Learning Concepts and Semantic Contexts from Domain Text Corpora and Terminologies

Svitlana Moiseyenko [0000-0002-7300-8818]
svitlana.moiseyenko@gmail.com

Abstract. This paper outlines a PhD proposal focused on the generation of semantic contexts, as knowledge graph (KG) fragments around the concepts within a subject domain. These contexts are regarded as formalized definitions of the properties of these concepts to be further fused in one graph – a domain ontology. This domain ontology could be used as a namespace for completing knowledge graphs for the domain. The data for these knowledge graphs is expected to be taken from natural language texts describing the domain, like scholarly publications. The envisioned approach incorporates the novel methods of concept or property identification and relationship extraction. The input data for these methods is a representative domain-bounded document collection and the terminology extracted from this collection. The objective is to provide a solution that enhances knowledge extraction and integrates seamlessly with existing approaches and solutions. This goal is set to be achieved using the State-of-the-Art (SotA) Deep Learning frameworks. The proposed work is deemed to contribute to the provision of instrumental software for semi-automatically developing and completing Scientific KGs.

Keywords: Relation Extraction, Concept Extraction, Semantic Context, Ontology Engineering, Knowledge Graph, Deep Learning.

1 Introduction

This paper presents the motivation for, and vision of a Ph.D. project focused on building the semantic contexts of the concepts for a domain or topic ontology. This proposed work is inspired by [1] and based on the previous research [2, 3] by the author. It is envisioned that these contexts will be built in a partially automated manner with a human expert in the loop. The role of the human expert is to validate, revise, and refine the fragments of a domain ontology suggested by the tool to be developed in the proposed work.

Building semantic contexts of the concepts for an ontology is, in fact, the completion of concept models in the form of the sub-graphs of the schema of the developed ontology. Hence, several techniques from ontology learning and knowledge discovery are relevant: concept or property identification, taxonomy extraction, relationship extraction. In the proposed work, we focus on: (i) putting all the relevant SotA component methods together to form a holistic baseline framework for concept sub-graph discovery; and (ii) developing deep learning models to enhance the efficacy of the component methods in domain-neutral settings. (iii) conduct rigorous evaluations of our integrated

system and the deep learning models, comparing their performance against baseline solutions.

The remainder of this paper is structured as follows: Section 2 reviews the current state of research and examines relevant articles in the field. Section 3 defines the problem, explaining the motivation to find a solution. Then introduce of envisioned approach to solution and methodology provided in Section 4, detailing the steps taken to address the problem. Finally, Section 5 provides a conclusion.

2 Related Work and Open Issues

Numerous techniques have been used to pull out taxonomies, or classifications, from various sources of knowledge. The section is structured to cover: (i) baseline and DL-based approaches; and (ii) component methods.

2.1 Baseline and DL-based Approaches

Historically, the methods for extracting concepts and semantic contexts from text have evolved as follows. Initially, pattern-based methods had been proposed. Then, hybrid methods have been introduced in reply to the need for improved quality. Unsatisfied with the existing limitations, the field began involving machine learning approaches. Recently, advanced Deep Learning approaches came into play.

Pattern-based approaches. The first approaches for concept and relationship extraction were pattern-based. Initial studies [4, 5] developed specific patterns; this specificity was the limit, however. It restricted their application to covering all possible situations that could be met in natural language texts. This direction of work has been further continued in [6, 7, 8].

Hybrid Approaches. To improve precision, linguistic and statistical methods were bundled together [9]. An approach that combined Natural Language Processing (NLP) and pattern-based techniques was proposed, leading to the use of lexico-syntactic patterns (LSP) for matching concepts and relationships [10].

Machine Learning Approaches. Extraction methods were classified into unsupervised, semi-supervised, or supervised techniques. Unsupervised methods, like the feature grouping technique [11], did not require domain-specific training sets or expert rules. Supervised methods, like dynamic clustering [12], were proposed to identify taxonomic relationships and understand term relationships. Word2Vec [13] used neural networks for understanding word relationships. Semi-supervised methods were used to extract concepts and relationships from web pages [14, 5].

Deep Learning-based Approaches. These methods further improved the performance of machine learning-based approaches by increasing the precision of concept and relationship extraction in NLP. Models such as GloVe [15] aggregated word co-occurrence statistics for robust word representation. SensEmbed [16] introduced sense-disambiguated word embeddings, furthering advancements in the field. BERT [17] and GPT [18], based on the transformer architecture [19], achieved SotA performance on a wide range of NLP tasks. In difference to BERT and GPT, LLaMA [20] is a

fully open framework. It was based on the use of a collection of foundation language models ranging from 7B to 65B parameters and was trained on trillions of tokens.

2.2 Component Methods

Concept identification and relationship extraction are complex tasks that require several operations to be performed in a pipeline. Hence, the related work also focused on the development of the component methods for these operations. The key components and corresponding related publications are as follows.

Concept or Property Identification. One of the mainstream approaches [21] was based on annotation, identifying potential domain-independent concepts, and using an annotated corpus for baseline classification experiments and training a domain-neutral classifier.

Taxonomy Extraction. A widely used supervised approach [12] was based on dynamic clustering to find taxonomic relationships.

Relationship Extraction. Relationship extraction is often enabled by using additional knowledge from terms extracted from texts or text collections [22, 7, 11]. Several approaches also exploited post-processing methods to boost accuracy, such as statistics-based cuts [23] or domain significance scores [1, 24].

Framework Consolidation. The Plumber framework [25] consolidates many third-party research and development efforts regarding scientific KG completion into a unified and tunable pipeline. It incorporates circa 40 reusable components for various KG completion subtasks and dynamically constructs the most suitable pipeline for completing a KG in a specific domain.

2.3 Open Issues in State of the Art

The analysis of existing approaches and methods for the possible re-use in the concepts extraction and semantic contexts construction revealed several potential research gaps that motivate the envisioned research project.

Pragmatics. Existing SotA approaches, including LLM, often **lack context sensitivity**, meaning they might not fully consider the specific context in which a term is used. As a result, they may misidentify terms or fail to capture their correct meaning in the given context.

Narrow scope. The specificity of the reviewed **pattern-based methods** is that these methods are too specific. Hence, their scope is too limited for covering all the required aspects in natural language texts or different domains with sufficient quality.

Quality of relationship discovery. The precision/recall balance of relationship extraction remains a significant issue, with current machine learning approaches often failing to accurately identify all possible relationships. The problem is more complex in scientific domains described by complex texts using domain-specific terminologies. One more complication is the presence of alternative descriptions for the same concepts, which adds pragmatic and scholastic ambiguity on the top of domain specifics.

Despite notable advancements in semantic context creation and concept extraction, significant challenges remain in the field in terms of context adaptability and relationship extraction performance.

3 Motivation

Creating relevant contexts across different domains is indeed a considerable challenge. This encourages to delve deeper into this area of study. After analyzing the open issues, two major problems have come to the fore.

Term Definition Identification and Understanding. Existing methods for term identification and understanding across domains mainly involve lexicographic approaches and domain-specific dictionaries. However, these might only partially encapsulate the richness and ambiguity of natural language, particularly in specialized scholarly fields. New methodologies could harness the power of machine learning and NLP to create more dynamic and context-sensitive models of term identification and understanding.

Extracting Relationships from Term Definitions. There remains a notable research gap in the quality of relationship extraction, that is often carried out through rule-based systems or manually curated knowledge graphs. These approaches have limitations, including scalability and adaptability to new, unseen relationships. The open issue could be addressed by developing automated, machine learning-powered relationship extraction systems. These systems could learn to identify and categorize relationships from term definitions by being trained on representative annotated datasets. A key to this will be the creation of rich and representatively complete training data that captures a wide array of relationships in varied contexts. Additionally, attention should be paid to creating robust evaluation metrics that capture the system's ability to recognize and understand relationships.

By effectively tackling these problems, we would be on the track to devise solutions for the outlined open issues. The innovations derived from this research will not only enrich our understanding of term definitions and their relationships but also pave the way for more sophisticated tools for knowledge extraction, KG completion, data curation, and scientific knowledge dissemination.

4 Envisioned Approach to Solution and Methodology

As a result of this research work, we anticipate having a solution that is derived from the extraction of terms [26]. This solution will enable us to identify term definitions and extract relationships from them, which will be used to build learning concepts and semantic contexts.

4.1 Approach to Solution

The proposed **approach to solution** aims to provide a systematic method based on the insights derived from our previous work [3]. This background has, however, to be refined to encompass the following steps.

Term Definition Identification and Understanding is seen as an iterative process including the following steps:

- **Extracted Terms Ingestion**. The terms have been derived from a domain-specific collection, a result of the Saturated Terminology Extraction and Analysis [27]. These terms will form the primary input for the pipeline to process.
- **Term Definition Embedding**. This step is aimed for enriching the terms that have been extracted from the text collection with their respective definitions. The objective here is to identify whether a piece of text serves as a definition for a term. The text collections previously used for term extraction will serve as the primary source for obtaining term definitions. Given the complexity of natural language, a simple pattern matching, or keyword-search approach may not be sufficient. This is why more sophisticated techniques, that can handle various linguistic nuances, need to be incorporated.

Extracting Relationships from Term Definitions. Building a relationship extraction system involves applying part-of-speech tagging and dependency parsing to understand sentence structure, using pattern matching to identify potential relationships, extracting relevant features from the text, and finally training a classifier to accurately identify and categorize these relationships.

POS Tagging and Dependency Parsing. Using a POS tagger (e.g. NLTK[1], SpaCy[2], Stanford NLP[3]) to tag each token in your text with its corresponding part of speech and the relationships between words. This will label each token as a noun, verb, or adjective, which can be valuable information when identifying relationships.
Relations Identification: To identify relationships, we can apply 2 approaches:

- Applying the LLM Technique to convert words or phrases into vectors (embeddings). These embeddings can capture the semantic meaning of the words to identify relationships.
- Pattern Matching: Using Lexico-Syntactic Patterns (based on the approach of Hearst [4] patterns) to extract relationships from sentences (e.g: "is-a" will be identified as hypernymy).

Feature Extraction. From the above steps, extract features that would be proper for building a relation extraction classifier. These might include POS tags, parse tree structures, word embeddings, etc.

[1] https://www.nltk.org/
[2] https://spacy.io/
[3] https://nlp.stanford.edu/

Training a Classifier. Using POS tags to train a machine learning model can improve accuracy in identifying relationships in text. Different models like decision tree, logistic regression, SVM, random forest, or neural network can be used.

4.2 Methodology

The envisioned methodology for the proposed research needs to comprise the following key stages.

Background Examination and Candidate Selection. In this initial stage, we will focus on experimental study for existing solutions, selecting the most capable model candidate addressing the outlined problems in our research. This selection will be based on various factors such as the model's accuracy, computational efficiency, and compatibility with our problem domain. Specifically, Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM) networks, or Transformer-based models could be a good choice, given their proven effectiveness in handling sequential data like text.

Developing the Foreground. Upon selection, the chosen candidate will be integrated as a component of our solution. This phase includes the preparation of data, the training of the model, and the tuning of hyperparameters. Simultaneously, we will assess how well the implemented model aligns with our research objectives. Additionally, this step will also encompass the identification and implementation of any additional components necessary to improve the performance and fit of the model for our specific requirements.

The **Evaluation process** is composed of the two distinct phases.

(i) **Evaluation of Improvement.** In this phase, we will evaluate how much our solution (using the chosen and implemented model) improves upon the existing SotA. The envisioned evaluation aspects include accuracy, precision, recall, F1 score, computational efficiency, and general usability.

(ii) **Iterative Evaluation.** It is planned consistently Evaluation each step of our research and check if the outcome improves upon the background state. This continuous evaluation will ensure the progress and effectiveness of our research.

4.3 Evaluation

The evaluation process will initially involve a comparative analysis. It will be divided into distinct phases, each serving a unique purpose. These phases include:

(i) **Selection Phase Evaluation.** At the selection phase, each candidate model or solution will be evaluated regarding their fit for purpose and quality. This will result in selecting the best fitting 3-d party component for our background.

(ii) **Interim Assessment of the Progress.** Evaluations will be performed for partial solutions throughout the iterations of the research following the Scientific Method [28]. These iterative evaluations will facilitate to monitoring the progress, detect potential issues promptly, and guide necessary optimizations to refine our path toward the final solution.

(iii) **Final Evaluation**: At the major project milestones, a more comprehensive evaluation will be conducted for a major version of the solution. This assessment aims to gauge the efficacy of our solution, check the fit with the research objectives and cross-evaluate against the competing SotA solutions. It is also expected that the insights into potential improvements for future implementation will emerge after these evaluation iterations.

Together, these phases will form a structured evaluation program that facilitates informed decision-making, monitors ongoing progress, and maximizes the efficacy of our envisioned solution.

5 Conclusion

In conclusion, the envisioned methodology and approach to solution aim to solve the current problems of: (i) term definition identification and (ii) relationship extraction for constructing the semantic contexts of discovered concepts. Notably, the solution could also interlock effectively with existing systems, such as PLUMBER [25], and augment their functionality by seamlessly integrating with its array of components. Ultimately, the method, that will be researched and developed, aims to achieve higher precision of term definition identification and relationship extraction.

References

1. Ermolayev, V.: OntoElecting requirements for domain ontologies. The case of time domain. EMISA Int J of Conceptual Modeling 13(Sp. Issue), 86–109 (2018). doi: 10.18417/emisa.si.hcm.9
2. Moiseyenko, S., Ermolayev, V.: Conceptualizing and formalizing requirements for ontology engineering. In: Antoniou, G., Zholtkevych, G. (eds.) PhD Symposium at ICTERI 2018, CEUR-WS, vol. 2122, pp. 35–44 (2018).
3. Moiseyenko, S., Vasileyko, A., Ermolayev, V.: Building a feature taxonomy of the terms extracted from a text collection. In: Proc. MS-AMLV 2019, CEUR-WS vol. 2566, 59–70 (2020)
4. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: 14th Conf on Computational Linguistics, pp. 539–545 (1992)
5. Kozareva, Z., Hov, E.: A semi-supervised method to learn and construct taxonomies using the web. In: Proc. 2010 Conf on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 1110–1118, MIT, Massachusetts, USA (2010)
6. Ritter, A., Soderland, S., Etzioni, O.: What is this, anyway: automatic hypernym discovery. In: AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read, pp. 88–93 (2009)
7. Tuan, L.A., Kim, J., Ng, S.K.: Taxonomy construction using syntactic contextual evidence. In: 2014 Conf on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 810–819 (2014)
8. Snow, R., Jurafsky, D., Ng. A.: Learning syntactic patterns for automatic hypernym discovery. In: 17th Annual Conf on Neural Information Processing Systems, pp. 1297–1304 (2005)
9. Wu, W., Li, H., Wang, H.,Zhu, K.O.: Probase: a probabilistic taxonomy for text understanding. In: ACM SIGMOD Int Conf on Management of Data, pp. 481–492 (2012)

10. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: a look back and into the future. ACM Comput. Surv. 44(4), 20:1–20:36 (2012)
11. Kosa, V., Chaves-Fraga, D., Keberle, N., Birukou, A.: Similar terms grouping yields faster terminological saturation. In: Ermolayev, V. et al. (eds.) ICTERI 2018. Revised Selected Papers. CCIS, vol. 1007, pp. 43–70 (2019). doi: 10.1007/978-3-030-13929-2_3
12. Yamane, J., Takatani, T., Yamada, H., Miwa, M., Sasaki, Y.: Distributional hypernym generation by jointly learning clusters and projections. In: 26th Int Conf on Computational Linguistics, pp. 1871–1879 (2016)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: 27th Annual Conf on Neural Information Processing Systems, pp. 3111–3119 (2013)
14. Carlson, A., Betteridge, J., Wang, R.C., Hruschka Jr., E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: 3d Int Conf on Web Search and Web Data Mining, pp. 101–110 (2010)
15. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: 2014 Conf on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 1532–1543 (2014)
16. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: learning sense embeddings for word and relational similarity. In: 53d Annual Meeting of the Association for Computational Linguistics and 7th Int Joint Conf on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 95–105 (2015)
17. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805 (2019)
18. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. (2018)
19. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017) Attention is All You Need. In: Advances in Neural Information Processing Systems.
20. Touvron, H., Lavril, T., Izacard, G., et al.: LLaMA: Open and Efficient Foundation Language Models. ArXiv, abs/2302.13971 (2023)
21. Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R.: Domain-Independent Extraction of Scientific Concepts from Research Articles. In: Jose, J.M. et al. (eds.) Advances in Information Retrieval. ECIR 2020. LNCS, vol. 12035. Springer, Cham. doi: 10.1007/978-3-030-45439-5_17 (2020)
22. Yang, H.: Constructing task-specific taxonomies for document collection browsing. In: 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1278–1289 (2012)
23. de Knijff, J., Frasincar, F., Hogenboom, F.: Domain taxonomy learning from text: the subsumption method versus hierarchical clustering. Data & Knowledge Engineering 83, 54–69 (2013)
24. Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W.-E.: Quantifying ontology fitness in OntoElect using saturation- and vote-based metrics. In: Ermolayev, V., et al. (eds.) Revised Selected Papers of ICTERI 2013, CCIS, vol. 412, pp. 136–162 (2013). doi: 10.1007/978-3-319-03998-5_8
25. Jaradeh, M.Y., Singh, K., Stocker, M., Roth, A., Auer, S.: Information extraction pipelines for knowledge graphs. *Knowledge and Information Systems* 65: 1989–2016 (2023) doi: 10.1007/s10115-022-01826-x
26. Kosa, V., Ermolayev, V.: Terminology Saturation: Detection, Measurement, and Use. Cognitive Science and Technology, Springer Singapore (2022)

27. Kosa, V., Ermolayev, V.: Saturated Terminology Extraction and Analysis in Use. In: Terminology Saturation: Detection, Measurement, and Use. Cognitive Science and Technology, Springer Singapore, p. 155--170 (2022)
28. Gower, B.: Scientific Method. A Historical and Philosophical Introduction. 1st Ed. Taylor & Francis (1996) DOI: 10.4324/9780203046128