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

Xinhao Zhu, Xuan Wu, Ruiqing Zhao, Yu Dong, Yizheng Zhao*

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

School of Artificial Intelligence, Nanjing University, China

1. Introduction

- Ontologies are typically monolithic, and knowledge therein is rich and covers multiple topics. To improve the **reusability** and **interoperability** of ontologies, a straightforward way is to extract **ontology fragments** that can behave in the same way as the original ontology in a specific context, but significantly smaller.
- ***Modularization*** [1] and ***uniform interpolation*** *(UI)* [2] are logic-based approaches developed for computing ontology fragments, both of which preserve all the logical consequences of a **seed signature**, which is a sub-vocabulary of the ontology.
- Nevertheless, very little attention has been paid to the problem of **term selection** (seed signature selection) for ontology extraction.
- We argue that both the **logical information** and **lexical information** (e.g., **annotations**) of an ontology can help in establishing the relevance between terms. Fig. 1 shows that the *MentholSpray* concept could be relevant to the Sports domain under certain circumstances (e.g., when answering queries regarding the treatment of an injury in a football match).

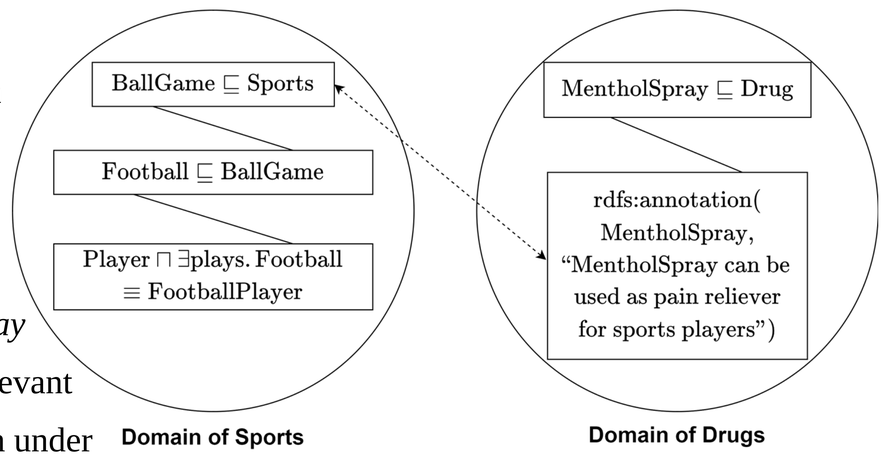

Fig. 1: A snippet of a multi-domain ontology

- In this work, we propose a novel term selection approach to discovering semantic relationships between two isolated groups of terms utilizing ontology embedding techniques.

2. The Proposed Method

- The signature $\mathrm{sig}(O)$ of an ontology O is the set of all concept names in O. Given an ontology O and a **primitive seed signature** $\Sigma \subseteq \mathrm{sig}(O)$ containing few concept names suggested by domain experts or simply selected by users, which are believed to be terms that can best summarize the topic of interest, our approach computes an extension Σ' of Σ in **three steps**.
- **Step 1** (*concept representation learning*): Transform all concept names A in O into ***D*-dimensional vectors** using the OWL2Vec* [3] model, the computation of which is based on all **metadata** (including **logical** axioms and **annotation** axioms) of O.
- **Step 2** (*computing relevance value*): The relevance value (ranged from 0 to 1, with 1 standing for the strongest relevance and 0 for the weakest relevance) is computed by a newly developed algorithm called Nearest Neighbor Ranking algorithm (NN-RANK), which is shown in Alg. 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 $d : \mathbb{R}^D \times \mathbb{R}^D \to [0, \infty]$.
Output: A relevance function $f : N_C \to [0, 1]$,

Let g be a mapping of $N_C \to [0, \infty]$.
for all $A \in N_C$ **do**
  $g(A) := \infty$
  for all $A' \in \Sigma$ **do**
    $g(A) := \min\left(d\left(\mathbf{e}_A, \mathbf{e}_{A'}\right), g(A)\right)$.
  end for
end for
Let f be a mapping of $N_C \to [0, 1]$.
for all $A \in N_C$ **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

Alg. 1: The Nearest Neighbour Ranking Algorithm

- **Step 3** (*Relevance-based Seed Signature Extension*): Given a threshold σ at the scale of 0 to 1, the **extended seed signature** Σ' is generated by $\Sigma' = \Sigma \cup \{A \mid A \in \mathrm{sig}(O) \ \wedge \ f(A, \Sigma) \ \geq \ \sigma\}$.

3. Empirical Evaluation on SNOMED CT

- In this experiment, our task was to predict concepts in **SNOMED CT Refsets** (which contain terms in a specific clinical domain) based on a seed signature (randomly or manually) selected from these Refsets.

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	$\mathbf{80.39 \pm 12.02}$	$\mathbf{84.41 \pm 10.95}$	$\mathbf{85.53 \pm 10.20}$	$\mathbf{94.65 \pm 5.01}$	$\mathbf{96.49 \pm 3.73}$	$\mathbf{96.97 \pm 3.06}$

Tab. 1: Metrics on the NRC Refset using manually selected seed (the higher the better).

- Tab. 1 shows that the proposed method (NN-RANK) outperformed two logic-based term selection baselines, especially when fine-tuning is done on the concept embedding. Besides, NN-RANK slightly outperforms another embedding-based method.
- NN-RANK was designed to fit in the multi-clusters pattern of topic words in the embedding space (see Fig. 2), which made it more effective.

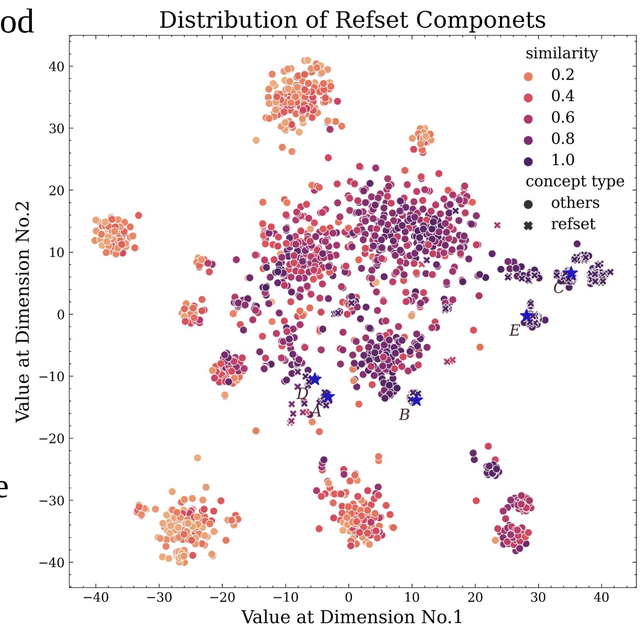

Fig. 2: Concept distribution on the malaria Refset

4. Case Study: Ontology Abstraction

- Moreover, we evaluated how *modularization* and *uniform interpolation* benefit from the seed signature extended by NN-RANK in the OWL ontology abstraction task on the HeLiS ontology (shown in Fig. 3).
- It is shown in Tab. 2 that, with the help of NN-RANK, both methods output fragments that cover more concepts *relevant* to the topic, while UI fragments cover fewer *irrelevant* concepts and have better compactness.

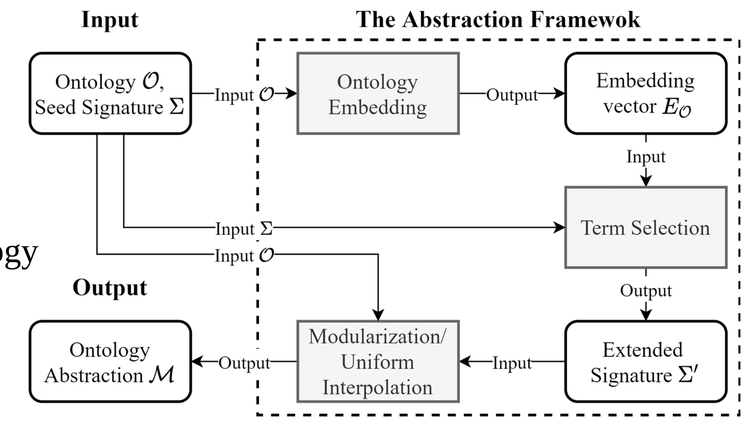

Fig. 3: The abstraction framework

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

Tab. 2: Fragment compactness evaluation

5. Conclusion

- This work made a preliminary attempt to address the problem of extending a given seed signature with new terms selected sophisticatedly through embedding-based computation of important metadata of an OWL ontology.
- The proposed method outperformed other baseline models in selecting terms relevant to a given seed signature on the SNOMED CT ontology, and proved to be helpful in enabling ontology abstraction techniques to cover more terms relevant to the topic.

6. References

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

[2] Lutz, C., Wolter, F.: Foundations for Uniform Interpolation and Forgetting in Expressive Description Logics. In: Proc. IJCAI'11. pp. 989–995. IJCAI/AAAI Press (2011)

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