How to classify domain entities into top-level ontology concepts using large language models: A study across multiple labels, resources, and languages

Alcides Lopes[1,*], Joel Carbonera[1], Fabricio Rodrigues[1], Luan Garcia[1,2] and Mara Abel[1]

[1]Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre, 15064, Brazil
[2]Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga 6681, Porto Alegre, Brazil

Abstract

Classifying domain entities into their respective top-level ontology concepts is a complex problem that typically demands manual analysis and deep expertise in the domain of interest and ontology engineering. Using an efficient approach to classify domain entities enhances data integration, interoperability, and the semantic clarity of ontologies, which are crucial for structured knowledge representation and modeling. Based on this, our main motivation is to help an ontology engineer with an automated approach to classify domain entities into top-level ontology concepts using informal definitions of these domain entities during the ontology development process. In this context, we hypothesize that the informal definitions encapsulate semantic information crucial for associating domain entities with specific top-level ontology concepts. Our approach leverages state-of-the-art language models to explore our hypothesis across multiple languages and informal definitions from different knowledge resources. In order to evaluate our proposal, we extracted multi-label datasets from the alignment of the OntoWordNet ontology and the BabelNet semantic network, covering the entire structure of the Dolce-Lite-Plus top-level ontology from most generic to most specific concepts. These datasets contain several different textual representation approaches of domain entities, including terms, example sentences, and informal definitions. Our experiments conducted 3 study cases, investigating the effectiveness of our proposal across different textual representation approaches, languages, and knowledge resources. We demonstrate that the best results are achieved using a classification pipeline with a K-Nearest Neighbor (KNN) method to classify the embedding representation of informal definitions from the Mistral large language model. The findings underscore the potential of informal definitions in reflecting top-level ontology concepts and point towards developing automated tools that could significantly aid ontology engineers during the ontology development process.

Keywords

Top-level ontology classification, Informal definition, Ontology learning, Language Model

1. Introduction

In recent years, two significant areas in Artificial Intelligence (AI) have collided again. On the one hand, we have ontologies, defined as *a formal and explicit specification of a shared conceptualization* [1]. On the other hand, we have the advances in natural language processing (NLP) for text representation and classification with Large Language Models (LLMs) [2, 3, 4, 5]. This convergence has ushered in a new renaissance in ontology learning from text, in which the goal is to generate ontologies in an automatic or semi-automatic way, using text as input and state-of-the-art NLP techniques. While there are several approaches to tackle this challenge [6, 7, 8], the field is still rife with open questions and opportunities for exploration.

Among various uses of ontologies, knowledge representation and data management are two usages that stand out because ontologies provide a structured and standardized way to represent and orga-

Proceedings of the Joint Ontology Workshops (JOWO) - Episode X: The Tukker Zomer of Ontology, and satellite events co-located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), July 15-19, 2024, Enschede, The Netherlands.

*Corresponding author.
agljunior@inf.ufrgs.br (A. Lopes); jlcarbonera@inf.ufrgs.br (J. Carbonera); fabricio.rodrigues@inf.ufrgs.br (F. Rodrigues); lfgarcia@inf.ufrgs.br (L. Garcia); marabel@inf.ufrgs.br (M. Abel)
0000-0003-0622-6847 (A. Lopes); 0000-0002-4499-3601 (J. Carbonera); 0000-0002-0615-8306 (F. Rodrigues); 0000-0001-9328-9007 (L. Garcia); 0000-0002-9589-2616 (M. Abel)

nize complex information, allowing for improved data integration and interoperability and enabling information retrieval and knowledge sharing across different systems or resources [9, 10, 11]. Also, ontologies enhance semantic clarity by precisely defining domain entities and their relationships, reducing ambiguity, and ensuring a common understanding of complex knowledge domains. In this context, state-of-the-art ontology engineering methodologies, such as NeOn [12], recommend using a top-level ontology to ensure these benefits when using domain ontology.

A top-level ontology defines a foundational and high-level structure for categorizing and organizing knowledge across various domains [13, 14, 15]. It serves as a broad and abstract framework for representing fundamental concepts and relationships that are universally applicable and not tied to any specific field. Thus, top-level ontologies are the starting point for organizing and categorizing domain-specific knowledge represented in domain ontologies and act as a common framework for the semantic integration of domain ontologies or data assets from distinct sources. However, identifying which top-level ontology concept a domain entity specializes in is laborious and time-consuming. This task is usually performed manually and requires expertise in the target domain and ontology engineering. Although ontology learning approaches often focus on tasks like entity recognition, extraction, and relationship extraction, which are crucial for building ontologies from text sources [6, 7, 8], classifying domain entities into top-level ontology concepts is particularly challenging, involving not only entity identification and extraction but also understanding the theoretical foundation and implications of choosing one top-level concept over another.

In this work, we addressed the problem of classifying domain entities into top-level ontology concepts using informal definitions to represent domain entities textually. From that, our central hypothesis is that the informal definitions represent semantic information that allows domain entities to be related to top-level ontology concepts. In this context, we leveraged two crucial ideas. Firstly, since top-level ontologies also contain a taxonomy of concepts, i.e., concepts related through hierarchical relationships (e.g., “is-a" and “subclass-of"), we used the notion of similarity in taxonomies, which more specific entities grouped in the same more general entity share common features or attributes making them more similar to each other. Secondly, since our hypothesis uses a textual representation for domain entities, we leverage the Distributional Hypothesis [16], which suggests that words that occur in similar contexts are likely to have related meanings, i.e., they are closer in the distributional space. Therefore, since informal definitions provide the intended meaning of a domain entity in a particular domain [17] and possibly similar domain entities have similar informal definitions, we can use these definitions to find the top-level ontology concepts of domain entities.

In order to validate our hypothesis, we proposed the extraction of multi-label, multi-language, and multi-resource datasets from the alignment between OntoWordNet ontology [18] and the BabelNet semantic network [19, 20]. From this alignment, we extract datasets in 291 languages and with various informal definitions resources, such as WordNet [21], and Wikipedia, and their variations. Also, we proposed two supervised classification pipelines for classifying domain entities into top-level concepts using text as input. The first pipeline utilizes a pre-trained language model that we fine-tuned to our target task. The second involves using a pre-trained language model to generate the embedding representation of the input text. From the embedding, we used classical machine learning classifiers (e.g., K-Nearest Neighbors, Decision Tree, and Support Vector Machine) to classify the embedding into top-level ontology concepts.

In the experiments covered in this work, we evaluated the performance of our proposal by training and testing the proposed pipelines using informal definitions from different knowledge resources, such as WordNet and Wikipedia[1]. The results suggest using informal definitions is the best way to represent domain entities textually and classify them into top-level ontology concepts. Also, our results show that our hypothesis is valid for informal definitions from different resources. Furthermore, although fine-tuning a classification model for a specific task has promising results in many fields, our work shows that employing a K-Nearest Neighbor (KNN) approach using the embedding representation of informal definitions as input has better results and avoids the computation costs of fine-tuning.

[1]For the full experimental evaluation, read the original paper in [22]

We achieved our best result using the Mistral7B language model in the second pipeline with a KNN classifier.

The paper is organized as follows. In section 2, we present the notion of ontologies in Computer Science and top-level ontologies. We also describe the Distributional Hypothesis in computational linguistics. In section 3, we introduce our proposed approach of using informal definitions for classifying domain entities into top-level ontology concepts, detailing dataset extraction, classification pipelines, and training methodologies. In section 4, we discuss the practical results and effectiveness of the proposed classification pipelines over three study cases. Finally, in section 5, we offer concluding remarks on our work[2].

2. Background

In this section, we discuss the use of ontologies in computer science, contrasting their philosophical origins and highlighting their role in knowledge modeling. We focus mainly on top-level ontologies because they are general frameworks for knowledge representation across various domains. Additionally, we review the Distributional Hypothesis, illustrating its importance in linguistics and computational linguistics, which influenced the development of word embedding and state-of-the-art language models.

2.1. Ontologies and Top-level Ontologies

Ontologies are tools to support knowledge modeling. In Computer Science, ontology developers commonly use structured languages such as the Web Ontology Language (OWL) to implement ontologies. This approach allows computers to interpret and manage organized data effectively. From that, ontologies bridge human conceptual understanding and machine readability, thus facilitating more efficient and accurate data processing, analysis, and decision-making in complex systems and enabling semantic interoperability between different resources. In this context, a top-level ontology (a.k.a, upper-level, foundational, or general ontologies) is a kind of ontology that facilitates the interoperability of knowledge across different domains [23]. This type of ontology encapsulates fundamental concepts and principles that are universally applicable, regardless of the domain, including general concepts like time, space, events, objects, relationships, and qualities. From that, the primary purpose of top-level ontologies is to establish a shared understanding and a universal vocabulary from which the entities in the other types of ontologies can subsume and ensure consistency and interoperability between them. In this context, several top-level ontologies are proposed, for example, BFO [24, 15], DOLCE [25, 13], SUMO [26], UFO [14], among others.

In this work, we focused only on DOLCE's top-level ontology structure, particularly on the DOLCE-Lite-Plus (DLP)[3]. The DLP top-level ontology represents the DOLCE [25] foundations using an OWL representation approach, extending the original DOLCE structure with other top-level concepts for representing descriptions, situations, temporal relations, information objects, actions, agents, social units, collections, and collectives. The whole taxonomic structure of DLP contains a total of 244 top-level concepts.

2.2. Distributional Hypothesis

The Distributional Hypothesis [16] presents a foundational idea in linguistics, suggesting that words that occur in similar contexts are likely to have related meanings. This hypothesis proposes a method for inferring the semantics of words by analyzing their distributional patterns across texts. The approach proposed by Harris [16] to structural linguistics emphasized the importance of context in understanding linguistic meaning, suggesting that the semantic attributes of words could be uncovered through a systematic examination of their usage in various linguistic environments. By focusing on empirical language analysis, Harris established a semantic analysis framework that relies on observable,

[2]The source code and data are available at https://github.com/BDI-UFRGS/ALopes-AO2024
[3]http://www.ontologydesignpatterns.org/ont/dlp/DLP_397.owl

quantitative data, shifting away from more introspective or purely theoretical methods. This framework has profoundly influenced how linguists and computational linguists approach language study by suggesting that words' meanings can be deduced from their use patterns.

This hypothesis also had a significant impact on the development of computational linguistics, particularly in creating technologies like word embedding [27, 28, 29], and language models [3, 2, 4, 5]. These models represent words as vectors in a high-dimensional space, where the proximity between vectors indicates semantic similarity based on their distributional properties. This approach has enabled natural language processing (NLP) advances, allowing for a more nuanced and effective machine understanding of human language. The practical applications of the Distributional Hypothesis are evident in various NLP tasks, including machine translation, information retrieval, and sentiment analysis, demonstrating its vital role in bridging linguistic theory and computational applications.

3. Proposed Approach

This section details an approach for classifying domain entities into top-level ontology concepts using informal definitions as input. In this context, we present a methodology to extract multi-label, multi-language, and multi-resource datasets from the alignment between OntoWordNet and BabelNet. Then, we advocate for using informal definitions as the optimal textual representation for domain entities, proposing two classification pipelines leveraging language models to predict top-level ontology concepts. Finally, we present training methodologies for the data augmentation technique called the "explode" approach.

3.1. Dataset Extraction

The OntoWordNet project plays a crucial role in aligning WordNet synsets with the top-level concepts of the Dolce-Lite-Plus (DLP) ontology, effectively transforming WordNet from a lexical database into a structured ontology. This alignment process categorizes synsets into different groups, including concept-synsets, relation-synsets, meta-property-synsets, and individual-synsets. From that, OntoWordNet not only organizes 65,973 domain entities but also links these entities to 120 top-level concepts from the structure of DLP, which consists of 244 concepts. This structured alignment highlights the uneven distribution of synsets across DLP concepts, revealing areas within the ontology where certain concepts are overrepresented while others are underrepresented. Understanding this distribution is essential for identifying the richness or sparsity of domain entities within the ontology, which has implications for the usability and coverage of OntoWordNet in real-world applications.

WordNet divides the informal definitions of its synsets into two components: definiendum and definiens. The definiendum refers to each term within a synset, indicating the subject of the definiens, which provides the explanation or descriptive content of the synset's meaning. The goal of this division is to maintain clarity and precision in defining synsets. Also, this division has been adopted for the OntoWordNet domain entities. In addition, this structured organization facilitates mapping OntoWordNet entities to other semantic networks, such as BabelNet. BabelNet extends the coverage of WordNet by integrating its lexical database with encyclopedic knowledge from Wikipedia, spanning multiple languages. In this work, we take advantage of the shared use of English WordNet's definienda and definientia to align OntoWordNet and BabelNet entities.

The alignment between OntoWordNet and BabelNet results in a dataset containing 65,018 domain entities. This dataset marks a transition from a single-language resource, which relied solely on English definitions from WordNet, to a multi-language and multi-resource dataset enriched by the diverse elements of BabelNet. This expanded dataset allows for the inclusion of various forms of data, such as example sentences, Wikipedia pages, images, and more. By integrating these different types of data, the dataset gains a greater depth and breadth of linguistic and semantic information, which enhances its value for a wide range of applications, particularly in natural language processing and ontology-based research.

To further enhance the dataset, a transformation is applied to convert it from a multi-class to a multi-label format. In the multi-class format, each domain entity is associated with only one top-level concept, which can limit the ability to represent the multiple inheritances that domain entities may have. The multi-label transformation assigns a branch of the top-level ontology, from the most generic to the most specific concept, to each domain entity. For instance, a domain entity like “rock" might be classified under “Amount of Matter" in a multi-class format. In the multi-label format, “rock" would also be categorized under “Endurant," a more general concept in the DLP hierarchy, capturing its broader classification. This transformation addresses the class imbalance issue [30, 31] by ensuring that more general concepts, which have more instances, are represented, thereby improving the robustness of classification models.

The multi-label approach also accounts for the multiple inheritance characteristic of domain entities, where an entity belongs to more than one parent concept within the ontology. For example, a domain entity such as “river" could be classified under both “Physical Object" and “Geographical Feature" due to its dual characteristics. In a multi-class dataset, this dual classification would require two separate examples with different labels, which could negatively impact the accuracy of classification models. In contrast, the multi-label dataset merges these labels into a single example, reducing the total number of examples from 65,018 to 61,483 after processing multiple inheritances. This reduction not only streamlines the dataset but also improves the accuracy and reliability of classification outcomes by reducing the noise introduced by conflicting labels.

However, the process of merging labels in a multi-label dataset introduces a challenge related to disjoint labels, which occurs when a domain entity is mistakenly assigned to incompatible categories. Disjoint labels are problematic because they contradict the logical structure of the ontology. For example, assigning both “Abstract" and “Physical Object" labels to a domain entity like “justice" would create a conceptual inconsistency, as these top-level concepts are fundamentally incompatible within the DLP hierarchy. To address this issue, the algorithm filters out examples with disjoint labels, further refining the dataset. This filtering reduces the number of examples from 61,483 to 59,666, resulting in a final dataset that maintains the integrity of the ontology’s logical structure while providing a more accurate representation of domain entities.

OntoWordNet’s alignment with DLP top-level concepts, while extensive, does not cover the entire ontology. Specifically, OntoWordNet maps WordNet synsets to only 120 of the 244 top-level concepts in DLP, covering less than half of the concepts. Despite this limitation, OntoWordNet effectively captures the most generic concepts in DLP, making it a valuable resource for ontology-based research. The coverage gap between OntoWordNet and DLP becomes more pronounced at deeper levels of the ontology hierarchy, particularly after level 2, where the number of top-level concepts in OntoWordNet begins to diverge from those in DLP. Nevertheless, OntoWordNet remains a robust tool for researchers, providing a well-organized structure for linking lexical data to top-level ontology concepts.

Finally, the inherent imbalance in top-level ontologies like DLP, due to their complexity and diversity, presents challenges for dataset creation and classification. This imbalance manifests in the uneven distribution of instances across different categories, reflecting the varying emphasis placed on different areas or categories by ontology designers. This imbalance can negatively affect the performance of classification models, as less populated classes are often underrepresented in the training data. Addressing this imbalance requires careful consideration, including the potential use of data augmentation techniques, which can be complex and difficult to implement effectively. However, recognizing and addressing this imbalance is crucial for developing robust and reliable classification models that can accurately categorize domain entities within a top-level ontology framework.

3.2. Textual Representation of Domain Entities

In this work, we propose an approach to classify domain entities into top-level ontology concepts using the textual representation of these domain entities. In this context, we hypothesize that informal definitions represent semantic information that allows domain entities to be related to top-level ontology concepts. This hypothesis comes from two crucial ideas: similarity in taxonomies and distributional

Table 1
Examples of domain entities and their informal definitions extracted from BabelNet.

Definiendum	Informal Definition
Gold	Gold is a soft yellow malleable ductile (trivalent and univalent) metallic element; occurs mainly as nuggets in rocks and alluvial deposits; does not react with most chemicals but is attacked by chlorine and aqua regia.
Iron	Iron is a heavy ductile magnetic metallic element; is silver-white in pure form but readily rusts; used in construction and tools and armament; plays a role in the transport of oxygen by the blood.
Silver	Silver is a soft white precious univalent metallic element having the highest electrical and thermal conductivity of any metal; occurs in argentite and in free form; used in coins and jewelry and tableware and photography.
Day	Day is the time after sunrise and before sunset while it is light outside.
Night	Night is the time after sunset and before sunrise while it is dark outside.
Evening	Evening is the latter part of the day (the period of decreasing daylight from late afternoon until nightfall).

hypothesis.

A taxonomy is a classification system that organizes concepts, entities, or information into categories hierarchically based on common characteristics or criteria. Also, taxonomies are the base structure for ontologies and top-level ontologies, facilitating knowledge organization. Based on that, domain entities grouped in the same top-level ontology concept mean that they share common features or attributes, i.e., they are more similar to each other than those of other domain entities under another top-level concept. For example, all domain entities grouped under the "Amount of Matter" top-level concept, such as “gold," “iron," and “silver," have a greater similarity score because they share common characteristics about physical substances or materials. These entities are inherently more similar to one another when contrasted with entities classified under a completely different concept, such as “evening," “night," and “day" under the “Temporal Region" top-level concept. In this context, since informal definitions provide the intended meaning of a domain entity in a particular context, they also contain attributes and features of a given definiendum described textually in their definiens.

Usually, informal definitions are written using the Aristotelian format (“X is a Y that Z"), containing the definiendum “X," the copula “is" and the definiens “Y that Z." As discussed before, the definiendum is the defined term, the copula is the relationship between the definiendum and the definiens, and the definiens is the explanatory part of the informal definition. In this context, we expected that the definiens include the information necessary to guarantee the intended meaning of the definiendum in a particular context to guarantee a clear and precise interpretation of the defined domain entity. Table 1 describes examples of informal definitions from the dataset described in Section 3.1 for the domain entities “gold," “iron," “silver," “evening," “night," and “day." As we can see, these informal definitions contain unique characteristics that guarantee the meaning of each domain entity in an unambiguous way. However, even without explicit intention, we can informally categorize these domain entities into two groups, i.e., we can create a group with “gold," “iron," and “silver," and another group with “evening," “night," and “day." This classification can be justified because the first group contains informal definitions with terms related to matter and its composition, reactions, and practical usages. On the other hand, the second group contains informal definitions with terms related to time, period of the day, and certain aspects of the temporal order in which they occur.

Since the informal definitions of similar domain entities contain related definientia, we can take advantage of the distributional hypothesis, which says that words that occur in similar contexts are likely to have related meanings. The assumption is that in distributional space, the domain entities “gold," “iron," “silver," “evening," “night," and “day" are closer to each other in their respective group based on their informal definitions. Based on that, we can use state-of-the-art language models founded in the distributional hypothesis to encode the informal definitions in an embedded form. Figure 1 describes an example of the distribution of the domain entities using the embedding of their informal

definition with the BERT-Base language model. In this figure, we employed Principal Component Analysis (PCA) to reduce the length of the output embedding from BERT to a 2D vector for better visualization. As presented, our assumption is valid for this example, i.e., based on the embedding from BERT, we obtained the same groups as the ones we supposed based on our interpretation of the characteristics of the informal definitions of each domain entity of Table 1.

Following the line of reasoning line that top-level ontologies group domain entities that have common attributes or properties, the informal definitions are textual representations of domain entities that encapsulate some attributes or properties that express the intended meaning of the domain entities in a particular context and, based on the distributional hypothesis, we can group similar domain entities based on the embedding representation of their informal definitions, we support our original hypothesis, which says that informal definitions represent semantic information that allows domain entities to be related to top-level ontology concepts.

The datasets extracted through the alignment of OntoWordNet and BabelNet (as given in Section 3.1) present other ways of textually representing domain entities, such as using only the definiendum, only definiens, or an example sentence. As we discussed, the definiendum and the definiens are the parts of the informal definitions. Definiendum (or term) is the shortest way to represent a domain entity since the definiendum names a domain entity through a combination of a few words. Although definienda with similar meanings are closer because of the distributional hypothesis, they can be polysemous, i.e., a single definiendum can have multiple meanings. Using the examples in Table 1, the definienda “gold" and “silver" could also represent domain entities about color. However, colors are generally considered qualities in top-level ontologies rather than the amount of matter, like in the examples. From that, polysemy is certainly a problem that would cause unwanted effects on an approach to classifying domain entities into top-level ontology concepts. On the other hand, although the definiens is the explanatory part of an informal definition and it is expected that the definiens does not have problems with polysemy, a knowledge resource (e.g., WordNet) has only one definiens per domain entity to ensure its precision, which implies few instances in a dataset that contains only the definitions of the domain entities. Therefore, since a domain entity contains many definienda and only one definiens, the informal definitions discussed in this work combine each definienda with its respective definiens. From this, we can generate more dataset instances without the polysemy problem.

The example sentences are a good candidate against informal definitions for representing domain entities textually. These example sentences are pieces of text where the definiendum of the domain entity occurs in plain text. For example, an example sentence for the “gold" domain entity should be “The necklace, made of pure gold, gleamed brilliantly under the soft light of the chandelier." However, this kind of representation for domain entities faces some challenges. The first challenge regards the polysemy problem of the target definiendum in the example sentence. In this situation, if the example

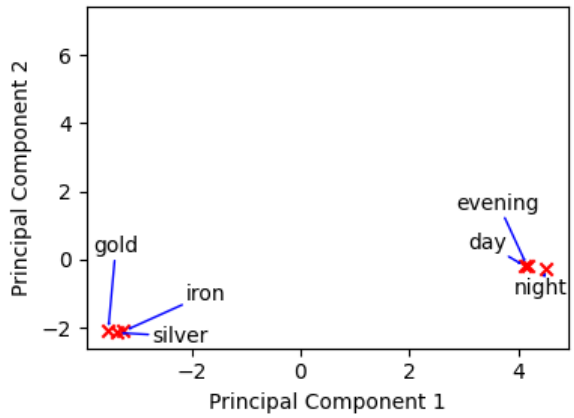

Figure 1: Distribution of domain entities based on the embedding of their informal definitions using BERT-Base language model.

sentence is not previously curated, its extraction requires using a word sense disambiguation technique to ensure the right sense of the target definiendum. Another challenge regards the length of an example sentence since longer sentences can provide more context. However, longer example sentences require language models with larger sequence lengths, which increases the time needed to convert the text into an embedding vector. In contrast, shorter example sentences often need more detail to convey the complete example of using a domain entity. As a last challenge, an example sentence always has the exact embedding representation for all the definienda. For example, we can use the example sentence for the “gold" domain entity to represent other domain entities like “light," “necklace," or “chandelier," even though they are different domain entities with different top-level ontology concepts. We can mitigate this latter challenge by combining the target definiendum with the example sentence, like in informal definitions, where we combine the definiendum with the definiens. However, in the embedding space, the example sentences are still very close due to the distributional hypothesis since they have only the definiendum as the difference between them.

Based on the approaches for textually representing domain entities that we described throughout this section, we assume that the informal definition is the best option for several reasons: a domain expert curates informal definitions before or during the ontology development process; we can extract the informal definitions from existing knowledge resources, like WordNet or Wikipedia; informal definitions are provided early in the ontology development process; the length of informal definitions is relatively small, with an average of a few dozen words, which requires a smaller sequence length for language models, reducing the computational costs; informal definitions are free of polysemy problems since they aim to clearly and precisely explain a domain entity’s meaning in a specific context. In contrast, the most significant disadvantage of informal definitions is that we depend on their clarity and precision, i.e., the informal definition needs to express sufficient characteristics to ensure the meaning of a domain entity.

3.3. Classification Pipelines

In this work, our task is to classify domain entities into top-level concepts using the informal definitions of these domain entities. So, we took advantage of the similarity in taxonomies and distributional hypothesis to support our hypothesis that informal definitions represent semantic information that allows domain entities to be related to top-level ontology concepts. In this context, we employed Language Models (LMs), since they are state-of-the-art models based on distributional hypothesis, as classifiers or embedding providers in two distinct classification pipelines (Figure 2). In Pipeline 1 of Figure 2, we aimed to fine-tune a pre-trained LM for the desired task. On the other hand, in Pipeline 2 of Figure 2, our goal is to use the LM to generate the embedding representation of the input text and then use classical machine learning approaches (e.g., K-Nearest Neighbor, Decision Tree, Support Vector Machine) to classify them into the top-level ontology concepts.

We draw inspiration from state-of-the-art models for text classification for pipelines 1 and 2. In Pipeline 1, fine-tuning a pre-trained language model for a specific task means updating the model’s internal weights to reduce the classification error during the training stage and have a better-adjusted model for predicting novel inputs. However, the training cost of Pipeline 1 tends to be higher than that of Pipeline 2, depending on the number of training epochs used, i.e., the number of times the entire training samples pass through the model. Also, observing the decision process made by Pipeline 1 when classifying a new input tends to be complex due to the characteristics of the language models. Based on that, Pipeline 2 employed traditional machine-learning approaches over the output embedding from a language model. In this case, we pass the training samples only once through the model to generate the embedding representation. However, we have the additional computational cost of the machine-learning classifier. Also, the main benefit of Pipeline 2 is the explainability of the output results. For example, we can see the closest training samples of a new input using a K-Nearest Neighbor as the machine-learning classifier.

Pipelines 1 and 2 in Figure 2 use the textual representation of the domain entities as input. As discussed in Section 3.2, we can access several textual representations of the domain entities from our

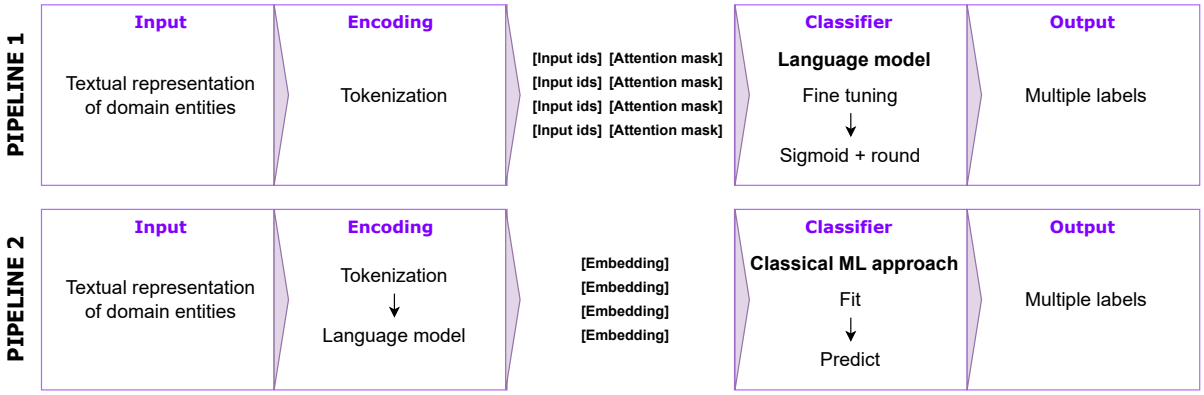

Figure 2: Proposed pipelines for classifying domain entities into top-level ontology concepts using the textual representation of these domain entities.

alignment between OntoWordNet and BabelNet. Based on that, we can consider the informal definition, definiendum, definiens, example sentence, and the combination of definiendum and example sentence as possible candidates in the Input step. Also, in both pipelines, we choose not to apply text preprocessing techniques, such as lowercasing, stop word removal, and lemmatization, in the input text. In our view, these techniques remove the structural integrity of the text, which plays a crucial role in understanding the nuanced meanings and contextual cues inherent in language.

In the Encoding step, we used a pre-trained tokenizer to break the input text into smaller units (tokens). This process is fundamental to both pipelines, transforming unstructured text into a structured form that an LM can interpret. Also, the outputs of the tokenizer are the Input IDs and the Attention Mask values. In Pipeline 1, we fed and fine-tuned the LM with these values. On the other hand, in Pipeline 2, we directly transform the Input IDs and the Attention Mask values into embedding arrays. These embedding arrays comprehend dense and high-dimensional vectors encapsulating word meanings and their relationships in a vector space (which is particularly advantageous for classical machine learning models that use numerical input to find underlying patterns in the data).

The desired output of both classification pipelines is multiple top-level ontology concepts in which a given input subsumes. In this context, we represent all the possible top-level concepts in a one-hot encoded vector, i.e., each concept is represented by a binary value, with 1 indicating the presence of the concept and 0 indicating its absence. This vector serves as the target for the Classification step of both pipelines. In this context, for Pipeline 1, we used a sigmoid activation to transform the average pooled output from the language model into probability values. After that, we rounded these probability values to handle the one-hot encoding representation of the top-level concepts. On the other hand, for Pipeline 2, we do not need any transformation from a mathematical function since traditional machine learning approaches, such as K-Nearest Neighbor and Decision Tree, can handle binary decisions directly.

3.4. Training Methodology

This section discussed dataset extraction and the relation between informal definition, distributional hypothesis, and top-level ontology concepts. From the dataset perspective, we represented each domain entity by a row containing three columns: the definienda, the definiens, and the labels. Since each knowledge resource, such as WordNet or Wikipedia, only includes one definiens per domain entity and a domain entity can have more than one definiendum, we can use an approach to augment the dataset (increase the dataset size) by creating rows for each domain entity definiendum. For each new row of a domain entity, we kept the same definiens and labels. This technique is also called the “explode" approach. For example, the “gold" domain entity in Table 1 has three different definienda in BabelNet: gold, Au, and atomic number 79. From that, we create a new row in the dataset for these three definienda and keep the same definiens and labels.

Figure 3 presents the number of definienda of the domain entities in the dataset from the alignment

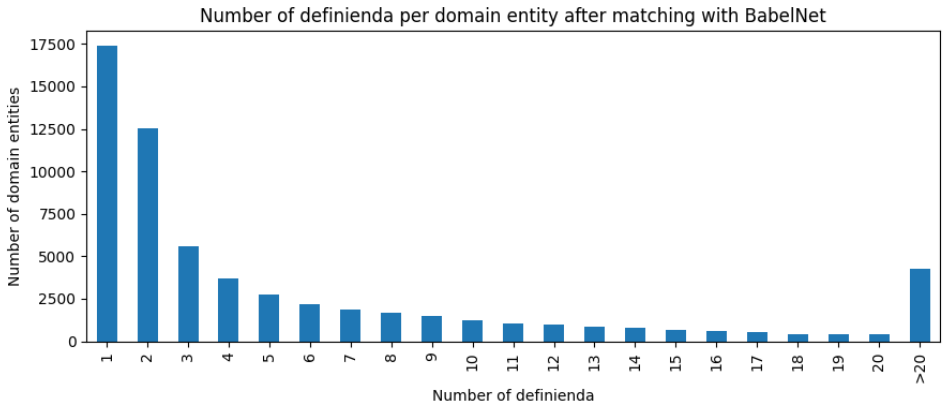

Figure 3: The number of definienda of the domain entities in the dataset.

between OntoWordNet and BabelNet. Considering that most dataset domain entities have more than one definiendum, the “explode" approach helps increase the number of examples without artificially generating data. For instance, we can expand the dataset from 59,666 to 387,814 examples. However, this process has implications for the performance of a classification model. Considering the example of the “gold" domain entity (from Table 1) and that we followed this sequence of decisions, used Pipeline 2 (described in the previous section) with a KNN machine-learning model with $k = 3$, trained this model with two of the three new rows generated from “gold" domain entity, and predict the labels of the last remaining “gold" row. Based on the distributional hypothesis, the three closest examples predicted by KNN also include the two variations of the “gold" domain entity because they share almost the same informal definition with just a change in the definiendum part. This fact also occurs in Pipeline 1 but is more easily explainable using Pipeline 2. Although this characteristic appears after using the “explode" approach, this may not be seen as a problem but should be considered when evaluating classifier performance, especially in evaluations using training and test sets of the same dataset, like in k-fold cross-validation.

In this work, we explored three training methodologies to deal with the consequences of the “explode" approach. Considering that the training and test samples are from the same dataset, the first method ignores this consequence, i.e., we apply the “explode" approach before dividing the dataset into training and test samples. However, suppose that we aim to evaluate the performance of a classification model with completely novel domain entities. In this case, as our second method, we must apply the "explode" approach after dividing the dataset into training and test samples, ensuring that no test domain entity is present in the training samples. As a third method, we considered that the training and test samples are from different knowledge resources, such as the training samples from WordNet and the test samples from Wikipedia, or vice-versa. In this case, although we can have the same domain entity in multiple resources, they usually do not have the same definiens. For example, in Wikipedia, “gold" is defined as “a chemical element; it has symbol Au and atomic number 79," an almost entirely different definiens from the one described in Table 1. So, the embedding representation of the informal definitions of these domain entities is more discrepant than changing just the definiendum. Thus, the “explode" approach does not generate such perceptible consequences in this training methodology.

4. Experiments

In this section, we present a study case performed to validate our hypothesis that the informal definitions represent semantic information that allows domain entities to be related to top-level ontology concepts. This study case addresses the task of classifying domain entities into top-level ontology concepts. We conducted experiments with several different Language Models (LMs) and Large Language Models (LLMs) in the proposed pipelines to validate our hypothesis using training and test samples from

Table 2
The number of examples in each resource dataset.

Dataset	Original size	After preprocessing	After “explode"
WordNet 3.0	65,018	59,666	387,814
Wikipedia	33,547	30,603	371,766

different knowledge resources.

In order to evaluate the results, we used the macro F1-score (Equation 1) to ensure a better classification analysis in the unbalanced scenario. We also conducted the study cases and experiments on a machine equipped with an Intel i7-10700 CPU (4.8GHz), 32 GB of RAM, and a GeForce RTX 3060 GPU with 12GB of VRAM.

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} 2 \cdot \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i} \tag{1}$$

where $F1_{macro}$ is the macro F1-score; N is the number of labels; i is the ith class in N; precision_i is the precision for the ith class, defined as the number of true positive predictions divided by the total number of positive predictions (true positives plus false positives); recall_i is the recall for the ith class, defined as the number of true positive predictions divided by the total number of actual positives (true positives plus false negatives).

From that, for Pipeline 1, we employed the BERT-Base and GPT2, fine-tuned using 10 epochs. In Pipeline 2, we utilized the BERT-Base, BERT-Large, ROBERTA-Base, ALBERT-Base, GPT2, T5, BART, Gemma2B, Mistral7B, and Lamma7B, where 2B and 7B means the number of parameters of the respective language model. Also, for Pipeline 2, we only used the KNN classifier because it is the one that best fits our hypothesis, as presented in the previous study cases. In addition, Table 2 presents the number of examples of each evaluated dataset after performing preprocessing and "explode" steps. From that, we used the informal definitions from WordNet to train the pipelines and the informal definitions from Wikipedia to evaluate the pipelines.

Figure 4 presents the average macro F1-score of each evaluated pipeline across all DLP depths. Our best pipeline combined the Mistral7B large language model with a KNN classifier, achieving 86.6% of the macro F1-score. Following this kind of pipeline with a KNN classifier, other large language models, such as Llama7B and Gemma2B, also achieved results close to 80% of the average macro F1-score. Interestingly, although BERT-Large has more parameters than BERT-Base, they presented similar results. The pipelines with fine-tuned versions of BERT-Based and GPT2 achieved around 75% of the average macro F1-score, showing distinct results compared with their same version with a KNN classifier. In this case, the pipeline with GPT2 and KNN achieved the worst results in our experiments. Other pipelines with KNN and language models, such as ROBERTA, ALBERT, BART, and T5, showed results ranging from 60% to 75% of the average macro F1 scores.

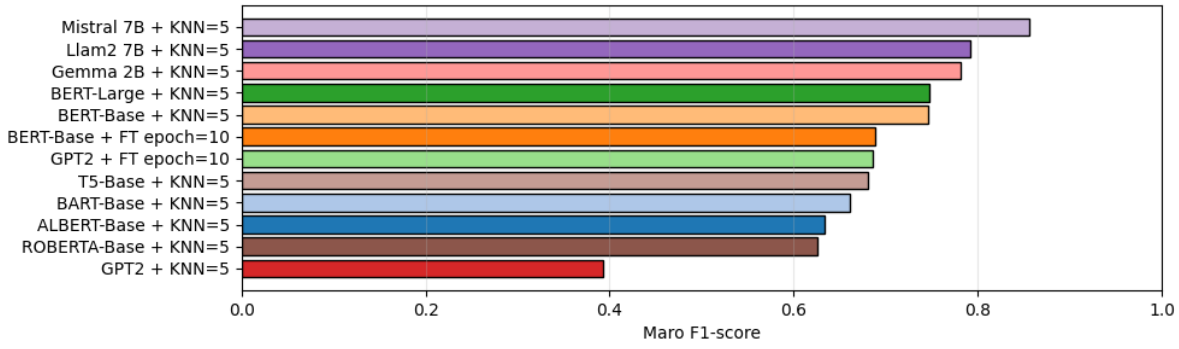

Figure 4: The macro F1-score results for each evaluated pipeline across all DLP levels.

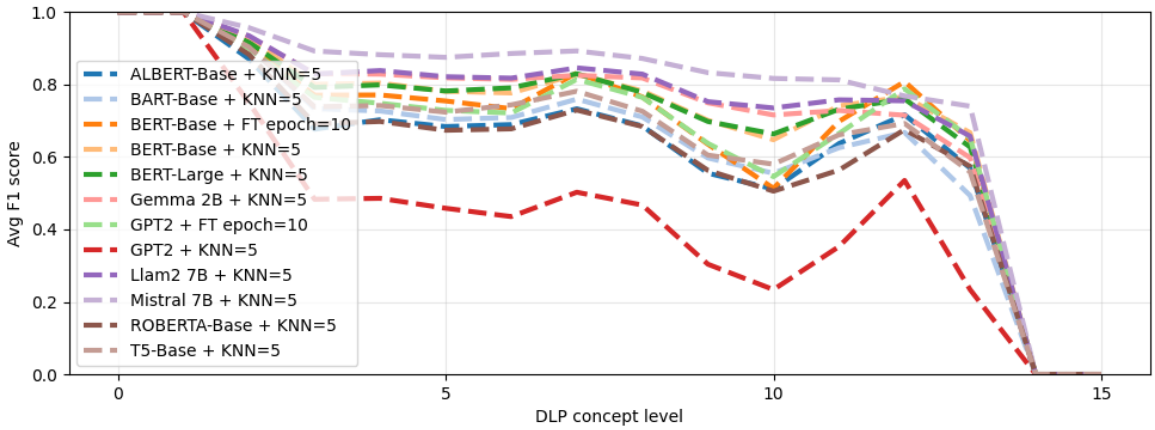

Figure 5: The macro F1-score results for each evaluated pipeline across each DLP level.

Figure 5 presents in detail the average macro F1-score of each evaluated pipeline for each DLP depth. As in the other study cases, the scores appear to decrease as the level of DLP increases. Also, all pipelines generally maintain their performances until level 10, which can be associated with the 10th level having the more significant number of top-level concepts regarding all DLP levels. Also, we can justify the better performance of the pipeline with the Mistral7B large language model and the KNN classifier with their stability regarding a better classification performance from the more generic to the more specific top-level ontology concepts.

This case study presented a rich evaluation of our hypothesis that the informal definitions represent semantic information that allows domain entities to be related to top-level ontology concepts. From that, we employed various language models and state-of-the-art large language models in the proposed classification pipelines. Also, the results suggest that we can avoid the computational cost of fine-tuning by employing the embedding from a pre-trained language model and using classical machine learning approaches, like KNN. In this context, we showed that using a KNN classifier with contextual embedding derived from informal definitions yields better results than fine-tuning methods. This outcome supports the notion that the distributional properties of informal definitions encapsulate the top-level ontology concepts associated with domain entities. Furthermore, the enhancement of this process by state-of-the-art models such as Mistral, Llama, and Gemma indicates that the classification of domain entities into top-level ontology concepts via informal definitions and contextual embedding benefits from advancements in language model improvements.

5. Conclusion

This study explains a novel approach leveraging the similarity in taxonomies and the distributional hypothesis to classify domain entities into top-level ontology concepts using the informal definitions of these domain entities. Firstly, we proposed a method to align the OntoWordNet ontology and Babel semantic network. From this alignment, we extracted multi-label datasets to encompass all the top-level ontology structures. Also, this alignment allows us to extract multi-language and multi-resource datasets, from which we extracted several text representation approaches for domain entities. Also, by harnessing state-of-the-art language models, we developed two distinct pipelines: one focusing on fine-tuning a pre-trained language model for our classification task and another utilizing a classical machine learning algorithm to classify the embedding generated from these models. We used these two pipelines to validate our hypothesis that the informal definitions represent semantic information that allows domain entities to be related to top-level ontology concepts. In addition, our findings underscore the significant impact of using informal definitions to represent domain entities concerning other representation approaches, demonstrating their intrinsic value in capturing the nuances necessary for

accurately representing domain entities through contextual embedding from language models. Also, we proposed an approach for augmenting the extracted datasets without the necessity of generating artificial examples, underscoring its consequences during training and validation. In our experiments, we conducted a study case to compare the performance of our hypothesis and pipelines with informal definitions from different knowledge resources. In this study case, we achieved our best results using a classification pipeline with the K-Nearest Neighbor approach and the Mistral7B large language model, highlighting the effectiveness of this classical machine-learning method over the expensive computational cost of fine-tuning. From the results analysis, we supported our hypothesis, demonstrated the versatility and robustness of our approach across different resources, and emphasized the potential for automated approaches to assist ontology engineers during ontology development. In future works, we aim to expand the datasets for other top-level concepts not covered by OntoWordNet. Also, we will perform new experiments by translating the informal definitions in English to different languages to use the same data across all multi-language experiments.

Acknowledgments

Research supported by Higher Education Personnel Improvement Coordination (CAPES), code 0001, Brazilian National Council for Scientific and Technological Development (CNPq), and Petrobras.

References

[1] R. Studer, V. R. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data & knowledge engineering 25 (1998) 161–197.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[3] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language understanding paper.pdf (2018).

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[5] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. arXiv:2401.04088.

[6] Y. He, J. Chen, D. Antonyrajah, I. Horrocks, Bertmap: a bert-based ontology alignment system, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 5684–5691.

[7] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, B. Sapkota, Deeponto: A python package for ontology engineering with deep learning, 2023. arXiv:2307.03067.

[8] H. Babaei Giglou, J. D'Souza, S. Auer, Llms4ol: Large language models for ontology learning, in: International Semantic Web Conference, Springer, 2023, pp. 408–427.

[9] F. Cicconeto, L. V. Vieira, M. Abel, R. dos Santos Alvarenga, J. L. Carbonera, L. F. Garcia, Geo-reservoir: An ontology for deep-marine depositional system geometry description, Computers & Geosciences 159 (2022) 105005.

[10] B. Kulvatunyou, M. Drobnjakovic, F. Ameri, C. Will, B. Smith, The industrial ontologies foundry (iof) core ontology, Formal Ontologies Meet Industry (FOMI) 2022, Tarbes, FR, 2022. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=935068.

[11] Y. Qu, M. Perrin, A. Torabi, M. Abel, M. Giese, Geofault: A well-founded fault ontology for interoperability in geological modeling, Computers & Geosciences 182 (2024) 105478.

[12] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The neon methodology for ontology engineering, in: Ontology engineering in a networked world, Springer, 2011, pp. 9–34.

[13] S. Borgo, R. Ferrario, A. Gangemi, N. Guarino, C. Masolo, D. Porello, E. M. Sanfilippo, L. Vieu, Dolce: A descriptive ontology for linguistic and cognitive engineering, Applied ontology 17 (2022) 45–69.
[14] G. Guizzardi, A. Botti Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. Prince Sales, Ufo: Unified foundational ontology, Applied ontology 17 (2022) 167–210.
[15] J. N. Otte, J. Beverley, A. Ruttenberg, Bfo: Basic formal ontology, Applied ontology 17 (2022) 17–43.
[16] Z. S. Harris, Distributional structure, Word 10 (1954) 146–162.
[17] S. Seppälä, A. Ruttenberg, Y. Schreiber, B. Smith, Definitions in ontologies, Cahiers de Lexicologie 2016 (2016) 173–205.
[18] 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.
[19] R. Navigli, P. Velardi, Learning domain ontologies from document warehouses and dedicated web sites, Computational Linguistics 30 (2004) 151–179. URL: https://api.semanticscholar.org/CorpusID:2453822.
[20] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, F. Cecconi, Ten years of babelnet: A survey., in: IJCAI, 2021, pp. 4559–4567.
[21] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.
[22] A. Lopes, J. Carbonera, F. Rodrigues, L. Garcia, M. Abel, How to classify domain entities into top-level ontology concepts using large language models, Applied Ontology (2024) 1–29.
[23] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernandez-Lopez, The neon methodology framework: A scenario-based methodology for ontology development, Applied ontology 10 (2015) 107–145.
[24] R. Arp, B. Smith, A. D. Spear, Building ontologies with basic formal ontology, Mit Press, 2015.
[25] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, L. Schneider, Sweetening ontologies with dolce, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2002, pp. 166–181.
[26] I. Niles, A. Pease, Towards a standard upper ontology, in: Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, 2001, pp. 2–9.
[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.
[28] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of the International Conference on Learning Representations (ICLR), 2013.
[29] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[30] A. Lopes, 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.
[31] A. Lopes, J. Carbonera, D. Schmidt, L. Garcia, F. Rodrigues, M. Abel, Using terms and informal definitions to classify domain entities into top-level ontology concepts: An approach based on language models, Knowledge-Based Systems 265 (2023) 110385.