Language models for extracting Wikidata statements and generating Wikipedia content

Ta Hoang Thang
Instituto Politecnico Nacional (IPN), Mexico

Abstract

In this project, I plan to create language models specifically designed for extracting statements from Wikidata and generating content for Wikipedia in English. These models will operate in a cyclical training process, where the output of one model serves as the input for another, enhancing overall performance. The objective of this project is to enhance the content of Wikidata and English Wikipedia by integrating the outputs of these models with results from ChatGPT. The anticipated outcome is an accelerated development of content for Wikidata and Wikipedia.

Introduction

To propel the growth of Wikipedia and Wikidata, we must integrate new AI technologies, specifically in content generation. Advanced AI bots, like ChatGPT, excel in natural language processing challenges, including data-to-text (D2T) and text-to-data (T2D) tasks. Utilizing language models becomes crucial for enhancing and expanding the content of these projects.

The current problems of Wikipedia and Wikidata:

- The average number of statements of each item in Wikidata is still NOT enough for data-to-text problem in generating long texts.
- No existing language models are tailored for generating ***long texts*** like Wikipedia content from Wikidata statements.

The two problems are described in Figure 1 and Figure 2. In Figure 1, a triple (data) can only generate a single sentence (reference). In Figure 2, from a set of triples (source), the model can generate a short paragraph (target). This project aims to create longer texts.

properties

reference	category	webnlg_id

reference	The leader of Aarhus is Jacob Bundsgaard.
category	Airport
webnlg_id	dev/Airport/1/Id1

data

subject	predicate	object
Aarhus	leader	Jacob Bundsgaard

Figure 1. A random example of WebNLG.

Wida2WL **source:** item title : Bowmans Rail | headquarters location : Bowmans > country : Australia | headquarters location : Bowmans > located in the administrative territorial entity : South Australia | country : Australia | legal form : company | inception : 2003

target: *Bowmans Rail is an Australian railway operator hauling freight in South Australia. Bowmans Rail was founded as Bowmans Intermodal in 2003. The company adopted its current name in 2016, when it transported over 36,000 containers annually.*

Figure 2. A random example of Wikida2WL.

Language models are crucial for enhancing Wikidata items, minimizing human efforts, and prioritizing content validation. As the number of statements per item increases, the system generates longer text. This project streamlines tasks for Wikipedia editors, enabling a focus on content validation rather than starting anew. This shift motivates editors and researchers, contributing to accumulating more data on Wikidata and Wikipedia for further development. The proposed loop involves extracting Wikidata statements from Wikipedia text and generating text, maximizing synergies between the two platforms.

Date: June 1, 2024 to June 15, 2025.

Related work

Transformer models, especially in transfer learning, excel in NLP tasks like data-to-text and text-to-data [16]. Researchers commonly use pre-trained models from HuggingFace, such as T5 [14] and BART [5], for creating language models.

D2T and T2D, two language models, can be independently trained on collected data or undergo cycle training, where the output of D2T becomes the input for T2D, and vice versa, until both models converge [7], as shown in Figure 3.

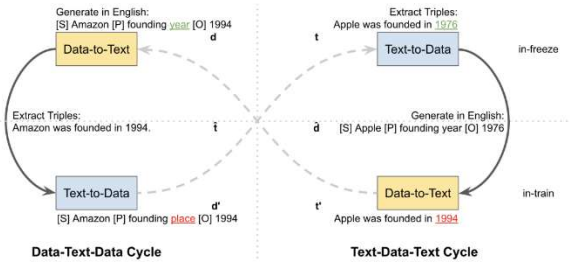

Figure 3. Cycle Training of the Data-to-Text model and Text-to-Data model.

In addition to the mentioned methods, I plan to explore alternatives like self-training, contrastive learning, and reinforcement learning. This allows for model comparisons, and I intend to use adapter models to reduce training time for large language models.

Methods

Data Collection: I developed a crawler to extract Wikidata statements, aligning them with short paragraphs on Wikipedia. I've created a dataset called Wida2WL (Wikidata to Weak Labels) with 40,000 source-target pairs, shown in Figure 2.

ChatGPT data: It is used to extract “source” or “target” results from a given “source” or “target” text. It also helps to improve the content quality instead of weak labels in Figure 2.

Methods: There are two models, D2T and T2D and both uses Transformer models with sequence-to-sequence architecture. I also use cycle-training [7, 8, 9], contrastive learning in the training to improve the quality.

Model training facilities:
CPU GeForce RTX 3060 Ti: BART-base, T5-base, etc.
CPU Nvidia RTX A6000: Alpaca, Llama, FLAN, etc.

Participants: Due to my research relationship, I will invite several researcher from some universities.

Discussions & Surveys: I open discussions on Meta and invite participants from Wikidata, English Wikipedia, and research groups, prioritizing active, experienced editors. Participants are selected based on contributions, and random invitations are sent, mainly on Wikidata and English Wikipedia.

Expected output

For this work, I offer these outputs:

- *Scientific publications*: At least two scientific papers that will be published in top conferences (ACM, ACL, NAACL, EMNLP etc) or top journals (Q1 and Q2). They help to attract the attention from the scientific community.
- *Source code, datasets and pre-trained modes*: Offer a repository on GitHub and HuggingFace, containing source code, datasets, pre-trained models, and guidelines for everyone can download and reuse.

Risks

Like text generation, D2T models may introduce hallucinations—outputs with incorrect facts, redundant information, or missing statements. Proposed solutions include a "risk" metric [10], word-level labels [11], token-level reference-free detection [12], and human evaluation. T2D models often identify incorrect statements due to similarities, mitigated by techniques like contrastive learning and reinforcement learning.

Community impact plan

As a Wikipedian, I will create my project space on Meta for opening discussions to receive community suggestions or opinions. Wikipedians can ask for evaluating outputs to see their requirements so that I can update the language models.

Evaluation

Since the main outcome is language models, the evaluation should be:

- Automatic metrics: For the D2T (data to text) model, they are F1 and ROC AUC score. For T2D (text to data), they are BLEU, METEOR, ROUGE and other string metrics.
- Human concensus: Measure the inter-rater reliability by annotators over outputs several criteria such as correctness, adequacy, and naturalness [5, 6].

Budget

1 year salary: 36000 USD (3000 USD/per month)
CPU with Nvidia RTX A6000: 6000 USD
ChatGPT API: 4000 USD
Publication fee: 3000 USD (2 papers)
5-10 annotators: 1000 USD

In total: **50000 USD**

Prior contributions

A paper enhances Wikidata descriptions using a two-stage summarization method on short Wikipedia paragraphs [3]. Another paper, on Arxiv, explores mapping sentences to Wikidata statements with triples and qualifiers.

I participated in SMART TASK 2022, submitting two papers on using BERT and data oversampling for answer type and set relation prediction [1, 2]. Recently, I was invited as a reviewer of Wikidata Workshop 2023 (https://wikidataworkshop.github.io/2023/).

References

[1]. Thang, T. H., Ojo, O. E., Adebanji, O. O., Calvo, H., & Gelbukh, A. (2022). The combination of BERT and data oversampling for answer type prediction. In *CEUR Workshop Proceedings* (Vol. 3119). CEUR-WS.

[2]. Thang, T. H., Ojo, O. E., Adebanji, O. O., Calvo, H., & Gelbukh, A. (2022). The combination of BERT and data oversampling for relation set prediction. In *CEUR Workshop Proceedings* (Vol. 3119). CEUR-WS.

[3]. Ta, H. T., Rahman, A. B. S., Majumder, N., Hussain, A., Najjar, L., Howard, N., ... & Gelbukh, A. (2023). WikiDes: A Wikipedia-based dataset for generating short descriptions from paragraphs. *Information Fusion*, *90*, 265-282.

[4]. Ta, H. T., Gelbukha, A., & Sidorov, G. (2022). Mapping Process for the Task: Wikidata Statements to Text as Wikipedia Sentences. *arXiv preprint arXiv:2210.12659*.

[5]. Ta, T. H., & Anutariya, C. (2015). A model for enriching multilingual Wikipedias using infobox and Wikidata property alignment. In *Semantic Technology: 4th Joint International Conference, JIST 2014, Chiang Mai, Thailand, November 9-11, 2014. Revised Selected Papers 4* (pp. 335-350). Springer International Publishing.

[6]. van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, *67*, 101151.

[7]. Wang, Z., Collins, M., Vedula, N., Filice, S., Malmasi, S., & Rokhlenko, O. (2023). Faithful Low-Resource Data-to-Text Generation through Cycle Training. *arXiv preprint arXiv:2305.14793*.

[8]. Polat, F., Tiddi, I., Groth, P., & Vossen, P. (2023, September). Improving Graph-to-Text Generation Using Cycle Training. In *Proceedings of the 4th Conference on Language, Data and Knowledge* (pp. 256-261).

[9]. Guo, Q., Jin, Z., Qiu, X., Zhang, W., Wipf, D., & Zhang, Z. (2020). Cyclegt: Unsupervised graph-to-text and text-to-graph generation via cycle training. *arXiv preprint arXiv:2006.04702*.

[10]. Akani, E., Favre, B., Bechet, F., & Gemignani, R. (2023, September). Reducing named entity hallucination risk to ensure faithful summary generation. In *Proceedings of the 16th International Natural Language Generation Conference* (pp. 437-442).

[11]. Rebuffel, C., Roberti, M., Soulier, L., Scoutheeten, G., Cancelliere, R., & Gallinari, P. (2022). Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 1-37.

[12]. Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., & Dolan, B. (2021). A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.

[13]. Munoz, E., Hogan, A., & Mileo, A. (2013, October). DRETa: Extracting RDF from Wikitables. In *ISWC (Posters & Demos)* (pp. 89-92).

[14]. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485-5551.

[15]. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[16]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.