yasaminal at Semeval2026: Constraint-Aware Humor Generation with Knowledge Graph Guidance

Yasamin Aali
Brock University / St. Catharines, ON
yaali@brocku.ca

Abstract

This paper presents a knowledge-guided humor generation system, which involves generating humorous text from either a pair of words or a news headline. The proposed approach integrates structured semantic reasoning derived from a knowledge graph (KG) with controlled generation using large language models (LLMs). The system constructs an intermediate KG hint consisting of related concepts retrieved in the target language, which is appended to the prompt to guide the generation process in a structured manner. A single candidate joke is generated per input using controlled top-p decoding. Experimental results show that incorporating KG reasoning improves relevance and constraint satisfaction, while LLM-based generation ensures fluency and creativity. Overall, the proposed method offers a structured and interpretable framework for humor generation across multiple languages.

1 Introduction

Humor generation is a challenging problem in natural language processing, as it requires not only linguistic fluency but also creativity, world knowledge, and an understanding of incongruity. In this task, systems must generate humorous text based on structured inputs, including either a pair of words or a news headline, across multiple languages such as English, Spanish, and Chinese. The evaluation is based on human preference, where outputs are judged by how funny they are rather than how correct they are. We follow the task definition and evaluation framework introduced in the shared task overview paper (Castro et al., 2026).

The proposed system follows a hybrid approach that combines KG reasoning with LLM generation. Instead of relying only on LLMs, we first retrieve related concepts from a knowledge graph (WordNet for English, ConceptNet for all three languages) to identify meaningful connections between input concepts. These connections are appended to the prompt as optional knowledge, which guides the generation process. By integrating symbolic semantic reasoning with neural text generation, the system aims to produce humor that is both coherent and creative.

Our code is publicly available at: https://github.com/yasaminaali/HumorKG.

2 Related Works

Recent work in computational humor focuses on structured reasoning, theoretical grounding, and LLM-based generation. Chen et al. (Chen et al., 2024) introduce chain-of-humor and humor mind maps to explicitly model how jokes are constructed step by step. Similarly, Dubey (Dubey, 2025) proposes a planning-based framework that selects humor strategies and leverages knowledge graphs along with iterative refinement to improve joke quality. These approaches highlight the importance of structured reasoning, which motivates our use of semantic anchors and twist planning.

From a theoretical perspective, De et al. (De Marez et al., 2024) incorporate classical humor theories such as incongruity and superiority into interpretable models, while (Kim and Chilton, 2025) emphasizes the role of reasoning, creativity, and social context in humor. These works support the need to combine symbolic knowledge with generative models.

Recent studies also examine LLM-based humor generation. Evstafev et al. (Evstafev, 2025) show that decoding strategies, particularly lower temperature, significantly impact humor quality, while He et al. (He and Mei, 2025) demonstrate the effectiveness of fine-tuned LLMs for humor generation. Additionally, Horvitz et al. (Horvitz et al., 2024) and Jain et al. (Jain, 2025) highlight the importance of evaluation and the subtle distinctions between humorous and non-humorous text.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2409–2412

Compared to prior works, our approach combines KGs with LLM-based generation, injecting retrieved concepts into the prompt while preserving the fluency of the underlying LLM.

3 System Overview

The system adopts a multi-stage pipeline that integrates lightweight knowledge graph retrieval with LLM-based generation. The overall design decomposes humor generation into a small number of well-defined stages: input classification, knowledge graph retrieval, and LLM-based generation.

Input Processing

Each input is first classified into one of two constraint types: (i) two-word inputs, where the generated joke must explicitly include both words, and (ii) headline-based inputs, where the humor should be derived from a given news headline. This distinction is important because it determines how semantic information is extracted and how constraints are handled during generation. For two-word inputs, the system focuses on identifying meaningful relationships between the words, while for headline inputs, it aims to capture the main topic and contextual meaning of the headline.

Knowledge Graph Retrieval

We use two knowledge graphs: WordNet (via NLTK, English only) and ConceptNet (via its REST API, queried in the target language for all three languages). For two-word inputs, we extract hypernyms, antonyms, and shared-neighbor concepts of each word. For headline inputs, we tokenize the headline (with language-specific rules, including CJK bigrams) and retrieve the top-weight neighbors of the most salient content tokens. The retrieved concepts are returned as a short hint string. When no useful edges are available, the hint is empty.

LLM-Based Generation

The retrieved concepts are appended to a language-specific prompt as "Optional Knowledge" and passed to a LLaMA-based language model, which generates a single candidate joke with top-p sampling. The output is then normalized: Unicode NFKC, whitespace cleanup, removal of any <think> reasoning blocks left by reasoning-mode models, and defensive stripping of any leaked "assistant" role prefix.

Here is an example of how the model works. For the headline "Ryanair to cut 1 million more passenger seats," ConceptNet retrieval returns neighbors of passenger and seat including chair, plane, and flight. These are appended to the prompt, and the model produces: "Ryanair cuts a million seats again – soon passengers will have to bring their own chairs." The retrieved hint nudges the model toward the seat -> chair pun.

4 Experimental Setup

We use the official data splits provided by the shared task, which include training, development, and test partitions for each language. Task A contains 300 evaluation instances per language. As the task focuses on humor generation rather than supervised classification, our approach does not involve additional parameter training on labeled humor data. Instead, it relies on prompt-based LLM generation augmented with lightweight knowledge-graph retrieval.

Preprocessing

All input instances undergo light preprocessing to ensure consistency and improve downstream generation quality. This includes basic text normalization, such as removing formatting artifacts and standardizing whitespace. For headline-based inputs, keyword extraction is performed to identify salient entities and thematic concepts, which guide the subsequent knowledge graph retrieval stage.

Model Configuration

We employ Llama-3.3-70B-Versatile, served via the Groq Cloud inference API. Text generation is performed using controlled decoding to balance creativity and coherence. The primary hyperparameters are: temperature = 0.7, top-p = 0.9, no beam search, and a maximum of 400 new tokens. For each input instance, a single candidate joke is generated.

External Resources

Our system uses two knowledge graphs: WordNet, accessed via the NLTK library for English, and ConceptNet, queried via its public REST API in the target language for all three languages. All LLMs are served by the Groq Cloud inference API in their pretrained form without additional fine-tuning, ensuring a computationally efficient and reproducible setup.

Evaluation Protocol

Evaluation follows the official shared task setup, which is based on human preference judgments rather than automated metrics. Systems are compared based on which generated text is considered funnier by human annotators. This evalua-

tion paradigm emphasizes subjective qualities such as creativity, relevance, and clarity. In addition to the shared-task evaluation, we perform an internal LLM-as-judge evaluation using Llama-4-Scout-17B-16E-Instruct as the judge. Each output is scored 1-5 on four dimensions – funny, relevant, creative, fluent – and we report the unweighted mean (AVG). For our final configuration we also report a pairwise win rate, in which the judge picks the funnier and more relevant of two position-randomised candidates on a per-item basis.

5 Results

Our final configuration (Llama-3.3-70B-Versatile with ConceptNet) achieves strong performance in terms of constraint satisfaction and linguistic quality. Across all three languages, nearly all generated outputs incorporate the required input elements – the central topic of the headline – and outputs are generally fluent and grammatically well-formed (Fluent scores pegged at or near 5.0 out of 5 in the matrix of Table 1).

Table 1 reports the full 4 x 3 screening matrix per language (n = 30). The top three configurations are tightly clustered in every language (AVG within 0.06), which indicates that the choice of generator is the dominant factor and the choice of knowledge graph is a second-order effect. Scaling the generator from 8B (Llama-3.1-8B-Instant) to 120B (GPT-OSS-120B) lifts the average by roughly 0.2, while moving from no KG to the best available KG contributes between 0.03 and 0.08. ConceptNet matches or beats WordNet in every language, with the gap largest on Spanish and Chinese where WordNet is English-only. The smallest model is in the bottom three of every language regardless of KG.

Pairwise comparison. Table 2 reports the pairwise win rate of the final configuration (n = 300) against our initial shared-task submission: 100.0%, 95.3%, and 96.0% on English, Spanish, and Chinese respectively. The judge preferred the new system on 874 of 900 items overall. Two factors contribute. First, the new pipeline produces cleaner jokes on the merits. Second, our initial submission contained a chat-template decoding bug that leaked an "assistant" role token at the start of roughly 78% of rows; the judge correctly penalises these as off-topic or disfluent.

Despite these strengths, several limitations remain. The system struggles with culturally dependent or highly contextual humor that requires deeper socio-cultural knowledge, most visibly on Spanish and Chinese headlines that mention local entities. Some outputs are also moderately generic: the Creative dimension averages 3.0 - 4.1 across cells, noticeably below Funny and Fluent, indicating that the mode often reuses the most obvious pun or anthropomorphism available rather than finding a less predictable twist. These limitations are consistent with prior observations that humor evaluation is dominated by creativity rather than surface form.

6 Conclusion

We presented a hybrid humor generation system that combines knowledge graph retrieval with LLM-based generation. Our final configuration (Llama-3.3-70B-Versatile with ConceptNet) wins 100.0% / 95.3% / 96.0% of pairwise judgments against our initial shared-task submission on English, Spanish, and Chinese; across the full 4x3x3 comparison, model choice matters more than knowledge-graph choice, and ConceptNet is the preferable default for multilingual humor. Future work includes incorporating larger knowledge graphs, improving humor evaluation, and adapting to cultural contexts.

References

Santiago Castro, Luis Chiruzzo, Santiago Góngora, Salar Rahili, Naihao Deng, Ignacio Sastre, Victoria Amoroso, Guillermo Rey, Aiala Rosá, Guillermo Moncecchi, J. A. Meaney, Juan José Prada, and Rada Mihalcea. 2026. SemEval-2026 Task 1: MWAHAHA, Models Write Automatic Humor And Humans Annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17826–17834.

Victor De Marez, Thomas Winters, and Ayla Rigouts Terryn. 2024. Thinc: A theory-driven framework for computational humor detection. *arXiv preprint arXiv:2409.01232*.

Shivam Dubey. 2025. Humorplansearch: Structured planning and hucot for contextual ai humor. *arXiv preprint arXiv:2508.11429*.

Lang	Model	KG	Funny	Rel.	Creat.	Fluent	AVG
EN	**gpt-oss-120b**	**conceptnet**	**4.07**	**5.00**	**4.03**	**5.00**	**4.53**
	gpt-oss-120b	none	4.03	4.93	4.00	5.00	4.49
	qwen3-32b	conceptnet	4.03	4.97	3.93	5.00	4.48
	gpt-oss-120b	wordnet	4.03	4.97	3.87	5.00	4.47
	qwen3-32b	none	4.00	4.87	4.00	5.00	4.47
	llama-3.3-70b	none	4.00	5.00	3.80	5.00	4.45
	llama-3.1-8b	conceptnet	3.97	5.00	3.73	5.00	4.43
	llama-3.3-70b	conceptnet	4.00	4.97	3.67	5.00	4.41
	llama-3.1-8b	none	4.00	4.93	3.70	5.00	4.41
	llama-3.3-70b	wordnet	3.90	4.93	3.63	5.00	4.37
	llama-3.1-8b	wordnet	3.83	4.87	3.40	5.00	4.28
	qwen3-32b	wordnet	2.90	4.37	2.97	4.73	3.74
ES	**gpt-oss-120b**	**none**	**4.00**	**5.00**	**3.83**	**5.00**	**4.46**
	gpt-oss-120b	conceptnet	4.00	5.00	3.80	5.00	4.45
	qwen3-32b	none	4.03	5.00	3.73	5.00	4.44
	gpt-oss-120b	wordnet	3.97	4.93	3.83	5.00	4.43
	llama-3.3-70b	wordnet	4.03	5.00	3.63	5.00	4.42
	qwen3-32b	conceptnet	4.00	5.00	3.63	5.00	4.41
	llama-3.3-70b	conceptnet	4.00	4.97	3.50	5.00	4.37
	llama-3.3-70b	none	3.97	4.93	3.53	5.00	4.36
	llama-3.1-8b	none	4.00	4.93	3.47	5.00	4.35
	llama-3.1-8b	conceptnet	3.90	4.87	3.53	5.00	4.33
	llama-3.1-8b	wordnet	3.97	4.80	3.33	5.00	4.28
	qwen3-32b	wordnet	2.87	4.33	3.03	4.33	3.64
ZH	**gpt-oss-120b**	**conceptnet**	**3.97**	**4.97**	**3.77**	**4.97**	**4.42**
	qwen3-32b	none	4.03	5.00	3.57	5.00	4.40
	qwen3-32b	conceptnet	4.03	5.00	3.53	5.00	4.39
	gpt-oss-120b	wordnet	3.90	4.87	3.67	5.00	4.36
	llama-3.3-70b	conceptnet	4.03	5.00	3.37	5.00	4.35
	gpt-oss-120b	none	3.90	4.87	3.63	4.97	4.34
	llama-3.3-70b	wordnet	4.00	5.00	3.33	5.00	4.33
	llama-3.3-70b	none	3.97	4.97	3.17	4.93	4.26
	llama-3.1-8b	none	3.83	4.93	3.03	4.83	4.16
	llama-3.1-8b	conceptnet	3.77	4.80	2.97	4.80	4.08
	llama-3.1-8b	wordnet	3.80	4.73	2.90	4.77	4.05
	qwen3-32b	wordnet	3.00	4.50	3.27	4.67	3.86

Table 1: Screening matrix across three languages ($n = 30$ per cell). Rubric dimensions are scored 1–5 by LLAMA-4-SCOUT-17B-16E-INSTRUCT; AVG is the unweighted mean. Best cell per language in bold.

Language	n	Win	Loss	Tie	Win rate
English	300	300	0	0	1.000
Spanish	300	286	13	1	0.953
Chinese	300	288	11	1	0.960

Table 2: Pairwise win rate of our final configuration (Llama-3.3-70B-Versatile with ConceptNet) against our initial shared-task submission, judged by LLAMA-4-SCOUT ($n = 300$ per language).

Evgenii Evstafev. 2025. Optimizing humor generation in large language models: Temperature configurations and architectural trade-offs. *arXiv preprint arXiv:2504.02858*.

Jinliang He and Aohan Mei. 2025. Advancing computational humor: Llama-3 based generation with distilbert evaluation framework. In *ITM Web of Conferences*, volume 70, page 03024. EDP Sciences.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869.

Sulbha Jain. 2025. Jokeeval: Are the jokes funny? review of computational evaluation techniques to improve joke generation.

Sean Kim and Lydia B Chilton. 2025. Ai humor generation: Cognitive, social and creative skills for effective humor. *arXiv preprint arXiv:2502.07981*.