Integrating News Headlines with Wikidata for Enhanced Knowledge Graph Querying

Sebastián Ferrada
Department of Computer Science
Universidad de Chile, Chile

Hernán Sarmiento
IMFD Chile, Chile

Abstract

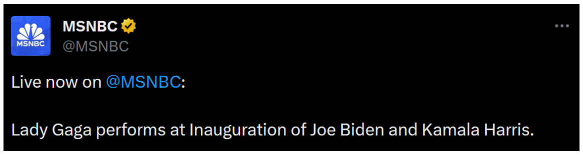
This research proposes constructing a Wikidata-based Knowledge Graph (KG) using Twitter data and news headlines. Our proposal focuses on semantic querying of events, linking news entities to Wikidata, and automatically extracting relationships.

Introduction

Social media platforms are valuable for research in social sciences and journalism as they capture extensive discussions among users [1]. Research has demonstrated that integrating these platforms with KGs for depicting real-world events results in clear and organized data representations [2]. This structure enhances the capability of computers to comprehend and efficiently process the information.

We propose utilizing Twitter data to construct a Wikidata-based KG of news events, collected from headlines of news outlets' accounts through Galean [3]. Integrating factual statements from these news headlines into the KG aims to facilitate semantic querying of events, e.g., who has performed the national anthem in each presidential inauguration?

Example tweet:



MSNBC
@MSNBC

Live now on @MSNBC:

Lady Gaga performs at Inauguration of Joe Biden and Kamala Harris.

Some statements to be extracted from the tweet:

```
(Lady Gaga, performed at, Biden’s Inauguration),
(Biden’s Inauguration, participant, Joe Biden),
(Biden’s Inauguration, participant, Kamala
Harris).
```

Building this KG is relevant because:

A. It would supplement Wikidata with current affairs data.

B. The KG benefits from linking to Wikidata to be able to respond to queries requesting demographic and geographic data.

C. The KG could work with news from WikiNews both to extend the KG and the automatic generation of newsworthy articles.

The overarching goal of building this KG is to enable the public to build narratives and engage in journalistic investigation by querying the KG.

Research Questions:

RQ1: How can entities mentioned in news headlines be effectively linked to Wikidata entities?

RQ2: What automated techniques can be employed to extract relationships among these detected entities?
RQ3: To what extent can a Wikidata-enriched knowledge graph of news events respond to diverse and complex queries?

Date: From June 1, 2024 until June 30, 2025.

Related work

A Knowledge Graph (KG) is a collection of entities, described in terms of their attributes and the relationships among them [4].

Named entity recognition (NER) identifies real-world entities in unstructured text. Linking these entities to a KG (entity linking or NEL) is essential. Various works, notably using Deep Learning [5] and Large Language Models (LLMs) [6], address NER. Open-tapioca [7], a lightweight entity linker trained on Wikidata, provides Q-code, label, and metadata for recognized entities in the input text.

Extracting relationships among entities is a more complex problem, with initial work in leveraging LLMs for this purpose [8].

In news KGs or semantic news querying, NewsLink [9] produces news embeddings and enables semantic search by computing query embeddings. Some works focus on reasoning over news KGs [10] and fact-checking using facts in news KGs [11]. Notably, there is no known work that automatically extracts a knowledge graph from news events.

Our proposal involves a combination of existing approaches and the development of novel models to extract statements from news event headlines, ensuring easy linkage to Wikidata. The ultimate goal is to publish this data as a queryable KG accessible to the public.

Methods

The key steps are the following:

- Explore Galean's data to understand its contents, and manually tag specific tweets to establish a benchmark.
- Test NER and NEL techniques to obtain Wikidata entities mentioned in each news headline.
- Design and evaluate a method to extract and assess relationships between identified entities, utilizing a new or existing benchmark [12].
- Produce a pipeline to go from Social Media news extraction to KG construction.
- Publish the resulting KG on a public endpoint.

Expected output

- A SOTA method to extract relationships among Named Entities in news events headlines, beneficial for the AI research community, and to anyone looking to build KGs from text.
- A publicly available KG of news events based upon Wikidata accessible through a public endpoint.
- At least two scientific publications can stem from this work.

Risks

The most challenging aspect of the proposed research is extracting relationships from the text, given limited existing literature. However, we anticipate that LLMs will provide a strong initial approximation. Furthermore, changes in Twitter's data access policies could present a challenge to our research. Nevertheless, our proposal can be adapted to include any type of textual content on other platforms or social networks.

Community impact plan

We aim for journalists and social scientists to use our KG, gathering feedback to ensure the graph's meaningfulness. Additionally, we can collaborate with WikiNews editors to explore the value of automatically generating news articles or templates for their efforts. We aim to enrich Wikidata by expanding global event coverage, including non-US Presidential Inaugurations, and incorporating vocabulary for event-related predicates (e.g., performed at, attended, invited).

Evaluation

1. Evaluate statement extraction methods by comparing results to a self-generated gold-standard or existing benchmarks [12].
2. Measure scientific success by counting accepted or submitted publications, with a target of at least two.
3. Gauge outreach success through the number of queries to the KG and WikiNews articles generated.

Budget

Our budget supports a team collecting and organizing data, as well as developing and deploying specialized software for our project. This plan includes collaboration with social and political science experts to assess our findings. Additionally, we allocate funds for advanced computational resources and software tools. The budget includes the publication of OA papers. The amount requested is $38,200.

- Salary or stipend: $25,200
- Software: $4,000
- OA publishing: $3,000
- Institutional overhead: $3,000
- Others: $3,000 (communications and advertising)

Prior contributions

Sebastián Ferrada has leveraged Wikimedia Commons data and Wikidata to develop IMGpedia [13,14], a multimodal Knowledge Graph integrating visual and semantic queries. IMGpedia was presented at the WikiWorkshop in 2018. Sebastián has also designed benchmarks using the Wikidata Query Logs in his publications, showcasing his proficiency in querying and extracting value from Wikidata.

Hernán Sarmiento specializes in social network analysis with a focus on exploring the dynamics of social media. His expertise includes examining content during crises and socio-political events [15,16,17]. Additionally, he has led R&D projects in the fields of transportation and climate, incorporating user-generated content as an additional data source.

References

[1] Bruns, A., & Weller, K. (2016, May). Twitter as a first draft of the present: And the challenges of preserving it for the future. In Proceedings of the 8th ACM Conference on Web Science (pp. 183-189).
[2] Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review, 1-32.
[3] Peña-Araya, V., Quezada, M., Poblete, B., & Parra, D. (2017). Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using Twitter. *EPJ Data Science*, *6*(1), 1-35.
[4] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys (Csur)*, *54*(4), 1-37.
[5] Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50-70.

[6] Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., ... & Wang, G. (2023). GPT-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

[7] Delpeuch, A. (2019). Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*.

[8] Dunn, A., Dagdelen, J., Walker, N., Lee, S., Rosen, A. S., Ceder, G., ... & Jain, A. (2022). Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

[9] Yang, Y., Li, Y., & Tung, A. K. (2021, April). NewsLink: Empowering intuitive news search with knowledge graphs. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (pp. 876-887). IEEE.

[10] Liu, D., Lian, J., Liu, Z., Wang, X., Sun, G., & Xie, X. (2021, August). Reinforced anchor knowledge graph generation for news recommendation reasoning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 1055-1065).

[11] Zhu, B., Zhang, X., Gu, M., & Deng, Y. (2021). Knowledge enhanced fact checking and verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29,* 3132-3143.

[12] Mihindukulasooriya, N., Tiwari, S., Enguix, C. F., & Lata, K. (2023, October). Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In *International Semantic Web Conference* (pp. 247-265). Cham: Springer Nature Switzerland.

[13] Ferrada, S., Bustos, B., & Hogan, A. (2017). IMGpedia: a linked dataset with content-based analysis of Wikimedia images. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16* (pp. 84-93). Springer International Publishing.

[14] Ferrada, S., Bravo, N., Bustos, B., & Hogan, A. (2018, April). Querying Wikimedia images using Wikidata facts. In *Companion Proceedings of The Web Conference 2018* (pp. 1815-1821).

[15] Sarmiento, H., & Poblete, B. (2021, March). Crisis communication: A comparative study of communication patterns across crisis events in social media. In Proceedings of the 36th Annual ACM Symposium on Applied Computing (pp. 1711-1720).

[16] Sarmiento, H., Poblete, B., & Campos, J. (2018, May). Domain-Independent detection of emergency situations based on social activity related to geolocations. In Proceedings of the 10th ACM Conference on Web Science (pp. 245-254).

[17] Sarmiento, H., Bravo-Marquez, F., Graells-Garrido, E., & Poblete, B. (2022, May). Identifying and Characterizing New Expressions of Community Framing during Polarization. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 841-851).