DaSeLab at LLMs4OL 2024 Task A: Towards Term Typing in Ontology Learning

Adrita Barua*⬮, Sanaz Saki Norouzi*⬮, and Pascal Hitzler⬮

Kansas State University, USA

*Correspondence: Adrita Barua, adrita@ksu.edu

Abstract: The report presents the evaluation results of our approach in the LLM4OL Challenge, where we fine-tuned GPT-3.5 for Task A (Term Typing) across three different datasets. Our approach demonstrated consistent and robust performance during few-shot testing, achieving top rankings in several datasets and sub-datasets, proving the potential of fine-tuning LLMs for ontology creation tasks.

Keywords: LLM, Term Typing, Ontology Learning

1 Introduction

Large language models (LLMs) have made notable advancements in various natural language processing (NLP) tasks. In a recent study [1], the performance of LLMs was evaluated specifically for Ontology Learning (OL) using zero-shot prompting method. OL refers to the process of creating an ontology—a structured representation of knowledge in a particular domain, consisting of concepts, relationships, and categories. Researchers tackling the challenge of creating ontologies from text are essentially leveraging a broad range of methodologies developed in computational linguistics. By carefully selecting different NLP techniques, they address the three key issues in ontology construction: term association, the creation of term and concept hierarchies, and the identification and labeling of ontological relationships [2].

In [1], they provided an evaluation of three key OL tasks, one of which is term typing. Term typing aims to identify relevant terms from the text that will form the basic vocabulary of the ontology. This task is crucial because it determines the basic building blocks that will be used to construct the ontology. In many respects, ontology learning is a specialized extension of fundamental computational linguistics goals like automatic lexicon construction and semantic text labeling.

As part of the LLM4OL challenge[1] at the International Semantic Web Conference (ISWC) 2024 [3], we focused on fine-tuning GPT-3.5 for term typing across different datasets. The goal was to evaluate the performance of these models during the few-shot testing phase, where the testing dataset includes data from the same ontology domain that the model was trained on. This approach aims to enhance the model's

---

*These authors contributed equally to this work.

[1] https://sites.google.com/view/llms4ol/home?authuser=0

ability to accurately identify and categorize relevant terms, thereby improving the overall quality and utility of the created ontology. In the following, first, we talk about the approach and the datasets we use for this task, and then we go through the results and challenge leaderboard, and the conclusion.

2 Approach

Our approach involved using three different datasets to individually fine-tune the gpt-3.5-turbo-0125 model[2], training it to identify term types specific to each dataset. The three fine-tuned models were then evaluated during the few-shot testing phase using their respective test datasets.

2.1 Datasets

As part of the LLM4OL challenge [4], we used three datasets to fine-tune the GPT model: WordNet, GeoNames, and UMLS.

WordNet: The WordNet dataset is a lexicosemantic dataset derived from the original WordNet. The training set contains 40,559 terms, and the test set has 9,470 terms, covering 18 relation types and four term types: nouns, verbs, adverbs, and adjectives [5].

GeoNames: GeoNames consists of geographical locations that comprise 680 categories of geographical locations (e.g., streams, lakes, seas, roads, railroads, etc.). The training set contains 8,078,865 terms, and the test set has 702,510 terms. We used only the first 10 percent (approximately 878,137 terms) of the training dataset to fine-tune our model due to OpenAI's fine-tuning restrictions on the size of the training dataset [6].

UMLS: The UMLS (Unified Medical Language System) dataset integrates various biomedical terminologies and standards to support interoperability between different health information systems [7]. Three subcategories of this dataset have been used:

NCI: NCI is a UMLS subontology from NCI Enterprise Vocabulary Services (EVS), standardizing terminology for clinical care, research, and public information. It provides reference terminology for NCI and other systems. The training set contains 96,177 terms, and the test set has 24,045 terms, containing 125 term types.

MEDCIN: MEDCIN is a UMLS subontology that includes medical components like symptoms and treatments. It uses clinical hierarchies to link data elements, emphasizing relationships within diagnostic contexts. The training set contains 277,028 terms, and the test set has 69,258 terms, containing 87 term types.

SNOMEDCT_US: SNOMEDCT_US is a UMLS subontology foundational for electronic health records (EHRs), providing concepts with distinct meanings and formal definitions structured hierarchically. The training set contains 278,374 terms, and the test set has 69,594 terms, containing 125 term types.

A detailed discussion of the datasets can be found on the challenge website[3] and and the dataset statistics are presented in Table 1.

---

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo
[3] https://sites.google.com/view/llms4ol/task-a-term-typing?authuser=0

***Table 1.** Dataset Statistics for Fine-Tuning*

Dataset	Training Set	Test Set	No of Term Types	Term Types
WordNet	40,559 terms	9,470 terms	4	Nouns, verbs, adverbs, adjectives
GeoNames	878,137 terms	702,510 terms	680	Geographical locations
NCI	96,177 terms	24,045 terms	125	Clinical care, research, public information
MEDCIN	277,028 terms	69,258 terms	87	Symptoms, treatments
SNOMEDCT US	278,374 terms	69,594 terms	125	Electronic health records (EHRs)

2.2 Model Fine-Tuning

In our work, we fine-tuned the gpt-3.5-turbo-0125 model for each dataset using the OpenAI API. Fine-tuning OpenAI's text generation models is a powerful method to tailor them to specific needs, but it requires significant time and resources [8]. In previous work [1], the authors used a zero-shot prompting method for the term typing task on the aforementioned datasets. We built upon their work by incorporating the optimized prompts from their study into our fine-tuning process to achieve the best results.

To fine-tune the model, we had to prepare the training dataset. For the OpenAI API, the data must be stored in JSONL format, which is a text format where each line is a separate JSON object. This format is ideal for processing large datasets line by line. To prepare the dataset, we created a diverse set of demonstration conversations that closely resemble the interactions the model will encounter during inference in production. Each example in the dataset is formatted as a conversation in the same style as required by the Chat Completions API. Specifically, each example is a list of messages where each message has a role and content. The prompt template can be found in Table 2.

***Table 2.** Prompt template for each dataset*

Dataset	Prompt Template
WordNet	Perform a sentence completion on the following sentence: The part of speech of the term "Term" in the sentence "Sentence" is —
GeoNames	Perform a sentence completion on the following sentence: "Place Name/Location" geographically is a ___
UMLS	Perform a sentence completion on the following sentence: "Medical related term" in medicine can be described as ___

The example formats used for generating all three datasets for fine-tuning are given in Tables 3 to 6.

Each entry in the training datasets is formatted according to the example format, using the respective prompt for that dataset to prepare the final JSONL file. After creating all three training datasets, we uploaded the training files to fine-tune the gpt-3.5-turbo-0125 model. The Later, the fine-tuned models for each dataset were evaluated using the test datasets. The Table 7 highlights the training details of five datasets used for fine-tuning. GeoNames has the largest number of trained tokens (37.7 million) but also the highest training loss (0.0603), indicating that it was more challenging for the model

Table 3. *Data Format with an Example Sentence (WordNet)*

Role	Content
user	Perform a sentence completion on the following sentence: The part of speech of the term "cover" in the sentence "cover her face with a handkerchief" is ___
assistant	The part of speech of the term "cover" in the sentence "cover her face with a handkerchief" is verb.

Table 4. *Data Format without an Example Sentence (WordNet)*

Role	Content
user	Perform a sentence completion on the following sentence: The part of speech of the term "land reform" is ___
assistant	The part of speech of the term "land reform" is noun.

Table 5. *GeoNames Example*

Role	Content
user	Perform a sentence completion on the following sentence: "Pic de Font Blanca" geographically is a ___
assistant	"Pic de Font Blanca" geographically is a peak.

Table 6. *UMLS Example*

Role	Content
user	Perform a sentence completion on the following sentence: "1,2-Dihydro-3-methyl-benz(j)aceanthrylene" in medicine can be described as ___
assistant	The type of "1,2-Dihydro-3-methyl-benz(j)aceanthrylene" in medicine can be described as: ['organic chemical', 'hazardous or poisonous substance'].

to learn from this dataset. WordNet, with the smallest dataset (2.2 million tokens) and a smaller batch size, shows a relatively high training loss (0.0413). In contrast, MEDCIN and SNOMEDCT exhibit the lowest training losses (0.0055 and 0.0086, respectively), suggesting better model performance during training. All datasets were trained for just one epoch, and the varying batch sizes (from 27 for WordNet to 128 for GeoNames, SNOMEDCT, and MEDCIN) reflect the differences in dataset sizes and computational strategies used. Overall, the datasets with larger token counts and higher batch sizes performed well, but some (like GeoNames) may require further tuning to improve performance.

Table 7. *Training Information for Datasets*

Dataset	Trained Tokens	Epochs	Batch Size	LR Multiplier	Training Loss
WordNet	2,208,173	1	27	2	0.0413
GeoNames	37,737,184	1	128	2	0.0603
NCI	6,109,613	1	64	2	0.0273
SNOMEDCT	18,533,107	1	128	2	0.0086
MEDCIN	19,256,674	1	128	2	0.0055

3 Evaluation Results

The performance of our fine-tuned models was evaluated across five different datasets. We used the OpenAI API for evaluation, employing the same prompts that were used during the training phase (e.g., as mentioned in the example format for the user's role 2.2). Each of the three fine-tuned models was assessed using the few-shot testing dataset specific to that model. The results, as provided by the challenge organizers, are summarized in the following tables, which show the leaderboard rankings and the corresponding performance metrics for each dataset. The source code for training and evaluating the models is available online.[4]

3.1 WordNet

Table 8 shows the leaderboard rankings and performance metrics for the WordNet dataset, Our model achieved a top-3 ranking, demonstrating competitive performance in terms of accuracy and other relevant metrics. Here, our model's performance highlights its effectiveness in achieving balanced precision and recall.

Table 8. *SubTask A.1 (FS) – Term Typing – WordNet*

	Team Name	F1	P	R
1	TSOTSALearning	0.9938	0.9938	0.9938
2	DSTI	0.9716	0.9716	0.9716
3	**DaSeLab**	**0.9697**	**0.9689**	**0.9704**
4	RWTH-DBIS	0.9446	0.9446	0.9446
5	TheGhost	0.9392	0.9389	0.9395
6	Silp_nlp	0.9037	0.9037	0.9037
7	Phoenixes	0.8158	0.7689	0.8687

3.2 GeoNames

Table 9 presents the leaderboard for GeoNames. Our model secured the first position indicating its superior performance. It's important to note that our model was evaluated on a portion of the test data, which highlights its robustness and effectiveness even with partial data.

Table 9. *SubTask A.2 (FS) – Term Typing – GeoNames*

	Team Name	F1	P	R
1	**DaSeLab**	**0.5906**	**0.5906**	**0.5906**
2	Silp_nlp	0.4433	0.7503	0.3146
3	RWTH-DBIS	0.4355	0.4355	0.4355
4	TSOTSALearning	0.2937	0.2937	0.2937
5	TheGhost	0.1489	0.1461	0.1519

3.3 UMLS

As mentioned, this dataset consists of three sub-datasets, and our model demonstrated outstanding performance, ranking first in two of the sub-datasets and second in the other one. The detailed results are presented in the following.

---

[4] https://github.com/AdritaBarua/LLMs4OL-2024-Task-A-Term-Typing

3.3.1 NCI

In this sub-dataset our model achieved the top ranking, significantly outperforming other models in terms of precision, recall, and F1-score that are shown in Table 10.

Table 10. *SubTask A.3 (FS) – Term Typing – NCI subontological source from UMLS*

	Team Name	F1	P	R
1	**DaSeLab**	**0.8249**	**0.8161**	**0.8340**
2	Silp_nlp	0.6974	0.8792	0.5779
3	TheGhost	0.5370	0.4450	0.6769
4	RWTH-DBIS	0.1691	0.1821	0.1579
5	Phoenixes	0.0737	0.0562	0.1070

3.3.2 SNOMEDCT_US

Our model also ranked first here, demonstrating its robustness and consistent high performance. The leaderboard is shown in Table 11.

Table 11. *SubTask A.3 (FS) – Term Typing – SNOMEDCT_US subontological source from UMLS*

	Team Name	F1	P	R
1	**DaSeLab**	**0.8829**	**0.8810**	**0.8848**
2	Silp_nlp	0.7552	0.8583	0.6742
3	TheGhost	0.5275	0.4266	0.6910
4	RWTH-DBIS	0.4747	0.4888	0.4613

3.3.3 MEDCIN

As shown in Table 12, in this sub-dataset, we were ranked as the second one, closely following the top-ranked model. These results indicate that our model maintains a strong balance between precision and recall.

Table 12. *SubTask A.3 (FS) – Term Typing – MEDCIN subontological source from UMLS*

	Team Name	F1	P	R
1	Silp_nlp	0.9382	0.9591	0.9181
2	**DaSeLab**	**0.9373**	**0.9379**	**0.9366**
3	TheGhost	0.5328	0.4183	0.7336
4	RWTH-DBIS	0.4566	0.4607	0.4526

Analysis of the evaluation shows that the model exhibits significant performance variation across different datasets, particularly with GeoNames demonstrating substantially lower scores compared to WordNet and UMLS datasets. The model achieved an F1 score of 0.5906 on GeoNames, which may be due to the complexity and ambiguity associated with geographical locations and the high number of term types (680), showing a more significant challenge in classification. This ambiguity may refer to the same geographical term representing different places, such as cities with identical names in different countries, or it may arise from varying interpretations of boundaries and regions across cultures and languages. In contrast, WordNet, with its limited scope of four grammatical term types, allowed the model to perform much better, with an F1 score of 0.9697. UMLS datasets, with term types ranging from 87 to 125, still show relatively high scores due to medical terminology's structured and specialized nature.

4 Conclusion

In this paper, we presented the results of our approach to the challenge on different datasets: WordNet, GeoNames, and UMLS. Our models consistently demonstrated robust and competitive performance, achieving top rankings in several datasets and sub-datasets highlighting their strength and potential for practical applications. We are optimistic about the future development and improvement of our approach by utilizing different prompting methods and LLMs.

Acknowledgements

References

[1] H. B. Giglou, J. D'Souza, and S. Auer, "LLMs4OL: Large language models for ontology learning," in *The Semantic Web – ISWC 2023 – 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I*, T. R. Payne, V. Presutti, G. Qi, *et al.*, Eds., ser. Lecture Notes in Computer Science, vol. 14265, Springer, 2023, pp. 408–427.

[2] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology Learning from Text: Methods, Evaluation and Applications* (Frontiers in Artificial Intelligence and Applications). IOS Press, Amsterdam, 2005, vol. 123.

[3] H. Babaei Giglou, J. D'Souza, and S. Auer, "Llms4ol 2024 overview: The 1st large language models for ontology learning challenge," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[4] H. Babaei Giglou, J. D'Souza, S. Sadruddin, and S. Auer, "Llms4ol 2024 datasets: Toward ontology learning with large language models," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[5] G. A. Miller, "WordNet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[6] GeoNames, *Geonames geographical database*, http://www.geonames.org/, 2024.

[7] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[8] J. Wei, M. Bosma, V. Y. Zhao, *et al.*, "Finetuned language models are zero-shot learners," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=gEZrGCozdqR.