

Metadata-based Term Selection for Modularization and Uniform Interpolation of OWL Ontologies

Xinhao Zhu², Xuan Wu^{1,2}, Ruiqing Zhao², Yu Dong², and Yizheng Zhao^{1,2}

¹ National Key Laboratory for Novel Software Technology, Nanjing University, China

² School of Artificial Intelligence, Nanjing University, China

Abstract. This paper explores the problem of selecting good terms as seed signature for abstraction of OWL ontologies. Existing methods generate seed signatures based on geographic connections, which is far from sufficient to produce a satisfactory abstract. This restricts the reusability of OWL ontologies from the aspect of knowledge management. In this paper, we propose a signature extension approach to generate seed signatures for modularization and uniform interpolation of OWL ontologies, both of which are ontology abstraction techniques. The approach establishes the semantic relevance of terms by taking into account as much as possible metadata information of an OWL ontology, and computes a numerical value to measure the relevance of terms using their embedding transformed based on a so-called OWL2Vec* framework. An empirical evaluation of the approach shows that the proposed method significantly outperforms other term selection baselines in making accurate selections. Besides, a case study on ontology abstraction tasks shows that modularization tools can make more complete and precise abstractions using the signature extended by our method.

Keywords: OWL Ontology · Term Selection · Modularization · Uniform Interpolation

1 Introduction

Because of the heterogeneous nature of web resources, ontologies developed for the semantic web are typically large, sometimes monolithic, and knowledge modelled therein is rich and covers multiple topics. This may however restrict the reusability and interoperability of ontologies in real-world application scenarios, since large ontologies can be difficult to manage, unwieldy to manipulate, and moreover costly to reason about.

Consider an *ontology reuse* use case where an ontologist wants to import a football ontology into a growing sports knowledge base. Currently the only well-established ontology concerning football is the BBC Sports Ontology³, which, however, publishes data about all types of competitive physical activities, pertaining not only to the topic of football. Importing the whole ontology into the knowledge base is not difficult from an engineering perspective, but as one can expect, many web services upon the knowledge base such as *search*, *querying*, *retrieval*, which typically involve extensive reasoning,

³ <https://www.bbc.co.uk/ontologies/sport>

may become problematic, as too much irrelevant information has been, automatically yet unnecessarily, introduced. Such information makes no contribution to the formalization of the information about football but increases the computational cost.

A straightforward way to tackle these challenges of reusability and interoperability is to extract a fragment of an ontology that can behave in the same way as the original ontology in a specific context, but is significantly smaller. In the above case, this means to extract from the **BBC Sports Ontology** a fragment that contains sufficiently many logical statements to summarize all knowledge about football. Ideally, this fragment should be as small as possible.

Two logic-based approaches have been developed for computing fragments of ontologies. One is based on *modularization* [5,9,12,8,2,14], which seeks to identify from an ontology a subset (module) that preserves several reasoning tasks for a sub-vocabulary of the ontology, namely a *seed signature*.⁴ The other is *uniform interpolation* [17,15,6,16], which computes a more compact representation of a module of an ontology which preserves the underlying logical definitions of the terms in the seed signature.

As one could expect, the quality of extracted fragments depends largely on the seed signature fed to modularization and uniform interpolation procedures. We may say that a fragment is *complete* if it covers all essential information about the topic of interest, and a fragment is *precise* if it is complete and in addition, it does not include too much irrelevant information about the topic of interest. More specifically, if we selected as seed signature too few terms to summarize all materials of the topic, we would lose important information that a user may be interested in, and if we selected as seed signature too many terms with some of them not strongly relevant to the topic, we would include too much additional information. Importing more information can also change the definitions of the terms in the original ontology [9].

Nevertheless, very little attention has been paid to the problem of term selection for ontology extraction. Chen et al. [2] have proposed a *signature extension* algorithm to generate seed signatures for ontology modularization. The idea is to (1) fix a *primitive seed signature* Σ , often containing several domain expert-suggested terms, and (2) extend Σ with new terms collected from the axioms which contain the current Σ -terms. This step is iterated until no new terms can be added to Σ . One may understand this as: if two people p_1 and p_2 live together in a house h_1 on an island, then they are *relevant* and team up as $\Sigma = \{p_1, p_2\}$, and if there exists a road connecting h_1 with another house h_2 , then the people living in h_2 are collected into Σ . Iteratively, the same strategy applies to the entire island, and in the end, Σ will probably have collected all habitants on the island. However, a person who lives on another island will never be collected by Σ since there is no road connecting two islands; islands are *geographically isolated*.

Evidently, following this signature extension strategy one must obtain a larger seed signature with which, a more informative fragment will be produced, but we may argue that the seed signature obtained in this way, i.e., using a signature extension algorithm based merely on geographic connections, could hardly yield a *complete* fragment. Our argument is that: the *relevance* between a term and the expanding seed signature should be evaluated based on a consideration of all metadata of the participating terms in the context of the host ontology, rather than based merely on their geographic connections.

⁴ A signature of an ontology is the set of all concept and role names in the ontology.

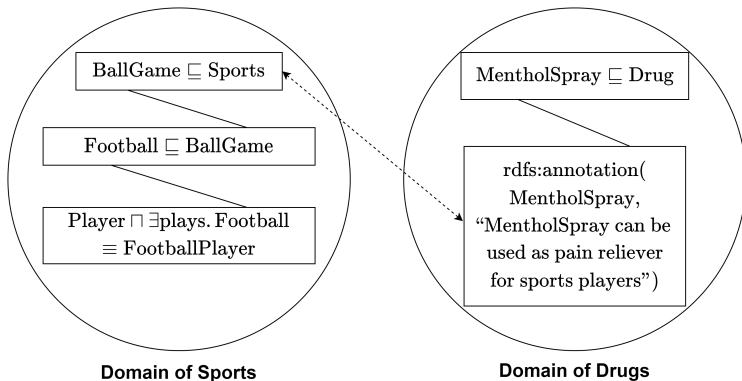


Fig. 1: A snippet of a multi-domain ontology

Consider a scenario where an ontologist wants to extract from a multi-domain ontology a fragment that describes football and closely related information; see Figure 1. With the central term “Football” being selected as a single seed in the primitive signature, an extension $\Sigma = \{\text{Football}, \text{BallGame}, \text{Sports}, \text{Player}, \text{FootballPlayer}\}$ is obtained using the above signature extension algorithm. Terms in other domains such as `MentholSpray` will not be collected in Σ , because it is geographically isolated from the domain of Sports. However, the annotated information of `MentholSpray` explains that “MentholSpray can be used as pain reliever for sports players”. In this sense the term `MentholSpray` is supposed to be strongly relevant to the topic. Collecting `MentholSpray` in the extended signature may enable the expanded knowledge base to answer queries regarding the treatment of an injury in a football match. This is a good example showing that the relevance between a term and the expanding seed signature in the context of the host ontology could be established based on important metadata of the participating terms, for example, based on their lexical information.

In this paper, we propose a novel term selection approach to discovering semantic relationships between two isolated groups of terms. The idea is to measure the relevance of non- Σ terms with Σ terms based on their D-dimensional vector representation computed from important metadata of the ontology using OWL2Vec* [3], a random walk- and word embedding-based OWL ontology embedding framework that encodes the semantics of OWL ontologies in a vector space by taking into account their graph structure, lexical information, as well as the logical constructors used therein. The work is intended to enhance existing logic-based ontology abstraction techniques as practical tools for many ontology-based knowledge processing tasks by exploiting non-logical approaches to facilitate this transfer. Previously, not much work has considered tightly coupled logical and data-driven techniques and exploited the complementary strengths of them to open up an application pipeline. Our empirical evaluation showed that the proposed approach significantly outperformed other term selection baselines in recommending good seed signatures, and with this approach, more precise fragments could be produced using one modularization and one uniform interpolation tool.

2 Metadata-based Term Selection

For space reasons, we have to assume readers' familiarity with the notions of ontology modularization [9] and uniform interpolation [16]. Our term selection approach accommodates ontologies described in OWL 2, which are based on the description logic *SR \mathcal{OIQ}* [11]; see the Description Logic Handbook [1] for a detailed description of the syntax and semantics of description logics.

Arguably, most topics can satisfactorily be summarized or defined by a set of concept names, but do not depend too much on role names. Hence, in this paper, we only consider the seed signature to be a set of concept names.

The signature $\text{sig}(\mathcal{O})$ of an ontology \mathcal{O} is the set of all concept names in \mathcal{O} . Given an ontology \mathcal{O} and a seed signature $\Sigma \subseteq \text{sig}(\mathcal{O})$ containing a single or a few concept names suggested by domain experts or simply selected by users, which are believed to be the central term or terms that can best summarize the topic of interest, our approach computes an extension Σ' of Σ in three steps, namely *concept representation learning*, *computing relevance value*, and *signature extension based on relevance value*. Σ' is the seed signature to be fed to modularization and uniform interpolation procedures.

2.1 Concept Representation Learning

The first step is to transform all concept names A in \mathcal{O} into D -dimensional vectors in a vector space where the *relevance* of each concept name (to Σ) is computed based on important metadata of \mathcal{O} .

Our concept representation learning model is based on OWL2Vec* [3], an ontology embedding framework, which computes the vector representations for concept names in OWL ontologies as expressive as *SR \mathcal{OIQ}* . OWL2Vec* computes the embedding of an OWL ontology based on a corpus of sequences of tokens, which are encoded from the metadata of the ontology. Such metadata includes the graph structure of the ontology, i.e., an RDF graph (a set of RDF triples) converted from the OWL ontology by OWL2Vec*, the so-called lexical information about the ontology, i.e., annotations, and the so-called logical information about the concepts and roles in the ontology, i.e., subsumption, equivalence, disjointness, etc.

We note that OWL2Vec* was not meant for term selection tasks, so we make modifications to the original OWL2Vec* model to maximize the performance of the downstream term selection models. In particular, we designed a fine-tuning process to further improve ontology embedding, which was task-specific and further discussed in section 3. In the end, every concept name A is represented as a D -dimensional vector e_A .

2.2 Computing Relevance of Concept Names w.r.t Σ

The second step is to compute the relevance value of every (non- Σ) concept name A in \mathcal{O} w.r.t. Σ . The computation is based on the relative distance of e_A to its nearest seed neighbour (the nearest seed name) in the vector space. The range of the relevance value is $[0, 1]$ with 1 standing for the strongest relevance and 0 for the weakest relevance. The relevance value is computed by a newly developed algorithm called Nearest Neighbor Ranking algorithm (NN-RANK), shown in Algorithm 1.

Algorithm 1 Nearest Neighbour Ranking

Input: A set of concepts N_C , 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 $f : N_C \rightarrow [0, 1]$,

```

1: Let  $g$  be a mapping of  $N_C \rightarrow [0, \infty]$ .
2: for all  $A \in N_C$  do
3:    $g(A) := \infty$ 
4:   for all  $A' \in \Sigma$  do
5:      $g(A) := \min(d(\mathbf{e}_A, \mathbf{e}_{A'}), g(A))$ .
6:   end for
7: end for
8: Let  $f$  be a mapping of  $N_C \rightarrow [0, 1]$ .
9: for all  $A \in N_C$  do
10:   Find  $i$ , s.t.  $A$  has the  $i$ -th smallest  $g(A)$  in  $N_C$ .
11:    $f(A) := 1 - (i - 1)/|N_C|$ .
12: end for
13: return  $f$ 

```

NN-RANK first computes the *distance* from each concept name to each seed name in the vector space. In principle, many distance functions $d : \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, \infty]$ can be used to achieve this, but 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 made the best measure of relevance in our experiments. $|\Sigma|$ distance values are computed in this way for each concept name A , while the smallest distance value, which denotes the shortest distance, is identified as a *valid distance value* of A to Σ . NN-RANK then sorts all concepts names in \mathcal{O} by their valid distance value. Concept names with smaller valid distance values are considered to be semantically more relevant to the seed signature, and thus to the central topic. These valid distance values (and the corresponding concept names) are then uniformly distributed between 0 and 1. The result is the *relevance value* of each A w.r.t. Σ .

2.3 Relevance-based Seed Signature Extension

A natural question arises: how to use the computed relevance values to guide the selection of terms for ontology abstraction? Upon different application demands, the strategies may vary. Without a well-acknowledged gold standard, a feasible solution could be to measure the “degree” of relevance and define to what degree the relevance is a concept name can be thought of as “relevant” to the seeds in Σ . In this work, we use a threshold σ at the scale of 0 to 1 to denote the “degree” of relevance. Our approach extends the primitive seed signature Σ by adding to Σ the concept names with relevance value no less than σ . The result is $\Sigma' = \Sigma \cup \{A \mid A \in \text{sig}(\mathcal{O}) \wedge f(A, \Sigma) \geq \sigma\}$.

Computing $|\text{sig}(\mathcal{O})| \times |\Sigma|$ distances requires linear time to $|\text{sig}(\mathcal{O})|$, and the subsequent sorting requires linear time to $|\text{sig}(\mathcal{O})| \cdot \log(|\text{sig}(\mathcal{O})|)$. Hence, we have the following lemma regarding the time complexity of NN-RANK.

Lemma 1. *Given any OWL ontology \mathcal{O} in $SR\mathcal{O}IQ$ and a primitive seed signature $\Sigma \subseteq \text{sig}(\mathcal{O})$ with $n = |\text{sig}(\mathcal{O})|$ and $k = |\Sigma|$, our term selection approach always computes an extended seed signature Σ' such that $\Sigma \subseteq \Sigma'$ in $O(n \log n + kn)$ time.*

3 Empirical Evaluation of NN-RANK

In this experiment, we used NN-RANK to predict SNOMED CT Refset components. The aim was to show that the algorithm could enrich a given primitive seed signature Σ with concept names highly relevant to the initial seeds (in a vector space). The experiment was conducted on a work station with an Intel Xeon CPU @2.60GHz and 32 GB memory.

SNOMED CT⁵ is currently the most comprehensive, multilingual clinical health-care ontology in the world. A SNOMED CT Refset⁶ is a collection of SNOMED CT components sharing specific characteristics (e.g., a specific domain). An example of SNOMED CT Refset is the Malaria refset released by the National Resource Centre for EHR Standards in India, which includes findings, disorders, and organisms related to Malaria. Arguably, the refset published officially by a group of ontology engineers and domain experts, can be considered as a *complete* and *precise* standard of an Malaria abstract of SNOMED CT.

Our task was to predict concepts in SNOMED CT Refsets based on a seed signature (randomly or manually) selected from the refsets. This task was designed to fit with realistic scenarios where we needed to develop a new refset with least intervention from domain experts. We assumed that refsets developed by the domain experts were *complete* and *precise* fragments, containing concepts that were highly interconnected on the semantic level (e.g., in the same clinical domain). Therefore, the task of predicting SNOMED CT Refset components could be used to evaluate the performance of term selection models.

To better position our algorithm, we compared NN-RANK with two other term selection strategies, namely, a strategy adapted from locality-based modularization [10] (denoted as *Star-modularization*), and the signature-extension based on geographic connections [2] (denoted as *Sig-Ext*, configured with depth d). We treated them as baselines. The idea of the locality-based modularity strategy was to take all concept names in the computed module as the extended signature of the seed. This may not be ideal but was nevertheless a means to extend the seed signature. In this way, the relevance value $f(A, \Sigma)$ of A was 1 if A was in the signature of the computed module, and 0 otherwise. We also considered a comparison of NN-RANK with Meta-SVDD [7], a model designed for few-shot one-class-classification problems. Using Meta-SVDD, we learnt patterns about refsets from existing refsets, in order to enhance its performance in predicting new refset components.

⁵ <https://www.snomed.org/>

⁶ <https://confluence.ihtsdotools.org/display/DOCGLOSS/refset>

We considered the International Edition of SNOMED CT (version July 2020), which contains 354,256 concepts, 355,214 logical axioms, and 1,506,185 description axioms. We used two sets of publicly accessible and in-use term collections, *NHS refsets*⁷ and *NRC refsets*⁸, as the target refsets.

The NHS refsets, issued by the National Health Service (NHS) in the UK, offered from the full Edition of SNOMED CT a set of components defined by a particular requirement. The NRC refsets were released by the National Resource Centre for EHR Standards (NRCeS) in India, which contained 30 standalone refsets covering concepts related to common diseases.

We adopted two metrics widely used in classification and ranking tasks, namely the Normalized Discounted Cumulative Gain (NDCG) and the Area under the ROC Curve (AUC), to evaluate the performance of term selection models. Both measures returned high values if a model made accurate predictions, i.e. they measured the similarity between the approximations and the refset components.

Ontology embedding generated by OWL2Vec* on SNOMED CT was used for the concept embedding, where each concept was represented by a 200-dimensional vector. Different from the original OWL2Vec* model, we used a fine-tuning process specially designed for this task, to further improve the ontology embedding. Specifically, refsets in this process were transformed to documents containing (*concept_uri*, *refset_identifier*, *concept_uri*) triples, then a Word2Vec model was used to fine-tune the pre-computed concept embedding on these documents. The fine-tuning process was done in a 10-fold cross validation manner, which meant that evaluations on any refset is based on a concept embedding fine-tuned on 90% refsets other than itself.

For NRC refsets, two seed signatures Σ_r and Σ_s consisted of K concepts respectively were used throughout the experiment. Σ_r was randomly selected among all the refset concepts, while Σ_s was manually selected with the aim that the K concepts it contained could describe the topic from different aspects. For NHS refsets, we only used a different set of Σ_r generated in the same way. It was crucial to be able to set the size of the primitive seed signature K accordingly to the application. In realistic use cases, the seed signature may be manually selected, where smaller K means less manual cost, so $K = 5$ is used in the experiments.

We used the OWL API syntactic locality module extraction tool⁹ as the implementation of the locality-based module, and the official implementation of Sig-Ext. For Meta-SVDD, our implementation was based on the source code provided by [4].

3.1 Results and Analysis

The results (mean value \pm standard deviation of the two measures) in Table 1 and 2 show that embedding-based methods outperformed logical approaches in the above settings. This was because logical methods were not designed for this task, and it did not capture lexical information of the ontology, which was crucial in determining the semantic relevance between concepts.

⁷ <https://dd4c.digital.nhs.uk/dd4c/>

⁸ https://www.nrces.in/resources#snomedct_releases

⁹ <https://github.com/owlcs/owlapi>

Besides, NN-RANK slightly outperformed Meta-SVDD, particularly when using Σ_s . We will conduct a case study on the aforementioned Malaria refset to explain the mechanism and effectiveness of NN-RANK in this task.

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

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 2: Results on NRC refset using Σ_s (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

Figure 2 shows the distribution of the Malaria refset components and other SNOMED CT concepts in a 2-dimensional vector space. As illustrated in the figure, refset components tended to form a number of minor clusters, with each containing some highly semantically relevant concepts. The whole refset was composed of several concept clusters instead of a giant cluster. This meant that when two seed concepts A_1 and A_2 were given, any concept A that was similar to A_1 or A_2 , i.e. $d(\mathbf{e}_A, \mathbf{e}_{A_1}) < \epsilon$ or $d(\mathbf{e}_A, \mathbf{e}_{A_2}) < \epsilon$ with ϵ being a small value greater than 0, were more likely to be a refset component compared to another A which was similar to the average of \mathbf{e}_{A_1} and \mathbf{e}_{A_2} , i.e., $d(\mathbf{e}_A, (\mathbf{e}_{A_1} + \mathbf{e}_{A_2})/2) < \epsilon$. NN-RANK was designed to fit in this multi-clusters pattern, and achieved better performance compared to other models utilizing concept embedding.

The performance of NN-RANK could be significantly enhanced when seed signatures described the topic from different aspects. For a high quality primitive seed signature like Σ_s , an increased seed signature size would generally led to more accurate selection results.

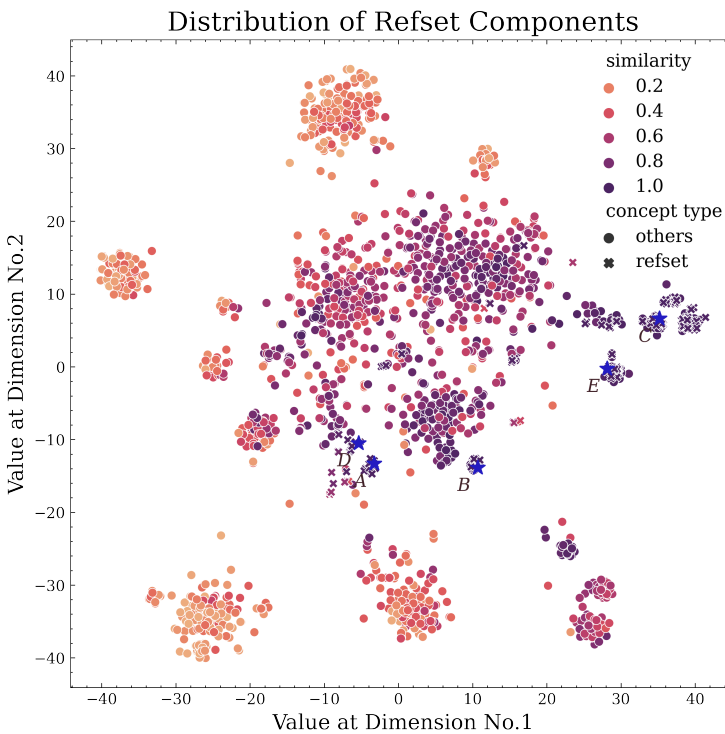
3.2 Time Efficiency

For the current setting of $N = 354, 256$, $K = 5$, $D = 200$ and using Cosine distance as the distance function, NN-RANK generated Σ' within 5 seconds. For comparison,

it usually takes minutes to hours for other approaches (e.g., Star-modularization and Sig-Ext) to compute on a large-scale ontology like SNOMED CT, and five minutes for the Meta-SVDD model to converge in the same setting.

It is true that our approach takes hours to build embedding vectors on SNOMED CT, but this cost is acceptable in real-life scenarios since the training is conducted only once but can be meaningfully used many times and forever. Also, the training time can be adjusted. When the ontology contains less than 100K logical and annotation axioms, it is typically less than one hour.

Fig. 2: Distribution of malaria refset components and other SNOMED CT concepts (170 concepts from the malaria refset and 1700 random concepts outside of the refset). Each point corresponds to a SNOMED CT concept, whose colour shows its relevance with the seed signatures computed by NN-RANK (the higher, the deeper), and shape denotes its type (cross for being refset components, and circles for not). Seed concepts are depicted as blue stars accompanied with tags. The mappings between tag and label are: A - Malaria (disorder), B - Allergy to primaquine (finding), C - Accidental pyrimethamine poisoning (disorder), D - Malaria outbreak education (procedure), E - Antimalarial drug adverse reaction (disorder)



4 Case Study: Ontology Abstraction

In this part, we explored how input signature extended by NN-RANK benefits differently between modularization and uniform interpolation in the OWL ontology abstraction task.

As we need ontology having enough metadata to test the effectiveness of term selection method, we considered HeLiS¹⁰, an $\mathcal{ALCHIQ}(\mathcal{D})$ ontology integrating knowledge about food and activity from a nutritional point of view. The experiment was based on HeLiS v1.10 which has 172,213 axioms, 277 concepts, and 50 roles.

4.1 Setup Details

First, we randomly generated 10 concept subsets from $\text{sig}(\mathcal{O}_{\text{HeLiS}})$ with the size of subsets ranged from 1 to 5. These randomly generated concept sets, denoted as Σ_r , could be the approximations of seed signatures around random topics. Then NN-RANK returned the ordered sets Σ' .

As the abstractions in real-life are usually small in size, we chose the top 10% of Σ' (i.e., set the threshold as 0.9) to be the input signature for modularization and uniform interpolation. We used UI-FAME [18] to compute uniform interpolants, and Star-modularization to compute locality-based modules as they are publicly accessible. Both preserved full logical entailments of the input signature Σ' in $\mathcal{O}_{\text{HeLiS}}$ [10,13]. Then the abstraction results computed by these two tools with the input of Σ' (denoted as $\Sigma'+\text{UI-FAME}$, $\Sigma'+\text{Star-modularization}$) were assessed with four metrics: module size $|\mathcal{M}|$, module inherent richness *InhRich*, module intra distance *IntraDist* and module cohesion *Cohesion*. A module with relative smaller size, higher inherent richness, relative smaller intra distance, and higher cohesion was said to be more *compact*. We also test $\Sigma_r+\text{Star-modularization}$ and compared it with $\Sigma'+\text{Star-modularization}$.

4.2 Results and Analysis

We compared $\Sigma'+\text{UI-FAME}$ and $\Sigma'+\text{Star-modularization}$ to see the effectiveness of NN-RANK to different abstraction methods. From table 3, we can see that UI-FAME generated more *compact* abstractions. Besides, UI-FAME was sensitive to the input signature. These results make sense because locality-based modularization introduced other terms which were not in Σ' but uniform interpolation stuck to Σ' . Experiments with thresholds setting as 0.3, 0.5, and 0.7 show that the size of Σ' did not affect the compactness of the locality-based module abstraction.

Term selection allowed users to extend the seed signature in an adjustable way. For uniform interpolation, it is a key step to select suitable terms for the specified topic, because the semantics of the topic is mainly captured by the input terms. We observe that once if the input terms were not sufficient enough for uniform interpolation, the module could be very small, containing many meaningless axioms like $A \sqsubseteq \top$ or concept assertion axioms. NN-RANK+UI-FAME generated knowledge highly relative to the topic. For instance, in Table 4, the topic was “SpecialBread”. The related axioms in $\mathcal{O}_{\text{HeLiS}}$

¹⁰ <https://horus-ai.fbk.eu/helis/>

Table 3: Module compactness Evaluation (Use top 10% Σ' as input. $|\mathcal{M}|$ is the sum of quantities 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 extent to which entities are related to each other in the module.)

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 4: Term selection for SpecialBread topic in HeLiS

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

were contained in $\mathcal{O}_{fragment}$. Clearly, “SpecialBread” had five individuals. Besides, these individuals had no other super-classes except “SpecialBread”. As commonsense knowledge, “OliveBread” can be “OlivesAndOliveProducts”, “SoyBread” can be “SoyProducts”, “MilkBread” can be “MilkAndDairyProducts”, which were missing in \mathcal{O}_{HeLiS} . So without the extension of NN-RANK, these related concepts could not be preserved in $\Sigma_r + \text{Star-modularization}$ or $\Sigma_r + \text{UI-FAME}$. While NN-RANK could preserve them according to that “OlivesAndOliveProducts”, “SoyProducts”, and “MilkAndDairyProducts” were lexically close to the individuals of the topic concept “SpecialBread”.

To sum up, with NN-RANK modules and uniform interpolants produced more *complete* fragments. In addition, Σ' +uniform interpolation produced more *precise* fragments than Σ' +modularization.

5 Conclusion and Future Work

This paper makes a preliminary attempt to address the problem of extending the given seed signature with new terms selected sophisticatedly through embedding-based computation of important metadata of an OWL ontology. An evaluation of the approach on a predication task of a SNOMED CT refset shows that our approach makes accurate selections compared with other term selection baselines. A case study shows that high-quality modules and uniform interpolants of OWL ontologies can be produced using our term selection approach.

The absence of standardized benchmarks remains the main bottleneck in evaluating the performance of term selection methods. Hence, a number of pre-defined question answering instances that are generated based on the input ontology might be helpful in deciding 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} regarding an input signature Σ should be able to answer Q if Q is relevant to Σ , and should not be able to answer Q if Q is not relevant to Σ .

Besides, the current experiments only considered concepts. Roles will also be considered in the future work.

References

1. F. Baader, I. Horrocks, C. Lutz, and U. Sattler. *An Introduction to Description Logic*. Cambridge University Press, 2017.
2. 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.
3. 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.
4. G. Dahia and M. P. Segundo. Meta learning for few-shot one-class classification. *arXiv preprint arXiv:2009.05353*, 2020.
5. M. d'Aquin. Modularizing ontologies. In *Ontology Engineering in a Networked World*, pages 213–233. Springer, 2012.
6. T. Eiter, G. Ianni, R. Schindlauer, H. Tompits, and K. Wang. Forgetting in managing rules and ontologies. In *Web Intelligence*, pages 411–419. IEEE Computer Society, 2006.
7. 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.
8. 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.
9. 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.
10. B. C. Grau, B. Parsia, E. Sirin, and A. Kalyanpur. Modularity and web ontologies. In *KR*, pages 198–209, 2006.
11. I. Horrocks, O. Kutz, and U. Sattler. The even more irresistible *SRQLQ*. In *Proc. KR'06*, pages 57–67. AAAI Press, 2006.
12. B. Konev, C. Lutz, D. Walther, and F. Wolter. Model-theoretic inseparability and modularity of description logic ontologies. *Artif. Intell.*, 203:66–103, 2013.
13. R. Kontchakov, F. Wolter, and M. Zakharyashev. Logic-based ontology comparison and module extraction, with an application to dl-lite. *Artificial Intelligence*, 174(15):1093–1141, 2010.
14. P. Koopmann and J. Chen. Deductive Module Extraction for Expressive Description Logics. In *Proc. IJCAI'20*, pages 1636–1643. ijcai.org, 2020.
15. J. Lang, P. Liberatore, and P. Marquis. Propositional independence: Formula-variable independence and forgetting. *J. Artif. Intell. Res.*, 18:391–443, 2003.
16. 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.

17. A. Visser. *Bisimulations, Model Descriptions and Propositional Quantifiers*. Logic Group Preprint Series. Utrecht University, 1996.
18. Y. Zhao, G. Alghamdi, R. A. Schmidt, H. Feng, G. Stoilos, D. Juric, and M. Khodadadi. Tracking logical difference in large-scale ontologies: a forgetting-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3116–3124, 2019.