# Analysing the Problem of Automatic Evaluation of Language Generation Systems

Analizando el Problema de la Evaluación Automática de los Sistemas de Generación de Lenguaje

> Iván Martínez-Murillo, Paloma Moreda, Elena Lloret Department of Language and Computing System University of Alicante {ivan.martinezmurillo, moreda, elena.lloret}@ua.es

**Abstract:** Automatic text evaluation metrics are widely used to measure the performance of a Natural Language Generation (NLG) system. However, these metrics have several limitations. This article empirically analyses the problem with current evaluation metrics, such as their lack of ability to measure the semantic quality of a text or their high dependence on the texts they are compared against. Additionally, traditional NLG systems are compared against more recent systems based on neural networks. Finally, an experiment with GPT-4 is proposed to determine if it is a reliable source for evaluating the validity of a text. From the results obtained, it can be concluded that with the current automatic metrics, the improvement of neural systems compared to traditional ones is not so significant. On the other hand, if we analyse the qualitative aspects of the texts generated, this improvement is reflected. **Keywords:** Natural Language Generation, evaluation metrics, NLG architectures, language models.

**Resumen:** Las métricas automáticas de evaluación de texto se utilizan ampliamente para medir el rendimiento de un sistema de Generación de Lenguaje Natural (GLN). Sin embargo, estas métricas tienen varias limitaciones. Este artículo propone un estudio empírico donde se analiza el problema que tienen las métricas de evaluación actuales, como la falta capacidad que tienen estos sistemas de medir la calidad semántica de un texto, o la alta dependencia que tienen estas métricas sobre los textos contra los que se comparan. Además, se comparan sistemas de GLN tradicionales contra sistemas más actuales basados en redes neuronales. Finalmente, se propone una experimentación con GPT-4 para determinar si es una fuente fiable para evaluar la calidad de un texto. A partir de los resultados obtenidos, se puede concluir que con las métricas automáticas actuales la mejora de los sistemas neuronales frente a los tradicionales no es tan significativa. En cambio, si se analizan los aspectos cualitativos de los textos generados, si que se refleja esa mejora.

**Palabras clave:** Generación de Lenguaje Natural, métricas de evaluación, arquitecturas de generación, modelos de lenguaje.

#### 1 Introduction

Natural Language Generation (NLG) is a sub-field within the Natural Language Processing (NLP) field that has rapidly evolved in recent years, attracting the interest of the scientific community (Ji et al., 2023). Its evolution and the subsequent advancements have provoked a change in the NLG architectures paradigm, going from architectures that split the generation of language into different sub-tasks to architectures that perform all the generation in just a single task (Gatt and Krahmer, 2018). Two breakthroughs can be highlighted to understand this paradigm shift. On the one hand, the development of deep learning methods improved the state of the art of NLG, generating more coherent and natural text by capturing complex language patterns and context (Gatt and Krahmer, 2018). On the other hand, the proposal of the Transformers architecture (Vaswani et al., 2017) increased considerably the performance of the models, thanks to its selfattention mechanism.

However, these developments have not been reflected in the automatic evaluation of texts. Evaluating the quality of NLG systems' output still remains a challenge (Dong et al., 2023). There is a lack of standard automated evaluation metrics, as the traditional metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), OR SPICE (Anderson et al., 2016) tend not to correlate well with human judgements (Khapra and Sai, 2021). Most of these metrics rely on comparing a candidate sentence to one or several target sentences, based on a feature overlap (i.e., words or fragments). Thus, the resulting score for the candidate sentence largely depends on how it aligns with the reference sentence (Tang et al., 2023).

This issue can be accentuated in some tasks where the output can vary in content and does not need to follow a predefined structure. Therefore, these texts can also be valid, although quite different from the target sentences. For instance, consider the CommonGen task (Lin et al., 2020), where the model is given a tuple of words and a sentence that incorporates all those words must be generated. Then, that generated sentence is compared against a set of target sentences. In this task, there can be various candidate sentences that can be valid, although they differ from the target sentences. Consequently, when relying on classical evaluation metrics, a model that produces such different sentences might be penalised, even though the generated sentence is both syntactically and semantically correct. Figure 1 shows an example of the CommonGen task, for different models.

To address this, researchers have explored the hypothesis that Large Language Models (LLMs) can exhibit strong correlations with human judgements, being a more suitable approach for evaluating text (Tang et al., 2023). In this line, novel metrics based on LLMs have been proposed, including BARTScore (Yuan, Neubig, and Liu, 2021) and GPTScore (Fu et al., 2023). However, despite efforts to find more suitable metrics, the most popular metrics to evaluate text are still the traditional evaluation metrics.

Given the issues that may arise when evaluating NLG models, this paper aims to empirically analyse and compare different evaluation metrics in the context of a relatively recent task, commonsense generation. Comonsense generation is the task of reasoning about the commonsense while generating coherent text. This task can favour the generation of texts that are not only correct but also diverse, as including commonsense in the generation can produce multiple valid outputs while being semantically different (Yu et al., 2022). Specifically, this work focuses on the CommonGen shared task, previously mentioned. Sentences will be generated using different NLG architectures (traditional and recent) and evaluated based on different metrics during the experimentation. Our goal is to address the following three research questions: (1) Which type of NLG architectures perform best for commonsense generation? (2) How good are the generated sentences of the best-performing system from a qualitative perspective? (3) To what extent can LLMs-based tools, such as ChatGPT, be a good alternative or complement to automatic evaluation metrics?

## 2 Related Work

### 2.1 NLG Architectures

The first NLG works date from the decade of 1970 (McDonald, 2010). Since that time, numerous NLG approaches have been proposed, and they can be broadly categorised into three distinct groups according to their architecture (Gatt and Krahmer, 2018):

- Modular architectures: This group of approaches considers language generation as a process of three well-differentiated stages. (1) Macroplanning, which includes all the tasks that select what information should be included in the generated text, (2) Microplanning, which includes all the tasks related to the parsing of that information, selecting how to say the information selected previously and (3) Realisation, which perform the tasks of generating the selected information by applying the correct syntactical and grammatical rules. Reiter proposed the standard architecture of this group (Reiter, 1994), consisting of a sequential pipeline of those previously mentioned stages. Other examples of this architecture can be found in (Mann and Moore, 1981), (Hovy, 1987), (Levelt, 1989), and (Nirenburg, Lesser, and Nyberg, 1989).
- *Planning perspectives:* This group of approaches still considered language generation as a process of different stages, but they needed a smaller number of tasks

Analysing the Problem of Automatic Evaluation of Language Generation Systems



Figure 1: Ilustrative example of the CommonGen task.

to perform the generation. Some examples of approaches within this group are: (Appelt, 1985), (Bateman, 1997),(Koller and Stone, 2007), (Rieser and Lemon, 2009), (Nakatsu and White, 2010) and (Lemon, 2011).

• Global approaches: This group is the most predominant in recent years. They do not distinguish between tasks, performing the generation process in one An important architecture in step. this group is the Transformers proposed by (Vaswani et al., 2017), which significantly improved the performance of the NLP field with the concept of selfattention. Other research works using alternative architectures to Transformers are: Graph Neural Networks (Scarselli et al., 2008), Generative Adversarial Nets (Mirza et al., 2014), Recurrent Neural Networks (Sutskever, Vinyals, and Le, 2014), and Memory Networks (Sukhbaatar et al., 2015).

### 2.2 Commonsense Generation

The task we want to focus on in this empirical analysis is the commonsense generation: therefore, we will briefly describe and contextualise it. LLMs tend to base their predictions on the likelihood of relationships between words, so they lack a fundamental characteristic in human communication, the commonsense. Commonsense knowledge refers to the information that is widely accepted in everyday life (Bhargava and Ng, 2022). Integrating commonsense knowledge in the human language has been recognised as an important and challenging task in the NLG field (Wang et al., 2021), as there is a need to enhance the capability of NLG systems of integrating it in their outputs. Therefore, some shared tasks have been proposed to advance the state of the art of the commonsense generation. In the Aviccenna (Aghahadi and Talebpour, 2022) task, a model is provided with two premises containing a syllogistic relation. The objective is to generate a conclusion that completes that relation. Integrating commonsense in keyword-to-text task have also been studied. For instance, SituatedGen task (Zhang and Wan, 2024) involves generating a pair of contrastive sentences, given a group of concepts that includes temporal or geographical entities. CommonGen (Lin et al., 2020) and  $C^2 Gen$  (Carlsson et al., 2022) tasks consist of generating a logical sentence describing an everyday scenario given a set of words. Additionally, the  $C^2 Gen$  task also gives as input a context to which the generated text has to adhere.

# 2.3 NLG Evaluation Metrics

To evaluate NLG systems is essential to be able to compare and monitor the advancements in the field. While human evaluators would be the most accurate, this type of evaluation is usually impractical due to the temporal and economic cost it involves. Consequently, automatic evaluation metrics can be a viable alternative (Khapra and Sai, 2021). In that line, rule-based metrics have been used for many years, and widely adopted for evaluating different NLG tasks (Sai, Mohankumar, and Khapra, 2022). These metrics tend to evaluate the quality of a text by comparing it against a reference text based on features such as words, characters or embedding. Word-based metrics are the most employed, measuring the word overlapping between a candidate sentence and a reference sentence. This group include metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh, 2015), or SPICE (Anderson et

al., 2016). Character-based metrics tend to align better with morphologically rich languages comparing the characters of a text instead of the words. Extended Edit Distance (Stanchev, Wang, and Ney, 2019) or chrF (Popović, 2015) are metrics from this group. Finally, embedding-based metrics capture better the semantic similarity in some contexts, by comparing the similarity between the embeddings. Some metrics within this group are: MEANT (Lo, Tumuluru, and Wu, 2012), BERTscore (Zhang et al., 2020), and Word Mover-Distance (Kusner et al., 2015). Other recent types of approaches have studied to use of LLMs to evaluate semantic similarity, usually with a higher correlation with human evaluation from a deep semantic perspective. Some works have explored the accuracy of using ChatGPT as an NLG evaluator (Wang et al., 2023). Inside this group of approaches some metrics have been proposed, e.g. BARTScore (Yuan, Neubig, and Liu, 2021) and GPTScore (Fu et al., 2023) that use BART or GPT models to evaluate the generated text.

### 3 CommonGen Task

We are going to compare the performance of different NLG architectures and study the suitability of some state-of-the-art metrics in the *CommonGen* task<sup>1</sup> (Lin et al., 2020). The objective is to create a coherent sentence describing an everyday situation using a set of given concepts. This task assesses both the capacity to apply commonsense knowledge and the ability to generalise when working with unfamiliar combinations of concepts.

CommonGen released a dataset to address this task, split into three subsets: training, validation and testing. For every set of concepts, several target sentences are provided. Table 1 shows the total number of concept sets and target sentences for each subset.

Corpus	Train	Dev	Test
Concept sets	32 651	993	1497 N/A
Target sentences	67 389	4018	N/A

Table 1: Commongen Dataset distribution.

To train our models we used the train subset. As reference sentences in the test subset are not available, to test the models' performance we used the evaluation subset.

## 4 Experimental Setup

To analyse the efficacy of diverse NLG automatic evaluation metrics, we have formulated a meticulously controlled scenario where the generated text can be evaluated. This scenario centres around addressing the research questions outlined in Section 1.

### 4.1 NLG Models

We wanted to compare the performance of classical architectures against the recent architectures; thus, we selected three different models (one modular architecture and two global approaches) to conduct the experimentation. We omitted to explore planning perspective architectures, which share similarities with modular architectures. The key distinction lies in the varying number of steps addressed during generation.

The first model we tested is  $SimpleNLG^2$ (Gatt and Reiter, 2009), which handles the final step of language generation in the traditional modular NLG architecture, surface realisation. It is one of the most popular traditional NLG systems, and due to its popularity, it has been adapted to other languages, such as German (Braun et al., 2019), Mandarin (Chen, van Deemter, and Lin, 2018), Spanish (Ramos-Soto, Janeiro-Gallardo, and Bugarín, 2017) or Galician (Cascallar-Fuentes, Ramos-Soto, and Bugarín Diz, 2018). Since we already knew what information to incorporate into the final text (specifically, the concepts among the concept set), we followed an overgeneration and ranking strategy to obtain the syntactic representation. We first systematically generated all the possible sentence combinations of the three concepts. Then, these combinations were passed to SimpleNLG to perform the generation. Finally, we evaluated and ranked the resulting sentences according to Rouge\_L (Lin, 2004) against the target sentences on the evaluation set to obtain the most suitable candidate sentence with a higher score.

The second model we used was trained using a fine-tuning strategy from a T5 model (Raffel et al., 2020). T5 is a pre-trained neural model that follows an encoder-decoder

 $<sup>^{1}{\</sup>rm This}$  task can be accessed on https://inklab.usc.edu/CommonGen/index.html.

<sup>&</sup>lt;sup>2</sup>This API is available at https://github.com/simplenlg/simplenlg.

architecture. This model can perform well on various NLG tasks, by adding a prefix to the input. Furthermore, this model has shown great results on concept-to-text tasks (Roos, 2022), similar to the *CommonGen* task. We fine-tuned a T5-small pre-trained model, adding the tag "*CommonGen*:" before every input concept-set. Table 2 shows the hyperparameter configuration we used to perform the training.

Parameters	Values
Number train epochs	4
Batch size	16
Dropout	0.1
Learning rate	1e - 4
Weight decay	0

Table 2: T5 fine-tuning hyperparameters.

The last model tested we was RMT (Zhang et al., 2023). We reproduced the model available at https://github.com/littlehacker26/ Residual\_Memory\_Transformer because it showed promising results in addressing the CommonGen task and was publicly avail-This model passes the probabilities able. of a GPT-2 decoder (Radford et al., 2019) to an encoder-decoder architecture with the novelty that this architecture incorporates three attention layers in the decoder, obtaining the last hidden states from the GPT-2 decoder, and the RMT encoder.

#### 4.2 Metrics

We used different metrics to measure the performance of the aforementioned models. On the one hand, as CommonGen is a shared task with a leaderboard<sup>3</sup>, we employed the same metrics as the leaderboard. This allowed us to somehow compare our results with those obtained by other participants. However, it is important to note that competitors' results were evaluated using the test subset (not publicly available), whereas the results from our experimentation are based on the evaluation subset, so a direct comparison is not possible. Those metrics are:

• *BLEU* (Papineni et al., 2002): This metric is commonly employed for machine translation tasks. It quantifies the word overlap between a candidate sentence and a target sentence, resulting in

a score between 0 and 1. A higher value indicates greater similarity between the candidate sentence and the target sentence. In the *CommonGen* task, generated sentences were evaluated using BLEU\_4. That means that it is calculated the precision of the word overlapping performs at a 4-gram level. This metric is configurable, making it possible to calculate the precision at 3-gram, 2-gram, and 1-gram level.

- *CIDEr* (Vedantam, Lawrence Zitnick, and Parikh, 2015): This metric was proposed to address the evaluation of image captioning<sup>4</sup>. Specifically, it evaluates the agreement between a candidate and a target sentence. To do so, first, perform the stemming of all words for both candidate and target sentence. Then, it measures the co-existence frequency of n-grams for both sentences, computing the weight for each n-gram using the Term Frequency Inverse Document Frequency (TF-IDF). Finally, it combines the scores of the different n-grams.
- SPICE (Anderson et al., 2016): This metric is commonly used on image captioning tasks as well. It measures the similarity between two sentences using the scene graph tuples parsed from the candidate sentence and the target sentences. Spice is computed based on the F1-Score between the tuples of candidate sentences and target sentences.

Moreover, to have a wider vision of the obtained results, we employed several additional metrics. These are:

• Cosine Similarity: This measurement quantifies the similarity between two non-zero vectors in an inner product space (Han, Kamber, and Pei, 2012). To do so, Candidate and target sentences are converted into a vector. Then, the cosine of the angle between these vectors is applied, resulting in a value ranging from -1 (indicating opposite directions)

 $<sup>\</sup>label{eq:available} \begin{array}{ccc} ^{3}\mbox{Leaderboard} & \mbox{is} & \mbox{available} & \mbox{at} \\ \mbox{https://inklab.usc.edu/CommonGen/leaderboard.html.} \end{array}$ 

<sup>&</sup>lt;sup>4</sup>Target sentences from the CommonGen dataset are extracted from image captioning datasets, and therefore, they consider that is more convenient to use metrics focused on image captioning tasks as they usually assume system generations and human references use similar concepts, and thus focus on evaluate the associations between mentioned concepts.

to 1 (representing highly similar directions). This metric can lead to more accurate results when comparing texts of varying sizes because it considers the angle between vectors in a dimensional space. (Guo, 2022) validates the suitability of cosine similarity to address the textual similarity evaluation.

- *ROUGE* (Lin, 2004): Originally, this metric was proposed to address text summarisation tasks. Despite its nature, ROUGE is one of the most employed metrics in NLG to measure the n-gram lexical overlap between the candidate and target sentences (Zhu and Bhat, 2020). ROUGE calculates the recall score of the candidate sentence corresponding to the target sentence. Specifically, Rouge-L identifies the longest cooccurring in sentence n-grams.
- Flesch Reading Ease (Kincaid et al., 1975):The evaluation criterion consists of determining how easy something is to read. Specifically, shorter words and shorter sentences will be easier to read. In the CommonGen task generated sentences should be dramatically simple and describe an everyday Thus, generating sentences scenario. should be easy to read. This metric produces a value between 0 and 121, being easier to read the higher the value is. In our experimentation, the results obtained within this metric are normalised in a range from 0 to 1.
- *BERTScore* (Zhang et al., 2020): This metric evaluates the quality of a text by performing the sum of cosine similarities between candidate and target sentences. Specifically, BERTScore obtains the contextual embeddings of those candidates and target sentences from BERT and calculates the cosine similarity across their tokens. This metric shows to correlate better with human judgements on sentence-level evaluation.

#### 5 Results and Discussion

Within this section, we will expose the results obtained in our experimentation, analysing them in detail.

### 5.1 CommonGen Leaderboard Results

Firstly, to provide a brief background of the overall results obtained in the Common-Gen task, we selected some of the models participating in it. Specifically, the following models are selected to report their performance according to their official results<sup>5</sup>: The best-performing model of the competition, DKMR2 (He et al., 2022). The worstperforming model of the competition, a finetuning of a T5-base model (Raffel et al., 2020). A model with intermediate results, a fine-tuning of a T5-large (Raffel et al., 2020). Table 3 shows the results obtained by these models on the test set.

Model	SPICE	CIDEr	BLEU_4
DKMR2 T5-Large T5-Base	$0.5243 \\ 0.2885 \\ 0.1987$	$\begin{array}{c} 0.3764 \\ 0.1512 \\ 0.0940 \end{array}$	$0.4649 \\ 0.3196 \\ 0.1854$

Table	3.	CommonGer	leader	hoard	results
rable	J.	CommonGer	i ieauei	Duaru	results.

Although DKMR2 is the top-performing model, it achieves discrete results across these metrics. Its highest score, 0.5243, is in the SPICE metric. However, DKMR2 significantly outperforms both T5-Large and T5-Base, nearly doubling their performance in all three metrics (SPICE, CIDEr, and BLEU).

Another aspect that needs to be taken into account is that these metrics seem to be aligned with the expected results achieved by these models. T5-Base obtains lower results compared to its larger version, T5-Large. This means that metrics are performing as expected for the entire test set.

### 5.2 NLG Architectures Results

This subsection aims to answer the research question: Which type of NLG architectures perform best for commonsense generation? To test the metrics evaluation effectiveness, we trained and reproduced the models explained in Section 4.1. With the models outlined in Section 4.1 we generated the sentences for the whole evaluation set, as that set contained a collection of target sentences to compare with. Table 4 shows the results obtained for the different metrics studied in this research work.

<sup>&</sup>lt;sup>5</sup>https://inklab.usc.edu/CommonGen/.

Model	SPICE	CIDEr	BLEU_1	ROUGE-L	Readability	Cosine	BERTScore
SimpleNLG T5-Small RMT	$\begin{array}{c} 0.183 \\ 0.256 \\ 0.215 \end{array}$	$\begin{array}{c} 0.059 \\ 0.109 \\ 0.024 \end{array}$	$0.129 \\ 0.600 \\ 0.371$	$\begin{array}{c} 0.310 \\ 0.444 \\ 0.345 \end{array}$	$0.139 \\ 0.215 \\ 0.343$	$\begin{array}{c} 0.374 \\ 0.287 \\ 0.302 \end{array}$	$0.900 \\ 0.914 \\ 0.889$

Table 4: Results obtained by the experimented models.

T5-Small achieves the best results of the three tested models with the metrics used in the competition, SPICE, CIDEr, and BLEU. They are based on a word, or character overlapping among the candidate and the reference texts. In the same line, BERTScore and ROUGE-L also measure the embedding and word overlapping of both texts respectively. Consequently, they produce the same score, being T5-Small the best-performing model. Nonetheless, when comparing the cosine similarity of the target text against reference texts, SimpleNLG achieves a better score, as this metric does not penalise the length difference between texts. That means a shorter sentence could obtain a good score in this metric. Furthermore, the best score on the readability metric is obtained by the RMT model. That could indicate that sentences generated by RMT use a simpler vocabulary.

Otherwise, although the testing set was different from the test set evaluated in the competition, the results achieved by our experimented models are far from the best-performing model in the competition, DKMR2. However, when comparing with the other two selected models (T5-Base and T5-Large), the results are not as far. T5-Small scores 0.256 and 0.109 in the SPICE and CIDEr metrics respectively, while T5-Large obtains 0.288 and 0.151 in these met-That indicates, that the results are rics. aligned with what is expected, being a little superior to the results obtained by a larger model. Differently, the results obtained by SimpleNLG —a traditional surface realisation engine (0.183 and 0.059 in SPICE and)CIDEr respectively), are relatively close to the results obtained by T5-Base (0.1987 and0.0940), a model that follows a Transformer architecture, being SimpleNLG easier to use and less costly to train than T5.

In general, neural models (T5-Small and RMT) have performed better than a traditional architecture (SimpleNLG). Among both neural models, the best results are achieved by the T5-Small model, which according to our experiments and results, may be the most appropriate model to use for commonsense generation. Nevertheless, the improvement just by using these metrics is not as significant as the computing expense these models consume against a traditional architecture. Notwithstanding, analysing the performance of these models just by evaluating the generated sentences globally with automatic metrics might be not enough to be able to compare that improvement. Therefore, we will show a more detailed analysis.

### 5.3 Qualitative Analysis

Within this section, the research question we want to answer is: "How good are the generated sentences of the best-performing system from a qualitative perspective?". For this, we conducted a manual analysis of the entire set of sentences generated by SimpleNLG, T5-Small, and RMT models. Sentences generated with SimpleNLG tend to be a short combination of three concepts. In contrast, sentences generated by T5-Small and RMT are syntactically correct sentences, but sometimes they are semantically not accurate. However, they often repeat information within the same sentence.

To analyse those results deeply, we selected several examples of the generated sentences for different concept sets to make a detailed qualitative and human analysis. This selection was made based on the model with better results from the previous quantitative analysis, conducted in Section 5.2, i.e., T5-Small. To illustrate its qualitative performance, we selected a semantically correct sentence, a sentence that contained a fragment that was nonsensical, and a sentence which has no semantically correct fragments. Then, we selected the corresponding sentences of the same concept set in the other two models, SimpleNLG and RMT. Figure 2 shows the selected concept set and the sentence each model has generated for that collection of words.

As can be seen, the generated sentences by SimpleNLG are short and formed just by combining the words of the concept set.



Figure 2: Results obtained for selected tuples.

These sentences are neither semantically nor syntactically correct, and a major part of the generated sentences by this model are non-In contrast, sentences generated sensical. by T5-Small are syntactically well formed and in most cases, they include commonsense knowledge. However, there are fragments in those sentences that are semantically incorrect, as the example "A man is eating ameal and preparing it for the meal." that the first part of the sentence, (a man is eating a meal) do not correlate well with the second part (preparing it for the meal). Finally, sentences generated by RMT are also generally syntactically correct. Even so, they tend to repeat already said information in the generated sentence, e.g. the sentence "The man standing in the pocket is making a hand gesture with his left hand and stands in a pocket." repeats the information that the man is standing in a pocket. Moreover, some of these sentences are also nonsensical, such as the fragment "qives him a look of lies".

Table 5 shows the results obtained in the NLG evaluation metrics for the aforementioned generated sentences. For the first concept set, "*[bed, lie, look]*", the model that better results achieves in most metrics is T5, as the sentence generated by it is better formed in terms of semantic and syntactic quality. Nevertheless, RMT and SimpleNLG sentences obtain better scores than T5 on the SPICE and BERTScore metrics, despite they are nonsensical. This suggests that these metrics may not be fully effective for evaluating this sentence.

Results obtained for the second concept set, "[hand, pocket, stand]" reveal a divergence among the generated sentences. While SPICE, BLEU\_1, and ROUGE\_L favour the sentence produced by T5, Cosine Similarity, readability and BERTScore indicate that the RMT-generated sentence is better. Upon closer analysis, both sentences exhibit correct syntax, but the sentence generated by RMT redundantly reiterates information already present in the sentence. So, this sentence is considered semantically worse than the sentence generated by T5.

Finally, all metrics suggest that the sentence generated by RMT achieves the best performance on the third concept set, "/eat, prepare, meal.". Nonetheless, all the sentences are semantically inaccurate. SimpleNLG's generated sentence comment that a meal prepares to eat when that is not possible. In this respect, it is worth noting that sentences generated by SimpleNLG were configured to be in present tense only for simplicity reasons, as it would be very difficult to know a priori which verb tense would be more appropriate for each sentence for each given set of concepts. However, if this information was known, it would be possible to configure SimpleNLG to adjust the sentence to a specific tense, so in this example, the same sentence put in passive voice "Meal is prepared to eat" would be correct. T5's

Concept set	Model	SPICE	BLEU_1	Rouge-L	Readability	Cosine	BERTScore
[bed, lie, look]	SimpleNLG T5-Small RMT	$\begin{array}{c} 0.133 \\ 0.105 \\ 0.154 \end{array}$	$0.264 \\ 0.500 \\ 0.391$	$\begin{array}{c} 0.306 \\ 0.405 \\ 0.311 \end{array}$	$0.167 \\ 0.232 \\ 0.159$	$\begin{array}{c} 0.293 \\ 0.348 \\ 0.254 \end{array}$	$0.896 \\ 0.893 \\ 0.895$
[hand,	SimpleNLG	0.095	0.189	0.280	0.226	0.323	0.876
$\mathbf{pocket},$	T5-Small	0.133	0.471	0.389	0.265	0.322	0.905
$\mathbf{stand}]$	RMT	0.129	0.450	0.310	0.315	0.368	0.908
[eat,	SimpleNLG	0.143	0.368	0.336	0.226	0.374	0.909
meal,	T5-Small	0.105	0.429	0.344	0.240	0.253	0.921
prepare]	RMT	0.211	0.500	0.482	0.366	0.368	0.928

Table 5: Results obtained for the selected concept sets.

sentence contains the actions of eating and preparing a meal at the same time when that action can not be taken simultaneously. Finally, RMT's sentence mentions that a man is eating a meal prepared for another group of people. This sentence is not completely semantically wrong but is difficult to understand. It also repeats some information, the action of preparing to eat, so this issue affects the quantitative results obtained by this sentence that may increase. One aspect to remark on within this concept set is that a confusing sentence such as the one generated by RMT, obtains the best score on the readability score, suggesting that this metric may be not accurate in this generation context.

Generally speaking, since almost all the evaluation metrics highly rely on the target sentences with which the generated sentences are compared, the best option to be confident in the quantitative results achieved by a model is to compare different metrics and check the correlation among them, not forgetting the qualitative evaluation of the generated sentences as well. As shown in our empirical analysis, T5-Small, the model with better quantitative results, generated syntactically correct sentences which sometimes contained semantically wrong fragments.

#### 5.4 ChatGPT as a Text Evaluator

To debate the research question "To what extent can LLMs-based tools, such as Chat-GPT, be a good alternative or complement to automatic evaluation metrics?" we conducted a preliminary experimentation using the same concept sets as in Section 5.3.

We prompted GPT-4 model (OpenAI, 2023) via Bing interface<sup>6</sup> the following text: "Given the following three sentences, which one does make sense?". After that, the gen-

erated sentences are written. GPT-4 gives as output a detailed explanation of why each of the sentences makes sense or not, and argues which sentence is the most appropriate.

Figure 3 shows the generated output for the sentences generated with the concept set *[bed, lie, look]*. GPT-4 argues that the sentence generated with SimpleNLG is not grammatically correct, correlating with human judgement. For the sentences generated with T5-Small and RMT respectively, it says that they are grammatically correct. However, analysing the RMT sentence, the fragment "gives him a look of lies" is not semantically accurate. Despite this, GPT-4 concludes that the third sentence including that fragment is the best choice.

Let's analyze the three sentences:
1. "Lie looks bed."
<ul> <li>This sentence is not grammatically correct. It lacks a subject and verb. It seems to be a fragment rather than a complete thought.</li> </ul>
<ul> <li>It does not make sense in standard English<sup>■</sup>.</li> </ul>
2. "A woman is lying on a bed and looking at the camera."
<ul> <li>This sentence is grammatically correct and clear. It describes a woman who is both reclining on a bed and directing her gaze toward the camera.</li> </ul>
<ul> <li>It makes sense and provides a coherent image of the situation</li> </ul>
${\mathfrak 3}.$ "The baby lies on a bed and looks at the camera while his mother looks on and gives him a look of lies."
<ul> <li>This sentence is quite complex and intriguing. It describes a baby lying on a bed, observing the camera, while the mother watches and conveys a mysterious expression.</li> </ul>
<ul> <li>The phrase "look of lies" adds an element of intrigue and curiosity, suggesting that there might be more to the story.</li> </ul>
<ul> <li>While unconventional, it does make sense in context and invites the reader to ponder the mother's intentions<sup>1</sup>.</li> </ul>
Therefore, the third sentence is the most imaginative and thought-provoking choice. 😂

Figure 3: Generated sentence for the concept set [bed, lie, look].

The output generated for the second concept set ([hand, pocket, stand]) can be seen in Figure 4. In this output, GPT-4 contradicts itself on the action of "a man standing in a pocket". While the second sentence says that this action is not correct (it is metaphorical), it argues that the third sentence is well-constructed and clear. More-

 $<sup>^{6}\</sup>mathrm{GPT}\text{-}4$  can be accessed at https://www.bing.com/chat

over, GPT-4 concludes that the third sentence is the most natural choice, when it repeats the metaphorical action of "standing in a pocket" twice. At last, the concept set

Certainly! Let's a	nalyze the three sentences:
1. "Stand hand	ls pocket."
<ul> <li>This sent than a co</li> </ul>	ence is not grammatically correct. It lacks a subject and verb. It seems to be a fragment rather implete thought.
<ul> <li>It does not</li> </ul>	ot make sense in standard English <sup>1</sup> .
2. "A man stan	ding in a pocket with his hand in the middle of a pocket."
<ul> <li>This sent pocket (v</li> </ul>	ence is grammatically correct, but it's quite unusual. It describes a man standing inside a which is metaphorical and not literal) while having his hand positioned within the pocket.
<ul> <li>While it's specific f</li> </ul>	technically correct, it's not a common expression, and the context would need to be very or it to make sense $^{1}$ .
3. "The man st pocket."	anding in the pocket is making a hand gesture with his left hand and stands in a
<ul> <li>This sent making a</li> </ul>	ence is well-constructed and clear. It describes a man who is both standing in a pocket and hand gesture with his left hand.
<ul> <li>It makes</li> </ul>	sense and provides a coherent image of the situation $^{lacksymbol{1}}$ .
Therefore, the th	ird sentence is the most sensible and natural choice. 😂

Figure 4: Generated sentence for the concept set [hand, pocket, stand].

*[eat, meal, prepare]*, although being accurate in the conclusion that the last sentence is the most appropriate, fails at evaluating the second sentence. GPT-4 says that this sentence makes sense. But eating a meal while preparing it, is not physically possible, so that reasoning is not valid. As we have seen, GPT-4

Let's a	inalyze the three sentences:
1. <b>"M</b>	leal prepares eat."
٥	This sentence is not grammatically correct. It lacks a subject and verb. It seems to be a fragment rather than a complete thought.
0	It does not make sense in standard English 1.
2. <b>"A</b>	man is eating a meal and preparing it for the meal."
٥	This sentence is grammatically correct and clear. It describes a man who is both eating a meal and getting it ready for consumption.
0	It makes sense and provides a coherent image of the situation <sup>1</sup> .
3. <b>"T</b>	he man is preparing to eat a meal prepared for a group of people preparing to go out to eat."
0	This sentence is quite complex, but it is logically consistent. It describes a man who is about to consume a meal that has been prepared for a larger group of people who are themselves getting ready to dine out.
0	While it may seem convoluted, it does make sense in context 1.
Theref	fore, the third sentence is the most sensible and comprehensive choice. 😂

Figure 5: Generated sentence for the concept set [eat, meal, prepare].

explains why a sentence is valid or not. Nevertheless, that explanation is not always correlated with human judgements, and neither is accurate with its explanations. Therefore there is still room for improvement in the way LLMs such as GPT-4 evaluate the text appropriateness.

#### 6 Conclusions and Future Work

This paper presented an empirical analysis of several NLG evaluation metrics and models for the commonsense generation task. The experiments conducted and the discussion of the results led us to conclude that current evaluation metrics highly depend on the set of target sentences a text is compared with.

As we have seen, the alignment between automated metrics and human evaluation is not always accurate. When comparing the performance of different models, the results obtained by SimpleNLG are similar to the results achieved by the other two models. However, it is important to note that, when analysing these sentences manually, the sentences generated by SimpleNLG are just a combination of concepts, often resulting in nonsensical and not syntactically correct sentences, while the sentences generated by the other two models are at least syntactically correct. This difference is not reflected in the results obtained in these metrics. This may indicate that using words that are contained in the target sentences in a random order, can produce similar results to syntactically correct sentences using these metrics.

Another important point to consider when evaluating using the standard NLG evaluation metrics is repetition. In natural language, repeating the same information can be redundant, making that sentence of inferior quality to one that avoids the repetition. Nevertheless, in the context of these metrics, the repetition not only fails to penalise the results obtained but raises the results obtained. Furthermore, we have seen that some sentences that are not semantically correct obtain better results than sentences that are semantically accurate, as most used metrics do not evaluate the semantic information.

Finally, evaluating a model with several metrics is more accurate than doing so with only one metric, as it gives a wider vision of how that model is performing.

One future line of work is to expand the preliminary analysis of LLMs as text evaluators, exploring different available LLMs, as they can capture and learn human patterns from a semantic perspective. Therefore, these models can indeed learn such patterns, they may serve as valuable tools for evaluating sentence correctness. Additionally, it would be interesting to explore how to enhance the performance of our top-performing model, T5-Small, also by incorporating commonsense knowledge into the model.

### Acknowledgements

The research work conducted is part of the R&D projects "CORTEX: Con-Generation" scious Text (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by "ERDF A way of making Europe": "CLEAR.TEXT:Enhancing the modernization public sector organizations deploying Natural Language by Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 "European Union NextGenerand ationEU/PRTR": and the project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21)" funded by the Generalitat Valenciana. Moreover, it has been also partially funded by the European Commission ICT COST Action "Multi-task, Multilingual, Multi-modal Language Generation" (CA18231).

## References

- Aghahadi, Z. and A. Talebpour. 2022. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning. *Journal of Applied Non-Classical Logics*, 32(1):55–71.
- Anderson, P., B. Fernando, M. Johnson, and S. Gould. 2016. SPICE: Semantic propositional image caption evaluation. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, pages 382–398. Springer.
- Appelt, D. 1985. Planning english sentences. cambridge university press.
- Banerjee, S. and A. Lavie. 2005. ME-TEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.
- Bateman, J. A. 1997. Enabling technology for multilingual natural language generation: the KPML development environment. *Natural Language Engineering*, 3(1):15–55.

- Bhargava, P. and V. Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12317–12325.
- Braun, D., K. Klimt, D. Schneider, and F. Matthes. 2019. SimpleNLG-DE: Adapting SimpleNLG 4 to German. In Proceedings of the 12th International Conference on Natural Language Generation, pages 415–420, Tokyo, Japan, October– November. Association for Computational Linguistics.
- Carlsson, F., J. Öhman, F. Liu, S. Verlinden, J. Nivre, and M. Sahlgren. 2022. Finegrained controllable text generation using non-residual prompting. In *Proceedings of* the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6837–6857.
- Cascallar-Fuentes, A., A. Ramos-Soto, and A. Bugarín Diz. 2018. Adapting SimpleNLG to Galician language. In Proceedings of the 11th International Conference on Natural Language Generation, pages 67–72, Tilburg University, The Netherlands, November. Association for Computational Linguistics.
- Chen, G., K. van Deemter, and C. Lin. 2018. SimpleNLG-ZH: a linguistic realisation engine for Mandarin. In Proceedings of the 11th International Conference on Natural Language Generation, pages 57– 66, Tilburg University, The Netherlands, November. Association for Computational Linguistics.
- Dong, C., Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang. 2023. A survey of natural language generation. ACM Computing Surveys, 55:1–38, 8.
- Fu, J., S.-K. Ng, Z. Jiang, and P. Liu. 2023. GPTScore: Evaluate as you desire. arXiv preprint arXiv:2302.04166.
- Gatt, A. and E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelli*gence Research, 61:65–170.
- Gatt, A. and E. Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European*

workshop on natural language generation (ENLG 2009), pages 90–93.

- Guo, K. 2022. Testing and validating the cosine similarity measure for textual analysis. Available at SSRN 4258463.
- Han, J., M. Kamber, and J. Pei. 2012. 2 getting to know your data. In *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Boston, third edition edition, pages 39–82.
- He, X., Y. Gong, A.-L. Jin, W. Qi, H. Zhang,
  J. Jiao, B. Zhou, B. Cheng, S. Yiu, and
  N. Duan. 2022. Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 839–852, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.
- Hovy, E. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.
- Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38.
- Khapra, M. M. and A. B. Sai. 2021. A tutorial on evaluation metrics used in natural language generation. NAACL-HLT 2021
  2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Tutorials, pages 15–19.
- Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Millington (Tenn.)*.
- Koller, A. and M. Stone. 2007. Sentence generation as a planning problem. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 336–343, Prague, Czech Republic, June. Association for Computational Linguistics.

- Kusner, M., Y. Sun, N. Kolkin, and K. Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957– 966. PMLR.
- Lemon, O. 2011. Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. Computer Speech & Language, 25(2):210–221.
- Levelt, W. 1989. Speaking: From intention to articulation MIT press. *Cambridge*, *MA*.
- Lin, B. Y., W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. 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, November. Association for Computational Linguistics.
- Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text* summarization branches out, pages 74–81.
- Lo, C.-k., A. K. Tumuluru, and D. Wu. 2012. Fully automatic semantic MT evaluation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 243–252.
- Mann, W. C. and J. A. Moore. 1981. Computer generation of multiparagraph english text. *American Journal of Computational Linguistics*, 7(1):17–29.
- McDonald, D. D. 2010. Natural language generation. *Handbook of natural language* processing, 2:121–144.
- Mirza, M., B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, I. J. Goodfellow, and J. Pouget-Abadie. 2014. Generative adversarial nets. Advances in neural information processing systems, 27:2672– 2680.
- Nakatsu, C. and M. White. 2010. Generating with discourse combinatory categorial grammar. *Linguistic Issues in Language Technology*, 4.
- Nirenburg, S., V. R. Lesser, and E. Nyberg. 1989. Controlling a language generation planner. In *IJCAI*, pages 1524–1530.

OpenAI. 2023. GPT-4 technical report.

- Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.
- Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the tenth workshop on statistical machine translation, pages 392– 395.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-totext Transformer. J. Mach. Learn. Res., 21(1), jan.
- Ramos-Soto, A., J. Janeiro-Gallardo, and A. Bugarín. 2017. Adapting SimpleNLG to spanish. pages 144–148. Association for Computational Linguistics.
- Reiter, E. 1994. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation.*
- Rieser, V. and O. Lemon. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation, pages 105–120.
- Roos, Q. 2022. Fine-tuning pre-trained language models for CEFR-level and keyword conditioned text generation: A comparison between google's t5 and openai's gpt-2.
- Sai, A. B., A. K. Mohankumar, and M. M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. ACM Comput. Surv., 55(2), jan.
- Scarselli, F., M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

- Stanchev, P., W. Wang, and H. Ney. 2019. EED: Extended edit distance measure for machine translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 514–520.
- Sukhbaatar, S., J. Weston, R. Fergus, et al. 2015. End-to-end memory networks. Advances in neural information processing systems, 28.
- Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.
- Tang, T., H. Lu, Y. E. Jiang, H. Huang, D. Zhang, W. X. Zhao, and F. Wei. 2023. Not all metrics are guilty: Improving NLG evaluation with LLM paraphrasing. arXiv preprint arXiv:2305.15067.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.
- Vedantam, R., C. Lawrence Zitnick, and D. Parikh. 2015. CIDEr: Consensusbased image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4566–4575.
- Wang, H., Y. Liu, C. Zhu, L. Shou, M. Gong, Y. Xu, and M. Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062, Online, August. Association for Computational Linguistics.
- Wang, J., Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore, December. Association for Computational Linguistics.
- Yu, W., C. Zhu, L. Qin, Z. Zhang, T. Zhao, and M. Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In NAACL 2022 Workshop on Deep

Learning on Graphs for Natural Language Processing.

- Yuan, W., G. Neubig, and P. Liu. 2021. BARTScore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, 34:27263– 27277.
- Zhang, H., S. Si, H. Wu, and D. Song. 2023. Controllable text generation with residual memory transformer. arXiv preprint arXiv:2309.16231.
- Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations.
- Zhang, Y. and X. Wan. 2024. Situated-Gen: Incorporating geographical and temporal contexts into generative commonsense reasoning. Advances in Neural Information Processing Systems, 36.
- Zhu, W. and S. Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association* for Computational Linguistics: EMNLP 2020, pages 94–108, Online, November. Association for Computational Linguistics.