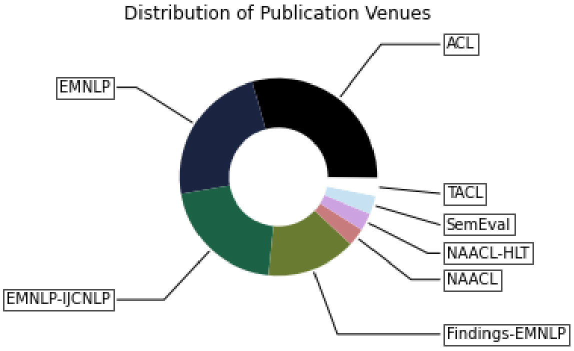

Figure 1: Distribution of publication venues across the commonsense paper dataset.

Due to the uncertainty of defining what implies commonsense knowledge in a natural language text, human evaluation by specialists or lay users might be the only way of providing a more comprehensive evaluation. On the other hand, human evaluation of commonsense knowledge can have some drawbacks as humans may have conflicting opinions and perspectives. In addition, the process of evaluating with humans can be time-consuming and costly.

Many papers report automatic evaluations of pre-trained models for specific commonsense knowledge tasks. However, based on a gold standard, natural language text annotated by humans as correct for a given task may not capture all of the commonsense knowledge nuances.

3 Paper Selection & Annotation

We used the PRISMA method (Moher et al., 2009) to select papers to be included in this study following Howcroft et al. (2020) and (Reiter, 2018). We began by considering all papers published in ACL venues (ACL, CL, CoNLL, EMNLP, Findings, NAACL, SemEval, *SEM, TACL and INLG) in the past three years (2018–2020). We screened the papers using the following search terms (in their title): commonsense, generation, reasoning, domain knowledge, expert, expertise, sensible, ontology, knowledge. This left us with 129 papers. From these, we randomly pick 55 papers that were annotated by the authors of this paper, following the annotation scheme proposed by Howcroft et al. (2020). Papers on commonsense reasoning can either focus on language generation or understanding. For instance, commonsense reasoning can be addressed as a classification task, where based on the context, a reasoning system can choose an option from a set of options (Talmor et al., 2019). During annotation, such papers were omitted.

Following Howcroft et al. (2020), papers were annotated using the three broad categories: (1) **system** attributes (input, output, task and language) which describe evaluated NLG systems, (2) **quality criterion** attributes (Verbatim Criterion Name, Definition and Paraphrase), and (3) **operationalisation** attributes (e.g. type of instruments, type of collected data etc.) which specify how evaluations are performed. In addition to these, we introduced a fourth category, **commonsense knowledge**, with five new annotation items which are relevant for commonsense-enhanced NLG, namely:

- *Definition of commonsense knowledge*: free text field. Here the annotators either copied the definition as provided in the paper or specified ”None”.
- *Type of commonsense knowledge*: free text field. Here the annotators had to specify the type of commonsense knowledge that the paper tried to address, for instance, sarcasm or reasoning about the order of events.
- *External knowledge:* free text field. Examples of external knowledge can include commonsense knowledge bases such as ConceptNet.
- *Was the knowledge evaluated in the generated text? (Yes/No)*: The annotators specified whether the underlying knowledge was evaluated.
- *Criterion name for evaluation of external knowledge*: The annotators could specify the criterion used to evaluate the knowledge base, for instance in terms of coverage or correctness.

These additional items were deemed important to investigate whether there is a relationship between the human evaluation criteria and the type of commonsense knowledge covered by the NLG system. In addition, when evaluating generated text, it is vital to know whether errors in the generated text arise from the underlying data or the text generator.

3.1 Inter-Annotator Agreement

Following (Howcroft et al., 2020), ten papers were annotated by all three annotators and Inter-Annotator Agreement (IAA) was calculated. The papers were randomly selected by proportionally accounting for the year and the publication venue.

Pre-processing: We pre-processed the annotations by normalising capitalisation, spelling and stripping extra spaces. We also removed papers that did not report a system that generates text.

Calculating agreement: The data resulted from the annotation process was a 10 (papers) $\times n$ (evaluation criteria identified by annotator for each paper) $\times 19$ (attribute value pairs) data frame, for each of the annotators. As such, IAA aims to measure the agreement across all annotators given the aforementioned data frames. The agreement was calculated using Krippendorff’s alpha with Jaccard as the distance measure (Artstein and Poesio, 2008).

Results are presented in Table 1. For system attributes (system input, system output and system task) IAA agreement is good, although the score for the system task is lower. The latter might be affected by the multitude of tasks presented in papers, as the evolution of NLG led to the need for proposing different tasks for generating text in new domains. Surprisingly, external knowledge attributes received a low IAA agreement which might indicate that there is vagueness in what constitutes external knowledge. Also, relatively low agreement scores were obtained for the two attributes **elicit form** and **instrument type**. The majority of the papers do not provide enough detail about the operationalisation attributes; our findings are not very different from the ones presented by Howcroft et al. (2020).

Attributes	IAA Test
System Input	0.70
External Knowledge	0.15
System Output	1.00
System task	0.37
Knowledge Evaluation	0.18
Paraphrase	0.39
Elicit form	0.05
Data type	0.25
Instrument type	0.07

Table 1: Krippendorff‘s alpha using Jaccard distance for closed class attributes.

4 Analysis and Results

In this section, we present the results from the analysis of the annotated papers. The annotations and the developed code can be found in the projects’ repository[1].

[1] https://github.com/nlgknowledge/commonsense

VERBATIM CRITERION NAME	Count
fluency	6
coherence	4
informativeness	3
grammaticality, correctness, diversity, appropriateness, accuracy	2
commonsense, topic-consistency, sarcasticness, interpretability, engagement, commonsense plausibility, commonsense reasoning, reasonability, novelty, usefulness, intention, information, naturalness, logicality, humour, relevance, common ground, answerability, plausible, effect, validity, quality, event-centered commonsense reasoning, best-worst scaling, consistency, attribute, creativity, effectiveness	1
mixed: grammatical correctness and fluency	2
none given	3

Table 2: The table presents all verbatim criterion names found in the annotated papers as mentioned by the authors. The only pre-processing applied is lower-casing.

NORMALISED CRITERION NAME	Count
text property	7
fluency	4
goodness of outputs relative to input	4
goodness of outputs relative to input (content)	4
coherence	4
information content of outputs	4
grammaticality	3
correctness of outputs in their own right	2
correctness of outputs relative to input (both form and content)	2
correctness of outputs relative to input (content)	2
naturalness (form)	2
appropriateness (content)	2
Goodness of outputs in their own right	1
Appropriateness	1
Appropriateness (both form and content)	1
Quality of outputs	1
Correctness of outputs relative to external frame of reference (content)	1
Goodness of outputs in their own right (both form and content)	1
Correctness of outputs relative to input	1
35a. Naturalness (both form and content)	1
Goodness of outputs relative to system use	1
Multiple (list all)	1

Table 3: The table presents occurrence counts for normalised criterion names.

The 34 papers in the dataset corresponded to 70 individual evaluations, amounting to 2.05 evaluations per paper. This dataset was annotated between three annotators taking approximately 20 minutes or more to annotate each paper.

In the following subsections we will first report the paper and system level statistics (Section 4.1), followed by evaluation-level statistics for the quality-criterion (Section 4.2), then the operationalisation attributes (Section 4.3), and finally the commonsense criteria findings (Section 4.4).

4.1 Papers and Systems

All the papers analysed reported English as the system language. Only two papers in our dataset reported Chinese as an additional system language to English. All the papers in our dataset were published recently between 2018-2020 with most being published in 2019 (58%). Figure 1 and Appendix A gives a break down of the publication venues for our dataset.

In terms of the system task attribute, our analysis reveals that *question answering* and *dialogue turn generation* are the top two system task types within our dataset. This differs from the findings made by Howcroft et al. (2020) who found that *data-to-text* generation as being the most frequent system task in their analysis leading to 50% more than second-placed *dialogue turn generation*. This difference may indicate that commonsense NLG is more focused on domain problems with direct applicability to general end-users. Appendix B shows the system input, Appendix C for system output, and Appendix D task frequencies in more detail.

4.2 Quality criteria

In this section, we present the results related to the quality criteria, focusing on the *verbatim criterion names* and *the paraphrase of criterion names* based on our annotation. Table 2 shows the verbatim criterion names, as mentioned in the papers by the authors. We found that although most papers mention the quality criterion used for human evaluation a small subset does not. These findings are on par with Howcroft et al. (2020), demonstrating that this is a common issue for NLG. We also found that only a subset of papers define the quality criteria used. The most cited criterion is *fluency*, followed by *coherence*.

We further examined how often the normalised criteria occurred in the annotations as shown in Table 3. Most commonly, the evaluations considered a specific *text property*. The type of properties that evaluations considered are the following: complexity/simplicity (mentioned twice), creativity, novelty, sarcasticness, diversity and humour.

Although there is a lot of variability within one category, it actually shows that commonsense is generally a vague term and it can be interpreted in a plethora of ways and hence it is evaluated differently. Using a text property as an evaluation metric is an interesting finding. In broad human NLG evaluations, this criterion is not very prevalent - in fact, it is one of the rarest criteria. However, other criteria such as *fluency*, *goodness of outputs*, *grammaticality* and *correctness* are equally found in both commonsense-enhanced NLG systems and broad NLG systems (as reported by Howcroft et al. (2020)).

Surprisingly, *commonsense*, *commonsense reasoning* and *commonsense plausibility* have only been named 4 times as criteria in the 34 annotated papers. We would expect to come across criteria names related to commonsense or reasoning more often, as we only examined papers reporting commonsense and reasoning NLG tasks. In Section 4.5, we discuss why this might be the case.

4.3 Operationalisation

Table 4 presents the most frequent forms used for response elicitation. Relative quality estimation was the most frequent form of response elicitation (21 times), followed by direct quality estimation (14 times). Unforeseen, as a reason for not providing enough details of how the evaluation was implemented, in the third place we have the value "unclear" (7 times). The most frequent values for the type of rating scale were numerical rating scale (12 times), rank-ordering (8 times), followed by the Likert scale (7 times).

In addition, nearly half of the investigated papers did not provide a verbatim question/prompt (30 out of 56 evaluation entries). This can be problematic for reproducibility, as results obtained with a different question cannot be directly compared to the original results if the same question hasn't been asked. In addition, this can also hinder the comparability of future work, since, for the same reason, results obtained on new systems cannot be meaningfully compared to previous work. Similar to Howcroft et al. (2020), we also found two cases where *fluency* and *grammaticality* were both mentioned in a question put to evaluators. van der Lee et al. (2021) discuss how this can lead to mixed results as evaluators may put more emphasis on one criterion over the other.

FORM	Count
relative quality estimation	21
direct quality estimation	14
unclear	7
(dis)agreement with quality statement	5
evaluation through post-editing/annotation	4
task performance measurements	2
classification	1

Table 4: Counts of values selected for form of response elicitation.

4.4 Commonsense criteria

The commonsense category includes the criteria defined in Section 3 namely, (1) definition of commonsense; (2) type of commonsense; (3) external knowledge; (4) whether the external knowledge was evaluated; and (5) the criterion name of the external knowledge evaluation.

Definition of Commonsense Unexpectedly, out of the 70 evaluations, only 4 provide a written definition of commonsense with the majority providing no definition whatsoever. Table 5 presents the verbatim definitions from these papers.

DEFINITIONS
"Commonsense reasoning, the ability to make acceptable and logical assumptions about ordinary scenes in our daily life" (Lin et al., 2020).
"Machine common sense, or the knowledge of and ability to reason about an open ended world" (Talmor et al., 2019).
"commonsense evidence is intuitive to humans, the agent's ability to select the right kind of commonsense evidence will allow the human and the agent to come to a common understanding of actions and their justifications, in other words, common ground" (Yang et al., 2018).
"counterfactual reasoning: the ability to predict causal changes in future events given a counterfactual condition applied to the original chain of events" (Qin et al., 2020).

Table 5: Definitions of Commonsense extracted from literature.

Type of commonsense Almost half of the papers did not contain a definition of commonsense neither mentioned the type of commonsense that their task was addressing ($n = 16$). The second most prevalent type of commonsense was reasoning - eight paper reported that the focus of the task is to perform some form of reasoning ($n = 8$). Other types of reported commonsense included temporal and spatial commonsense reasoning, social commonsense, and underlying commonsense abilities such as sarcasm and humour.

External knowledge External knowledge bases are usually incorporated into NLG systems in order to provide commonsense capabilities. As shown in Figure 2, the most used common knowledge base is ConceptNet (13 times), own developed KB most often in the form of triples that describe the connection between entities) (14 times), followed by ATOMIC (5 times), COMET (once) and Cosmos (once). Although pre-trained language models have been shown to encode commonsense knowledge in some situations, we did not consider them here as external knowledge. The most used pre-trained model though is GPT-2.

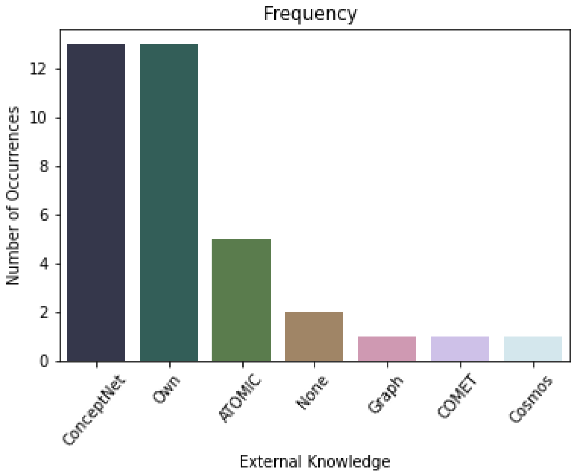

Figure 2: Frequency graph of external knowledge mentions in the commonsense dataset.

Was the external knowledge evaluated? External knowledge was evaluated less than half of the time (14 out of 34). An assumption for this is that authors might consider external knowledge bases such as ConceptNet and ATOMIC accurate and they do not normally evaluate them in their domains. Bauer et al. (2018) argue that even when using a large pre-trained dataset, it might be hard for a model to not only find but also look at the correct relationships between concepts and apply them in reasoning tasks. They further conducted a human evaluation where they report how many cases their system would require external knowledge and in what percentage of these cases, their system selected the relevant/correct commonsense knowledge. From their results, it can be inferred that in a small set of cases, some errors in the generated text can be a result of the underlying erroneously inferred commonsense relationships. Wang et al. (2020a) also report a human evaluation of their commonsense knowledge in terms of validity and relevance, where they also show that the extracted commonsense relationships might contain errors (or be irrelevant). As such, it is clear that there should be a distinction between errors resulting from the text generation models or the external knowledge bases (note that here we have used the term external knowledge bases to refer to any form of external knowledge, including graphs).

Criterion name of external knowledge evaluation External knowledge has been evaluated in a number of ways (the following is not an exhaustive but an indicative list): Bosselut et al. (2019) evaluate whether their model can adequately produce a triple of a subject, object and their relationship in terms of *plausibility*; Wang et al. (2020a) evaluate commonsense knowledge in terms of *validity* (”How valid are the paths?”) and *relevance* (”How relevant are the paths to the question?”); Bauer et al. (2018) evaluated the commonsense relationships between concepts. In other evaluation settings, evaluators are given the top related underlying concepts and are instructed to pick the ones that describe or explain the text better (e.g. (Sydorova et al., 2019)).

4.5 Discussion

From the evidence we gathered through our annotations, there are several key observations. Firstly, only a subset of authors actually provide definitions of the quality criteria used for human evaluations. As Howcroft et al. (2020) found in their survey, there can be a significant mismatch between what authors specify as the quality criterion name and definition provided. Therefore, there is a need for definitions to be included in papers to give readers an unambiguous understanding of the quality criterion being evaluated. Secondly, there is a need to provide complete and accurate information for reproducing the human evaluation. Our analysis has shown that nearly half of the papers did not provide the prompt with the verbatim question/prompt given to the human participants. Thirdly, and finally, our analysis has shown that very few papers investigate the correctness or plausibility of commonsense reasoning in their evaluations with humans.

This analysis has shown the need for better reporting of human evaluations. The low levels of inter-annotating agreement for annotating some of the attributes might be a strong indication of the challenges of how hard it is to locate information about evaluations in a given paper.

Given our experiences, we believe that researcher working on commonsense-enhanced NLG systems should go beyond evaluating their systems using standard NLG quality criteria such as naturalness, grammaticality etc. In addition, researchers should further:

- evaluate the generated text of a commonsense-enhanced NLG system in terms of commonsense or reasoning capabilities in order to verify that the system actually displays commonsense capabilities.
- make an effort to investigate the correctness or plausibility of the commonsense knowledge/reasoning implemented with human assessors. As discussed in Section 4.4, not always the external knowledge is useful and it might even contain erroneous information.

Our analysis has motivated the creation of the *Commonsense Evaluation Card* which serves two roles. It firstly aims to motivate researchers to evaluate their systems in terms of common sense (i.e. are they fit for purpose?) and secondly, it aims to promote better practices and evaluation standardisation by introducing reporting recommendations (i.e. how was the evaluation done?).

5 The Commonsense Evaluation Card

The *Commonsense Evaluation Card (CEC)* (Table 6) aims to standardise human evaluation and reporting of commonsense-enhanced NLG systems, enabling researchers to compare models not only in terms of classic NLG quality criteria, but also by focusing on the core capabilities of such models. CEC has been inspired by recent work on model reporting (Mitchell et al., 2019), datasheets for datasets (Gebru et al., 2018) and The Human Evaluation Datasheet 1.0 (Shimorina and Belz, 2021). It is not designed to replace these, but rather complement them.

CEC includes three main sections: (1) definition of common sense in the context of the reported work and the type of commonsense knowledge; (2) evaluation of the validity of external commonsense knowledge; and (3) evaluation of commonsense knowledge in a generated text.

Commonsense Evaluation Card (CEC)

Commonsense Knowledge Definition: Basic definition of commonsense knowledge in the reported work.
- Definition
- Type of commonsense
- Example output of generated text that displays the intended commonsense capabilities.

External Knowledge: Basic information regarding the use of external knowledge and its evaluation
- Structured Knowledge
- Pre-trained Language Models
- Other
- Metrics for Evaluation of External Knowledge

Commonsense Knowledge in Generated Text: Evaluation Settings
- Automatic Metrics for Evaluation of commonsense knowledge in generated text
- Human Evaluation of commonsense knowledge in generated text

Table 6: Summary of the commonsense evaluation card (CEC).

Next, we describe each of these sections in more details with guidelines on how to complete the evaluation card.

5.1 Definition of Common Sense

This section should answer basic questions regarding the presented work as follows:

How do you define commonsense knowledge in the context of this work? Here, researchers should provide a definition of commonsense knowledge that is relevant to their reported work. Our analysis showed that common sense is hard to define since its definition is highly dependent on the context. Providing a definition of common sense will help researchers better understand the setting in which work was evaluated.

What type of commonsense knowledge do you address? For standardisation reasons, choose one of the following high-level categories: (1) *Commonsense knowledge of entities* in the environment including their properties and the relationship between entities; (2) *Entities interactions and procedural knowledge*; (3) *Figurative language* such as irony, humour, sarcasm, emotion etc; (4) *Causal relationships*, e.g. X will cause Y; (5) *General knowledge* such as facts, e.g. the water boils at 100C; (6) *Reasoning*; or (7) *Other*, not covered by any of the categories above.

Example output of generated text that displays the intended commonsense capabilities: An example of the expected output with an explanation on why this constitutes commonsense knowledge, for instance, the information in the output is not represented in the input.

There are cases where commonsense might refer to more than one of the types mentioned above. The authors can specify more than one types of commonsense or create separate evaluation cards if it is more appropriate.

5.2 External Commonsense Knowledge

This section should provide information regarding external commonsense knowledge bases and their evaluation.

Structured Knowledge: Does the proposed work make any use of an external structured knowledge base such as ConceptNet? If yes, provide details on how to access the knowledge base and its version if public, or alternatively. If the external knowledge base is subjected to privacy concerns or is private, then provide a detailed description.

Pre-trained language models: Does the proposed work make use of any pre-trained language models? If yes, provide a detailed description, such as the version used, the API, hyperparameters etc.

Other: Was commonsense knowledge represented in any other way? How? If none of the above is applicable, explain how the system displays commonsense knowledge. For instance, knowledge might be encoded as rules or it might be inferred from the input training data.

Metrics for Evaluation of External Knowledge : Was the external knowledge evaluated? Describe whether the external knowledge was evaluated and in what way. Essentially this section should answer whether the external knowledge was fit for purpose.

5.3 Commonsense knowledge in generated text

Automatic Metrics for Evaluation of commonsense knowledge in generated text: Provide the metrics and the evaluation details such as the samples used for evaluation.

Human Evaluation of commonsense knowledge in generated text: Does your human evaluation include any metrics specifically related to commonsense knowledge? Provide their definition and include the evaluation details, including a detailed description of the experimental setup, the definition of the metric(s) and the questions asked to participants.

6 Conclusions

This paper presented a human evaluation analysis on works describing systems that incorporate commonsense knowledge or other external knowledge bases with the aim to enhance the reasoning abilities of NLG systems. We have utilised an annotation scheme that has been verified in previous work and we have enhanced it with five additional criteria relevant for commonsense-enhanced NLG systems and we have reported our analysis of the annotations.

Our analysis showed that there is a large variability on how such systems are evaluated, the type of evaluation criteria that are selected and we questioned whether standard NLG criteria are fit for purpose when evaluating reasoning abilities. We have therefore recommended that researchers should evaluate the reasoning ability of their systems (in addition to standard NLG metrics). We did not specify how these evaluations should be performed as this can vary depending on the task. We recommend nevertheless, that authors provide their definition(s) of commonsense knowledge to their evaluators. Additionally, we recommend that researchers validate their external knowledge bases to ensure that any errors present in generated output are not derived from the underlying knowledge.

Finally, as this field grows in the future and attracts further attention, it would be useful to document commonsense knowledge errors in a more structured way, as for instance in (Chen et al., 2019).

Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback. Gkatzia's contribution was supported under the EPSRC projects CiViL (EP/T014598/1) and Natural Language Generation for Low-resource Domains (EP/T024917/1). Clinciu's contribution is supported by the EPSRC Centre for Doctoral Training in Robotics and Autonomous Systems at Heriot-Watt University and the University of Edinburgh. Clinciu's PhD is funded by Schlumberger Cambridge Research Limited (EP/L016834/1, 2018-2021).

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2019. Bidirectional attentive memory networks for question answering over knowledge bases. 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)*, pages 2913–2923, Minneapolis, Minnesota. Association for Computational Linguistics.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 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*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 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)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Paul S. Jacobs. 1986. Knowledge structures for natural language generation. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, page 554–559, USA. Association for Computational Linguistics.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for 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)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT:

Enabling language representation with knowledge graph. In *Proceedings of AAAI 2020*.

John McCarthy. 1959. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*.

Marvin Minsky. 1991. Logical versus analogical or symbolic versus connection or neat versus scruffy. *AI Magazine*, 12(2).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *BMJ*, 339.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

François Portet, Albert Gatt, Jim Hunter, Ehud Reiter, Somayajulu Sripada, and Feng Gao. 2008. BabyTalk: A Core Architecture to Summarise ICU Data as Tailored Text. In *21st International Congress of the European Federation for Medical Informatics (MIE 2008)*, page 1, Göteborg, Sweden.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2020. Counterfactual story reasoning and generation. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Xi Peng Qiu, Tian Xiang Sun, Yi Ge Xu, Yun Fan Shao, Ning Dai, and Xuan Jing Huang. 2020. Pre-trained models for natural language processing: A survey.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Explain Yourself! Leveraging language models for commonsense reasoning. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

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

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*.

Ehud Reiter, Somayajulu G. Sripada, and Roma Robertson. 2003. Acquiring correct knowledge for natural language generation. *J. Artif. Int. Res.*, 18(1):491–516.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview.

Anastasia Shimorina and Anya Belz. 2021. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in nlp.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in concept net 5. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*.

Alona Sydorova, Nina Poerner, and Benjamin Roth. 2019. Interpretable question answering on knowledge bases and text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4943–4951, Florence, Italy. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Niket Tandon, Aparna S. Varde, and Gerard de Melo. 2018. Commonsense Knowledge in Machine Intelligence. *ACM SIGMOD Record*, 46(4).

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model Oriol Vinyals. *ICML Deep Learning Workshop*, 37.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020a. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2020b. K-ADAPTER: Infusing knowledge into pre-trained models with adapters.

Shaohua Yang, Qiaozi Gao, Sari Sadiya, and Joyce Chai. 2018. Commonsense justification for action explanation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2627–2637, Brussels, Belgium. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2018-July.

Appendices

A Publication Venue

VENUE	Total
EMNLP	11
EMNLP-IJCNLP	8
ACL	7
NAACL	5
SemEval	1
TACL	1
NAACL-HLT	1

Table 7: Publication venues for commonsense papers.

B System Input

INPUT TYPE	Total
text:sentence	9
text:multiple sentences	6
raw/structured data	6
text: subsentential units of text	3
visual	2
Others (8 Input Types)	8

Table 8: Types of system inputs for commonsense papers.

C System Output

OUTPUT TYPE	Total
text:sentence	17
text: subsentential units of text	4
text:multiple sentences	3
raw/structured data	2
text: variable-length	2
Others (6 Output Types)	6

Table 9: Types of system outputs for commonsense papers.

D System Task

TASK TYPE	Total
Question Answering	12
Dialogue Turn Generation	7
End-to-End Generation	3
Other: Story Ending Generation	2
Content Selection/Determination	2
Feature-Controlled Generation	2
Others (6 Task Types)	6

Table 10: Types of system tasks for commonsense papers.