Term Selection for Abstraction of OWL Ontologies

Xuan Wu
State Key Laboratory for Novel Software Technology, Nanjing University
China
wuxuan@smail.nju.edu.cn

Yu Dong
State Key Laboratory for Novel Software Technology, Nanjing University
China
dongy@smail.nju.edu.cn

Yizheng Zhao*
State Key Laboratory for Novel Software Technology, Nanjing University
China
zhaoyz@nju.edu.cn

ABSTRACT

This paper attempts to address the challenge of effectively selecting suitable terms (i.e., concept names) as seed signatures for the abstraction of OWL ontologies. Established methods for generating seed signatures rely predominantly on geographical connections, a practice that proves to be inadequate in yielding satisfactory abstractions. This limitation consequently curtails the practical reusability of OWL ontologies within the broader context of knowledge management. To overcome these limitations, this paper introduces a novel approach called "signature extension". This approach serves the dual purpose of generating seed signatures for "modularization" and "uniform interpolation" of OWL ontologies, both of which are pivotal ontology abstraction techniques. The signature extension approach is designed to establish the semantic relevance of concept names by harnessing the treasure trove of metadata information available within a given OWL ontology. Specifically, it quantifies the relevance of these names through a numerical measure, which is derived from their embeddings transformed using the OWL2Vec* framework. Empirical evaluations conducted on this approach unequivocally demonstrate its superior performance when compared to other established methods for term selection. In addition, a comprehensive case study on ontology abstraction tasks shows that modularization tools can create more comprehensive and precise abstractions when utilizing the signatures extended through our proposed approach.

CCS CONCEPTS

• **Theory of computation → Description logics**; **Automated reasoning**; • **Computing methodologies → Ontology engineering**.

KEYWORDS

Description Logics; Ontologies; Uniform Interpolation; Forgetting

ACM Reference Format:
Xuan Wu, Yu Dong, and Yizheng Zhao. 2023. Term Selection for Abstraction of OWL Ontologies. In *Proceedings of The 12th International Joint Conference on Knowledge Graphs (IJCKG '23), December 8–9, 2023, Tokyo, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3583780.3614771

*Yizheng Zhao is the corresponding author of this paper.

IJCKG'23, December 8–9, 2023, Tokyo, Japan

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614771

1 INTRODUCTION

Given the inherent diversity of web resources, ontologies designed for the semantic web—such as OWL ontologies [3]—tend to be extensive, capturing knowledge across a broad range of topics [13]. However, their scale and breadth can impede reusability and interoperability in practical applications. This challenge stems chiefly from the difficulties in managing and manipulating large, complex ontologies, which can become cumbersome and demand substantial computational resources during the reasoning process. A promising approach to mitigate these issues is to distill a fragment from an ontology that replicates the original ontology's functionality within a specific context, yet is significantly smaller in size. The ultimate goal is to optimize the size reduction of this fragment to the maximum extent possible.

Two logic-based approaches have been developed for extracting meaningful fragments of ontologies. One approach, known as *modularization* [5, 8, 10, 11, 15, 17], focuses on identifying a syntactic subset, namely a *module*, within an ontology. This module is carefully selected to maintain the validity of several reasoning tasks within a specific sub-signature of the ontology, referred to as a *seed signature*. The second approach, referred to as *uniform interpolation* [19, 23, 24], is concerned with the computation of a more compact representation of a module within an ontology. This refined module, called a *uniform interpolant*, retains the underlying logical definitions of the terms within the seed signature. To sum up, these two approaches offer distinct strategies for computing ontology fragments, and their utilization is contingent on specific use cases and requirements in the context of ontology engineering.

As one might expect, the quality of extracted fragments significantly relies on the seed signature used in the modularization and uniform interpolation procedures. We can define a fragment as "complete" if it encompasses all essential information pertaining to the topic of interest. A fragment is considered "precise" when it meets the criteria of completeness and, additionally, avoids incorporating excessive irrelevant information concerning the topic. To elaborate, selecting too few terms for the seed signature could result in the loss of vital information that a user might find valuable. Conversely, choosing too many terms, some of which are weakly related to the topic, would introduce excessive additional information. Furthermore, importing more information can alter the definitions of terms within the original ontology, potentially undermining its coherence and consistency [11].

Nonetheless, there has been limited focus on the problem of term selection when it comes to extracting fragments of ontologies. Chen

et al. [5] proposed a "signature extension" approach to generating seed signatures as input to ontology modularization procedures. The idea behind this approach is twofold:

Step (1) Start with a "primitive seed signature" Σ, typically containing several terms suggested by domain experts;
Step (2) Extend Σ by including new terms found in the axioms that involve the existing Σ-terms.

This process is repeated until no additional terms can be added to Σ. To illustrate this, consider the analogy of people living on an island: if two individuals, denoted as p_1 and p_2, reside together in a house h_1 on an island, they are considered "relevant" and are included in $\Sigma = \{p\_1, p\_2\}$. If there exists a road connecting h_1 to another house h_2, the individuals in h_2 are added to Σ. This iterative strategy is applied throughout the entire island, resulting in Σ eventually encompassing all residents on the island. However, it is important to note that a person living on a different island will never be included in Σ due to the absence of a connecting road; these islands are "geographically isolated".

Clearly, when applying this signature extension approach, one would acquire a larger seed signature, and consequently, a more informative fragment. However, it can be argued that the seed signature obtained from this approach, which relies solely on geographic connections, may struggle to produce a "complete" fragment. This argument stems from the belief that assessing the "relevance" between a term and the expanding seed signature should encompass a comprehensive evaluation of all metadata associated with the participating terms within the framework of the host ontology. This perspective emphasizes the need to consider factors beyond geographic connections when determining term relevance in this context.

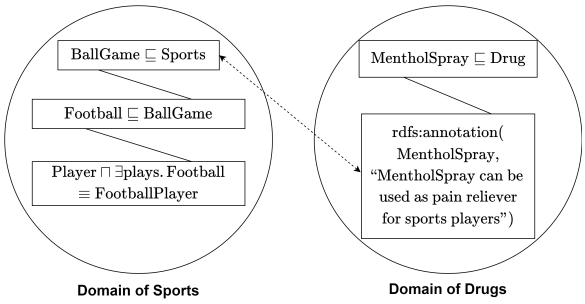

Figure 1: A snippet of a multi-domain ontology

Consider an ontologist working with a medical ontology that contains information about various medications and their applications. Within this ontology, there is a term called "MentholSpray" that is part of the medical domain. Initially, when selecting the seed signature for abstraction, the ontologist focuses on terms related to general healthcare but does not include "MentholSpray" in the primitive signature because it is not explicitly associated with sports or football. However, upon further examination of the ontology's metadata, it becomes evident that "MentholSpray" is often used to treat sports-related injuries, especially in the context of football. The metadata includes notes from healthcare professionals indicating that "MentholSpray" is commonly prescribed for pain relief in football players after matches or during recovery from sports-related injuries. In this scenario, despite the initial geographic isolation of "MentholSpray" from the sports domain, its strong relevance to the topic of football emerges when considering the comprehensive metadata. Consequently, including "MentholSpray" in the extended signature could significantly enhance the ability of the ontology fragment to answer queries related to pain management in football, showcasing the importance of metadata-driven term relevance assessment within the host ontology's context.

The metadata of OWL ontologies comprises a wealth of semantic information and other critical information, which may encompass potential semantic relationships among concepts within these ontologies. Effectively leveraging this information can assist in the establishment of connections within concepts that extend beyond the boundaries of the logical components defined in the ontologies. However, given the multi-dimensional nature of semantic information contained within metadata, the current consensus has shifted toward mapping this information into vector representations. During the process of term selection, the absence of supervision signals renders the similarity between vector representations contingent upon the choice of distance function [4]. It is crucial to stress that the main contribution of this work does not lie in delineating the similarity function. Rather, our focus is centered on integrating explicit metadata semantic information within logic-based ontology modularization frameworks, thereby enhancing their functionality.

In this paper, we introduce a novel approach to term selection aimed at uncovering semantic relationships between two distinct groups of terms. Our method involves assessing the relevance of non-Σ terms to Σ terms based on their D-dimensional vector representations, which are computed using essential metadata from the ontology. These vector representations are generated using OWL2Vec [6], an ontology embedding framework that leverages random walk and word embedding techniques. OWL2Vec encodes the semantics of OWL ontologies in a vector space, taking into account various aspects, including the ontologies' graph structure, lexical information, and logical constructs. The primary objective of this work is to enhance the practical utility of existing logic-based ontology abstraction methods for a wide range of ontology-based knowledge processing tasks. This enhancement is achieved by incorporating non-logical approaches to streamline the process. Historically, there has been limited exploration of the synergies between logical and data-driven techniques and how they can leverage each other's strengths to establish a robust application framework. Our empirical evaluation demonstrates that our proposed approach outperforms conventional term selection baselines when recommending suitable seed signatures. By adopting this approach, we can generate more precise ontology fragments using two established modularization and uniform interpolation tools.

2 PRELIMINARIES

2.1 Description Logic and OWL

An *ontology* fixes a vocabulary of *terms* (called classes, which are sets of instances characterized by some shared properties of its instances) relevant to a subject domain, and specifies (as a formal description of domain knowledge) constraints among the classes and

the relationships holding between classes by logical statements [22]. The logical statements on memberships of data elements (called individuals) in classes or relationships between individuals form a base of facts, i.e., a database [20]. When we speak of ontology, we think only of the knowledge specifying the constraints among the classes and the relationships. We include both the constraints and the facts under the term *knowledge base*. In DLs, classes are called *concepts* and relationships are called *roles*.

Modern ontologies are formulated in the OWL 2 Web Ontology Language (OWL 2 for short) [18] based on description logics (DLs) [1, 2], which are a successful family of knowledge representation languages. OWL 2 introduces different profiles, each with varying levels of expressiveness to cater to different use cases and computational complexities. Among the most commonly used profiles are OWL 2 EL (Description Logic EL++), OWL 2 QL (Query Language), and OWL 2 RL (Rule Language). These profiles allow ontology developers to choose the appropriate level of expressiveness and reasoning complexity for their specific applications, balancing computational efficiency with modeling capabilities. OWL 2 Full is by far the most expressive of the profiles in OWL 2, which corresponds to the DL $\mathcal{SROIQ}$ [14]. Our term selection approach does not require a specific level of expressiveness (i.e., logic-independent) and is universally applicable across all OWL 2 profiles.

Definition 1 ($\mathcal{SROIQ}$ **Concepts and Roles**). *Let* $\mathsf{N_C}$, $\mathsf{N_R}$ *and* $\mathsf{N_I}$ *be countably infinite and pairwise disjoint sets of respectively concept names, role names, and individual names, with the top concept* $\top \in \mathsf{N_C}$*, the bottom concept* $\bot \in \mathsf{N_C}$*, and the universal role* $U \in \mathsf{N_R}$*. An* $\mathcal{SROIQ}$ *role is either a role name or the inverse of a role name.* $\mathcal{SROIQ}$ *concepts are defined from* $\mathsf{N_C}$ *and* $\mathsf{N_I}$ *by induction with the constructs of* $\neg$ *(negation),* $\sqcap$ *(conjunction),* $\sqcup$ *(disjunction),* $\exists$ *(existential restriction),* $\forall$ *(value restriction),* $\geq$ *and* $\leq$ *(qualified number restriction) and* Self *(self restriction).*

The semantics of $\mathcal{SROIQ}$ is defined in terms of an *interpretation* $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$, where $\Delta^{\mathcal{I}}$ is the *domain of the interpretation* (a non-empty set), and $\cdot^{\mathcal{I}}$ denotes the *interpretation function*, which assigns to every individual $a \in \mathsf{N_I}$ an element $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$, to every concept name $A \in \mathsf{N_C}$ a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$, and to every role name $r \in \mathsf{N_R}$ a binary relation $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. The interpretation function $\cdot^{\mathcal{I}}$ is inductively extended to other $\mathcal{SROIQ}$ concepts as in [2].

Let $\mathcal{I}$ be an interpretation. $\mathcal{I}$ is a *model* of an ontology $\mathcal{O}$, written $\mathcal{I} \models \mathcal{O}$, iff every axiom in $\mathcal{O}$ is *true* in $\mathcal{I}$. An axiom α is a *logical consequence* of an ontology $\mathcal{O}$, written $\mathcal{O} \models \alpha$, iff α is true in every model $\mathcal{I}$ of $\mathcal{O}$. In this paper, a *signature* $\mathsf{sig} \subseteq \mathsf{N_C}$ is defined as a set of concept names. Let $\mathsf{sig}(X)$ denote the set of the concept names present in X, where X ranges over concepts, axioms, and sets of axioms (ontologies).

2.2 Metadata of OWL Ontologies

Metadata refers to information about the ontology or knowledge artifact itself, distinct from the domain knowledge it represents. For instance, if a class like "Electron" is authored by a specific individual, say "Woodstock" on a particular date, this constitutes an editorial statement about the class rather than its individual instances. Dublin Core[1] offers a set of annotations for such metadata, which are accessible within tools like Protégé[21], although they are not inherent to OWL 2. In OWL ontologies, metadata can be categorized into three main types:

- **Editorial/Provenance Meta Statements**: These encompass information about the knowledge acquisition process and its sources. This includes details about the author, date of entry, revision history, authority, and more. The Dublin Core subset is a widely adopted standard for editorial and provenance metadata, with some statements being temporary and geared towards development processes, while others are intended to be permanent.
- **Explanatory Statements**: These consist of text-based definitions, guidelines, comments, and other explanatory content. They often incorporate natural language definitions for classes, along with comments, evidence, and provenance information related to the conceptualization of the class.
- **Structural Information about the Artefact**: Some information artefacts contain meta-models or structural information that describes their own organization. For example, regular patterns of axioms may be employed, and templates for these patterns can be part of the meta-model. This aspect focuses on describing how the ontology is constructed, rather than the ontology model itself.

Metadata in OWL ontologies serves multiple purposes, including documentation, data integration, reasoning, and interoperability. It enhances the understanding of ontology content, facilitates ontology management, and supports various Semantic Web applications by providing context and context-awareness for ontology elements.

2.3 Modularization and Uniform Interpolation

Modularization can be defined in various ways depending on the properties of the computed modules. The definition given in this paper is a generalized one that collects common conditions necessary for an ontology to be recognized as a module. More specific definitions of modularization can be derived by introducing particular conditions into this generalized framework.

Definition 2 (**Module**). *Let* $\mathcal{O}$ *be an ontology in a DL* $\mathcal{L}$ *and* $\Sigma \subseteq sig(\mathcal{O})$ *be a set of concept names (i.e., the seed signature). An* $\mathcal{L}$*-ontology* $\mathcal{M}$ *is a* Σ*-module of* $\mathcal{O}$ *iff the following conditions hold:*
(1) $\mathcal{M} \subseteq \mathcal{O}$*, and*
(2) for any $\mathcal{L}$*-axiom* α *with* $sig(\alpha) \subseteq \Sigma$*,* $\mathcal{M} \models \alpha$ *iff* $\mathcal{O} \models \alpha$*.*

The definition of uniform interpolation is given as follows.

Definition 3 (**Uniform Interpolant**). *Let* $\mathcal{O}$ *be an ontology in a DL* $\mathcal{L}$ *and* $\Sigma \subseteq sig(\mathcal{O})$ *be a set of concept names (i.e., the seed signature). An* $\mathcal{L}$*-ontology* $\mathcal{V}$ *is a* Σ-uniform interpolant *of* $\mathcal{O}$ *iff the following conditions hold:*
(i) $sig(\mathcal{V}) \subseteq \Sigma$*, and*
(ii) for any $\mathcal{L}$*-axiom* α *with* $sig(\alpha) \subseteq \Sigma$*,* $\mathcal{V} \models \alpha$ *iff* $\mathcal{O} \models \alpha$*.*

Σ-modules and Σ-uniform interpolants have exactly the same logical consequences in the seed signature Σ. The difference is that, while a module is always a syntactic subset of the given ontology, and may use names outside of Σ, a uniform interpolant uses only the names in Σ. We therefore call the view computed by uniform interpolation a *signature-restricted abstract*. A large number of new axioms will be deduced from the given ontology in order to preserve the semantics of the Σ-names when eliminating the other

[1] https://www.dublincore.org/

names from the ontology, i.e., the computation of uniform interpolants relies on more reasoning than that of modules. This can lead to uniform interpolants containing substantially more complex axioms than the input ontology [19]. This is why the problem of uniform interpolation is generally believed to be computationally harder than modularization [15]. We regard modularization and uniform interpolation as 'non-standard" forms of reasoning, because it cannot be reduced to the "standard" satisfiability/entailment tests.

3 METADATA-BASED TERM SELECTION

Our term selection approach accommodates ontologies described in OWL 2, which are based on the description logic $\mathcal{SROIQ}$ [14]; see the Description Logic Handbook [2] for a detailed description of the syntax and semantics of description logics.

Arguably, most topics can be sufficiently summarized or defined by a set of keywords (i.e., key concept names), with less reliance on role names. Therefore, in this paper, we consider seed signatures solely as a set of concept names.

Given an ontology $\mathcal{O}$ and a seed signature $\Sigma \subseteq \text{sig}(\mathcal{O})$ consisting of either a single or a few concept names, normally suggested by a group of domain experts or simply chosen by users, as the most representative keywords for the topic of interest, our approach computes an extension Σ' of Σ through a three-step process involving *concept representation learning*, *computation of relevance values*, and *signature extension based on relevance values*. The resulting Σ' serves as the updated seed signature fed into subsequent modularization and uniform interpolation procedures.

3.1 Concept Representation Learning

Concept representation learning involves the conversion of every concept name A in $\mathcal{O}$ into a D-dimensional vector, denoted as $\mathbf{e}_A$, within a vector space. These vectors are constructed with consideration of the "relevance" of each concept name A to the seed signature Σ and are computed based on important metadata of $\mathcal{O}$.

Our concept representation learning model is based on the core OWL2Vec* [6], an ontology embedding framework for computing vector representations of concept names in OWL ontologies as expressive as $\mathcal{SROIQ}$. OWL2Vec* computes the embedding of an OWL ontology based on a corpus of token sequences, which are generated from the ontology's metadata. This metadata includes diverse components, encompassing the graph structure of the ontology represented as an RDF graph (a collection of RDF triples) derived from the OWL ontology through the OWL2Vec* transformation. Additionally, it incorporates what is commonly referred to as the "lexical information" about the ontology, which encompasses annotations. Furthermore, it encompasses the "logical information" concerning the concepts and roles within the ontology, encompassing relationships like subsumption, equivalence, and disjointness, among others.

It should be noted that OWL2Vec* was originally designed for purposes other than term selection tasks. Therefore, we have made modifications to the core OWL2Vec* model to optimize its performance for downstream term selection tasks. Specifically, we have introduced a fine-tuning process specifically tailored to enhance ontology embedding. Further elaboration on this fine-tuning process can be found in Section 4.

Algorithm 1 Nearest Neighbour Ranking

Input: A set N_C of concept names, A set of seed signatures Σ s.t. $\Sigma \subseteq N_C$,
A set of concept embedding $\{\mathbf{e}_A : A \in N_C\}$,
A distance function $d : \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, \infty]$.
Output: A relevance function $f : N_C \rightarrow [0, 1]$,

```
Let g be a mapping of N_C → [0, ∞].
for all A ∈ N_C do
    g(A) := ∞
    for all A' ∈ Σ do
        g(A) := min (d (e_A, e_A') , g(A)).
    end for
end for
Let f be a mapping of N_C → [0, 1].
for all A ∈ N_C do
    Find i, s.t. A has the i-th smallest g(A) in N_C.
    f(A) := 1 − (i − 1)/|N_C|.
end for
return f
```

3.2 Computing Relevance Values

This step computes the relevance value of every non-seed concept name B in $\mathcal{O}$ to Σ. This computation is based on the relative distance between B and its nearest seed neighbor within the vector space. The relevance value is a scalar within the range of $[0, 1]$, where a value of 1 indicates the highest level of relevance, while that of 0 is the lowest. The computation of this relevance value is performed using our newly developed algorithm, known as the Nearest Neighbor Ranking algorithm (NN-RANK), which is shown in Algorithm 1.

More specifically, NN-RANK computes the *distance* from each non-seed name to each seed name in the vector space. In principle, several distance functions $d : \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, \infty]$ can be employed for this purpose. However, in our experiments, the *Cosine distance*, formulated as:

$$d(\mathbf{e}_A, \mathbf{e}_B) = 1 - \frac{\mathbf{e}_A \cdot \mathbf{e}_B}{\|\mathbf{e}_A\|_2 \, \|\mathbf{e}_B\|_2}$$

has proven to be the most effective measure of relevance compared to euclidean distance and dot product similarity. Based on this formula, $|\Sigma|$ distance values are computed for each non-seed name A. Among these values, the smallest one is identified as the *valid distance value* of A to Σ. NN-RANK subsequently arranges all concept names in $\mathcal{O}$ in ascending order based on their valid distance values. Concept names with smaller valid distance values are considered to be semantically more relevant to the seed signature, and consequently, to the central topic. These valid distance values (along with the corresponding concept names) are then uniformly mapped to a range between 0 and 1. The outcome represents the *relevance value* of each A with respect to Σ.

3.3 Relevance-Based Seed Signature Extension

A question naturally emerges at this stage: how can the computed relevance values be effectively employed to guide the selection of meaningful terms for ontology abstraction? Different application

demands may warrant varying strategies. In the absence of a well-established gold standard, a practical approach is to quantify the "degree" of relevance and determine to what extent a concept name can be considered "relevant" to the seeds in Σ. In this work, we set a threshold σ within the range of 0 to 1 to represent the "degree" of relevance. Our approach extends the primitive seed signature Σ by including concept names with a relevance value greater than or equal to σ. The result is $\Sigma' = \Sigma \cup \{A \mid A \in \text{sig}(\mathcal{O}) \wedge f(A, \Sigma) \geq \sigma\}$. The higher the threshold σ, the smaller the extension obtained.

Computing $|\text{sig}(\mathcal{O})| \times |\Sigma|$ distances has a linear time complexity w.r.t. $|\text{sig}(\mathcal{O})|$, and the subsequent sorting operation has a log-linear time complexity w.r.t. $|\text{sig}(\mathcal{O})|$.

4 EMPIRICAL STUDY OF NN-RANK

In this study, we employed NN-RANK to predict SNOMED CT Refset components. The objective was to demonstrate the algorithm's capability to augment a given seed signature Σ with concept names that exhibit high relevance to the initial seeds within a vector space. All evaluations were conducted on a server equipped with an Intel(R) Xeon(R) Gold 5117 CPU @ 2.00GHz and 128 GB of memory.

SNOMED CT[2] is currently the most comprehensive, multilingual clinical healthcare ontology in the world. A SNOMED CT Refset[3] constitutes a compilation of SNOMED CT components that share specific attributes or characteristics, often related to a particular domain or field. As an illustrative example, consider the Malaria Refset, curated by the National Resource Centre for EHR Standards in India. This Refset includes various findings, disorders, and organisms directly associated with Malaria. Arguably, Refsets officially published by a group of ontology engineers and domain experts can be regarded as *complete* and *precise* standards for abstracting specific domains within SNOMED CT, such as Malaria.

The task was to predict concepts within SNOMED CT Refsets based on a seed signature. This seed signature could be either randomly generated or manually selected from the Refsets. The task was meticulously designed to align with real-world scenarios in which we aimed to create new Refsets with minimal involvement from domain experts. We operated under the assumption that Refsets developed by domain experts represented "complete" and "precise" fragments, containing concepts intricately linked at the semantic level, often within the same clinical domain. This predictive task of SNOMED CT Refset components served as a crucial evaluation metric for the performance of term selection models.

To accurately position our algorithm within the landscape of existing approaches, we conducted a comparative analysis of NN-RANK against two alternative term selection approaches. These approaches, which we considered as baselines, are as follows:

- star-modularization: An approach adapted from locality-based modularization [12]. It involves considering all concept names within the computed module as an extension of the seed signature. While not necessarily ideal, this method serves as a means to expand the seed signature. Under this approach, the relevance value $f(A, \Sigma)$ of concept name A is assigned a value of 1 if A is present in the signature of the computed module, and 0 otherwise.
- Sig-Ext (Signature Extension): This approach, based solely on geographic connections [5], is configured with a specified depth parameter d.

Additionally, we also conducted a comparative evaluation of NN-RANK against Meta-SVDD [9], a model designed for few-shot one-class classification problems. Using Meta-SVDD, we extracted patterns from existing Refsets to improve its predictive performance when applied to new Refset components.

We utilized the International Edition of SNOMED CT (version July 2020) for our experiments. This edition contains 354,256 concepts, 355,214 logical axioms, and 1,506,185 description axioms. Our study focused on two sets of publicly accessible and actively used term collections, namely the *NHS Refsets*[4] and the *NRC Refsets*[5]. The NHS Refsets, curated by the National Health Service (NHS) in the UK, provide a subset of components derived from the full SNOMED CT Edition, specifically tailored to meet specific requirements. The NRC Refsets, on the other hand, were created and released by the National Resource Centre for EHR Standards (NRCeS) in India. This collection comprises 30 distinct Refsets, each focusing on concepts related to common diseases.

We adopted two established metrics that were widely used in classification and ranking tasks, namely the Normalized Discounted Cumulative Gain (NDCG) and Area under the ROC Curve (AUC), to assess the performance of our term selection models. Both metrics yield higher (lower) values when a model makes more (less) accurate predictions, effectively quantifying the similarity between the model's approximations and the Refset components.

The ontology embedding generated by OWL2Vec* for SNOMED CT was applied to create the concept embeddings, with each concept represented as a 200-dimensional vector. Differing from the original OWL2Vec* model, we implemented a fine-tuning process tailored for this task to further enhance the ontology embedding. Specifically, Refsets in this procedure were transformed into documents containing *(concept_uri, refset_identifier, concept_uri)* triples. Subsequently, we employed a Word2Vec model to refine the pre-computed concept embedding based on these documents. The fine-tuning process followed a 10-fold cross-validation approach. This means that the evaluations for any Refset were based on a concept embedding fine-tuned on 90% Refsets other than itself.

For NRC Refsets, we used two distinct seed signatures Σ_r and Σ_s throughout the experiment, each containing K concepts. Σ_r was randomly chosen from the entire pool of the Refset concepts, while Σ_s was carefully crafted to ensure that the K seeds it encompassed could provide multifaceted coverage of the topic. For NHS Refsets, we used a separate set of Σ_r generated following the same strategy. It was imperative to have the flexibility to adjust the size K of the initial seed signature in accordance with the specific application requirements. In real-world scenarios, the seed signature may be manually selected, and a smaller value of K translates to reduced manual effort. Consequently, we adopted $K = 5$ in our experiments.

We employed the SyntacticLocalityModuleExtractor class from the OWL API[6] to generate star-modules. As previously mentioned,

[2] https://www.snomed.org/
[3] https://confluence.ihtsdotools.org/display/DOCGLOSS/refset
[4] https://dd4c.digital.nhs.uk/dd4c/
[5] https://www.nrces.in/resources#snomedct_releases
[6] https://owlcs.github.io/owlapi/

Methods	NHS Refsets				NRC Refsets			
	NDCG		AUC		NDCG		AUC	
	K=1	K=5	K=1	K=5	K=1	K=5	K=1	K=5
Star-modularization	40.93 ± 14.61	47.33 ± 13.36	50.84 ± 1.10	54.73 ± 5.56	49.10 ± 16.23	51.83 ± 14.62	50.64 ± 0.98	54.58 ± 7.82
Sig-Ext (d=1)	-	49.14 ± 10.92	-	54.31 ± 5.58	-	55.68 ± 11.06	-	53.60 ± 6.73
Sig-Ext (d=2)	-	47.99 ± 11.66	-	54.31 ± 5.58	-	54.31 ± 11.81	-	53.78 ± 6.91
Meta-SVDD	-	67.72 ± 23.26	-	91.55 ± 10.43	-	71.65 ± 16.62	-	88.81 ± 8.87
NN-RANK	68.57 ± 20.36	77.93 ± 14.91	92.19 ± 11.19	96.49 ± 5.11	71.32 ± 14.33	77.25 ± 10.03	89.66 ± 8.42	94.29 ± 5.51
NN-RANK + fine-tuning	**69.50 ± 20.13**	**78.76 ± 14.62**	**93.33 ± 9.57**	**96.98 ± 4.62**	**73.57 ± 12.54**	**80.19 ± 9.14**	**90.40 ± 8.43**	**94.79 ± 5.69**

Table 1: Results on NHS, NRC Refsets using Σ_r (the higher the better).

Methods	NDCG			AUC		
	K=1	K=3	K=5	K=1	K=3	K=5
Star-modularization	48.85 ± 16.68	50.65 ± 15.26	52.25 ± 14.42	50.82 ± 2.01	53.21 ± 5.53	54.68 ± 7.50
Sig-Ext (d=1)	49.97 ± 15.36	53.42 ± 12.16	55.76 ± 10.48	50.80 ± 1.53	52.14 ± 3.89	53.34 ± 5.81
Sig-Ext (d=2)	49.56 ± 15.82	52.33 ± 13.06	54.38 ± 11.36	50.87 ± 1.60	52.28 ± 4.01	53.48 ± 5.93
Meta-SVDD	71.28 ± 12.25	74.91 ± 16.48	75.24 ± 13.73	72.4 ± 16.23	86.83 ± 10.05	92.01 ± 6.45
NN-RANK	79.77 ± 11.79	83.67 ± 10.74	84.83 ± 9.95	94.07 ± 5.11	96.09 ± 3.73	96.64 ± 3.09
NN-RANK + fine-tuning	**80.39 ± 12.02**	**84.41 ± 10.95**	**85.53 ± 10.20**	**94.65 ± 5.01**	**96.49 ± 3.73**	**96.97 ± 3.06**

Table 2: Results for NRC Refset using Σ_s (the higher the better).

all concept names contained in the star-modules collectively constitute the extended seed signature. To generate the extended seed signature using the Sig-Ext method, we utilized the official implementation[7]. As for Meta-SVDD, our implementation was developed based on the source code provided by [7].

4.1 Results and Analysis

The results (mean value ± standard deviation of the two measures) presented in Tables 1 and 2 show that embedding-based methods outperformed their logical counterparts in the above scenarios. This is because, the logical approaches, not being initially designed for this particular task, omitted the ontology's lexical information. As it turned out, this lexical information was of key importance when it came to evaluating the semantic relevance between concepts.

Furthermore, NN-RANK exhibited a slight performance advantage over Meta-SVDD, especially when taking Σ_s as the seed signature. To provide a more in-depth understanding of the workings and effectiveness of NN-RANK in this context, we conducted a case study using the Malaria Refset mentioned earlier. Figure 2 illustrates the distribution of the Malaria Refset components and other SNOMED CT concepts in a 2-dimensional vector space. As depicted in the figure, the Refset components exhibited a tendency to form several smaller clusters, each containing highly semantically relevant concepts. Instead of being a single large cluster, the entire Refset comprised several concept clusters. This implies that when two seed concepts A_1 and A_2 were provided, any concept A exhibiting similarity to either A_1 or A_2, quantified by $d(\mathbf{e}A, \mathbf{e}A_1) < \epsilon$ or $d(\mathbf{e}A, \mathbf{e}A_2) < \epsilon$ with ϵ representing a small value greater than 0, had a higher likelihood of being a component of the Refset compared to another concept A that resembled the average of $\mathbf{e}A_1$ and $\mathbf{e}A_2$, denoted as $d\left(\mathbf{e}A, (\mathbf{e}A_1 + \mathbf{e}_{A_2})/2\right) < \epsilon$. NN-RANK was specifically designed to accommodate this multi-cluster pattern and demonstrated superior performance compared to other models that utilized concept embeddings. The performance of NN-RANK could be significantly enhanced when the seed signatures described the topic from various perspectives. In the case of a high-quality primitive seed signature such as Σ_s, increasing the size of the seed signature typically led to more accurate selection results.

[7] http://bit.ly/2JEaraz

4.2 Time Efficiency

In the current configuration with parameters set to $|N_C| = 354,256$, $K = 5$, $D = 200$ and Cosine distance serving as the distance metric, NN-RANK successfully computed Σ' in just five seconds. In contrast, alternative approaches such as Star-modularization and Sig-Ext necessitated a timeframe ranging from several minutes to even hours for performing similar computations on large-scale ontologies like SNOMED CT. The Meta-SVDD model, in the same configuration, converged within five minutes. While acknowledging that our approach does incur a more substantial training duration—approximately 4 hours—for generating embedding vectors on SNOMED CT, this investment in time is deemed justifiable within practical applications. The rationale behind this is that the training phase is conducted a single time, yet it has the potential to produce significant and relevant outcomes across a multitude of applications, extending indefinitely into future applications. Furthermore, it is worth noting that the required training time is subject to adjustment contingent upon the scale of the ontology, with typical instances requiring less than one hour when the ontology comprises fewer than 100,000 logical and annotation axioms. This capability to tailor the training time according to the ontology's scale enhances the adaptability and efficiency of our approach.

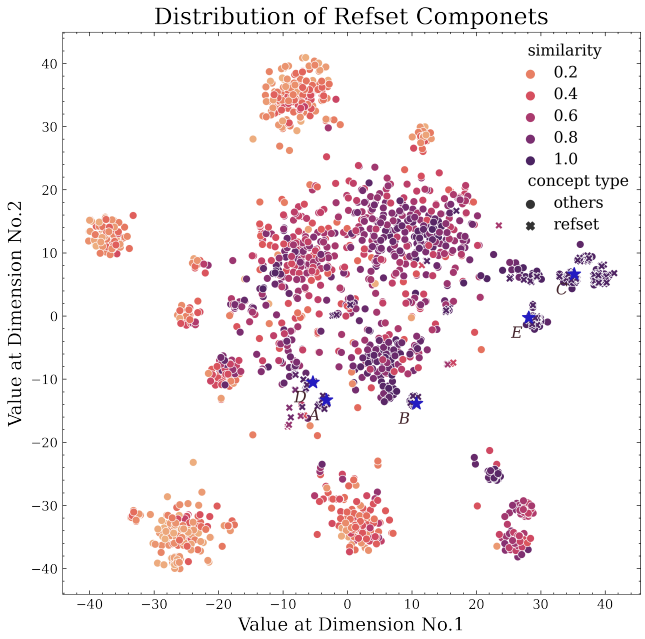

Figure 2: The distribution of components within the malaria Refset and other SNOMED CT concepts is illustrated in this figure, encompassing 170 concepts from the malaria Refset and 1700 random concepts outside the Refset. Each data point on the graph represents a SNOMED CT concept, with its color indicating its relevance to the seed signatures computed by NN-RANK (deeper colors signifying higher relevance). The shape of each point distinguishes its type, with a cross denoting Refset components and circles representing concepts outside the Refset. Seed concepts are denoted by blue stars, accompanied by corresponding tags. These tags correspond to labels as follows: A - Malaria (disorder), B - Allergy to primaquine (finding), C - Accidental pyrimethamine poisoning (disorder), D - Malaria outbreak education (procedure), E - Antimalarial drug adverse reaction (disorder).

5 CASE STUDY: ONTOLOGY ABSTRACTION

In this section, we understand how extending the input signature with NN-RANK brings distinct advantages for the two ontology abstraction approaches — modularization and uniform interpolation. To evaluate the effectiveness of our term selection approach, we required a test ontology with sufficient meaningful metadata. To this end, we selected HeLiS[8], an $\mathcal{ALCHIQ}(\mathcal{D})$ ontology that describes knowledge related to food and activity from a nutritional perspective. The experiment was conducted using HeLiS version 1.10, which consisted of 172,213 axioms, 277 concepts, and 50 roles.

5.1 Setup Details

First, we created 10 concept subsets from $\text{sig}(\mathcal{O}_{HeLiS})$ to serve as the initial seed signatures, denoted as Σ_r, with set sizes ranging from 1 to 5. As the selection process was random, the chosen concept

[8]https://horus-ai.fbk.eu/helis/

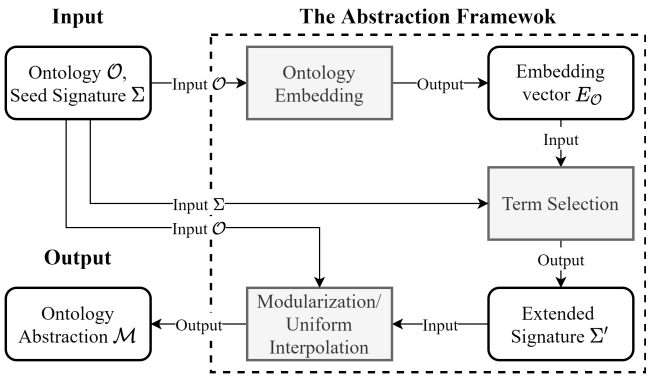

Figure 3: The ontology abstraction framework

names might pertain to different topics encompassed by the Helis ontology. NN-RANK then computed the extended seed signatures based on these initial selections.

Given that most real-world applications typically require smaller ontology abstracts, the size of the seed signature is expected to be relatively small. This correlation arises because the size of generated ontology abstracts is directly proportional to the size of the input seed signature. Consequently, we only included concept names from Σ' whose relevance values to the original Σ_r fell within the top 10% (i.e., by setting the threshold at 0.9). These selected concept names were collected into the final seed signature, serving as input to the subsequent modularization and uniform interpolation procedures. Figure 3 illustrates the ontology abstraction pipeline employing the term selection approach.

For our computations, we employed UI-FAME [24] to compute Σ-uniform interpolants and OWL API to compute Σ-star-modules, as these tools are publicly accessible. Notably, both methods preserved the complete logical consequences of the input signature Σ' within $\mathcal{O}_{HeLiS}$ [12, 16]. Subsequently, the abstraction results generated by these two tools using the input of Σ' (denoted as Σ'+UI-FAME, Σ'+Star-modularization) were evaluated using four metrics: module size $|\mathcal{M}|$, module inherent richness *InhRich*, module intra distance *IntraDist*, and module cohesion *Cohesion*. A module characterized by relatively smaller size, higher inherent richness, relatively smaller intra distance, and higher cohesion was considered more *compact*. Additionally, we conducted a comparative analysis between Σ_r+Star-modularization and Σ'+Star-modularization.

5.2 Results and Analysis

We conducted a comparison between Σ'+UI-FAME and Σ'+Star-modularization to assess the effectiveness of NN-RANK with different abstraction methods. Table 3 provides insights into this comparison, revealing that UI-FAME generated more *compact* abstractions. Additionally, it was observed that UI-FAME exhibited sensitivity to the input signature. These findings align with expectations since locality-based modularization introduced additional terms not present in Σ', whereas uniform interpolation adhered to Σ'. Further experiments with thresholds set at 0.3, 0.5, and 0.7 indicated that the size of Σ' did not significantly impact the compactness of the locality-based module abstraction.

Metrics	K=1		K=5	
	Star-modularization	UI-FAME	Star-modularization	UI-FAME
$\|\mathcal{M}\|$	171 ± 14	**20 ± 7**	174 ± 15	**18 ± 8**
InhRich	**2.92 ± 0.12**	2.1 ± 1.25	**4.08 ± 0.17**	3.75 ± 0.49
IntraDist	49683.90 ± 94.61	**618.75 ± 617.87**	49798.70 ± 278.77	**289.50 ± 344.26**
Cohesion	0.08 ± 0.01	**0.19 ± 0.09**	0.08 ± 0.00	**0.15 ± 0.10**

Table 3: Module Compactness Evaluation (Using the top 10% of Σ' as input). $|\mathcal{M}|$: The total number of concepts, roles, and individuals in $\mathcal{M}$. *InhRich*: The average number of subclasses per class. *IntraDist*: The overall distance between the entities in the module. *Cohesion*: The degree to which entities are interrelated within the module.)

Term selection plays a crucial role in allowing users to expand seed signatures in a customizable manner. In the context of uniform interpolation, selecting appropriate terms for a given topic is pivotal since the semantics of the topic heavily relies on the input terms. We have observed that insufficient input terms for uniform interpolation can result in very small abstracts, often containing numerous trivial axioms like $A \sqsubseteq \top$ or concept assertion axioms.

NN-RANK+UI-FAME demonstrates a significant ability to generate knowledge highly relevant to the specified topic. For example, as shown in Table 4, let us consider the topic "SpecialBread". The relevant axioms from $\mathcal{O}_{HeLiS}$ were found in $\mathcal{O}_{fragment}$. Notably, "SpecialBread" had five individuals, and these individuals had no super-classes other than "SpecialBread". Applying common sense, we can infer that "OliveBread" can be linked to "OlivesAndOliveProducts", "SoyBread" to "SoyProducts", and "MilkBread" to "MilkAndDairyProducts". These links were absent however in the logical components of $\mathcal{O}_{HeLiS}$.

Without NN-RANK's extension, these concepts, which are highly relevant to the central topic "SpecialBread", could not be preserved in Σ_r+Star-modularization or Σ_r+UI-FAME. In contrast, NN-RANK effectively preserved them in the seed signature based on the lexical proximity of "OlivesAndOliveProducts", "SoyProducts", and "MilkAndDairyProducts" to the individuals of "SpecialBread".

In summary, NN-RANK can serve as an optimization booster for modularization and uniform interpolation, aiding them in producing more complete abstracts. In addition, Σ'+uniform interpolation resulted in more precise abstracts compared to Σ'+modularization.

Σ_r	{SpecialBread}
$\mathcal{O}_{\texttt{fragment}}$	SpecialBread ⊑ Bread {SoyBread, OliveBread, MilkBread, OilBread, RyeBread}⊑ SpecialBread
Σ'@10	SpecialBread Bread WhiteBread PizzaAndFocacciaBread OlivesAndOliveProducts SoyProducts LegumesAndLegumeProducts WheatFlour WholeWheatFlour MilkAndDairyProducts

Table 4: Term seletion for SpecialBread topic in HeLiS

6 CONCLUSION AND FUTURE WORK

While modularization and uniform interpolation provide effective means to the abstraction of OWL ontologies, the process of selecting relevant terms—designated as seed signatures— for these abstraction approaches has often posed a significant challenge, hindering users from generating more meaningful ontology abstracts. This paper presents an initial effort to tackle this challenge by extending the given seed signature with carefully selected new terms, identified through embedding-based analysis of crucial metadata within an OWL ontology. An evaluation of this approach, conducted on a predication task involving a SNOMED CT Refset, demonstrated that our method consistently makes accurate selections when compared to other term selection baselines. Finally, a case study illustrates that our term selection approach is capable of producing high-quality modules and uniform interpolants for OWL ontologies.

The absence of standardized benchmarks remains a primary challenge when assessing the performance of term selection methods. Therefore, it would be beneficial to create predefined query answering instances generated from the input ontology. These instances can help verify the completeness and precision of the generated abstracts of OWL ontologies. For a problem Q that can be answered by querying an ontology $\mathcal{O}$, a satisfactory abstract $\mathcal{M}$ of $\mathcal{O}$, given an input signature Σ, should be capable of answering Q when Q is relevant to Σ. Conversely, it should not be able to answer Q when Q is not relevant to Σ.

Furthermore, the quality of term selection results heavily relies on the complexity of the OWL ontology embedding method. One major limitation of using OWL2Vec* for term selection is its inability to effectively capture logical information. Therefore, our current focus is on finding ways to map OWL ontologies into vector spaces with minimal loss of information. It's important to note that our current experiments have only considered concepts, but we plan to incorporate roles in future research.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their insightful comments and good suggestions. This work was supported by the National Natural Science Foundation of China (grant 62006114), the Natural Science Foundation of Jiangsu Province (grant BK20211150) and the Xiaomi Foundation.

REFERENCES

[1] G. Antoniou and F. van Harmelen. *Web Ontology Language: OWL*, pages 67–92. Springer Berlin Heidelberg, 2004.

[2] F. Baader, I. Horrocks, C. Lutz, and U. Sattler. *An Introduction to Description Logic*. Cambridge University Press, 2017.

[3] S. Bechhofer. OWL: web ontology language. In *Encyclopedia of Database Systems, Second Edition*. Springer, 2018.

[4] D. Chandrasekaran and V. Mago. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37, 2021.

[5] J. Chen, G. Alghamdi, R. A. Schmidt, D. Walther, and Y. Gao. Ontology Extraction for Large Ontologies via Modularity and Forgetting. In M. Kejriwal, P. A. Szekely, and R. Troncy, editors, *Proc. K-CAP'19*, pages 45–52. ACM, 2019.

[6] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, and I. Horrocks. Owl2vec*: Embedding of owl ontologies. *arXiv preprint arXiv:2009.14654*, 2020.

[7] G. Dahia and M. P. Segundo. Meta learning for few-shot one-class classification. *arXiv preprint arXiv:2009.05353*, 2020.

[8] M. d'Aquin. Modularizing ontologies. In *Ontology Engineering in a Networked World*, pages 213–233. Springer, 2012.

[9] J. Gamper, B. Chan, Y. W. Tsang, D. Snead, and N. Rajpoot. Meta-svdd: Probabilistic meta-learning for one-class classification in cancer histology images. *arXiv preprint arXiv:2003.03109*, 2020.

[10] W. Gatens, B. Konev, and F. Wolter. Lower and upper approximations for depleting modules of description logic ontologies. In *Proc. ECAI'14*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 345–350. IOS Press, 2014.

[11] B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Modular Reuse of Ontologies: Theory and Practice. *J. Artif. Intell. Res.*, 31:273–318, 2008.

[12] B. C. Grau, B. Parsia, E. Sirin, and A. Kalyanpur. Modularity and web ontologies. In *KR*, pages 198–209, 2006.

[13] I. Horrocks. Ontologies and the semantic web. *Commun. ACM*, 51(12):58–67, 2008.

[14] I. Horrocks, O. Kutz, and U. Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proc. KR'06*, pages 57–67. AAAI Press, 2006.

[15] B. Konev, C. Lutz, D. Walther, and F. Wolter. Model-theoretic inseparability and modularity of description logic ontologies. *Artif. Intell.*, 203:66–103, 2013.

[16] R. Kontchakov, F. Wolter, and M. Zakharyaschev. Logic-based ontology comparison and module extraction, with an application to dl-lite. *Artificial Intelligence*, 174(15):1093–1141, 2010.

[17] P. Koopmann and J. Chen. Deductive Module Extraction for Expressive Description Logics. In *Proc. IJCAI'20*, pages 1636–1643. ijcai.org, 2020.

[18] M. Krötzsch. OWL 2 profiles: An introduction to lightweight ontology languages. In *Proc. Reasoning Web'12*, volume 7487 of *Lecture Notes in Computer Science*, pages 112–183. Springer, 2012.

[19] C. Lutz and F. Wolter. Foundations for Uniform Interpolation and Forgetting in Expressive Description Logics. In *Proc. IJCAI'11*, pages 989–995. IJCAI/AAAI Press, 2011.

[20] B. Motik, I. Horrocks, and U. Sattler. Bridging the gap between OWL and relational databases. In *Proc. WWW'07*, pages 807–816. ACM, 2007.

[21] M. A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, 2015.

[22] S. Staab and R. Studer, editors. *Handbook on Ontologies*, International Handbooks on Information Systems. Springer, 2009.

[23] A. Visser. *Bisimulations, Model Descriptions and Propositional Quantifiers*. Logic Group Preprint Series. Utrecht University, 1996.

[24] X. Wu, W. Deng, C. Lu, H. Feng, and Y. Zhao. UI-FAME: A High-Performance Forgetting System for Creating Views of Ontologies. In *Proc. CIKM'20*. ACM, 2020.