Lib2Life – Domain Categorization of Books using BERT Language Models and Knowledge Graph Population

Melania Nițu, Mihai Dascălu, Gabriel Guțu-Robu, Maria-Iuliana Dascălu
University Politehnica of Bucharest
Splaiul Independentei 313, 060042, Bucharest, Romania
{suzana_melania.nitu, mihai.dascalu, gabriel.gutu, maria.dascalu}@upb.ro

Silvia Tomescu
Carol I Central University Library Bucharest
Boteanu 1, 010027, Bucharest, Romania
silvia.tomescu@bcub.ro

ABSTRACT
Ontologies provide domain conceptualizations for storing knowledge in a unified and structured machine-readable format. Following the initial conceptualization of the domain centered on defining classes and corresponding properties, the process of properly instantiating and keeping up-to-date individuals can become tedious. Thus, the need for knowledge processing systems that automatically populate and update their knowledge representations increases. This paper introduces an approach to manage, organize, and automatically populate the Lib2Life ontology containing historical documents provided by the Central University Libraries in Romania. Our method integrated in the Lib2Life platform instantiates individuals in the RDF knowledge graph following book categories that are inferred using state-of-the-art language models. As such, new documents to be included in the Lib2Life platform are automatically categorized with an F1-score of .73 using a fine-tuned Romanian BERT model.

Author Keywords
Digital library; Language models; Domain classification; Ontology population.

ACM Classification Keywords
I.2.7. Computing Methodologies: Artificial intelligence: Natural Language Processing: Text analysis.

I.7.1 Computing Methodologies: Document and Text Processing: Document and Text Editing: Document management.

General Terms
Design, Experimentation

DOI: 10.37789/rochi.2021.1.1.27

INTRODUCTION

In the information era, Library Management Systems (LMS) must provide intelligent platforms integrating information retrieval techniques for dealing with a massive number of documents. Lib2Life [14] is a smart digital platform that integrates a semantic search engine to provide access to old documents, such as books, manuscripts, or newspapers that are no longer under copyright protection. Its main goal is to increase the audience for these types of documents and to preserve the national heritage of Romania. The collection of documents hosted in the digital library is described using metadata to facilitate access to data, proper storage, and retrieval of relevant documents. The Lib2Life ontology is modeled as an RDF knowledge graph which defines the domains of the documents included in the scope of the platform [7].

This paper addresses two previously identified issues in Lib2Life. The first one refers to the manual categorization of each document introduced in the platform, while the second one considers the tedious task of manually populating the ontology with newly introduced documents. As such, we introduce a mechanism to insert, update, and populate the ontology based on the automated categorization of books into their corresponding domain.

In terms of structure, the second section presents a state-of-the-art on conventional and modern approaches to ontology population methods. The third section proposes a solution based on language models to categorize documents and afterwards instantiate corresponding individuals into the Lib2Life ontology. Afterwards, results and conclusions of the research are presented.

RELATED WORK

A similar ontology system for educational notions was developed by Amarnath et al. [1] on top of a digital library. The system compares various curricula and provides basic semantic search capabilities, as well as comparisons of multiple curricula using partial graph matching.

More recent studies leverage deep learning techniques to populate ontologies. Such an example is the framework proposed by Su et al. [16] that uses deep neural networks for triplets (subject, predicate, object) extraction. The system is structured in two modules: term extraction model using Bidirectional LSTM, and a multilayer perceptron-based extraction model. The approach used character-level embedding methods for Chinese language. Extracted terms were then used for ontology population. The accuracy of

their method for populating triplets was under 75% for the best performing model.

Bravo et al. [2] present an ontology population method for scientific publications and implement semantic similarity measures between publications to correlate the article title with an ACM classification category.

A different approach, implemented in the MIREL project [15], uses named entity recognition and classification for curriculum learning and ontology population. The method connects legal texts to ontology concepts and instances.

An extended research on semantic pattern-based ontology population was proposed by Maynard et al. [12; 13]. Their system considers the terms as contextual information, thus building the terms hierarchy using the Hieron classification algorithm [4], which is essentially a framework for supervised classification where labels are hierarchically structured, encoded by a rooted tree. In their study, lexico-syntactic, lexico-semantic, and named contextual patterns were used to identify new classes and extract relations between them.

Cruz et al. [3] show an innovative method to populate ontologies from XML data. The method defines mapping rules for properties conversion between OWL and XML schemas. Data population is performed by searching for new elements in the RDF mapping graph. The process requires manual definition of mapping rules, while the XML data extraction is the only automatized component. Ontology population is a modular activity, which is manually performed in a dedicated step and used in follow-up experiments.

Ontologies are widely used for semantic representation and various methods were proposed to automate the process of ontology population. In the current research, we propose a different approach for populating the Lib2Life ontology using a pre-trained language model to extract and classify documents into knowledge domains, followed by the update of corresponding individuals into the Lib2Life knowledge graph.

METHOD

Our solution consists of two modules: a) domain prediction of documents using a BERT language model [5], followed by b) an automated ontology population based on the predicted domains. The workflow consists of extracting document metadata, which represents a prerequisite in feeding labeled data to RoBERT model [11]. Then, the predicted domains are mapped to the ontology knowledge domains and are automatically inserted in the ontology as class individuals.

Corpus

Our corpus of 844 Romanian documents was provided by the Central University Libraries in Romania and represents a collection of scanned and OCRed historical documents and books, mostly dated in the 19th or the 20th century. A visualization of documents' distribution based on region at writing time is shown in Figure 37. Most writings originate from Bucharest and Iași. Each document contains corresponding metadata, such as book title, author(s), description, and domain (see Table 11 for domain distribution).

Table 11. Distribution of domains corresponding to the historical documents.

Domain Name	Document Count	Percentage
Literature	293	34.7%
History	198	23.4%
Pedagogy	39	4.6%
Juridical Science	45	5.3%
Generalities	33	3.9%
Theology	34	4.0%
Politics	28	3.3%
Applied Science	8	0.9%
Arts	14	1.6%
Linguistics/Philology	23	2.7%
Philosophy/Psychology	13	1.5%
Exact Science	16	1.8%
Economic Science	16	1.8%
Social Science	19	2.2%
Public Policy	21	2.4%
Natural Science	29	3.4%
Ethnography/Folklore	15	1.7%

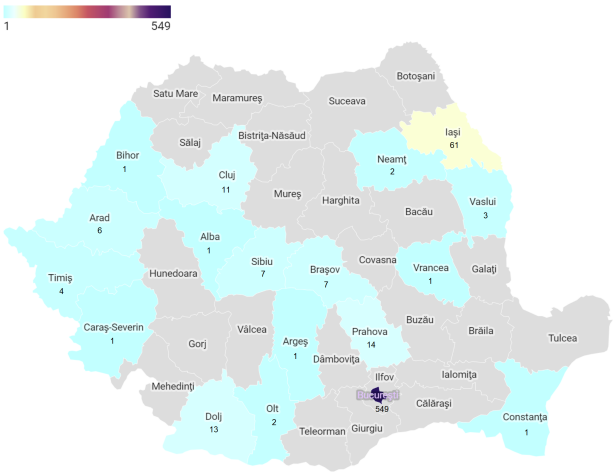

Figure 37. Regional distribution of historical documents.

The model for domain prediction detailed in the following subsection considers each document's title, its description, and the domain label. The distribution of the length of document descriptions is depicted in Figure 38.

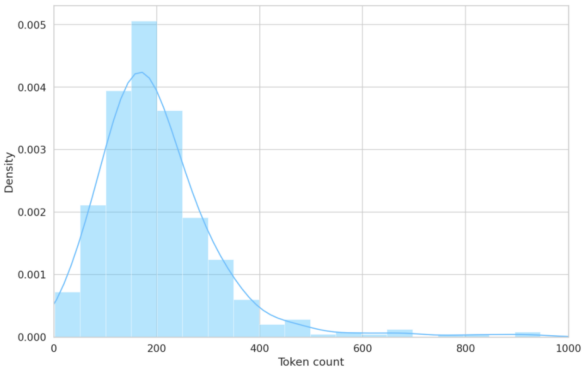

Figure 38. Token distribution for document descriptions.

Domain Prediction Method

The prediction model leverages a Romanian-only pre-trained Bidirectional Encoder Representation from Transformers (BERT) model [5], namely RoBERT [11]. RoBERT was pre-trained on a Romanian corpus that consists of text collected from a variety of sources, including the Romanian Wikipedia dump, the RoTex corpus [17], and the Oscar [8] collections. RoBERT followed the original BERT methodology for the training, mainly composed of two supervised tasks. In the first task, the model used a masked language model (MLM) to randomly predict masked tokens, while in the second task, the model used next sentence prediction (NSP) to learn if two sentences are randomly sampled for the dataset or follow each other. Also, the tokenization model was adapted to consider diacritics, which are very important for Romanian. Three language models were released (RoBERT-small, RoBERT-base and RoBERT-large) that achieved state-of-the art performance results on sentiment analysis, cross-dialect topic identification, and automatic diacritics restoration for Romanian language. Among the three available models, RoBERT-base was used for this experiment, whose architecture consists of 12 Transformer layers, 12 attention heads, a hidden size of 768 and a vocabulary size of around 38k words.

Domain categorization of documents is built as a multiclass text classification task. Given a new document's description as input, the deep neural model predicts its associated domain. Domains are structured as classes, such that each document has a feature consisting of its associated class that will serve as target in our model. A document can only belong to one out of 17 categories. If document belongs to the domain, its class label is 1, otherwise 0.

Our neural architecture model consists of a BERT encoder with 12 layers from which the embedding of the '[CLS]' token is passed through a dense layer, followed by a dropout layer. We considered categorical cross-entropy typically used for multiclass classification problems, as well as the softmax activation which outputs a probability between 0 and 1 for each class domain, as illustrated in Figure 39. The model uses a single categorical feature as target. The categorical features are hot encoded, automatically creating a one hot vector from all the domains; as such, each vector can be considered as a probability distribution.

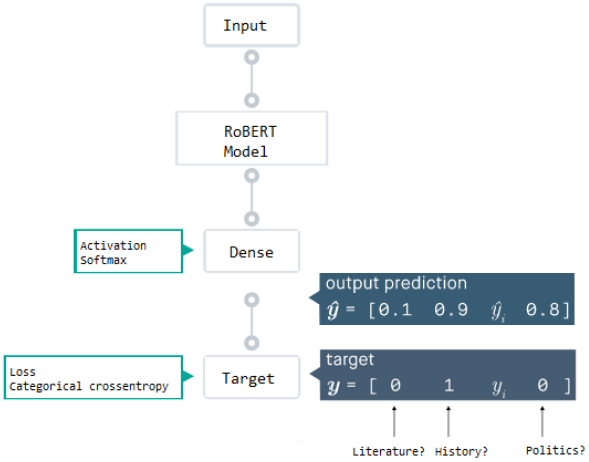

Figure 39. Model architecture.

We train the model for a batch size of 8 (8 sequences x 512 tokens = 4,096 tokens per batch) during 5 epochs, using a sequence length of 256 for training, which is required to learn positional embeddings. The model was trained on a GPU-based virtual machine with the maximum batch size and a data split ratio of 60-20-20 for train-test-validation, in an attempt to address the class imbalance illustrated in Table 11.

The model was fine-tuned leveraging Adam optimizer [10], with linear decay of the learning rate and the following parametrization: dropout=0.1; learning rate=2e-5; betas=(0.9, 0.999); Adam's epsilon for numerical stability=1e-6; and weight decay = 0. In our experiment, there are no warmup steps; normally, the learning rate increases linearly from 0 to the initial rate set in the optimizer during the warmup period.

Automatic Ontology Population Method

The Lib2Life ontology [7] considers the Dublin Core (DC) [18] and Friend-of-a-Friend (FOAF) [6] ontologies to model the 17 main domains as classes, some with specific sub-classes. The ontology can be accessed, visualized, but also queried using the Lib2Life platform [14].

The automatic population workflow of the ontology with class individuals is described in Figure 40.

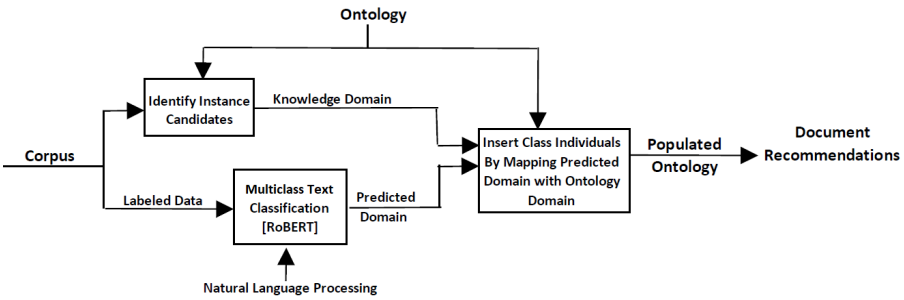

Figure 40. Workflow of automated ontology population.

The predicted domains are stored and mapped with the existing ontology domains and corresponding classes; afterwards, documents identifiers such as the author(s) and document title are used to insert new individuals into the ontology. The ontology is directly loaded from its Ontology Web Language (OWL) file and data is programmatically read using the Owlready library [9]. If the document is not already present in the ontology, the class individual will be inserted in the corresponding domain instance.

RESULTS

The RoBERT neural network was fine-tuned to perform multiclass text classification and predict a document's domain. The results reported after training the fine-tuned transformer network for 5 epochs are presented in Figure 42. Confusion matrix presented in Figure 5 highlights the model performance and we can notice a direct correlation between the prediction's accuracy and the number of samples in the test dataset for each domain. In general, the better a domain is represented in the test set, the better its classification.

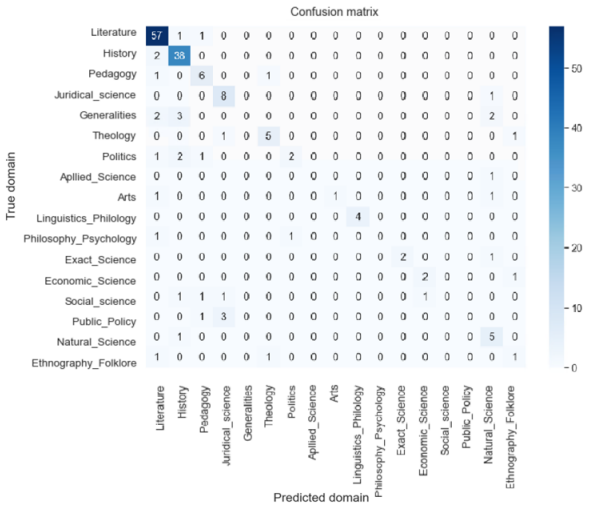

Figure 41. Confusion matrix.

Table 12 depicts the classification performance for each domain in terms of precision, recall and F1 scores. Within the test documents, the model is incapable of distinguishing books from the following five domains: generalities, applied science, philosophy/psychology, social science, and public policy. A determinant factor is represented by the number of documents in the test dataset. Moreover, when inspecting the document descriptions, text contained keywords that are present in several domains. For example, public policy samples contain terms that are representative for the juridical science domain (such as "politic", "jurist/juridic", "drept"); thus, the model encounters problems when distinguish between the two categories.

Table 12. Domain classification results

Domain	P	R	F1-score
Literature	.86	.97	.91
History	.83	.95	.88
Pedagogy	.60	.75	.67
Juridical Science	.62	.89	.73
Generalities	0	0	0
Theology	.71	.71	.71
Politics	.67	.33	.44
Applied Science	0	0	0
Arts	1.00	.33	.50
Linguistics/Philology	1.00	1.00	1.00
Philosophy/Psychology	0	0	0
Exact Science	1.00	.67	.80
Economic Science	.67	.67	.67
Social Science	0	0	0
Public Policy	0	0	0
Natural Science	.45	.83	.59
Ethnography/Folklore	.33	.33	.33
Weighted average	*.70*	*.78*	*.73*

Predictions are further mapped with knowledge domains. The ontology is automatically populated with corresponding RDF triplets, as illustrated in Figure 42. Categories such as Literature with associated individuals are exemplified, with the numbers of inserted documents displayed below the class name.

CONCLUSION

This paper describes a mechanism integrated in the Lib2Life's platform to automatically populate a web

ontology using Romanian BERT embeddings, which refines the current mechanism of semantic representation. This eliminates the tedious process of manually building these knowledge representations by librarians and overcomes the limitation of un-categorized newly added documents.

A fine-tuned RoBERT model was used for document category classification. The problem was modeled as a multiclass text classification task. Given a document's description, the model predicts the category associated with the document. The predictions are stored and mapped on existing ontology classes, each representing a specific knowledge domain; newly created individuals automatically populate the ontology, and their title is used for disambiguation. Our method can be further improved by training the model on a larger corpus, once new documents are added to the Lib2Life platform.

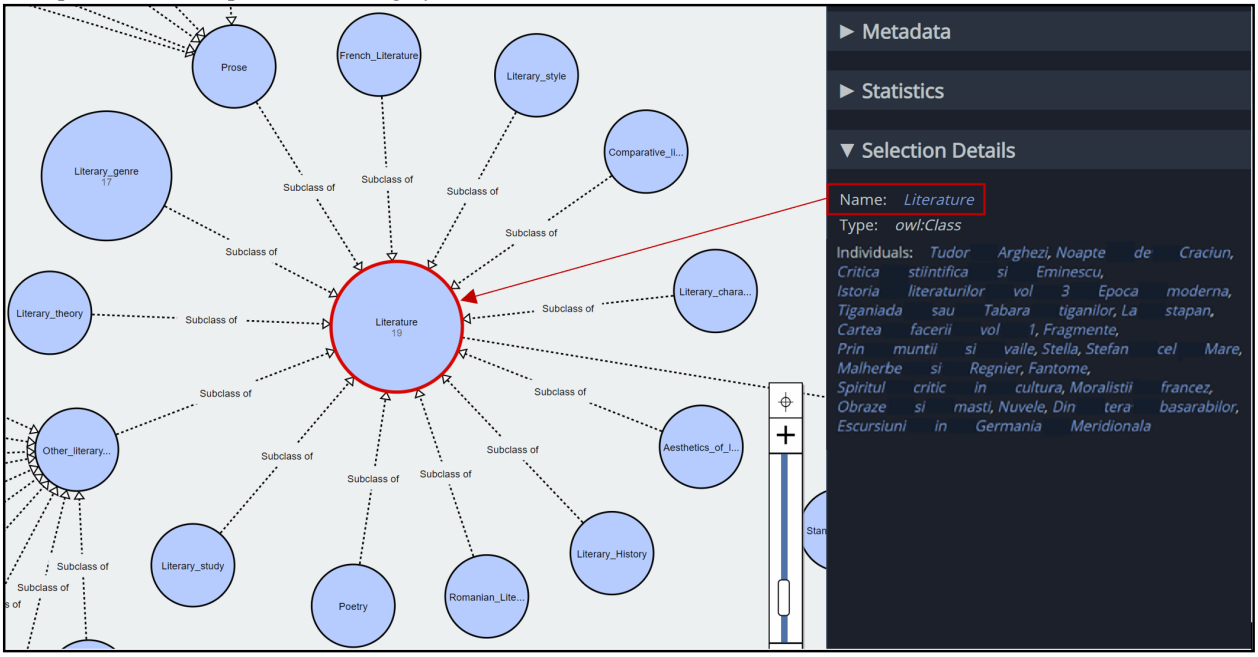

Figure 42. The Lib2Life ontology showing automatically added categories.

ACKNOWLEDGMENTS

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-III 69PCCDI/2018, Lib2Life – "Revitalizing libraries and cultural heritage through advanced technologies".

REFERENCES

1. Amarnath, G., Bertram, L., and Moore, R.W., 2002. Ontology services for curriculum development in NSDL. In Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries (Portland, OR, USA), ACM, 219–220.
2. Bravo, M., Aldea, A., and Hoyou-Reyes, L.F., 2021. Automated Ontology Population and Enrichment of Scientific Publications. *Journal of Physics: Conference Series 1828*, 012139. DOI= http://dx.doi.org/10.1088/1742-6596/1828/1/012139.
3. Cruz, C. and Nicolle, C., 2008. Ontology Enrichment and Automatic Population From XML Data. In Proceedings of the VLDB '08 (Auckland, New Zealand), ACM.
4. Dekel, O., Keshet, J., and Singer, Y., 2004. Large Margin Hierarchical Classification. In Proceedings of the 21st International Conference on Machine Learning (ICML) (Banff, Canada).
5. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACL 2018) (Minneapolis, Minnesota, USA), ACL.
6. Golbeck, J. and Rothstein, M., 2008. Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (Chicago, IL, USA), AAAI, 1138-1143.
7. Gutu-Robu, G., Ruseti, S., Tomescu, S., Dascalu, M., and Trausan-Matu, S., 2020. Designing an Ontology for Knowledge-based Processing in Romanina University Libraries. In Proceedings of the 8th Int. Workshop on Semantic and Collaborative Technologies for the Web, in

conjunction with the 16th Int. Conf. on eLearning and Software for Education (eLSE 2020) (Online), "CAROL I" National Defence University Publishing House, 119–126.

8. Javier Ortiz Suarez, P., Sagot, B., Romary, L., and Sagot, B.B., 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7) (Cardiff, UK).
9. Lamy, J.B., 2017. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence In Medicine*, 80, 11–28.
10. Loshchilov, I., and Hutter, F., 2019. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR 2019) (New Orleans, USA).
11. Masala, M., Ruseti, S., and Dascalu, M., 2020. RoBERT – A Romanian BERT Model. In Proceedings of the 28th Int. Conf. on Computational Linguistics (COLING) (online), ACL.
12. Maynard, D., Funk, A., and Peters, W., 2009. SPRAT: A Tool for Automatic Semantic Pattern-Based Ontology Population. *International conference for digital libraries and the semantic web 71*.
13. Maynard, D., Li, Y., and Peters, W., 2008. NLP Techniques for Term Extraction and Ontology Population. In *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, P. Buitelaar and P. Cimiano Eds. IOS Press, Amsterdam, Netherlands, 107–127.
14. Mitocaru, I., Gutu-Robu, G., Nitu, M., Dascalu, M., Trausan-Matu, S., Tomescu, S., and Florescu, G., 2020. The Lib2Life Platform – Processing, Indexing and Semantic Search for Old Romanian Documents. In Proceedings of the International Conference on Human-Computer Interaction (RoCHI2020) (Online), MatrixRom.
15. Robaldo, L., Di Caro, L., Alonso Alemany, L., Palmirani, M., and Villata, S., 2018. Ontology population: connecting legal text to ontology concepts and instances. *PROJECT: MIning and REasoning with Legal texts (MIREL)*.
16. Su, M.-H., Wu, C.-H., and Shih, P.-C., 2019. Automatic Ontology Population Using Deep Learning for Triple Extraction. In Proceedings of the APSIPA Annual Summit and Conference (Lanzhou, China), IEEE, 262–267. DOI= http://dx.doi.org/978-988-14768-7-6.
17. Tosca, A., 2019. RoTex Corpus Builder. Retrieved July 4h 2021 from https://github.com/aleris/ReadME-RoTex-Corpus-Builder.
18. Weibel, S., Kunze, J., Lagoze, C., and Wolf, M., 1998. Dublin Core Metadata for Resource Discovery. Retrieved September 15th 2020 from https://tools.ietf.org/html/rfc2413.