Combining FCA-Map with Representation Learning for Aligning Large Biomedical Ontologies*

Guoxuan Li[1,2], Songmao Zhang[1], Jiayi Wei[3] and Wenqian Ye[4]

[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100190, China
[3] University of Pennsylvania, 3451 Walnut St., Philadelphia, PA, USA
[4] New York University, 251 Mercer St., New York, NY 10012, USA
liguoxuan18@mails.ucas.ac.cn, smzhang@math.ac.cn, weijiayi@sas.upenn.edu, wy2029@nyu.edu

Abstract. In our previous studies, we developed FCA-Map to utilize the Formal Concept Analysis (FCA) formalism for aligning ontologies in an incremental way. The approach has been shown to be effective by its performance in OAEI 2016, 2018 and 2019. With FCA being inherently a symbolic, logical reasoning theory, we attempt to combine FCA-Map with representation learning techniques so as to take advantage of the semantic representation in numerical, latent space. The resultant system, called SBERTAlignment, is built based on Siamese BERT and has obtained competitive results for matching large biomedical ontologies. Both advantages and limitations are analyzed so as to further our study in exploring ontology similarity from diverse yet complementary perspectives.

1 Introduction

In our previous studies, we developed FCA-Map to utilize the Formal Concept Analysis (FCA) formalism for aligning large and complex ontologies [1]. FCA-Map incrementally constructs formal contexts for specifying the commonality across ontologies at various levels, including lexical matching, structural validation, and structural matching. Mappings are extracted from the derived concept lattice at each level and then used to enable the next-level FCA construction and derivation. The purpose was to push the envelope of FCA in exploiting the ontological knowledge, and our approach has been shown to be effective by its performance in OAEI 2016, 2018 and 2019 on anatomical, biomedical ontologies and knowledge graphs tasks [2].

With FCA being inherently a symbolic, logical reasoning theory, we intend to augment FCA-Map from a diverse perspective and the representation learning technology [3] becomes the one that can hardly be missed in nowaday knowledge engineering research. Representation learning transforms symbolic knowledge base into numerical, low-dimensional space, so that the correlation among entities can be revealed by their vector values.

The work is supported by the Natural Science Foundation of China grant 61621003.

2 Method and Result

Combining FCA-Map and the representation learning system Siamese BERT [4], our ontology matching approach SBERTAlignment consists of three main steps as follows. Firstly, multiple and diverse ways are developed for constructing training samples: (1) using lexical descriptions of entities in ontologies (names, labels and synonyms) and the tokens they share to build a lexical formal context, deriving a lexical lattice of formal concepts, and extracting pairs of entities as positive match samples; (2) for each lexical description of entities, retrieving corresponding terms from external resources like ConceptNet, BableNet and WikiSynonyms, and thus forming pairs as positive match samples; (3) training a word2vec model from PubMed, PMC and Wikipedia and computing the similarity of embeddings of entities so as to obtain positive match samples; (4) using the *is-a* and *part-of* relations within ontologies to yield more positive match samples; and (5) for negative match samples, using the *disjoint-with* relations to generate conflicts between ontologies. Secondly, SBERTAlignment trains a Siamese BERT model which is more effective for similarity-related tasks, and the resultant embeddings are compared in order to decide a one-to-one alignment by stable marriage rationale. Lastly, these matches, together with the matches obtained in (1) above, are fed into a structural formal context, and those validated by the derived structural lattice are the final mappings.

We evaluated on the OAEI 2020 LargeBio small version tasks. SBERTAlignment outperforms FCA-Map in all aspects; and when compared with the state-of-the-art AML and LogMap, obtains highest recall and F-measure for FMA-NCI (92.3% and 93.9%) and FMA-SNOMED (83.1% and 87.4%). We also compared with two representation learning-based systems DOME and MultiOM, and for all the tasks our system outperforms except that DOME has higher precisions.

3 Discussion

We report the preliminary yet promising result of an attempt to take advantage of both symbolic deduction and numerical, latent semantic representation for the purpose of matching complex domain ontologies. Of note, neither the formal clustering in FCA nor the semantic correlation from deep training can decisively determine the equivalence across ontologies, thus comprehensive resources and methods shall be incorporated. We also notice that both FCA-Map and Siamese BERT can be used to align multiple ontologies simultaneously, making indirect alignments available. And our approach should be evaluated on more OAEI tracks like the Disease and Phenotype.

References

1. Zhao, M., Zhang, S., Li, W., Chen, G.: Matching biomedical ontologies based on formal concept analysis. Journal of Biomedical Semantics, 9(1), 1–27 (2018).
2. OAEI Homepage, http://oaei.ontologymatching.org/, last accessed 2021/08/19.
3. Bengio, Y., Courville, A.,Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828 (2013).
4. Reimers, N., Gurevych I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: EMNLP-IJCNLP 2019 Proceedings, pp. 3980–3990. Association for Computational Linguistics (2019).