EXPLORING THE INFLUENCE OF RELEVANT KNOWLEDGE FOR NATURAL LANGUAGE GENERATION INTERPRETABILITY

Iván Martínez-Murillo, Paloma Moreda, and Elena Lloret

University of Alicante, Carr. de San Vicente del Raspeig, s/n,
San Vicente del Raspeig, Alicante, Spain, 03690
ivan.martinezmurillo@ua.es

ABSTRACT

This paper explores the influence of external knowledge integration in Natural Language Generation (NLG), focusing on a commonsense generation task. We extend the CommonGen dataset by creating KITGI, a benchmark that pairs input concept sets with retrieved semantic relations from ConceptNet and includes manually annotated outputs. Using the T5-Large model, we compare sentence generation under two conditions: with full external knowledge and with filtered knowledge where highly relevant relations were deliberately removed. Our interpretability benchmark follows a three-stage method: (1) identifying and removing key knowledge, (2) regenerating sentences, and (3) manually assessing outputs for commonsense plausibility and concept coverage. Results show that sentences generated with full knowledge achieved 91% correctness across both criteria, while filtering reduced performance drastically to 6%. These findings demonstrate that relevant external knowledge is critical for maintaining both coherence and concept coverage in NLG. This work highlights the importance of designing interpretable, knowledge-enhanced NLG systems and calls for evaluation frameworks that capture the underlying reasoning beyond surface-level metrics.

KEYWORDS

Natural Language Generation, Interpretability, Knowledge-enhanced, Commonsense Generation.

1 INTRODUCTION

Natural Language Generation (NLG) models have witnessed substantial advancements with the emergence of Transformer-based architectures [1]. The scaling of

David C. Wyld et al. (Eds): MLNLP, ASOFT, CSITY, NWCOM, SIGPRO, AIFZ, ITCCMA – 2025
pp. 21-34, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.152002

these models in terms of both size and training data has led to significant improvements in their performance in a wide range of downstream tasks [2]. However, despite these advancements, recent research [3, 4, 5] has highlighted persistent limitations in the ability of Large Language Models (LLMs) to store and generate factually accurate information [6]. These deficiencies pose significant challenges, particularly in domains where factual correctness is crucial, such as scientific writing, medical documentation, or legal reasoning.

To address this issue, a growing research direction focuses on integrating external knowledge sources into NLG models to enhance their factual consistency [7]. By leveraging structured knowledge bases, retrieval mechanisms, or hybrid neural-symbolic approaches, researchers aim to supplement the intrinsic knowledge of these models with verifiable external facts. This strategy has the potential to improve factual accuracy and contextual coherence in generated text. However, a critical challenge arises in evaluating the effectiveness of such knowledge integration techniques.

Current knowledge-enhanced NLG methods lack transparency analyses that explain how external knowledge contributes to performance improvements. Most existing approaches rely on automatic evaluation metrics, which mainly measure surface-level lexical similarity and often fail to accurately assess factual correctness in open-ended text generation tasks [8]. Additionally, while some studies include manual evaluations, these typically focus only on the perceived quality of the generated texts and do not provide a deeper interpretability analysis of how external knowledge influences model behavior and output quality. This methodological gap highlights the need for a more comprehensive evaluation framework that extends beyond automated and manual metrics to incorporate interpretability analyses of the injected knowledge.

Therefore, this paper aims to address this gap by conducting a detailed interpretability analysis of how the quality of injected external knowledge influences NLG systems. The hypothesis is that enhancing NLG systems with non-related, or wrong external knowledge, critically affects their outputs. Specifically, we focus on a constrained commonsense generation task, enhanced with retrieved external knowledge, to evaluate commonsense reasoning in text generation. Our study investigates how knowledge integration affects the factual accuracy of generated text and examines the interpretability of these effects.

The contributions of this paper are twofold:

- To propose an extension to a widely-used commonsense reasoning dataset. We augment the dataset by incorporating: (1) external knowledge aligned with the input data, (2) automatically generated outputs conditioned on that knowledge, and (3) manually annotations of the generated sentences as either plausible or implausible. We named this resulting dataset as KITGI: Knowledge-Improved Text Generation and Interpretability.

– To propose a method and conduct a clear and detailed interpretability analysis of commonsense generation, demonstrating how the inclusion or removal of external knowledge influences the generated outputs.

By addressing these objectives, this work contributes to a more reliable assessment of knowledge-enhanced NLG models, offering insights into their factual generation capabilities and evaluation methodologies. It complements existing automatic and human evaluation practices commonly employed in the field.

2 RELATED WORK

NLG field has advanced significantly with the introduction of the Transformer architecture [1], greatly improving fluency and coherence. These models outperformed earlier approaches on complex language tasks, such as paraphrasing, question answering or machine translation [9]. As a result, NLG systems are now targeting more specific and demanding applications [7]. To support this, recent research focuses on integrating external knowledge to enhance factual accuracy and contextual relevance.

Knowledge-Enhanced NLG: It refers to the integration of external knowledge from diverse sources into NLG systems [7]. Techniques such as retrieval-augmented generation (RAG) [10], knowledge-graph based generation [11], or knowledge enhanced prompt tuning [12] have been shown to improve the comprehension and generative capabilities of NLG models. Those methods enrich the models with additional context or facts retrieved from external sources such as knowledge graphs, domain-specific databases or documents.

NLG interpretability: While knowledge-enhanced NLG improves the relevance and coherence of generated outputs, the mechanisms through which external information shapes the generation process remain insufficiently understood. Amnesic Probing [13] addresses this gap by using counterfactual examples to analyze the causal influence of injected knowledge on model predictions. ReX [14] extend local, model-agnostic explanation techniques with temporal information, enhancing the alignment between input content and model outputs. Similarly, another approach [15] improves performance and interpretability by incorporating structured domain knowledge directly into dialogue systems, thereby increasing model transparency. Despite these advancements, more systematic methods are still needed to trace and quantify the influence of external knowledge on generation decisions, especially in contexts requiring commonsense reasoning. This research aims to address to mitigate this limitation.

3 STARTING SETUP

To analyze the impact of integrating external knowledge into NLG systems, we propose and create the initial setup described below.

As we aim to analyze the effect of external knowledge on a constrained commonsense generation task, we modified and enriched a subset of the CommonGen dataset [16], which is the most widely used dataset for this task. It involves generating a sentence that incorporates a given set of concepts to describe an everyday scenario.

We began with 993 instances from the validation set of the CommonGen dataset and automatically generated sentences for each concept. These sentences were generated using T5-Large model [17]. T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks, and for which each task is converted into a text-to-text format. T5 works well on a variety of tasks out-of-the-box by prepending a different prefix to the input corresponding to each task. Furthermore, this model has shown remarkable performance on the CommonGen task[1], as many of the approaches using this model as the foundational model obtained a great score with the automatic evaluation metrics.

The sentences were generated under two conditions:

- No External Knowledge: The model generated a sentence using only the provided concept set, without any additional context.
- Enhanced with External Knowledge: For each concept, we retrieved the top five semantic relations from ConceptNet [18], a knowledge graph where nodes represent concepts and edges represent relations such as "is a part of", "used for", or "capable of". These relations were appended to each instance alongside the original concept set, and the model generated a sentence using both the concepts and the supplementary knowledge.

Each generated sentence was then manually annotated as correct and plausible (1) or incorrect/implausible (0). Furthermore, for the sentences enhanced with external knowledge, the subset also incorporated the relations extracted from ConceptNet. Figure 1 shows an example from the crafted dataset.

Notably, 121 sentences that were initially incorrect became correct after incorporating external knowledge. Therefore, our corpus for this study will be the 121 sentences that were improved after the injection of knowledge, alongside the corresponding concept sets and the retrieved knowledge from ConceptNet.

4 INTERPRETABILITY METHOD

The proposed interpretability benchmark consists of a three-stage method. First, key knowledge is removed; second, the model is retrained and the sentences are regenerated; and third, the results are manually labeled and evaluated. This process allows us to analyze the effects of different types of commonsense knowledge. The outcome is a final dataset of automatically generated sentences, labeled according to whether they contain commonsense or not. Each stage is explained in detail below.

[1] https://inklab.usc.edu/CommonGen/leaderboard.html

Concepts	External Knowledge	Knowledge-enhanced Sentences	Label	No Knowledg-enhanced Sentences	Label
['dance', 'kid', 'room']	dance relations are: 0. RelatedTo movement. 1. RelatedTo music. 2. RelatedTo ballet 3. HasPrerequisite turn on some music. 4. RelatedTo waltz. kid relations are: 0. Synonym child. room relations are: 0. RelatedTo house. 1. RelatedTo space. 2. RelatedTo walls. 3. RelatedTo living. 4. AtLocation a room.	A kid is dancing in a living room.	1	A kid is dancing in a room.	1
['bike', 'music', 'ride']	bike relations are: 0. HasPrerequisite a bike. 1. HasA two wheels. 2. Synonym motorcycle. 3. Synonym bicycle 4. UsedFor transport. Music relations are: 1. HasProperty soothing. 2. RelatedTo notes. 3. HasProperty relaxing. 4. RelatedTo melody. 5. RelatedTo art. ride relations are: 0. AtLocation a carnival. 1. RelatedTo car. 2. RelatedTo action. 3. RelatedTo horse. 4. Synonym depend on	A man is riding his bike and listening to music.	1	A man riding a bike to the music.	0
['cat', 'clip', 'hold']	cat relations are: 0. AtLocation my lap. 1. AtLocation a bed. 2. AtLocation the windowsill. 3. CapableOf hunt mice. 4. HasA four legs. clip relations are: 0. AtLocation office. 1. Synonym snip. 2. MannerOf shorten. 3. AtLocation an accessory store. hold relations are: 0. RelatedTo grasp. 1. hold RelatedTo grab. 2. RelatedTo keep	A cat is holding a clip.	0	A cat holding a clip.	0

Fig. 1: Samples from the crafted dataset.

4.1 Stage 1: Analysis and Removal of Key Knowledge

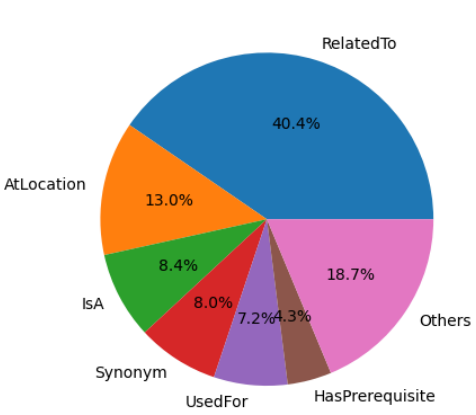
To assess the impact of external knowledge on the output of the NLG model, we focused on the subset of 121 sentences (i.e., evaluation dataset) for which the knowledge-enhanced model produced plausible outputs, which has been described in Section 3. Since the dataset includes the external knowledge used to enhance the model, we could directly determine its influence. To do so, we manually analyzed all the relations for each concept and discarded the ones that, according to human reasoning, could positively influence during the generation. With that, our goal was to corroborate that the injected knowledge really influenced the generation.

We calculated the external knowledge included in the 121 evaluation sentences and found that in total they contained 1635 relations corresponding to 121 concept sets, each containing from 3 to 5 words. However, not every word yielded 5 extractable relations. With respect to the existing relations for those 121 sentences of the dataset, the most common type found is the “RelatedTo” (40.4%), followed by the “AtLocation” relations (13.0%) and “IsA” (8.4%). The total distribution of the relations can be seen in Figure 2a.

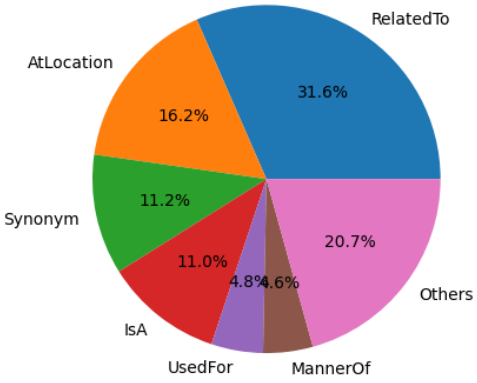
After calculating the total number of relations for the concepts, we carefully analyzed the relations related associated to the 121 concepts set. For each instance, we identified and removed the relations that, based on human judgment, seemed most useful for generating a sentence.

During this process, we removed 659 relations (40% of the initial knowledge) that were deemed highly relevant to the given keywords based on human judgment. This process resulted in a final set of 976 remaining relations. The distribution of relation types following this filtering step is shown in Figure 2b.

The most frequent relation type remained RelatedTo, which now accounts for 31.6% of the dataset. However, this represents a 10% decrease compared to its original proportion, suggesting that RelatedTo relations are particularly important from a human perspective when generating meaningful sentences. The second most common relation type is AtLocation, comprising 16.2% of the dataset. This is followed by the Synonym relation, which increased from 8.0% to 11.2% after filtering. This increase indicates that Synonym relations are less relevant to the specific contexts considered in our experiments and were therefore retained more frequently.



(a) Distribution of relation types in the initial evaluation dataset.



(b) Distribution of relation types in the initial evaluation dataset after filtering out relevant relations.

Fig. 2: Comparison of relation type distributions.

For example, consider the concept set {*"look", "watch", and "window"*}. The dataset includes the following relations: {*look relations are: 0. RelatedTo see. 1. RelatedTo glance. 2. RelatedTo eyes. 3. RelatedTo seeing. 4. RelatedTo view. watch relations are: 0. RelatedTo time. 1. RelatedTo wrist. 2. RelatedTo clock. 3. RelatedTo look 4. RelatedTo clook. window relations are: 0. RelatedTo glass. 1. RelatedTo opening. 2. RelatedTo looking. 3. RelatedTo house. 4. RelatedTo wall.*}. We evaluated which of these relations would strongly influence sentence generation. For "look", we removed relations like see and seeing because they are highly relevant to the intended meaning and could make sentence generation easier. In the case of "watch", most of its relations are associated with the concept of time, which is not relevant in this context. Therefore, we only removed the relation that connects

"watch" with "look". For "window", we found that the relation "looking" was most directly connected to the action we were focusing on, so we removed it as well. The remaining relations were kept to maintain background knowledge related to the object.

Once the relevant knowledge was filtered out, we generated the sentences again using the T5-Large model, incorporating the filtered knowledge as external input. Then, for this experiment, we analyze the following two sets:

- A set of 121 sentences generated using the complete set of relations as external knowledge. All of these sentences were manually labeled as correct, as stated in Section 3.
- Another set of 121 sentences generated using filtered knowledge, in which the most meaningful relations were intentionally excluded.

4.2 Stage 2: Assessment of Commonsense and Coverage

We conducted a manual assessment based on two distinct criteria to evaluate the generated sentences after removing key knowledge. Each criterion was rated on a binary scale: a score of 0 indicated inadequate performance, while a score of 1 signified correct execution. The evaluation was carried out according to the following criteria:

- **Commonsense:** We decide whether the generated sentence makes sense, or it does not make sense.
- **Coverage:** We analyze if the generated sentence contains all the concepts in the concept set.

Figure 3 presents an example of the evaluation process. The table is organized such that the vertical axis (rows) corresponds to the coverage criterion, while the horizontal axis (columns) reflects the commonsense criterion. We included the coverage dimension because, in many cases, the generated sentence appeared plausible yet failed to include all required keywords, thereby not fulfilling the task's objective properly.

For instance, as it is shown in the figure, given the concepts "dog", "pull", and "race", the sentence "A dog is racing against another dog in a race." is plausible and thus receives a commonsense score of 1. However, since it omits the keyword "pull", it scores 0 for coverage.

Some sentences fail on both criteria—lacking commonsense and omitting key terms. For example, for the concepts "car", "drive", and "phone", a sentence that excludes "phone" and is incoherent would receive a 0 for both criteria.

Conversely, a sentence might include all keywords (coverage score of 1) but still lack coherence, resulting in a commonsense score of 0. Thus, each criterion captures a distinct but complementary aspect of the sentence quality.

Coverage\CommonSense	0	1
0	['car', 'drive', 'phone']: A car is driving on a car.	['dog', 'pull', 'race']: A dog is racing against a dog in a race.
1	['cat', 'clip', 'hold']: A cat holding a clipped cat.	['chair', 'eat', 'toddler']: A toddler is eating in a chair.

Fig. 3: Representative samples of the criteria applied during evaluation.

4.3 Stage 3: KITGI Dataset Creation

The final dataset proposed for this experiment comprises 121 instances, each containing the following components:

- **Concept Set**: A group of 3 to 5 concepts.
- **Sentence with Full External Knowledge and Annotation**: A sentence generated by a T5-Large model, augmented with the complete set of retrieved knowledge, along with its annotation in terms of commonsense relevance and concept coverage.
- **Sentence with Filtered External Knowledge and Annotation**: A sentence generated by a T5-Large model, enhanced using only the filtered knowledge, with corresponding annotations for commonsense reasoning and concept coverage.
- **Retrieved Knowledge**: The set of relations retrieved from ConceptNet for each word in the concept set.
- **Filtered Knowledge**: A subset of the retrieved knowledge, containing only the relations that are not relevant to each specific concept set.

The dataset is available at `https://github.com/imm106/KITGI`.

5 RESULTS AND DISCUSSION

In this section, we present the scores from the manual evaluation conducted on the KITGI dataset, based on the criteria defined in Section 4.2, and provide an analysis of the results.

The outcomes of the manual evaluation are shown in Figure 4. Subfigure 4a (left side) displays the results for sentences generated using the full external knowledge

set, while Subfigure 4b the right side shows the results for sentences generated after removing relevant knowledge (i.e., Filtered External Knowledge).

The initial set of sentences enhanced with the full external knowledge contained all commonsense, as described in Section 3. However, as Subfigure 4a shows, 8% (10 sentences) of these did not include all the required keywords, meaning they did not fully meet the task's objective. In contrast, Subfigure 4b shows that only 42 (34+8) sentences generated with filtered knowledge were considered meaningful - this corresponds to 34% of the sentences. The performance is even lower when considering keyword coverage: only 8 out of the 42 meaningful sentences used all the words from the concept set. This represents just a 6% of the dataset—a considerably low result compared to the 91% of sentences that met both the coverage and commonsense criteria when enhanced with the full set of knowledge.

These results suggest that excluding relevant knowledge significantly impacts the coverage criterion. Specifically, in Subfigure 4b is shown that 88 (54 + 34) of the 121 sentences (72%) did not include at least one required concept word. Of the remaining 33 sentences that included all concept words, only 8 were complete and meaningful, thus having commonsense.

CommonSense Coverage\	0	1
0	0	10
1	0	111

(a) Manual analysis results for the sentences generated with all the knowledge.

CommonSense Coverage\	0	1
0	54	34
1	25	8

(b) Manual analysis results for the sentences generated with all the filtered knowledge.

Fig. 4: Manual analysis results.

Indeed, after carefully analyzing the generated sentences with the filtered knowledge, we detected three variants:

– **The external knowledge associated with certain words is misleading**: When this happens, the system often omits the problematic word in the generated sentence, sometimes still producing a sentence that is meaningful. This suggests that the model struggles to integrate the word with the context provided by the external knowledge and chooses to exclude it instead. For example, consider the concept set [“look”, “watch”, “window”]. The external knowledge provided for the word “watch” refers only to the object (e.g., a wristwatch), rather than its verb form. As a result, the model is unable to combine “watch” meaningfully with the other words, and generates the sentence “A man is looking at a window.” Another example involves the concept set [“fall”, “ground”, “jump”],

where the model generates "A man is jumping on the ground." In this case, the knowledge for the word "fall" is related to the season (Autumn), rather than the verb "to fall," preventing the model from using it correctly in the intended context.

- **The external knowledge is not helpful**: In many cases, although the provided knowledge corresponds to the correct meaning of each individual word in context, it does not support establishing meaningful connections between the words in the concept set. As a result, the system often generates nonsensical sentences. In some cases, it also fails to include all the words from the concept set. An example of this can be seen with the concept set ["attempt", "fence", "knife", "stick", "throw"]. The model generates the sentence "Someone throws a knife and attempts to throw it into the fence." While the external knowledge is relevant for each word—"attempt" is associated with trying, "fence" refers to a protective wall, "knife" is defined as a cutting object, "stick" as a small piece of wood, and "throw" as the action of launching something—there are no strong semantic relations linking these concepts together. As a result, the sentence lacks coherence despite the relevance of the individual word meanings.
- **The given knowledge establishes a slight connection among words**: In some cases, the provided knowledge does not establish a direct relationship among the concepts, but it still indirectly helps the model generate a coherent and accurate sentence. For example, consider the concepts ["boat", "sail", "day"]. The knowledge includes that a "boat" can travel on water and is located on a lake; "sail" is associated with wind and cloth; and "day" is defined as the antonym of night or related to time. Based on this, the system generates the sentence: "Boats sail on a sunny day." Here, the knowledge about boats supports the idea of traveling on water, which could help to connect it with the concept of sailing. Similarly, the knowledge about day helps form the phrase sunny day, drawing on its contrast with night. Thus, although the relationships are not explicitly defined, background knowledge still helps guide the model toward a meaningful output.

These findings demonstrate that the quality of external knowledge significantly impacts the model's performance. When the input includes misleading or ambiguous information, the model struggles to integrate it effectively, often resulting in incorrect sentence generation. This highlights the importance of considering the input context when retrieving relevant knowledge. Given the richness of language, many words have multiple meanings, and selecting the wrong one can negatively affect the output. Conversely, even a weak but relevant connection in the external knowledge can help the model produce a more accurate result.

6 CONCLUSIONS & FUTURE WORK

This study presents an in-depth interpretability method and analysis of how external knowledge enhances NLG, specifically in a constrained commonsense reasoning task. Using a controlled benchmark, we systematically eliminated highly relevant semantic relations to assess their impact. The results reveal that properly integrated external knowledge is essential for producing coherent and plausible sentences. We found that removing critical knowledge elements markedly reduces both the commonsense accuracy and conceptual coverage of the generated outputs, highlighting the vital role of external knowledge in ensuring factually grounded and logically consistent language generation.

For future work, this research can be extended in several directions. First, we plan to investigate the impact of external knowledge in multilingual settings to determine whether its influence is consistent across languages or language-dependent. Second, this approach could be applied to other NLP tasks to assess whether knowledge integration yields similar effects beyond text generation. Finally, it would be valuable to explore more advanced knowledge retrieval and integration methods to evaluate how different types and sources of knowledge influence model performance.

ACKNOWLEDGEMENTS

This research is part of the R&D projects “CORTEX: Conscious Text Generation” (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by “ERDF A way of making Europe”; QUMLAUDE: Mecánica cuántica para comprensión y generación del lenguaje” (PID2024-160791OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by “ERDF A way of making Europe; “CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities” (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and “European Union NextGenerationEU/PRTR”; and project “NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021)" funded by the Generalitat Valenciana. Additionally, this work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. Moreover, it has been also partially funded by the Ministry of Economic Affairs and Digital Transformation and “European Union NextGenerationEU/PRTR” through the “ILENIA” project (grant number 2022/TL22/00215337) and “VIVES” subproject (grant number 2022/TL22/00215334).

References

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[2] Linyao Yang et al. "Give us the Facts: Enhancing Large Language Models With Knowledge Graphs for Fact-Aware Language Modeling". In: *IEEE Transactions on Knowledge and Data Engineering* 36.7 (2024), pp. 3091–3110. DOI: 10.1109/TKDE.2024.3360454.

[3] Hanmeng Liu et al. "Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4". In: *CoRR* abs/2304.03439 (2023). URL: https://doi.org/10.48550/arXiv.2304.03439.

[4] Xuming Hu et al. "Towards Understanding Factual Knowledge of Large Language Models". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=9OevMUdods.

[5] Hoyeon Chang et al. "How Do Large Language Models Acquire Factual Knowledge During Pretraining?" In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: https://openreview.net/forum?id=TYdzj1EvBP.

[6] María Miró Maestre y Iván Martínez-Murillo y Tania J. Martin y Borja Navarro-Colorado y Antonio Ferrández y Armando Suárez Cueto y Elena Lloret. "Roadmap for Natural Language Generation: Challenges and Insights". In: *Procesamiento del Lenguaje Natural* 74.0 (2025), pp. 67–79. ISSN: 1989-7553. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6667.

[7] Meng Jiang et al. "Knowledge-augmented Methods for Natural Language Generation". In: *Knowledge-augmented Methods for Natural Language Processing*. Springer, 2024, pp. 41–63.

[8] Iván Martínez-Murillo y Paloma Moreda y Elena Lloret. "Analysing the Problem of Automatic Evaluation of Language Generation Systems". In: *Procesamiento del Lenguaje Natural* 72.0 (2024), pp. 123–136. ISSN: 1989-7553. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6583.

[9] Md Riyadh and M. Omair Shafiq. "Towards Automatic Evaluation of NLG Tasks Using Conversational Large Language Models". In: *Artificial Intelligence Applications and Innovations*. Ed. by Ilias Maglogiannis et al. Cham: Springer Nature Switzerland, 2023, pp. 425–437. ISBN: 978-3-031-34107-6.

[10] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[11] Haochen Liu et al. "Knowledge Graph-Enhanced Large Language Models via Path Selection". In: *Findings of the Association for Computational Lin-*

guistics: ACL 2024. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 6311–6321. DOI: 10.18653/v1/2024.findings-acl.376. URL: https://aclanthology.org/2024.findings-acl.376/.

[12] Hao An et al. "Knowledge-enhanced Prompt Tuning for Dialogue-based Relation Extraction with Trigger and Label Semantic". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, 2024, pp. 9822–9831. URL: https://aclanthology.org/2024.lrec-main.858/.

[13] Yanai Elazar et al. "Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 160–175. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00359. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00359/1924189/tacl_a_00359.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00359.

[14] Junhao Liu and Xin Zhang. "ReX: A Framework for Incorporating Temporal Information in Model-Agnostic Local Explanation Techniques". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.18 (2025), pp. 18888–18896. DOI: 10.1609/aaai.v39i18.34079. URL: https://ojs.aaai.org/index.php/AAAI/article/view/34079.

[15] Isabel Feustel et al. "Enhancing Model Transparency: A Dialogue System Approach to XAI with Domain Knowledge". In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara et al. Kyoto, Japan: Association for Computational Linguistics, 2024, pp. 248–258. DOI: 10.18653/v1/2024.sigdial-1.22. URL: https://aclanthology.org/2024.sigdial-1.22/.

[16] Bill Yuchen Lin et al. "CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, 2020, pp. 1823–1840. DOI: 10.18653/v1/2020.findings-emnlp.165. URL: https://aclanthology.org/2020.findings-emnlp.165/.

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

[18] Robyn Speer, Joshua Chin, and Catherine Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (2017). DOI: 10.1609/aaai.v31i1.11164. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11164.

AUTHORS

Iván Martínez-Murillo is a Ph.D. student in Computational Linguistics at the University of Alicante (Spain). His primary research interests lie in natural language generation, controllable generation methods and hallucination mitigation techniques. Martínez-Murillo received a Master's degree in Cybersecurity from the University of Alicante in 2022. In 2023, he joined the Language Processing and Information Systems Group from the University of Alicante.
Contact him at ivan.martinezmurillo@ua.es.

Paloma Moreda is a Lecturer in the Department of Languages and Computing Systems of the University of Alicante (Spain). His research interests include Text Simplification, Gender Bias and , Automatic Language Generation and some other Natural Language Processing topics. She received the Ph.D. degree in Natural Language Processing, and has been the PI of several research project. Contact her at moreda@dlsi.ua.es.

Elena Lloret is a Full Professor in the Department of Software and Computing Systems of the University of Alicante (Spain), teaching courses related to Databases and Natural Language Processing. Her research interests are in the field of Natural Language Processing, with special emphasis in Automatic Summarization and Natural Language Generation. Lloret received her Ph.D. degree in computer science from the University of Alicante in 2011. Contact her at elloret@dlsi.ua.es.