Noname manuscript No.
(will be inserted by the editor)

Machine Learning on Semantic Scientific Knowledge

Gollam Rabby
Faculty of Informatics and Statistics
University of Economics, Prague, Czech Republic

Received: date / Accepted: date

Abstract In this article, the author presents his dissertation project's description and status after the first year of his doctoral study program in Applied Informatics at the University of Economics, Prague. This document describes his primary topic as well as the specification of the research problem which involves the analysis and development a machine learning base method for the enhancement of the semantic scientific knowledge. Next, a literature review is listed to show that relevant resources and projects were found to explore the state of the art. The dissertation goals and objectives are then defined as the addressed problem. The author then explains his anticipated practical and scientific contributions that might come from his dissertation. Working research assumptions, scientific methods, and anticipated research artifact results are also displayed to address challenges and further clarify the intent of the dissertation which is to design and implement the mentioned support platform. Subsequently, the author also mentions his up-to-date research results which include 1 under review publication and several experimental results. Next, he also talks about his side projects that are focused on the same Semantic Web research area and are closely related to the dissertation research activities. Lastly, the author exposes his comprehensive-time plan for the rest of his Ph.D. study program.

Keywords Scientific Knowledge Graph · Machine Learning · Association rule learning · Bag-of-concept · RDF triple extraction .

1 Introduction

The author was enrolled in the doctoral study program Applied Informatics at the Department of Information and Knowledge Engineering in Autumn 2019 under the supervision of Dr. Tomáš Kliegr and Prof. Vojtěch Svátek(consultant). Together, they discussed potential plans for the future and agreed upon a dissertation project topic proposal with the goal of creating an efficient machine learning base knowledge graph completion method to the scientific knowledge. That goal has not changed since then and the plan has been, as a result of many discussions and research activities, gradually developed up until now. This document aims to bring in a comprehensive description of the project and its current status. For the main part, the author will talk about (1) related works, (2) specification of the dissertation topic, (3) goals and objectives, (4) anticipated practical and scientific contributions, (5) research methods, (6) expected outcomes, (7) up-to-date results including publications. In the last part of this document, (8) time plan for the rest of the study period is also presented to ensure the audience that the author will adhere to the plan and guarantee the dissertation's smooth progress.

Gollam Rabby
E-mail: rabg0@vse.cz

2 Related Works

It is obvious that data driven text representation learning requires sufficient human-annotated and high-quality training corpus, especially for deep compositional models. This is sometimes impractical due to the lexical, morphological, and syntactic variations of natural language. As a result, the learned representations may not be reliable and are normally domain-specific. Many works are proposed to incorporate existing knowledge from external knowledge bases as prior knowledge for machine learning systems, in order to reduce the reliance on training data and provide additional useful information for text representation. In NLP, external knowledge bases can be categorized into the lexical knowledge base and encyclopedic knowledge base.

2.1 Description of different types of Knowledge Graphs

knowledge graphs The most widely used lexical knowledge bases include WordNet[12], SentiWordNet [4], VerbNet[46], FrameNet[23] and ConceptNet[25]. Many works utilize the synonym sets and lexical categories of WordNet as concepts to improve text representation. In, knowledge from WordNet is utilized to perform word sense disambiguation (WSD) by modeling the semantic space and semantic path using LSA and PageRank respectively. Some works use the lexicons in SentiWordNet to perform sentiment classification of reviews [34]. In, knowledge from WordNet and FrameNet is incorporated into CNN for causal relation extraction from texts. To capture semantic and conceptual information of texts, many studies incorporate the knowledge from encyclopedic knowledge bases such as Wikipedia[28], DBpedia[3], Open Directory Project (ODP)[7], Yago[10] etc. In, both Wikipedia and WordNet are used to exploit useful features for short text representation. Some works use the units of knowledge from Wikipedia and DBpedia as concepts for text representation. Some works utilize knowledge base to learn concept embeddings which can be used for document representation. A few studies on short text understanding and classification use the conceptual knowledge in Probase to discover the semantics from short texts.

Research knowledge graphs Academic search engines (e.g. Google Scholar, Microsoft Academic, SemanticScholar) exploit graph structures such as the Microsoft Academic Knowledge Graph[11], SciGraph[52], or the Literature Graph[1]. These graphs interlink research articles through metadata, e.g. citations, authors, affiliations, grants, journals, or keywords. To help reproducing research results, initiatives such as Research Gap, Research Objects, and OpenAIRE [26] interlink research articles with research artefacts such as datasets, source code, software, and presentation videos. Scholarly Link Exchange (Scholix)[6], ORKG [18] etc aims to create a standardised ecosystem to collect and exchange links between research artefacts. Some approaches were proposed to interlink articles at a more semantic level: Paperswithcode.com is a community-driven effort to link machine learning articles with tasks, source code and evaluation results to construct leaderboards. Some people are interlink entity mentions in abstracts with DBpedia and Unified Medical Language System (UMLS)[5], and also extend the citation graph with semantic citation intents (e.g. cites as background or as used method)[5]. Various scholarly applications benefit from semantic content representation, e.g. academic search engines by exploiting general-purpose KGs, and graph-based research paper recommendation systems by utilising citation graphs and mentioned genes. However, the coverage of science-specific concepts in general-purpose KGs is rather low, e.g. the task "geolocation estimation of photos" from Computer Vision is neither present in Wikipedia nor in CSO (Computer Science Ontology)[45].

2.2 Research Ontologies

Various ontologies have been proposed to model metadata such as bibliographic resources and citations such as: FaBiO[36], CiTO[36], BiRO[37], OpenCitations[38]. Iniesta and Corcho[44] reviewed ontologies to describe scholarly articles. In the following, we describe some ontologies that conceptualise the semantic content in research articles. Several ontologies focus on rhetorical[39] (e.g. Background, Methods, Results, Conclusion), argumentative[47] (e.g. claims, contrastive and comparative statements about other work) or activity-based (e.g. sequence of research activities) aspects and elements of research articles. Others describe scholarly knowledge with interlinked entities such as problem, method, theory, statement[41], or focus on the main research findings and characteristics of research articles described in surveys with concepts such as problems, approaches, implementations, and evaluations. There are various domain-specific ontologies, for instance, mathematics[48] (e.g. definitions, assertions, proofs) and machine learning[8] (e.g. dataset, metric, model, experiment). The EXPeriments Ontology (EXPO) is a core ontology for scientific experiments conceptualising experimental design, methodology, and results[49].

2.3 Model Training and Algorithms

In rule-based approaches, one has to define a set of discourse features which can be relevant to characterize the sentences' semantics dedicated to different scopes. State-of-the-art method for rule-based information extraction applied to citation framing has been proposed, using pattern-based features, topic based features and prototypical argument features. As a final step, a training phase is used to weight the relevance of each of the available patterns depending on the class to predict. This is usually done through shallow machine learning models (for instance, k-nearest neighbors [13] or random forest [24]). Such models require smaller sizes of training datasets to provide satisfying results, compared to deep neural networks.

It should be noted that rule-based methods suffer only slightly from unbalanced datasets as the features are hand-crafted, and therefore inferred on wider knowledge and not limited to the sample in the training dataset. If a citation class is under-represented, the classifier could still capture part of the meaning, as the knowledge used for the capturing is provided by an expert. Thus, the lack of balance in the dataset, only slightly degrades the classifier learning.

On the other hand, Deep learning algorithms are artificial neural networks that learn to perform tasks by learning from samples. For the specific problem we address, the network takes as input some selected characteristics of the citation and learns to give as an output the appropriate prediction (citation class). The efficiency of such algorithms does not rely on any task-specific rules, but rather benefits from non linear functions dedicated to capture complex patterns during the learning phase in order to produce a model capable of categorizing new samples. Deep learning algorithms are highly sensitive to the quality of the training data as they do not rely on any external knowledge. As for any machine learning algorithm, the training data should be as balanced as possible, i.e. the variables have to be independent and identically distributed, and the training dataset should be large enough for the system to learn. For the so addressed problem, we need a dataset that is large and balanced across the different citation classes. In fact, if a citation class is underrepresented in the dataset, its characteristics will need to be extracted from a smaller number of samples and the inference mechanism will provide sub-optimal results. The BCN model (Biattentive Classification Network, [27]) designed to handle sentence classification tasks. ELMo (Embeddings from Language Models, [19]) is designed to extract word representations, and can be used to encode sentences to pass through classifiers.

Knowledge Graph specific modelling Linked Open Data has been recognized as a valuable source for background information in data mining. However, most data mining tools require features in propositional form, i.e., a vector of nominal or numerical features associated with an instance, while Linked Open Data sources are graphs by nature. RDF2Vec[43] is an approach that uses language modeling approaches for unsupervised feature extraction from sequences of words, and adapts them to RDF graphs. They also generate sequences by leveraging local information from graph substructures, harvested by Weisfeiler-Lehman Subtree RDF Graph Kernels and graph walks, and learn latent numerical representations of entities in RDF graphs.

RuDiK is a interesting idea for the discovery of declarative rules over knowledge-bases (KBs). RuDiK discovers both positive rules, which identify relationships between entities, e.g., "if coronavirus and alphacoronavirus have the same RNA pattern, they are homogeneous virus", and negative rules, which identify data contradictions, e.g., "if two virus are homogeneous virus, they don't belong to two different ranks in the category list".Rules help domain experts to curate data in large KBs. Positive rules suggest new facts to mitigate incompleteness and negative rules detect erroneous facts. Also, negative rules are useful to generate negative examples for learning algorithms. RuDiK goes beyond existing solutions since it discovers rules with a more expressive rule language w.r.t. previous approaches, which leads to wide coverage of the facts in the KB, and its mining is robust to existing errors and incompleteness in the KB.

Recent advances in information extraction have led to huge knowledge bases (KBs), which capture knowledge in a machine-readable format. Inductive Logic Programming (ILP) can be used to mine logical rules from the KB. These rules can help deduce and add missing knowledge to the KB. While ILP is a mature field, mining logical rules from KBs is different in two aspects: First, current rule mining systems are easily overwhelmed by the amount of data. Second, ILP usually requires counterexamples. KBs, however, implement the open world assumption (OWA), meaning that absent data cannot be used as counterexamples. They develop a rule mining model that is explicitly tailored to support the OWA scenario. It is inspired by association rule mining and introduces a measure for confidence. Our extensive experiments show that our approach outperforms state-of-the-art approaches in terms of precision and coverage. AMIE has shown how rules can be mined effectively from KBs even in the absence of counterexamples. This approach can be optimized to mine even larger KBs with more than 12M statements.

2.4 Knowledge Graph completion methods

Completion of knowledge graphs aims at increasing the coverage of a knowledge graph. Depending on the target information, methods for knowledge graph completion either predict missing entities, missing types for entities, and/or missing relations that hold between entities. Both internal and external methods are used for completing a knowledge graph.

2.4.1 Type Completion

Internal Methods Internal methods use only the knowledge contained in the knowledge graph itself to predict missing information. Binary and multi class link prediction are very common among KG completion tasks. Depending on the graph at hand, it might be worth while distinguishing multi-label classification, which allows for assigning more than one class to an instance (e.g., coronavirus being both related to alphacoronavirus and a rna virus), and single-label classification, which only assigns one class to an instance [50]. Probabilistic methods is used on DBpedia by [3] and Support Vector Machines is used both on DBpedia and Freebase by [31] to exploit interlinks between the knowledge graphs and classify instances in one knowledge graph based on properties present in the other, in order to increase coverage and precision. Besides, Nickel et al [33] propose the use of matrix factorization to predict entity types in YAGO. Whereas association rule mining was used by [42] to predict redundant information from DBpedia and from YAGO.

External Methods External methods use sources of knowledge – such as text corpora or other knowledge graphs – which are not part of the knowledge graph itself. Those external sources can be linked from the knowledge graph, such as knowledge graph interlinks or links to web pages, e.g., Wikipedia pages describing an entity, or exist without any relation to the knowledge graph at hand, such as large text corpora. Nuzzolese et al [40] propose the usage of the Wikipedia link graph to predict types in a knowledge graph using a k-nearest neighbors classifier. Given that a knowledge graph contains links to Wikipedia, interlinks between Wikipedia pages are exploited to create feature vectors, e.g., based on the categories of the related pages. Since links between Wikipedia pages are not constrained, there are typically more interlinks between Wikipedia pages than between the corresponding entities in the knowledge graph. Apriosio et al [2] use types of entities in different DBpedia language editions (each of which can be understood as a knowledge graph connected to the others) as features for predicting missing types. The authors also use a k-NN classifier with different distance measures, such as the overlap of two articles categories. In their setting, a combination of different distance measures is reported to provide the best results. Another set of approaches uses abstracts in DBpedia to extract definitionary clauses, e.g., using Hearst patterns [17]. Such approaches have been proposed by Gangemi et al[16] and Kliegr [20], where the latter uses abstracts in the different languages in order to increase coverage and precision.

2.4.2 Relation Prediction Methods

Internal Methods Among the popular internal methods, classification methods were used in [35] to predict missing relations where they train the system using tensor neural network and Association rule mining was used [35] to find meaningful chains of relations for relation prediction. Likewise, association rule mining was also used in predicting relations as well in [15], to predict relations between entities in DBpedia.

External Methods Like types, relations to other entities can also be predicted from textual sources, such as Wikipedia pages. Lange et al[22] learn patterns on Wikipedia abstracts using Conditional Random Fields [21]. A similar approach, but on entire Wikipedia articles, is proposed by [51]. Another common method for the prediction of a relation between two entities is distant supervision. Typically, such approaches use large text corpora. As a first step, entities in the knowledge graph are linked to the text corpus by means of Named Entity Recognition [30]. Then, based on the relations in the knowledge graph, those approaches seek for text patterns which correspond to relation types (such as: Y's book X being a pattern for the relation author holding between X and Y), and apply those patterns to find additional relations in the text corpus. Such methods have been proposed by Mintz et al[29] for Freebase, and by Aprosio et al[2] for DBpedia. In both cases, Wikipedia is used as a text corpus. A similar setting with DBpedia and two text corpora – the English Wikipedia and an English-language news corpus – is used, the latter showing less reliable results. A similar approach is followed in the RdfLiveNews prototype, where RSS feeds of news companies are used to address the aspect of timeliness in DBpedia, i.e., extracting new information that is either outdated or missing in DBpedia [2]. Nickel et al[32] propose the use of machine learning to fill gaps in knowledge graphs. Like in the works discussed above, they first discover lexicalizations for relations. Then, they use those lexicalizations to formulate search engine queries for filling missing relation values. Thus, they use the whole Web as a corpus, and combine information retrieval and extraction for knowledge graph completion. While text is unstructured, some approaches have been proposed that use semi-structured data for completing knowledge graphs. In particular, approaches leveraging structured data in Wikipedia are found in the literature. Those are most often used together with DBpedia, so that there are already links between the entities and the corpus of background knowledge, i.e., no Named Entity Recognition has to be performed, in contrast to the distant supervision approaches discussed above.

Once such patterns are found for the majority of the list items, they can be applied to the remaining ones to fill gaps in the knowledge graph. Many knowledge graphs contain links to other knowledge graphs. Those are often created automatically [35]. Interlinks between knowledge graphs can be used to fill gaps in one knowledge graph from information defined in another knowledge graph. If a mapping both on the instance and on the schema level is known, it can be exploited for filling gaps in knowledge graphs on both sides. Dutta et al[9] propose a probabilistic mapping between knowledge graphs. Based on distributions of types and properties, they create a mapping between knowledge graphs, which can then be used to derive additional, missing facts in the knowledge graphs. To that end, the type systems used by two knowledge graphs are mapped to one another. Then, types holding in one knowledge graph can be used to predict those that should hold in another.

3 Goals and objectives

The main goal of this dissertation is to propose a methodology for enhancement of the semantic scientific knowledge. This methodology should include the whole process of enhancement of the semantic scientific knowledge, starting with mining rule from knowledge graph and ending with the concept to vector representation. The methodology will be based on the already existing methodology for RDF2Vec[43] and it will be specialized for finding missing triples.

The methodology for RDF2Vec proposes only a general overview of the RDF data to word vector representation. But methodology for enhancement of the semantic scientific knowledge should capture the other knowledge base and find out the missing tuples by using machine learning techniques. Therefore, it is clear from the discussion that the primary objective of this research is to design an efficient machine learning base knowledge graph completion method for scientific knowledge. To attain this primary objective, following sub-objectives are are essential to accomplish.

- To design a concept representation method for scientific knowledge using the existing entity extraction techniques and apply supervised machine learning methods for predicting the impact of scientific publications.
- To enhance knowledge for a scientific knowledge graph (such as find missing triples) from the different Knowledge base (such as DBpedia, Wikidata, ConceptNet) using an impactful scientific publication and predict the relation in the newly added triples using different machine learning techniques.
- To apply rule learning tools/algorithms for knowledge graph completion on scientific knowledge KGs.
- To develop a tool that can assist the scientific knowledge publishers.

Semantic scientific knowledge graph could be essential to research community and knowledge seekers. Therefore, enhancing existing scientific KGs are now a demand of time and RdfRule mining and along with machine learning techniques could be an effective way to enhance and represent semantic scientific knowledge graphs.

4 Anticipated practical and scientific contributions

By solving the problem to find the impactful, reproducible, and important research from a large number of research documents and improving the semantic scientific knowledge, the anticipated practical contributions would bag-of-concept model and find important missing tuples, add them in semantic scientific knowledge. Although this idea is not only improving the semantic scientific knowledge, it also opens the possibility to help the researchers to find out the important knowledge from the published documents in the crisis moment of the society. The author strongly believes that by conducting this research and correctly evaluating it would increase the probable usability of the semantic scientific knowledge and create an effective way to learn about a new research idea and output for the research community.

The anticipated scientific contributions and solving the societal problem which involves the usage of Semantic Web technologies and methodologies. There could be a side goal of proposing best practices based on the process of creating the result application system. The results could be useful for knowledge engineering such as:

- Create a machine learning method that involves the enhancement of semantic scientific knowledge. Example: Find ”coronavirus” and ”coronaviral” in a research document as an entity but there are some other entities probably missing such as alphacoronavirus, blue comb, genome, RNA virus, etc. Also probably miss some useful information such as coronavirus RelatedTo severe acute respiratory syndrome, coronavirus DerivedFrom corona, coronavirus HasContext virology, coronavirus IsA species. It also helps you to get a proper understand of the concept of the research. Also using this information we will create a bag-of-concept model and predict missing subject or object, the relation between subject and object in specific scientific knowledge.

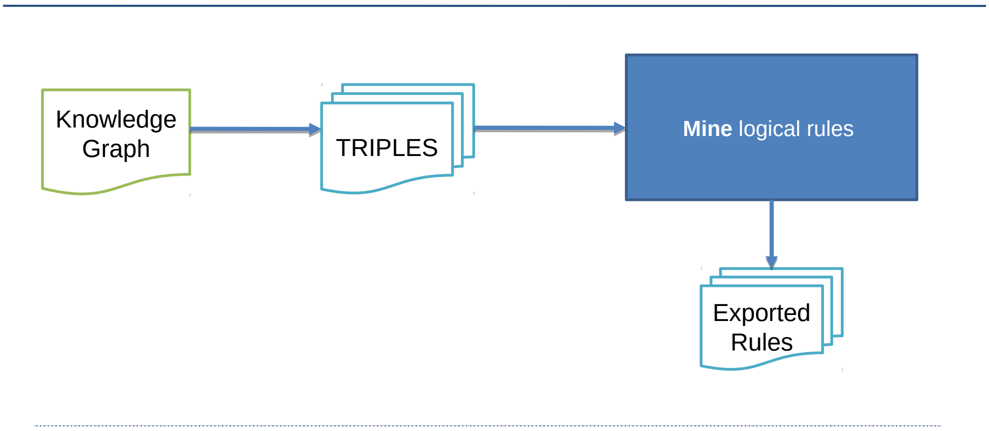

Fig. 1 Procedure for mining rules from KG.

- A potentially better and validated approach on working with heterogeneous, structured, and unstructured data sources. Example: Data sources such as different knowledge bases (such as ConceptNet, DBpedia) will be heterogeneous and structured but probably data sources such as various research documents will be heterogeneous and unstructured.
- A method or best practice on how to transform data and use them in the end-user application in a specific domain. Example: Concept to vector(C2V) will be an example for transform data from concept to vector and API will be helpful for various research document-based applications, such as understand the concept of a document.

This implies that by solving the societal context problem especially for the researcher and also full fill the research objectives for this research, there is a potential additional motivation to produce results that could be useful to the research community.

5 Research methodology

In our approach, we will adapt neural language models for scientific RDF graph embeddings. Such approaches take advantage of the word order in text documents, explicitly modeling the assumption that closer words in the word sequence are statistically more dependent. In the case of scientific RDF graphs, we consider entities and relations between entities instead of word sequences. Thus, in order to apply such approaches on RDF graph data, we first have to transform the graph data into sequences of entities, which can be considered as sentences. Using those sentences, we can train the same neural language models to represent each entity in the scientific RDF graph as a vector of numerical values in a latent feature space.

5.1 RDF Graph Sub-Structures Extraction

We propose a approach for converting graphs into a set of sequences of entities.

Definition 1 An RDF graph is a graph $G = (V, E)$, where V is a set of vertices, and E is a set of directed edges.

The objective of the conversion functions is for each vertex $v \in V$ to generate a set of sequences S_v, where the first token of each sequence $s \in S_v$ is the vertex v followed by a sequence of tokens, which might be edges, vertices, or any substructure extracted from the RDF graph, in an order that reflects the relations between the vertex v and the rest of the tokens, as well as among those tokens. In the first step of our approach is to convert the RDF graphs into a set of sequences. For each of the RDF scholarly datasets, we first build a corpora of sequences, i.e., set of sequences generated from rule base approach (marked as C2V).

Association Rule Mining Approach . Our target is to find all rules matching the specified minimum values of selected measures of significance. These are described in greater detail in the following. In order to mine logical rules from RdfStyle knowledge graphs a framework named ”RdfRules” [53] is used which has adopted and extended its rule mining algorithms from AMIE+ [14].

Rule Patterns . RdfRules allows the user to specify several rule patterns using a pre-defined grammar. All rules must match at least one pattern from the rule pattern list. Matching is performed during the mining phase and therefore the rules enumeration can be greatly sped up thanks to stricter pruning of the state space. Let us give an example to illustrate the of rules and its relevant features. Consider the following facts:

- $Coronavirus \Rightarrow causes \Rightarrow covid - 19$
- $Coronavirus \Rightarrow causes \Rightarrow flu$

Usually knowledge bases consist of the sorts of facts which can be quantified either by true or by false. In general, these facts have three basic parts: subject, predict and object. In both of the examples above ”coronavirus” is the subject, ”cause” is predicate whereas ”covid-19” and ”flu” are objects. From these facts we produce predicate named ”causes” which is shown below.

- $causes\ (coronavirus,covid - 19)$
- $causes\ (coronavirus,flu)$

These predicates are further used to make rules. Rules ultimately help to infer new facts. For example from the above mentioned predicate a rule ”isA” is made.

- $isA(X,Y) \Rightarrow causes(Z,X),\ causes(Z,Y)$

From this rule a new fact ”Covid-19 is a flu” can be produced. There can be a lot of patterns for rules. For our research purpose it is important to find the pattern of the rule.

Constraints Mining parameter specifies additional constraints and defines a way of mining. Here is a list of constraints that can be used:

- $OnlyPredicates(x)$: rules must contain only predicates defined in the set x.
- $WithoutPredicates(x)$: rules must not contain predicates defined in the set x.
- $WithInstances$: enable to mine rules with constants at the subject or object position.
- $WithObjectInstances$: enable to mine rules with constants only at the object position.
- $WithoutDuplicitPredicates$: rules that contain one predicate more than once will be removed.

The RuleSet object is on the output of the RdfRules workflow. It contains all discovered rules conforming to the restrictions.

Using machine learning methods for the knowledge graph is still a challenging task. The rule-based machine learning system gives interpretability, speed of the execution but the accuracy is still lowered than other machine learning techniques. Also, the rule base machine learning system creates a lot of complexity, when the system needs to handle a large number of of rules. On the other hand, a system like Neural Network, Deep Neural Network gives you more accuracy with less comprehensibility. But they do not provide you interpretability. Also, it takes a lot of time to train a model and execution. Also, it is difficult to get a lot of train data for problems like finding the missing triples or others in scientific knowledge. Also, it is a big problem for the train a model frequently with Neural Network or Deep Neural Network for the knowledge base because of computational resource and time complexity. For solving this problem, we will try to create training data depends on the true triples that we found from the rule base machine learning system, train a model and send the false triples for post evaluation. Depends on the post evolution result we will decide to add the rule base false triples in the knowledge base.

5.2 Neural Language Models –C2V

Neural language models have been developed in the NLP field as an alternative to represent texts as a bag of words, and hence, a binary feature vector, where each vector index represents one word. While such approaches are simple and robust, they suffer from several drawbacks, e.g., high dimensionality and severe data sparsity, which limits the performances of such techniques. To overcome such limitations, neural language models have been proposed, inducing low-dimensional, distributed embeddings of concepts by means of neural networks. The goal of such approaches is to estimate the likelihood of a specific sequence of concepts appearing in a corpus, explicitly modeling the assumption that closer concepts in the concepts sequence are statistically more dependent. While some of the initially proposed approaches suffered from inefficient training of the neural network models, with the recent advancements in the field several efficient approaches has been proposed. In this proposal, we also propose to implement C2V (concept to vector) language model for knowledge embedding which was never applied before in the RDF knowledge graph. One commonly used knowledge embedding approach is BOW which is a $n*n$ matrix of the words exist in the document. Whereas Bag-of-concept uses the entities of that document those are connected with other similar entities beyond the document. As a result a more robust matrix will be produced for bag of concepts which may increase probability for finding more relations between entities.The probability of its higher efficiency than BOW may be proved experimentally.

Bag-of-Concepts Model In this step, we already detect entities that may be relevant to the document. Here, we use a simple method based on an entity dictionary that maps an entity name (e.g., "coronavirus") to a set of possible referent entities (e.g., alphacoronavirus, severe acute respiratory syndrome, and RNA virus). In particular for the experiment, first, we use SciSpacy NER model, N-gram technique to find all the possible entities if they exist in the dictionary, and try to remove all possible referent entities for each detected entity name. Following past work, the boundary overlaps of the names are resolved by detecting only those that are the earliest and the longest. We use ConceptNet as the target Knowledge Base, and the entity dictionary is built by using the names and their referent entities of all internal anchor links in ConceptNet. We can also get two statistics from ConceptNet, namely link probability and commonness [25]. The former is the probability of a name being used as an anchor link in ConceptNet, whereas the latter is the probability of a name referring to an entity in ConceptNet. We generate a list of entities by concatenating all possible referent entities contained in the dictionary for each detected entity name and create a bag-of-concept (BoC) model presented in the next section. Note that we do not disambiguate entity names here, but detect all possible referent entities of the entity names.

After conceptualizing all the documents in the collection, we find the concepts vocabulary and their document frequencies. In bag-of-word (BoW) model, document frequency of a term t is the total number of documents containing t, as the term t either appears or not appears in a document (Boolean association). We discard the concepts whose do not exist at a minimun α time (a constant value depends on the dataset size) in the dataset, which means the concepts rarely appear in the documents and provide little information for document classification. Finally, we represent a document D_j as a distributed vector in the learned concept space: $d_j = (w1_j , w2_j , ... , wl_j)$, where l is the dimensionality of the concept space, and w_ij is the weight of concept c_i in document d_j. Compared with BoW representation, BoC has lower dimensionality due to the fact that concept space has much lower dimension than word or $n - gram$ space. Besides, BoC representation is less sparse than BoW representation, because the conceptualization process in BoC model maps words and phrases into diverse concepts in a probabilistic way instead the hard mapping of words as in BoW model. More importantly, BoC model is able to capture the semantic relatedness and conceptual information of words and phrases as well as higher level semantics of documents, which will benefit the document classification tasks.

We apply rule-based machine learning methods to find the combination of concepts that makes a document successful. We are using SBRL, CORELS, Random Forest algorithm for this purpose.

5.3 Classify important research papers or documents using OpenCitations ontology

For the improvement of the semantic scientific knowledge base, ontology is an impotent thing. Such as, we use OpenCitations ontology to find the impotent document in the scientific documents. We believe, if a scientific document gets more citations, the probability is so high the document is impact-full.

5.4 Develop a tool that can assist the scientific knowledge publishers

After enhancing the semantic scientific knowledge, we will introduce a novel scientific assist tool for the scientific knowledge publishers depends on the FCP method. The idea will deal with the newly formulated problem of estimating the categorization power of an ontology, expressed as the number of binary options offered for categorization of objects already assigned to a more general focus class. We will develop a formal basis for characterizing the focused categorization power (FCP) of an ontology, accounting for different description logic concept expression types and ontology patterns. Efficient algorithms for FCP calculation will be devised. The empirical analysis of large ontology collections will support both the construction and validation of the FCP models. Besides automatic analysis, feedback from users-ontologists will be exploited. The associated problem of transforming compound concept expressions to named classes.

6 Expected outcomes

The comprehensive type of this research is to plan for design research methods that focus on design and implement an efficient machine learning base knowledge graph completion method to the semantic scientific knowledge. The expected output of this research would be the following artifacts:

6.1 Knowledge artifacts

Which is represented by introducing a concept called bag-of-concepts and a machine learning method for enhancement knowledge of the semantic scientific knowledge.

6.2 Data artifacts

Which include information about research articles, citations, the impact of the research, research output, etc. Also, its represented by the enhancement of the semantic scientific knowledge (KGs) that includes find out the missing triples, ontologies, vocabularies, etc.

6.3 Software artifacts

Which include API and a tool that can assist the scientific knowledge base and publishers.

To validate these artifacts, the author looks forward to the use of qualitative methods after the development phase. The application would be tested thoroughly using other available data sets. Case studies will be carried out to verify and assess any discrepancies. It will be considered successful if it would successfully add huge information in the semantic scientific knowledge.

7 Current status and achievements

The current status and results of the dissertation progress after two semesters of study are the following facts.

The author has successfully submitted one research publication as a co-author about the project to a prestigious conference called "International Conference on Knowledge Engineering and Knowledge Management". This research paper is now under review process. Also one of the research papers called "Entity-based document classification" now in the final draft stage for submission. The author's target is to submit it in a prestigious journal called "Cognitive Computation" (Q1 journal, Impact factor: 4.29).

As for academic duties, the author currently working (Delay for the COVID-19 crisis) with three out of four subjects in his study program. Within the scope of working with research papers, the author has actively participated in some webinars (Such as CORD 19 Semantic Annotation Projects, Coronawhy) and present research works that can help the front-liner those who are fighting with COVID-19.

Furthermore, since March 2020, the author has started his first IGA project as a principal investigator called "Knowledge Engineering of Researcher Data (KNERD)" in which he aims to analyze the worldwide most renowned research effort in knowledge engineering for research data. The topic of this project is closely relevant to the dissertation since it focuses on semantic scientific knowledge. The author aims to publish two conference publications and one journal published by the end of this project.

The author is also working on the COVID-19 document classification task that depends on the entities which are also heavily related to the topic of his Ph.D. and IGA project. For the analysis and development of this work, the author is currently analyzing various well-known named-entity recognition and machine learning techniques. The author also aims to use this knowledge for knowledge graph completion on scientific knowledge KGs to find out the applicable semantic enrichment resources, such as ontologies and also to develop tools assisting the scientific knowledge KG publishers. The author expects to cooperate with his Ph.D. advisor and consultant in the development and validation of this project.

The author is considering collaboration options in the field of semantic scientific knowledge with experts such as Prof. Dr. Sören Auer from Leibniz Universität Hannover, Prof. Dr. Frank van Harmelen from Vrije Universiteit Amsterdam. In the scope of the author's internship (Open Research Knowledge Graph (ORKG) lab, German National Library of Science and Technology (TIB)), he plans to collaborate with local Semantic Web and Knowledge Graph experts and consult with them regarding the machine learning techniques in the semantic scientific knowledge. For the technical development of machine learning on semantic scientific knowledge, the author plans to consult with other engineering experts at his local working place such as Dr. Ondřej Zamazal.

Also, In March 2021, the author has the plan to submit his second IGA project (two years based) as a principal investigator called "Knowledge Base Enhancement of Researcher Data (KBERD)" in which he aims to analyze the worldwide most renowned scientific knowledge graph. The topic of this project is closely relevant to the dissertation since it focuses on the enhancement of semantic scientific knowledge. The author aims to publish two conference publications and two journals published by the end of that project.

Additionally, the author plans to collaborate with the semantic scientific collaboration in the recommendation system in UNICO.AI from September 2020. The author believes he can learn and enhance his knowledge and expertise in his research area. The author is planning to submit a technical research paper about the scientific collaboration in the recommendation system in UNICO.AI to any prestigious conferences.

8 Planning

This section shows the time plan for the rest of the author's study program until Summer, 2022.

For the rest of the 2nd semester and during the 3rd semester, the author plans to do a more comprehensive literature review on machine learning on semantic scientific knowledge and causality which proves to be another significant aspect of the project's objectives.

Requirements analysis and qualitative analysis should be carried out during a part of the third semester. The author will begin collecting data during this time as well. After selecting research related KG, the author will begin to model and apply various machine learning techniques. This activity will last until the third semester, during which he probably (depends on the COVID-19 crisis) doing an internship at a foreign university, where the potential choice is Leibniz Universität Hannover in Germany. During this internship, the author will intensively gather more data from the research group's developed KGs and other sources. The requirement analysis activity will consist of several steps listed below.

– The author will gather requirements from the various scientific KGs such as Open Research Knowledge Graph (ORKG) or Elsevier knowledge graphs by collecting KGs from the hosts. The collection techniques include collecting continuous development of KGs, brainstorming for the applicable ML techniques in the KGs and observation.

– The author will determine the quality of the gathered requirements to resolve any ambiguity and contradictions in the KGs.

– The author will document the gathered requirements using modeling languages and frameworks such as flow charts, use case diagrams in order to further communicate with target users.

The author will eventually begin to apply the machine learning techniques on semantic scientific knowledge, develop its API, and deploy in a server while continuously integrating and analyzing data he gathered previously. This process will take place during the 4th and 5th semesters including testing and evaluation.

Also, a more comprehensive evaluation will take place during the 5th semester after which the author makes a plan for the shadow defense of this thesis. The thesis write-up will be carried out during this semester until the 6th semester where it all ends with the thesis defense.

A visual chart for planning out the entire remaining study period is mapped out as Gantt's diagram accessible from this link: https://github.com/corei5/Gantt-s-diagram-PhD-

9 Conclusion

In this document, the author introduced the context problem to his dissertation and provided a descriptive insight into the proposed project. A literature review was conducted to explore the state of the art, revealing several projects and resources that are relevant to the dissertation project. Its goals and objectives were also described in the addressed research problem. Anticipated practical and scientific contributions were addressed as well, as they should be the main reason for conducting this research. Several working assumptions were displayed to address possible challenges during research. Basic research methods were introduced to tackle challenges and provide a cornerstone for the research. Next, the author described the anticipated result of the dissertation project and the classification of its composition based on design research. The author then mentioned the current status of his project and several preliminary up-to-date results. Last but not least, the author presented his time plan for the next 2 years of his study program to provide a guideline to ensure the best possible progression of his dissertation.

References

1. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al.: Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262 (2018)
2. Aprosio, A.P., Giuliano, C., Lavelli, A.: Extending the coverage of dbpedia properties using distant supervision over wikipedia. In: NLP-DBPEDIA@ ISWC (2013)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)
4. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Lrec, vol. 10, pp. 2200–2204 (2010)
5. Brack, A., Hoppe, A., Stocker, M., Auer, S., Ewerth, R.: Requirements analysis for an open research knowledge graph. arXiv preprint arXiv:2005.10334 (2020)
6. Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., Schindler, U.: The scholix framework for interoperability in data-literature information exchange. D-Lib Magazine **23**(1/2) (2017)
7. Chirita, P.A., Nejdl, W., Paiu, R., Kohlschütter, C.: Using odp metadata to personalize search. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 178–185 (2005)
8. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Ontology matching: A machine learning approach. In: Handbook on ontologies, pp. 385–403. Springer (2004)
9. Dutta, A., Meilicke, C., Stuckenschmidt, H.: Enriching structured knowledge with open information. In: Proceedings of the 24th international conference on world wide web, pp. 267–277 (2015)
10. Fabian, M., Gjergji, K., Gerhard, W., et al.: Yago: A core of semantic knowledge unifying wordnet and wikipedia. In: 16th international world wide web conference, www, pp. 697–706 (2007)
11. Färber, M.: The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data. In: International Semantic Web Conference, pp. 113–129. Springer (2019)
12. Fellbaum, C.: Wordnet. The encyclopedia of applied linguistics (2012)
13. Fukunaga, K., Narendra, P.M.: A branch and bound algorithm for computing k-nearest neighbors. IEEE transactions on computers **100**(7), 750–753 (1975)
14. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with amie+. The VLDB Journal **24**(6), 707–730 (2015)
15. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: Amie: association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the 22nd international conference on World Wide Web, pp. 413–422 (2013)
16. Gangemi, A., Nuzzolese, A.G., Presutti, V., Draicchio, F., Musetti, A., Ciancarini, P.: Automatic typing of dbpedia entities. In: International semantic web conference, pp. 65–81. Springer (2012)
17. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics-Volume 2, pp. 539–545. Association for Computational Linguistics (1992)
18. Jaradeh, M.Y., Oelen, A., Prinz, M., Stocker, M., Auer, S.: Open research knowledge graph: A system walkthrough. In: International Conference on Theory and Practice of Digital Libraries, pp. 348–351. Springer (2019)
19. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
20. Kliegr, T.: Linked hypernyms: Enriching dbpedia with targeted hypernym discovery. Journal of Web Semantics **31**, 59–69 (2015)
21. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
22. Lange, D., Böhm, C., Naumann, F.: Extracting structured information from wikipedia articles to populate infoboxes. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1661–1664 (2010)
23. Laparra, E., Rigau, G.: Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm. In: Proceedings of the International Conference RANLP-2009, pp. 208–213 (2009)
24. Liaw, A., Wiener, M., et al.: Classification and regression by randomforest. R news **2**(3), 18–22 (2002)
25. Liu, H., Singh, P.: Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal **22**(4), 211–226 (2004)
26. Manghi, P., Manola, N., Horstmann, W., Peters, D.: An infrastructure for managing ec funded research output-the openaire project. The Grey Journal (TGJ): An International Journal on Grey Literature **6**(1) (2010)
27. McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in translation: Contextualized word vectors. In: Advances in Neural Information Processing Systems, pp. 6294–6305 (2017)
28. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management, pp. 509–518 (2008)
29. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pp. 1003–1011. Association for Computational Linguistics (2009)
30. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)

31. Nguyen, P.T., Tomeo, P., Di Noia, T., Di Sciascio, E.: Content-based recommendations via dbpedia and freebase: a case study in the music domain. In: International Semantic Web Conference, pp. 605–621. Springer (2015)
32. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. Proceedings of the IEEE **104**(1), 11–33 (2015)
33. Nickel, M., Tresp, V., Kriegel, H.P.: Factorizing yago: scalable machine learning for linked data. In: Proceedings of the 21st international conference on World Wide Web, pp. 271–280 (2012)
34. Ohana, B., Tierney, B.: Sentiment classification of reviews using sentiwordnet. In: 9th. it & t conference, vol. 13, pp. 18–30 (2009)
35. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web **8**(3), 489–508 (2017)
36. Peroni, S., Shotton, D.: Fabio and cito: ontologies for describing bibliographic resources and citations. Journal of Web Semantics **17**, 33–43 (2012)
37. Peroni, S., Shotton, D.: The spar ontologies. In: International Semantic Web Conference, pp. 119–136. Springer (2018)
38. Peroni, S., Shotton, D., Vitali, F.: One year of the opencitations corpus. In: International Semantic Web Conference, pp. 184–192. Springer (2017)
39. Poster, C.: Being and becoming: Rhetorical ontology in early greek thought. Philosophy & rhetoric pp. 1–14 (1996)
40. Presutti, V., Consoli, S., Nuzzolese, A.G., Recupero, D.R., Gangemi, A., Bannour, I., Zargayouna, H.: Uncovering the semantics of wikipedia pagelinks. In: International Conference on Knowledge Engineering and Knowledge Management, pp. 413–428. Springer (2014)
41. Quan, T.T., Hui, S.C., Fong, A.C.M., Cao, T.H.: Automatic generation of ontology for scholarly semantic web. In: International Semantic Web Conference, pp. 726–740. Springer (2004)
42. Ringler, D., Paulheim, H.: One knowledge graph to rule them all? analyzing the differences between dbpedia, yago, wikidata & co. In: Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), pp. 366–372. Springer (2017)
43. Ristoski, P., Paulheim, H.: Rdf2vec: Rdf graph embeddings for data mining. In: International Semantic Web Conference, pp. 498–514. Springer (2016)
44. Ruiz-Iniesta, A., Corcho, O.: A review of ontologies for describing scholarly and scientific documents. In: SePublica (2014)
45. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: International Semantic Web Conference, pp. 187–205. Springer (2018)
46. Schuler, K.K.: Verbnet: A broad-coverage, comprehensive verb lexicon (2005)
47. Schwarz, B.B., Glassner, A.: The role of floor control and of ontology in argumentative activities with discussion-based tools. International Journal of Computer-Supported Collaborative Learning **2**(4), 449–478 (2007)
48. Shapiro, S.: Philosophy of mathematics: Structure and ontology. Oxford University Press on Demand (1997)
49. Soldatova, L.N., King, R.D.: An ontology of scientific experiments. Journal of the Royal Society Interface **3**(11), 795–803 (2006)
50. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3**(3), 1–13 (2007)
51. Watanabe, Y., Asahara, M., Matsumoto, Y.: A graph-based approach to named entity categorization in wikipedia using conditional random fields. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 649–657 (2007)
52. Yaman, B., Pasin, M., Freudenberg, M.: Interlinking scigraph and dbpedia datasets using link discovery and named entity recognition techniques. In: 2nd Conference on Language, Data and Knowledge (LDK 2019). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2019)
53. Zeman, V., Kliegr, T., Svátek, V.: Rdfrules preview: Towards an analytics engine for rule mining in rdf knowledge graphs. In: RuleML+ RR (Supplement) (2018)