LLMs4OL 2024: The 1st Large Language Models for Ontology Learning Challenge at the 23rd ISWC
LLMs4OL 2024 Task Participant Papers
https://doi.org/10.52825/ocp.v4i.2484

Published: 02 Oct. 2024

TSOTSALearning at LLMs4OL Tasks A and B : Combining rules to Large Language Model for Ontology learning

Carick Appolinaire Atezong Ymele [1] and Azanzi Jiomekong[12]

[1]Department of Computer Science, University of Yaounde I, Yaounde, Cameroon
[2]TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
*Correspondence: Carick Atezong, carick.atezong@facsciences-uy1.cm

Abstract: This paper presents our contribution to the Large Language Model For Ontology Learning (LLMs4OL) challenge hosted by ISWC conference. The challenge involves extracting and classifying various ontological components from multiple datasets. The organizers of the challenge provided us with the train set and the test set. Our goal consists of determining in which conditions foundation models such as BERT can be used for ontologies learning. To achieve this goal, we conducted a series of experiments on various datasets. Initially, GPT-4 was tested on the wordnet dataset, achieving an F1-score of 0.9264. Subsequently, we performed additional experiments on the same dataset using BERT. These experiments demonstrated that by combining BERT with rule-based methods, we achieved an F1-score of 0.9938, surpassing GPT-4 and securing the first place for term typing on the Wordnet dataset.

Keywords: Ontology Learning, Large Language Models, Rules, BERT

1 Introduction

Knowledge acquisition from scratch is costly in time and resources. Ontology learning aims to reduce this cost. Ontology learning is the extraction of ontological knowledge from unstructured, semi-structured or fully structured knowledge sources in order to build an ontology from them with little human intervention [1].

A lot of work has been done on the extraction of ontological knowledge from several data sources such as texts [2], databases [3], XML files [4], vocabularies [5], etc and several domain such as food information [6], food composition knowledge from scientifc literature [7], healthcare [8]. These works resulted into symbolic based techniques, statistical based techniques, and multi-strategy based techniques. Given that Large Language Models (LLMs) have shown significant advancements in natural language processing, Babaei et al. [9] proposed a Large Language Models for Ontology Learning (LLMs4OL) approach. The authors evaluated nine LLMs families on several datasets. These evaluations shows that foundational LLMs are not sufficiently suitable for ontology learning. However, in many context students, researchers, etc. do not always have enough resources to run LLMs such as LLaMA-7B or GPT-3.

The main goal of this study is to reply to the following research question: *"In which conditions foundations models can be used for ontology learning".* To reply to this question, we participate to LLMs4OL 2024 challenge [10]. This challenge aims to explore the intersection of LLMs and OL. The organizers of this challenge provided train and test datasets. The GPT-4 model was run and evaluate on four of the dataset. Thereafter, the BERT-Base uncased model was chosen and a set of experimentation was conducted. These experimentation's show that by merging the strengths of LLMs such as BERT with symbolic techniques such as rules, the model obtained can be as powerfully as GPT-4.

Before presenting the methodology in Section 2.2, we present the challenge in Section 2.1, followed by the evaluation in Paragraph 2.1, the approach we used in Section 2.2 with the results in Section 3. Finally, Section 4provides the conclusion. To facilitate the reproducibility of the results, the codes used in this study are available on our GitHub repository at https://github.com/sudo-001/LLMs4OL-2024.

2 An Approach Combining LLMs with Rules for Ontology Learning

Taking advantage of our experience in the field of ontology learning using symbolic approaches such as rules and LLMs such as BERT, we defined an approach combining LLMs and rules for ontology learning. This methodology was applied on the datasets provided by LLMs4OL challenge. Before we present this methodology in Section 2.2, the main ontology's components will be presented in Section 2.1.

2.1 LLMs4OL Challenge

LLMs4OL challenge aims for exploring the intersection of LLMs and OL. The following tasks were proposed by the organizers of this challenge:

- **Task A - Term Typing:** aims to discover the generalized type for a lexical term. This correspond to a concept or a class and aims to represent a category of object;
- **Task B - Taxonomy Discovery:** aims to discover the taxonomic hierarchy between type pairs;
- **Task C - Non-Taxonomic Relationship Extraction:** aims to identify non-taxonomic relation between types.

Evaluation

The organizers provided us for each dataset the train and the test dataset. To evaluate our system, we trained the model on the train data and we evaluated on the test data on the codalab platform. The evaluation was done using the Precision, Recall and F1-score.

2.2 Methodology

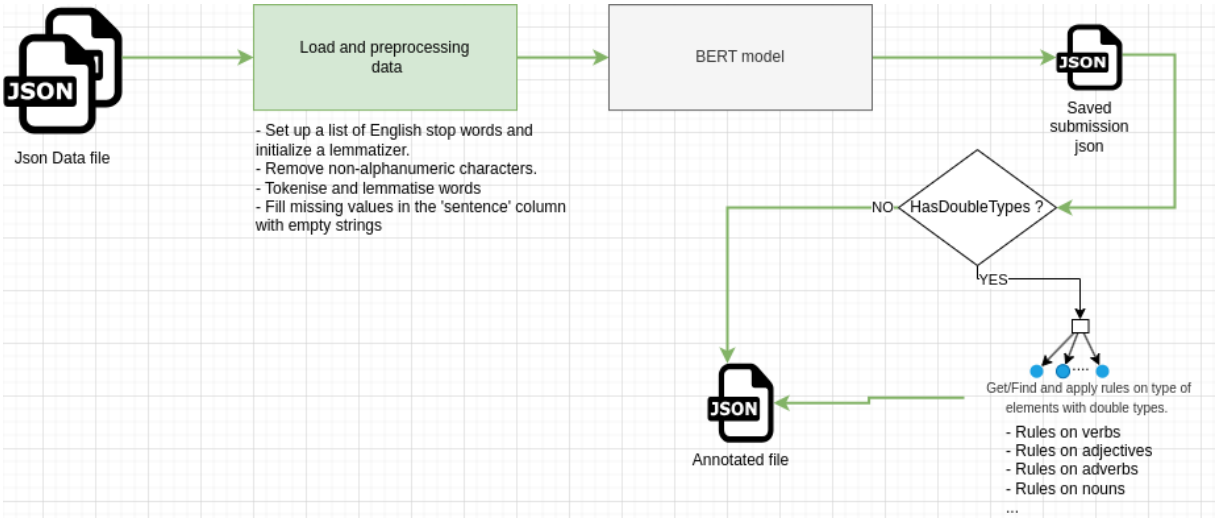

Figure 1. *An Approach Combining LLMs with Rules for Ontology Learning*

To enhance the process of ontology learning, we propose in this work a methodology (see Figure. 1) based on the combination of LLMs with rules derived from an in-depth analysis of the training data. This analysis involved identifying recurring patterns and contextual associations between terms and their corresponding types. This combination is described by the equations below (eq.1 and eq.2).

$$M = \{LLM_p, R_s\} \tag{1}$$

$$R_s = \{r_1, r_2, ..., r_n\} \tag{2}$$

- M : Represent our methodology.
- LLM_p: this is the pre-trained LLM on the trained dataset.
- R_s: this is the set of rules that characterises the dataset. An example of a rule found in the WordNet dataset is "if a a term ends with 'ly' and the predicted model predict two types or more then the type is 'adverb' ".

The workflow consists of the following key steps:

1. **Data Preprocessing:** This step consists of refining the dataset so as to assure that it is clean and properly structured. To this end, non-alphanumerical characters such as (!, #, -, _,...) are removed, text are converted into lowercase and tokenize, each token are reduce to it basical value, and the lemmatised token are combined to form the preprocessed sentence.
2. **Finetuning LLM for OL:** During this step we fine-tuned a pre-trained BERT model for our specific text classification task. Below are the additional details regarding the fine-tuning process : - **Model Choice**: We used the pre-trained BERT model (bert-base uncased) for its proven ability to capture rich contextual representations of text; - **Data preparation**: The data was pre-processed and encoded using the BERT tokenizer; - **Fine-tuning configuration**: we configured the fine-tuning with 3 epochs, a batch size of 16 for training, a warm-up steps of 500, a weight decay of 0.01.

3. **Evaluating the fine-tuned model:** In this step, the fine-tuned model is run on the test data. We do not use prompts or additional query formulations during this process. Instead, The evaluation was carried out by feeding the pre-processed test data directly into the model. The model results were then used to generate predictions for each test instance.;
4. **Evaluating the LLM output:** This step consists of evaluating the output of the test data using the precision, recall and F1-score. If the score is sufficiently high, one can stop the process. In our case we used the codalab platform to evaluate our results.
5. **Assessing the output:** This step consists of identifying the elements that are not well predicted;
6. **Complete the model with rules:** This step consists for each element identified in step 5 to defined a rule that allow us to predict the right output.

2.3 Experimentation environment

To evaluate the different systems for ontology learning, the organizers of the LLMs4OL provided several datasets [11]. The Table. 1 present a detailed description of the datasets for term typing and Table. 2 present the different datasets for taxonomy discovery.

Table 1. *Overview of the datasets used in this work for task A : Term typing*

Dataset	Train Size	Test Size	Number of Types
WordNet	40,559	9,470	4
GeoNames	8,078,865	702,510	680
GO-Biological Process	195,775	108,300	792
GO-Cellular Component	228,460	126,485	323
GO-Molecular Function	196,074	107,432	401

Table 2. *Overview of the datasets used for Task B : Taxonomy Discovery task*

Dataset	Train Size	Test Size
GeoNames	476	204
Schema.org	1,070	364
UMLS	74	45
GO	33,703	5,753

1. **Wordnet**: See table 1. The WordNet dataset is a large lexical database, where words are in english and organized into sets of synonyms called synsets. This dataset contains two types of entries: (1) Entry with term or group of terms accompanies with it's usage. For instance, "cover" as a term and "cover her face with a handkerchief" as the contextual sentence or the usage example. (2) Entry with terms or group of terms without example of usage. For this dataset, the task was to predict the type of terms (corresponding to Task A of the challenge).
2. **Geonames**: See table 1. GeoNames is a geographical database that contains over 8 million placenames and corresponding geographical information. It includes information such as location coordinates, population, and administrative divisions. Such as "Pic des Langounelles" a term or an entity with the type "peak". This dataset contains terms without context or usage sentence. This dataset was used for tasks A and B Taxonomy discovery.

3. **Gene Ontology (GO)**: The Gene Ontology dataset (see Table 1) provides a structured vocabulary for representing gene product attributes across species. This dataset includes three domains: **Biological Process**, **Molecular Function**, and **Cellular Component**. As WordNet and GeoName, this dataset contains terms with one or multiple words. An example is following: The term "Tetratricopeptide repeat protein 19, mitochondrial" with the type "mitotic cytokinesis" for biological process.

2.3.1 Hardware and software

The experimentation was conducted in a controlled environment to ensure the reproducibility and reliability of our results.

- The hardware used for our experiments was a laptop Dell Precision 5510, with an Intel Core i7-6820HQ CPU running at 2.70GHz with 8 cores, 16.0 GiB of RAM, and a disk capacity of 756.2 GB.
- The operating system was Ubuntu 22.04.4 LTS.

The BERT-Base uncased was chosen as the LLM to use. In addition, we have chosen GPT-4 as a very large LLM and our goal was to determine in which conditions the foundation model can beats an LLM such as GPT-4.

2.3.2 Experimentation processing

The first step of the experimentation consists of evaluating the performance of GPT-4 on the test data. Thereafter, we have chosen to use BERT-Base uncased as the foundation model. Once the pre-trained model is run on the test data, a manual assessment allow us to define the set of rules to combine with the pre-trained model and the model is tested once. For instance, a manual assessment of the WordNet dataset allowed us to realize that the terms without context was the one that was not well predicted. Thus, we defined a set of rules that we applied on verb, adjectives, and adverbs.

3 Results and Discussion

This section presents the results of the application of our methodology (see Section 2) for the term typing (see Section 3.1) and taxonomy discovery (see Section 3.2) on WordNet, GeoName, and GO datasets.

3.1 Term Typing Task

The following paragraphs presents the results (accompanied with ablation study) on WordNet and Geoname datasets.

3.1.1 Term Typing on WordNet Dataset

Concerning the WordNet Dataset, the BERT-Base uncased model [12] was combined with several rules obtained by assessing the dataset manually. Actually, the manual assessment allowed us to realize that when the context is not provided, BERT failed to identify the type. This allowed us to adapt the equations 1 and 2 in section 2 to the WordNet dataset and obtain the equation below.

$$R_s = \{verb_{rule}, adjective_{rule}, adverb_{rule}\} \quad (3)$$

The following equations describe the rules defined in equation 3.

$$verb\_rule = \{\ verb \text{ if term} \in \{\text{ate}, \text{ify}, \text{ize}\} \wedge |\text{obj\_type}| = 2\} \tag{4}$$

$$adjective\_rule = \{\ adjective \text{ if term} \in \{\text{ible}, \text{able}, \text{al}, \text{ic}, \text{ous}, \text{ful}, \text{ive}\} \wedge |\text{obj\_type}| = 2\} \tag{5}$$

$$adverb\_rule = \{\ adverb \text{ if term ends with "ly"} \wedge |\text{obj\_type}| = 2\} \tag{6}$$

This model was applied on the test data provided by the organizers. Figure. 2 presents the results obtained in comparison with the results of other systems. This figure shows that the system obtained using this model is the best system. It should be noted that this system was run on a simple laptop.

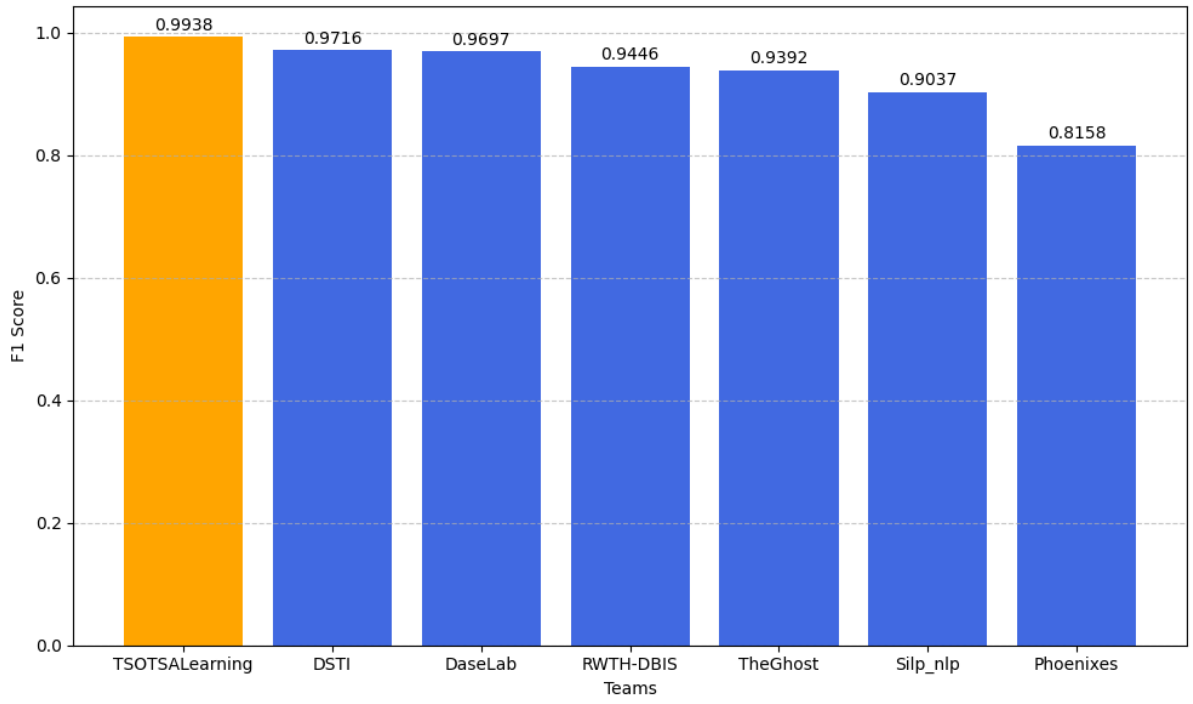

Figure 2. *Comparing the different score obtained per systems submitted to the challenge*

Ablation study

To study the impact of rules on the whole system, several parts of the rules were removed. The table 3 presents the results obtained from our approach, compared to ose achieved using the th obtained using the GPTGPT-4 model and -4 model and usithe rule-based method.ng different rules. This table shows that the performance of the model depends on the completeness of the rules identified.

The low performance compared to the performance when combining BERT-Base uncased with rules, suggest that rules can be an important component when learning ontology using LLMs.

In conclusion, when BERT-Base uncased is enhances with rules for ontology learning, the model obtained can be as powerful as the one obtained using LLMs such as GPT-4.

Table 3. *Results of the ablation study. (1) $BERT_{bu}$: BERT-Base uncased*

Method	Precision	Recall	F-score
$BERT_{bu}$	0.5994	0.9866	0.7457
GPT-4	0.9264	0.9264	0.9264
$BERT_{bu}$ + Verbs	0.9403	0.9403	0.9403
$BERT_{bu}$ + Adjectives	0.9332	0.9332	0.9332
$BERT_{bu}$ + Adverbs	0.9332	0.9332	0.9332
$BERT_{bu}$ + All Rules	**0.9938**	**0.9938**	**0.9938**

Given the results obtained after the experimentation on the WordNet dataset, we decided to adopt this approach for the other datasets. However, the GeoName and GO training datasets were too large and the time to finetune the model, test on the test data was not enough. It requires at least 6 days for all mollecular on our training environment (see Section 2.3.1) and at least 15 days for all geonames. We were able to finetuned the BERT-Base uncased model on only **16.67%** of data for GeoName, **16.67%** of data for Cellular, **16.67%** of data for molecular. During this process, a manual assessment of the dataset allowed us to identify several rules that can be used to enhance the LLM once finetuned.

3.1.2 Term Typing on GeoNames Dataset

The equation 1 presents the model used for term typing on GeoNames dataset.

This model was applied to the test data provided by the organizers. Figure. 3 presents the results obtained in comparison with the results of other systems. This figure shows that the system obtained using this model has the fourth position.

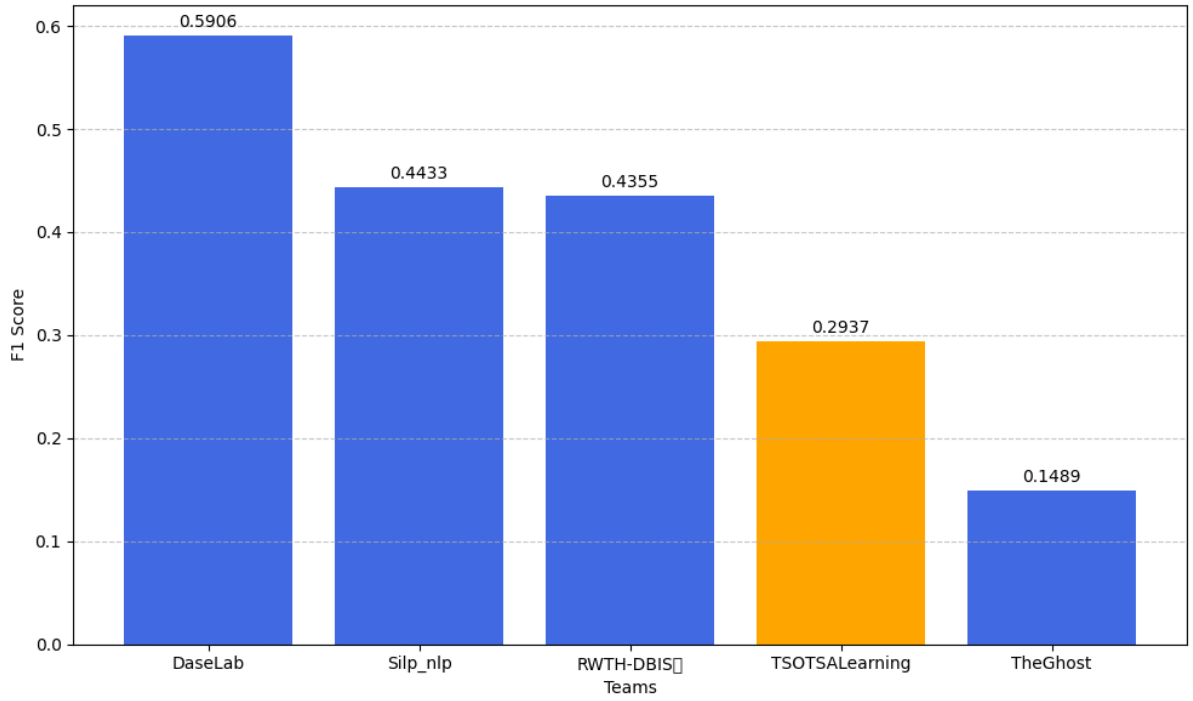

Figure 3. *Comparison of the F1-score of the TSOTSALearning system with other systems on the GeoNames dataset*

Ablation study

To study the impact of rules on the whole system, the rules applied on the GeoName dataset was removed and the system was evaluated on the test data. The results obtained are presented by table 4. This table shows that only rules allow to obtained the 0.2937 of F1-score. It should be noted that the model was finetuned on only **16.67%** of the training dataset.

Table 4. *Results of the ablation study. (1) BERT_bu: BERT-Base uncased*

	rules applied	BERT_bu	GPT-4
Precision	**0.2937**	0.0000	0.0000
Recall	**0.2937**	0.0000	0.0000
F-score	**0.2937**	0.0000	0.0000

3.1.3 Term Typing on Cellular Component Dataset

Similar to the WordNet and the GeoName dataset, the model defined (see equation 1) was applied on the "Cellular Component Dataset". The results obtained, compared with other systems are presented by the Figure. 4.

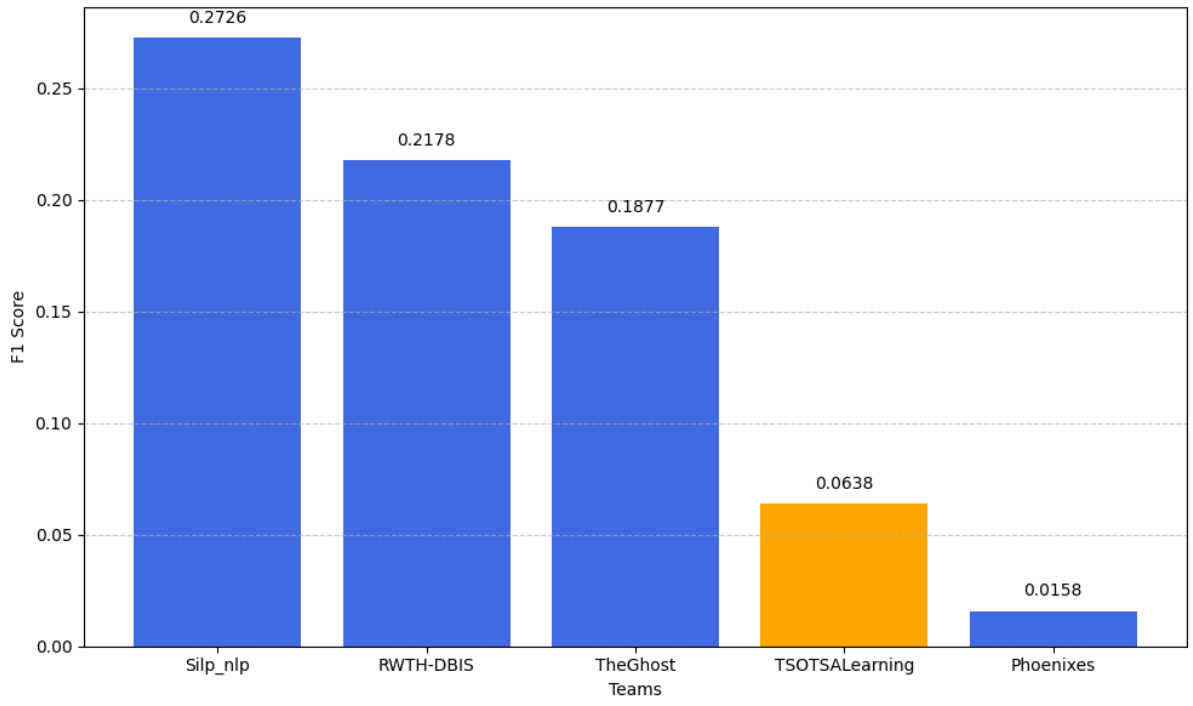

Figure 4. *Application of the TSOTSALearning system on test set of the Cellular dataset*

3.1.4 Term Typing on Biological Process Dataset

Concerning the Biological Process, the BERT-Base uncased model was pretrained, combined with rules (see Figure. 5) applied to the test data and submitted on the codalab platform for evaluation.

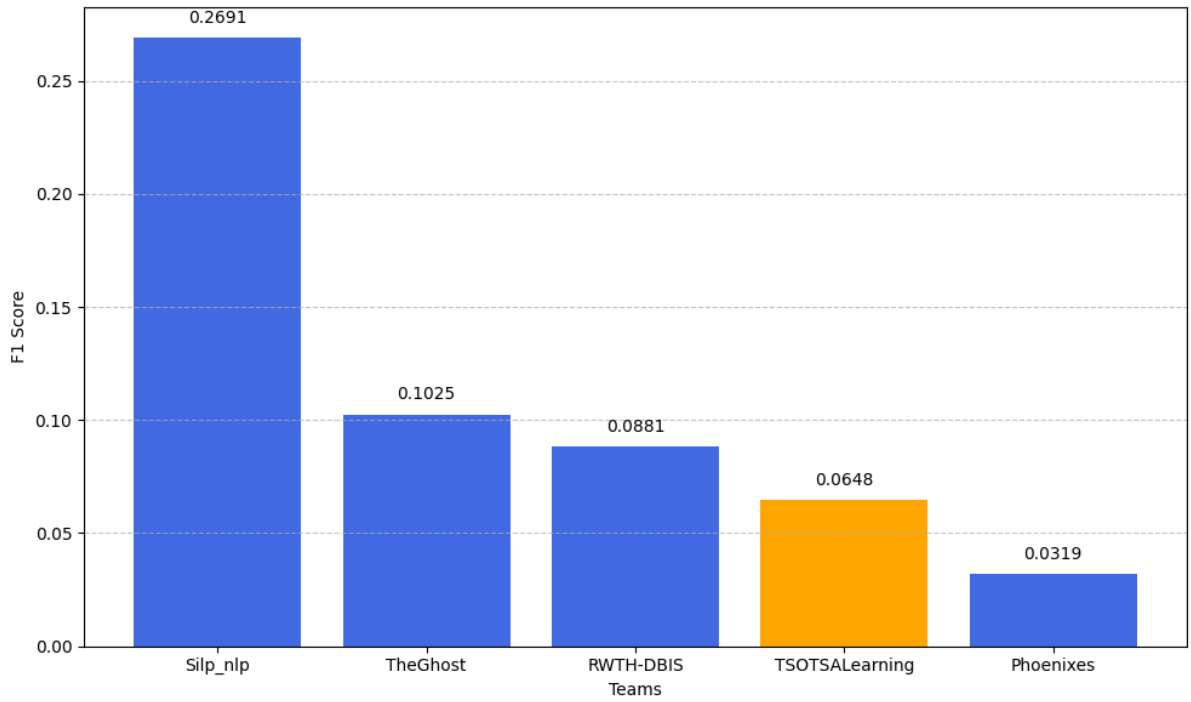

Figure 5. *Application of the BERT-Base uncased model on the Biological Process dataset*

3.1.5 Term Typing on Molecular Function Dataset

Concerning the Molecular Function, the BERT-Base uncased model was pre-trained, combined with rules and applied to the test data. Figure. 6 presents the results obtained compared to the results of other systems.

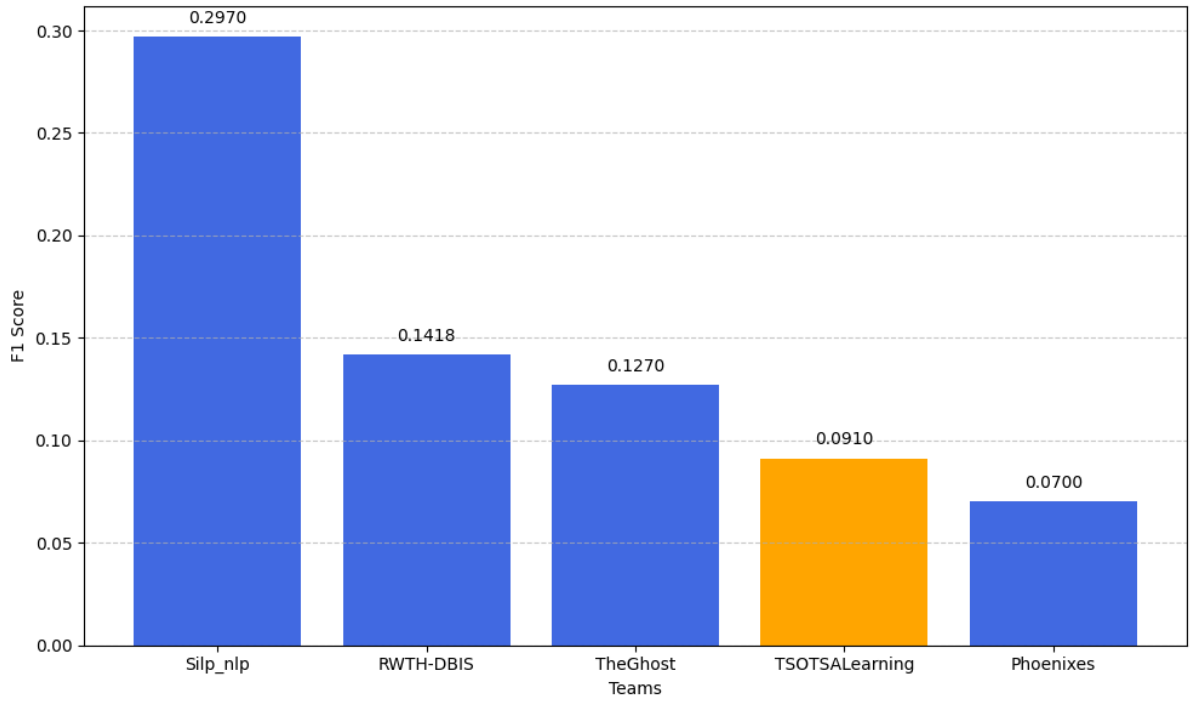

Figure 6. *Application of the BERT-Base uncased model on the Molecular dataset*

3.2 Taxonomy Discovery on GeoNames Dataset

During the taxonomy discovery task, given the time for submitting our results, only the BERT-Base uncased model was used on the GeoName dataset. Figure. 7 presents the results obtained compared to the results of other participants. This figure shows that the system proposed occupy the fourth position.

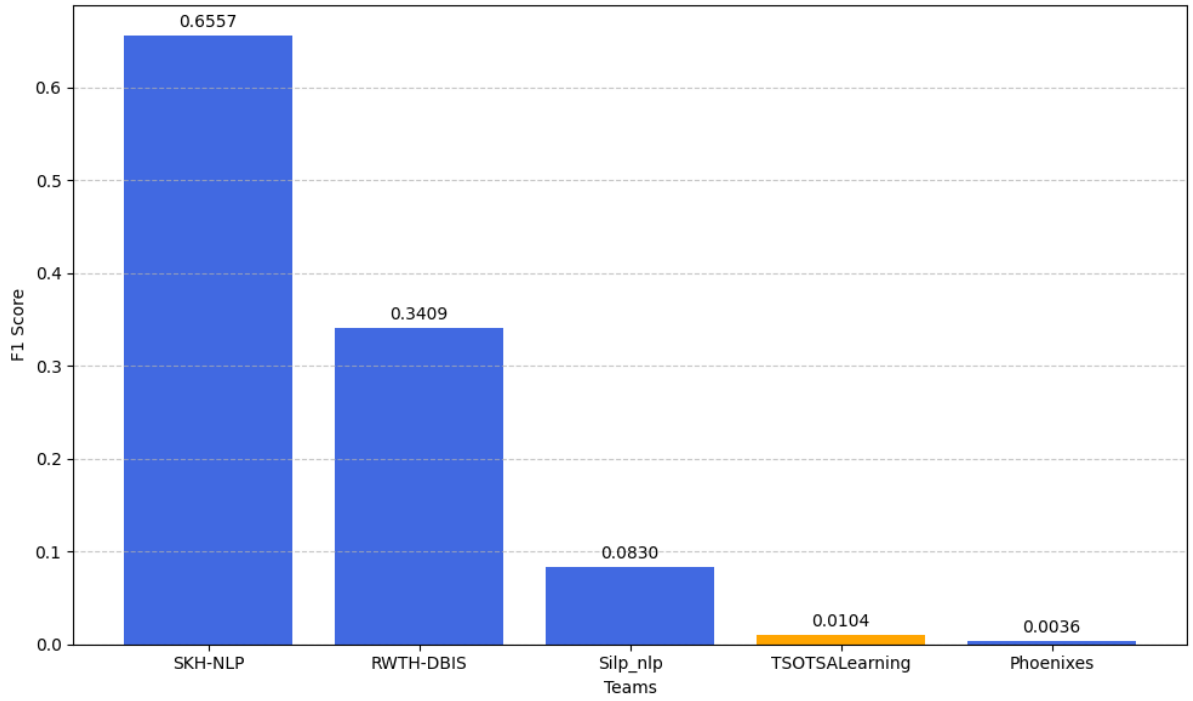

Figure 7. *Application of the BERT-Base uncased model on the GeoName dataset*

3.3 Conclusion

The results for the taxonomy discovery task, Fig 7 reveal considerable challenges, particularly relating to the accuracy of predictions. The low f1 score on Geonames, despite higher recall suggests that the model identifies many potentially relevant terms but has difficulty avoiding false positives. This highlights the complexity of taxonomic relationships and the importance of improving the accuracy of the model.

4 Conclusion

This research aims to determine in which conditions foundations models such as BERT can be used for ontology learning. A set of experimentation's was conducted using BERT and compared the results obtained to the results obtained using GPT-4. The results obtained on the WordNet dataset show that merging the strengths of LLMs with **rule-based strategies**, enhances the accuracy of ontology learning. The ablation study consists of comparing the performance of the LLM alone and the combination of the LLM with rules. This suggests that rules can be an important component when learning ontologies using LLMs. It should be noted that identifying rules to used is not an easy task. Future work consists of automatic detection of rules and the possibility to inject the rules in the LLM.

Author Contributions

Carick Appolinaire Atezong Ymele: Conceptualization, Methodology, Software, Writing - Original Draft, Writing - Review & Editing.
Azanzi Jiomekong: Conceptualization, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing, Supervision.

Competing interests

The authors declare that they have no competing interests.

References

[1] F. J. Azanzi, G. Camara, and M. Tchuente, "Extracting ontological knowledge from java source code using hidden markov models," *Open Computer Science*, vol. 9, no. 2, pp. 181–199, Aug. 2019. DOI: 10.1515/comp-2019-0013.

[2] H. Zaragoza, P. M. D. Cabeza, and J. R. Sanz, "Learning ontologies from text: A survey of approaches and techniques," *Journal of Computer Science and Technology*, vol. 32, no. 1, pp. 1–14, 2017. DOI: 10.1007/s11390-016-1662-0.

[3] P. F. Patel-Schneider, "A framework for ontology extraction from databases," in *Proceedings of the International Workshop on Ontology Learning*, Springer, 2005. DOI: 10.1007/11516172_12.

[4] R. Meersman, A. L. de Moor, and H. W. de Bruijn, "Ontology-based xml data management," *Data Knowledge Engineering*, vol. 55, no. 1, pp. 1–10, 2005. DOI: 10.1016/j.datak.2004.11.005.

[5] S. G. J. Zeng and H. M. Xie, "Ontology extraction from vocabularies and knowledge bases: A survey and new method," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 8, pp. 2267–2280, 2015. DOI: 10.1109/TKDE.2014.2345382.

[6] A. Jiomekong, A. Oelen, S. Auer, L. Anna-Lena, and V. Lars, "Food information engineering," *AI Magazine*, 2023. DOI: 10.1002/aaai.12185.

[7] H. T. Azanzi Jiomekong Martins Folefac, "Food composition knowledge extraction from scientific literature," in *Artificial Intelligence: Towards Sustainable Intelligence, AI4S 2023*, S. Tiwari, F. Ortiz-Rodríguez, S. Mishra, E. Vakaj, and K. Kotecha, Eds., ser. Communications in Computer and Information Science, vol. 1907, Springer, Cham, 2023, pp. 89–103, ISBN: 978-3-031-47996-0. DOI: 10.1007/978-3-031-47997-7_7. [Online]. Available: https://doi.org/10.1007/978-3-031-47997-7_7.

[8] G. C. Azanzi Jiomekong Hippolyte Tapamo, "An ontology for tuberculosis surveillance system," in *Iberoamerican Knowledge Graphs and Semantic Web Conference*, Springer Nature Switzerland, 2023, pp. 1–15.

[9] 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*, Springer, 2023, pp. 408–427. DOI: 10.1007/978-3-031-47240-4_22.

[10] 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.

[11] 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.

[12] D. Jacob, C. Ming-Wei, L. Kenton, and T. Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2019, pp. 4171–4186.