Learning Domain Ontologies Based On Top-Level Ontology Concepts Using Language Models And Informal Definitions

Alcides Lopes[1,*,†], Joel Carbonera[1,†] and Mara Abel[1,†]

[1]Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

Abstract

Ontology development is a challenging task that encompasses many time-consuming activities. One of these activities is the classification of the domain entities (concepts and instances) according to top-level concepts. This activity is usually performed manually by an ontology engineer. However, when the set of entities increases in size, associating each domain entity to the proper top-level ontological concept becomes challenging and requires a high level of expertise in both the target domain and ontology engineering. In this context, this work describes an approach for learning domain ontologies based on top-level ontology concepts using informal definitions as input. In our approach, we used informal definitions of the domain entities as text input of a language model that predicts their proper top-level concepts. Also, we present a methodology to extract datasets from existing domain ontologies to evaluate the proposed approach. Our experiments show that we have promising results in classifying domain entities into top-level ontology concepts.

Keywords

Ontology learning, Deep neural network, Text classification

1. Introduction

Over the years, ontologies have proved valuable in many domains, such as geology [1, 2] and biomedicine [3, 4]. In the literature, some methodologies for ontology development stands on a more abstract ontology, called top-level ontology [5, 6, 7], to explicitly define the ontological nature of the domain entities through the specialization of top-level concepts. The domain ontologies developed based on top-level ontologies have the advantage of adhering to a philosophically well-founded meaning. However, identifying which top-level concept generalizes a domain entity in complex domains is a laborious and time-consuming task that requires manual work and a high level of expertise in both the target domain and ontology engineering [8].

In this work, we proposed an approach to learning domain ontologies based on top-level ontology concepts using language models and the informal definitions of the domain entities. In our view, automatizing the task of learning which top-level concept a domain entity specializes

ONTOBRAS 2022 - 15th Seminar on Ontology Research in Brazil, 22–25 November 2022, Online

*Corresponding author.

†These authors contributed equally.

agljunior@inf.ufrgs.br (A. Lopes); jlcaronera@inf.ufrgs.br (J. Carbonera); marabel@inf.ufrgs.br (M. Abel)

0000-0003-0622-6847 (A. Lopes); 0000-0002-4499-3601 (J. Carbonera); 0000-0002-9589-2616 (M. Abel)

has many benefits to ontology engineering, mainly because allowing a more rational resource allocation in ontology development processes since the ontology engineer can invest more time in more complex tasks. Also, we hypothesize that knowing the top-level concept of the domain entities can help in automatically discovering the relationships between that domain entities. In order to evaluate our proposal, we proposed to extract datasets from existing domain ontologies developed based on two top-level ontologies: Dolce-Lite (DL) and Dolce-Lite-Plus (DLP) [9]. Our experiments show that language models have promising results in classifying domain entities into top-level ontology concepts, with 94% of micro F1-score.

The paper is organized as follows: firstly, we describe the background notions that support this proposal. Secondly, we present the current state of our research, describing the dataset extraction and the proposed approach for learning domain ontologies based on top-level ontology concepts using language models and informal definitions. After that, we show the current results achieved from our experiments. Finally, we present our future works.

2. Related Works

The manual development of ontologies is complex and laborious, bringing a significant challenge to the natural language processing area on automatically creating ontologies as sophisticated as those made by humans. In this context, several approaches propose automatizing specific activities in the ontology development process, such as domain entity identification and classification [10, 11, 8, 12, 13], semantic relation identification and classification [14, 12], etc.

In [11], the authors proposed an automatic methodology to extract domain entities from text corpora based on a domain-specific thesaurus and rank them based on their frequency in the corpora. Afterward, the geology engineers manually provided definitions to the selected entities, and domain experts manually classified them according to their respective concepts in the GeoCore ontology [15].

In [12], the authors proposed a framework to automatize domain ontology learning. Their framework has four steps: 1) Extract sentences from text corpora and extract their triples using NLP tools. 2) Matches the extracted sentences with public knowledge bases to extract the concept pairs. 3) Labels common concept pairs between the outputs of steps 1 and 2. 4) Learns the relationship between each concept pair using a bi-directional long-short term memory (Bi-LSTM) model and a convolutional neural network (CNN).

In [10], the authors proposed a methodology for ontology learning from big data. They used word embedding and hierarchical clustering to improve the quality of the ontology entities extraction from textual corpora and reduce the processing time. Their methodology started by extracting the most relevant word of the domain. After that, they identified the concepts and their properties using a set of pre-defined rules based on the Part-Of-Speech (POS) tag. The authors created the ontology hierarchy from the identified concepts by combining a word embedding model and a hierarchical clustering algorithm. Finally, they used the word embedding of the properties and the concepts to identify if the properties are data properties or object properties. The authors tested the proposed methodology using gold-standard datasets for ontology learning.

In [13], the authors proposed a semi-automatic methodology for ontology learning based

on a domain-specific corpus. The first step of this methodology is setting the main classes of the ontology and selecting the terms that specialize them. The approach accomplished the term selection by extracting all adjectives, nouns, and verbs from the domain corpus and requiring user intervention to classify some of these terms under the main classes. Thus, the methodology used a similarity function to classify the remaining entities according to the most similar previously classified terms. Finally, the authors used hierarchical clustering algorithms to build the final ontology hierarchy. The main drawback of this methodology is the necessity of human intervention in all of those steps.

Here, we revised some recent works on the ontology learning domain, but many more works propose solutions to learn ontologies from text [16, 17]. However, no one of them focuses on the task of learning the top-level concepts of the domain entities. In our view, this task is essential to create more powerful domain ontologies by adhering to a philosophically well-founded meaning. Also, in this context, there is a lack of datasets to evaluate new proposals on this line. Thus, in this work, in addition to proposing an approach to accomplish this learning task, we also propose several datasets extracted from existing domain ontologies developed under top-level concepts.

3. Current work

In this section, we first describe the methodology used to select the domain ontologies developed under top-level ontology concepts and the process performed to extract the datasets from these ontologies. After that, we present our proposal for learning domain ontologies based on top-level ontology concepts using language models and informal definitions.

3.1. Dataset Extraction

In our work, we aim to find domain ontologies developed under top-level ontologies containing informal definitions of their domain entities to build the datasets used for evaluating our proposed approach. In this context, in order to extract the datasets for predicting the top-level concepts of Dolce-Lite and Dolce-Lite-Plus ontologies, we select the OntoWordNet ontology [9]. This general domain ontology aligns 86,982 entities obtained from the WordNet synsets [18] with the Dolce-Lite-Plus (DLP) ontology structure. This top-level ontology is an extension of the Dolce-Lite (DL) ontology with several modules for representing information, communication, plans, and domain information, for example, legal and biomedical notions. Thus, for each domain entity extracted from OntoWordNet ontology, we took the lowest DLP and DL top-level concepts that it specializes, its informal definition, and the set of its labels. For each label in this set, we created an instance inside both DLP and DL datasets containing the label, the informal definition, and the respective top-level concept referent to the dataset. Finally, the Dolce-Lite-Plus dataset contains 90 classes, the Dolce-lite dataset contains 20 classes, and both datasets have 120,489 instances.

3.2. Learning the top-level concepts of domain entities

In our view, developing an approach that automatizes the prediction of the top-level concepts of the domain entities using only the informal definitions of these domain entities has many benefits for ontology engineering and artificial intelligence fields. For example, we can use this approach as a decision support system, thus allowing a more rational resource allocation in ontology development processes since the ontology engineer can invest more time in more complex tasks. Also, we can insert the notion of top-level ontology concepts in natural language processing tasks (e.g., named entity recognition, text classification, relationship prediction, etc.). As the last example, we can use this approach to predict the concepts of any top-level ontology, thus allowing the development of a classification system for multiple top-level ontologies.

We explored several machine-learning approaches to develop architectures for classifying domain entities into top-level ontology concepts using the informal definition of these domain entities. In this context, we have used three kinds of architecture. In the first kind, we explored the word embeddings of the terms that represent the domain entities and input these word embeddings in several machine learning models, such as Random Forest (RF), Linear Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Bernoulli Naive Bayes (BNB), Gaussian Naive Bayes (GNB), Feed-Forward neural network (FNN), and a bi-LSTM neural network. However, this architecture fails to deal with polysemic words (words with more than one sense) as the pre-trained word embedding models. In the second kind, we insert the informal definitions into a bi-LSTM neural network and concatenate its hidden layers with the hidden layer of the previous architecture before outputting the predicted class. Thus, we solve the problem of polysemic words. In the third architecture, we adopted language models (e.g., BERT, ELECTRA, ALBERT, etc.) and combined the term and the informal definitions in a single string before inputting it into the model. We achieved our best results using the third architecture.

4. Current experiments and results

We develop two experiments to evaluate the three proposed architectures in classifying domain entities into concepts specified by top-level ontologies. In both experiments, we applied the stratified k-fold cross-validation approach (with $k = 10$) to split the DLP dataset into train and test folds. In the first experiment, we compared the first proposed architecture against the second proposed architecture. Also, we selected the 30 most populated classes in the DLP dataset. From these classes, we downsample the training fold according to the size of the less populated class and insert the removed instances into the test fold. In the second experiment, we compared the second architecture with the proposed third architecture using the BERT-Base language model. In this experiment, we fined-tuned the BERT-Base model for our classification task and do not apply any sampling strategy in the train and test folds. Also, we selected the top 82 classes of the DLP dataset.

According to Table 1, in the first experiment, the second architecture achieved 59% of the F1-micro score against 57% of the first architecture with the SVM model. These results suggest that combining the informal definitions with the term representing the domain entities improves the performance of classifying domain entities into top-level concepts and also solves the first architecture's polysemy problem. Thus, making it possible to use more instances in the training

Experiment	Architecture	F1-micro
1	First architecture with SVM model	.57
	Second architecture	**.59**
2	Second architecture	.54
	Third architecture with BERT-Base model	**.94**

Table 1
The comparison between each architecture in the two performed experiments

and test folds. However, in the second experiment, the third architecture with the BERT-Base model reached an outstanding result compared with the second architecture. In this experiment, the third architecture achieves 94% of the micro F1-score against 54% of the second architecture. One reason for this is due to combining the term and the informal definition into a single textual sentence. Another reason is those language models are state-of-the-art for natural language processing tasks.

5. Future works

Extracting novel datasets. Nowadays, we have datasets for the Dolce-Lite, Dolce-Lite-Plus, and BFO top-level ontologies. Nevertheless, we aim to increase the number of instances in each dataset by exploring other domain ontologies developed under these top-level ontologies, or knowledge graphs developed from WordNet, such as BabelNet. From the BabelNet, we have access to multiple sources of informal definitions in many languages.

Learning the relationships between domain entities. From the presented architectures, we aim to combine the task of learning the top-level concepts of domain entities with the task of predicting the relationship between domain entities. We hypothesize that we can improve the results of the latter task by previously classifying the top-level concept of the evaluated domain entities.

Acknowledgements

The authors gratefully acknowledges the financial support of the Brazil Federal Agencies CAPES and CNPQ, and the grant and scientific cooperation of PETROBRAS company.

References

[1] L. F. Garcia, M. Abel, M. Perrin, R. dos Santos Alvarenga, The geocore ontology: A core ontology for general use in geology, Computers & Geosciences 135 (2020) 104387.
[2] F. Cicconeto, L. V. Vieira, M. Abel, R. dos Santos Alvarenga, J. L. Carbonera, L. F. Garcia, Georeservoir: An ontology for deep-marine depositional system geometry description, Computers & Geosciences 159 (2022) 105005.

[3] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, Chebi: a database and ontology for chemical entities of biological interest, Nucleic acids research 36 (2007) D344–D350.
[4] G. O. Consortium, The gene ontology resource: 20 years and still going strong, Nucleic acids research 47 (2019) D330–D338.
[5] N. Guarino, C. A. Welty, An overview of ontoclean, Handbook on ontologies (2004) 151–171.
[6] G. Guizzardi, Ontological foundations for structural conceptual models, Ph.D. thesis, University of Twente, 2005.
[7] R. Arp, B. Smith, A. D. Spear, Building ontologies with basic formal ontology, Mit Press, 2015.
[8] A. G. L. Junior, J. L. Carbonera, D. Schimidt, M. Abel, Predicting the top-level ontological concepts of domain entities using word embeddings, informal definitions, and deep learning, Expert Systems with Applications 203 (2022) 117291.
[9] A. Gangemi, R. Navigli, P. Velardi, The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet, in: R. Meersman, Z. Tari, D. C. Schmidt (Eds.), On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 820–838.
[10] N. Mahmoud, H. Elbeh, H. M. Abdlkader, Ontology learning based on word embeddings for text big data extraction, in: 2018 14th International Computer Engineering Conference (ICENCO), 2018, pp. 183–188.
[11] L. F. Garcia, F. H. Rodrigues, A. Lopes, R. d. S. A. Kuchle, M. Perrin, M. Abel, What geologists talk about: Towards a frequency-based ontological analysis of petroleum domain terms., in: ONTOBRAS, 2020, pp. 190–203.
[12] J. Chen, J. Gu, Adol: a novel framework for automatic domain ontology learning, The Journal of Supercomputing 77 (2021) 152–169.
[13] F. ten Haaf, C. Claassen, R. Eschauzier, J. Tjan, D. Buijs, F. Frasincar, K. Schouten, Web-soba: Word embeddings-based semi-automatic ontology building for aspect-based sentiment classification, in: R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, M. Alam (Eds.), The Semantic Web, Springer International Publishing, Cham, 2021, pp. 340–355.
[14] A. Lamurias, D. Sousa, L. A. Clarke, F. M. Couto, Bo-lstm: classifying relations via long short-term memory networks along biomedical ontologies, BMC bioinformatics 20 (2019) 1–12.
[15] L. F. Garcia, M. Abel, M. Perrin, R. dos Santos Alvarenga, The geocore ontology: a core ontology for general use in geology, Computers & Geosciences 135 (2020) 104387.
[16] A. Konys, Knowledge repository of ontology learning tools from text, Procedia Computer Science 159 (2019) 1614–1628.
[17] J. Watrobski, Ontology learning methods from text-an extensive knowledge-based approach, Procedia Computer Science 176 (2020) 3356–3368.
[18] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.