LamBERTa: Law article mining based on BERT architecture for the Italian Civil Code

(Extended Abstract)

Andrea Tagarelli[1], Andrea Simeri[1]

[1]*Dept. Computer Engineering, Modeling, Electronics, and Systems Engineering (DIMES), University of Calabria, 87036 Rende (CS), Italy*

Abstract

We present LamBERTa, a BERT-based framework for law article retrieval as a prediction problem, focusing on civil-law codes, and specifically trained on the Italian civil code. To the best of our knowledge, LamBERTa is the first advanced, deep learning approach to law article prediction for the Italian legal system. This paper is an extended abstract from our recent research work in [1].

Keywords

language models, deep learning, legal data, artificial intelligence and law

1. Introduction

Modeling law search and retrieval as prediction problems has recently emerged as a predominant approach in law intelligence [2]. Predictive tasks in legal information systems have often been addressed as text classification problems, ranging from case classification and legal judgment prediction, to legislation norm classification, and statute prediction. Early studies have focused on statistical textual features and machine learning methods, then the progress of *deep learning* methods for text classification has prompted the development of deep neural network frameworks for several learning tasks, such as charge prediction [3, 4], sentence modality classification [5, 6], legal question answering [7].

More recently, *deep pre-trained language models*, particularly the Bidirectional Encoder Representations from Transformers (BERT) [8], have emerged showing outstanding effectiveness in several NLP tasks. Thanks to their ability to learn a contextual language understanding model, they overcome the need for feature engineering (upon which classic, sparse vectorial representation models rely). Nonetheless, since these models are originally trained from generic domain corpora, they should not be directly applied to a specific domain corpus, as the distributional representation (embeddings) of their lexical units may significantly shift from the nuances and peculiarities expressed in domain-specific texts; this certainly holds for the legal domain as well, where language understanding is particularly challenging. In this respect, developing BERT models for legal documents has attracted increased attention, mostly concerning classification problems (e.g., [9, 10, 11, 12, 13, 14, 15]).

ICRDL'22: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padua, Italy

andrea.tagarelli@unical.it (A. Tagarelli); andrea.simeri@dimes.unical.it (A. Simeri)

CEUR Workshop Proceedings (CEUR-WS.org)

Motivations for BERT-based approach. Exploiting deep neural-network, pre-trained language modeling to solve the law article retrieval task has a number of key advantages that include the following. First, like any other deep neural network model, it totally avoids manual feature engineering, and hence the need for feature selection or relevance weighting methods (e.g., TF-IDF). Second, like sophisticated recurrent and convolutional neural networks, it models language semantics and non-linear relationships between terms; however, better than recurrent and convolutional neural networks, it is able to capture subtle and complex lexical patterns including the sequential structure and long-term dependencies, thus obtaining the most comprehensive local and global feature representations of a text sequence. Third, it incorporates the so-called *attention* mechanism, which allows a learning model to assign higher weight to text features according to their higher informativeness or relevance to the learning task. Fourth, being an effective bidirectional *Transformer* model, it overcomes the main limitations of early deep contextualized models (e.g., ELMO) [16] or decoder-based Transformer models (e.g., GPT) [17].

Research problem and challenges. The problem we tackle in this paper is law article retrieval, i.e., finding articles of interest out of a legal corpus that can be recommended as an appropriate response to a query expressing a legal matter. We assume that any query is expressed in natural language and discusses a legal subject that is in principle covered by the target law code corpus; moreover, a query is assumed to be free of references to any article identifier in the law code.

We address the law article retrieval problem based on the supervised machine learning paradigm: given a user-provided instance, i.e., a legal question, the goal is to automatically predict the category associated to the posed question, or more in general, to compute the probability distribution over all the predefined categories. In our context, the prediction is carried out by a machine learning system that must be trained on a target law code, in order to learn a classifier that will be used to perform the predictions against legal queries by exclusively utilizing the textual information contained in the law articles.

Like any other machine learning method, using deep pre-trained models like BERT for classification tasks requires the availability of data annotated with the class labels, so to design the independent training and testing phases for the classifier. However, we have to cope with a prediction task that is challenging from different perspectives, which are summarized as follows:

- The first challenge refers to the high number of classes, which are in the order of hundreds, or even thousands.
- The second challenge corresponds to the so-called *few-shot learning* problem, i.e., dealing with a small amount of per-class examples to train a machine learning model, which Bengio et al. recognize as one of the "extreme classification" scenarios [18]. Indeed, the number of classes are thousands, resp. hundreds, as they correspond to the number of articles in the law code, resp. portion of it, that is used to train the law article retrieval model. Also, all available articles must be used for training the model, therefore it is not straightforward to select a test set from the target corpus.
- The third challenge arises from our special interest in Italian law data, whereby we notice a lack of test query benchmarks for Italian legal article retrieval/prediction tasks.

2. Our proposed approach

Our research aims to address all the aforementioned challenges. To this purpose, in [1], we investigate on the modeling, learning and understanding of civil-law-based corpora, and we propose LamBERTa – **L**aw **a**rticle **m**ining based on **BERT a**rchitecture, a BERT-based framework for law article retrieval as a prediction problem. LamBERTa is designed to fine-tune an Italian pre-trained BERT on the Italian Civil Code (ICC) as the target law code, for law article retrieval as prediction, i.e., given a natural language query, predict the most relevant ICC article(s). Notably, few works have been developed for Italian BERT-based models [19], such as a retrained BERT for various NLP tasks on Italian tweets [20], and a BERT-based masked-language model for spell correction [21]; however, to the best of our knowledge, no study leveraging BERT for the Italian civil-law has been proposed until [1].

Data. The ICC is divided into six, logically coherent books, each in charge of providing rules for a particular civil law theme: *Book-1*, on Persons and the Family, articles 1-455 – contains the discipline of the juridical capacity of persons, of the rights of the personality, of collective organizations, of the family; *Book-2*, on Successions, articles 456-809 – contains the discipline of succession due to death and the donation contract; *Book-3*, on Property, articles 810-1172 – contains the discipline of ownership and other real rights; *Book-4*, on Obligations, articles 1173-2059 – contains the discipline of obligations and their sources, that is mainly of contracts and illicit facts (the so-called civil liability); *Book-5*, on Labor, articles 2060-2642 – contains the discipline of the company in general, of subordinate and self-employed work, of profit-making companies and of competition; *Book-6*, on the Protection of Rights, articles 2643-2969 – contains the discipline of the transcription, of the proofs, of the debtor's financial liability and of the causes of pre-emption, of the prescription.

Overview of the LamBERTa framework. Figure 1 shows the conceptual architecture of LamBERTa. The starting point is a pre-trained Italian BERT model whose source data consists of a recent Wikipedia dump, various texts from the OPUS corpora collection, and the Italian part of the OSCAR corpus; the final training corpus has a size of 81GB and 13 138 379 147 tokens.[1]

LamBERTa models are generated by fine-tuning the pre-trained Italian BERT model on a sequence classification task (i.e., BERT with a single linear classification layer on top) given in input the articles of the ICC or a portion of it. This fine-tuning is accomplished by using a typical configuration of BERT for masked language modeling, with 12 attention heads and 12 hidden layers, and initial (i.e., pre-trained) vocabulary of 32 102 tokens. Each model was trained for 10 epochs, using cross-entropy as loss function, Adam optimizer and initial learning rate selected within [1e-5, 5e-5] on batches of 256 examples.

The LamBERTa architecture can be configured w.r.t. two model aspects: (i) the learning approach and (ii) the training-instance labeling scheme for a given corpus of ICC articles. As for the former, we consider two learning approaches, here dubbed *global* and *local* learning, respectively. A **global model** is trained on the whole ICC, whereas a **local model** is trained on a particular book of the ICC, which is seen as a logically coherent subset of the whole civil code.

[1] bert-base-italian-xxl-uncased, available at https://huggingface.co/dbmdz/.

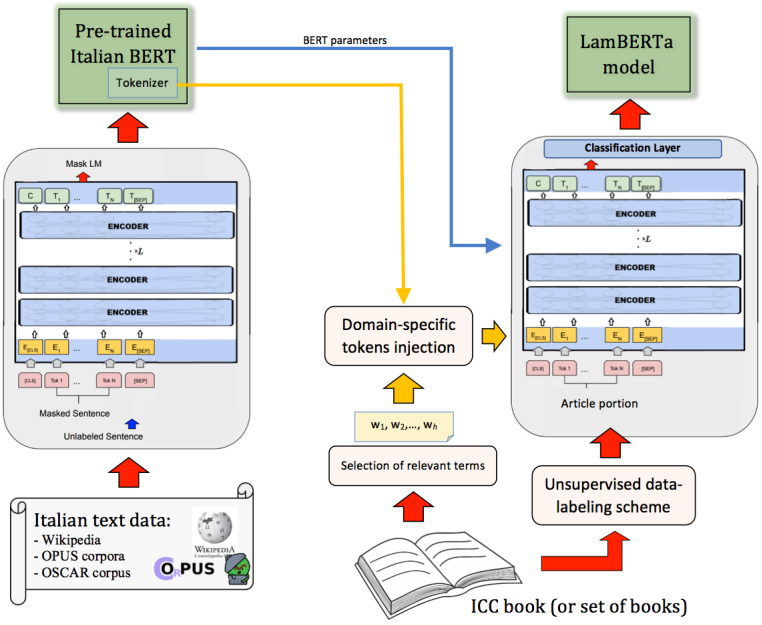

Figure 1: An illustration of the conceptual architecture of LamBERTa [1]

Either type of model is designed to be a classifier at article level, i.e., class labels correspond to the articles in the book(s) covered by the model.

LamBERTa models are trained using WordPiece tokenization of the article sentences. To avoid subwording domain-specific (i.e., legal) terms, thus disrupting their semantics, we enrich the BERT vocabulary with a selection of terms from the ICC articles, before tokenization.

Given the one-to-one association between classes and articles, and since the entire ICC must be used to embed the whole knowledge therein, a question becomes how to create as many training instances as possible for each article to make LamBERTa learn effectively. To this purpose, we devise various **unsupervised schemes of labeling of the ICC articles**, to create the training sets of LamBERTa models. These schemes adopt different strategies for selecting and combining portions from each article to derive the training set, while sharing the requirements of generating a minimum number of training units per article, here denoted as $minTU$; moreover, since each article is usually comprised of few sentences, and $minTU$ needs to be relatively large (we chose 32 as default value), each of the schemes implements a *round-robin* (RR) method that iterates over replicas of the same group of training units per article until at least $minTU$ are generated. The most effective scheme turned out to be the *unigram with parameterized emphasis on the title*, which builds the set of training units for each article as comprised of two subsets: the one containing the article's sentences with round-robin selection, and the other one containing only replicas of the article's title.

Experimental evaluation. In [1], LamBERTa models are evaluated through an extensive experimental analysis by considering **single-label** evaluation as well as **multi-label** evaluation tasks, based on **six different types of queries**, which vary by source, length and lexical

characteristics, and are summarized as follows: (Q1) Queries that correspond to randomly selected sentences from the articles of a book; (Q2) Same as Q1, but with *paraphrasing* of the queries; (Q3) Queries defined by *comments* on the ICC articles, i.e., annotations about the interpretation of the meanings and law implications associated to an article; (Q4) Same as Q3, but the comments are split into sentences; (Q5) Queries defined by *case law decisions* from the civil section of the Italian Court of Cassation that contains jurisprudential sentences associated with the ICC articles; (Q6) Queries defined by extracting the ICC metadata, i.e., headings of chapters, subchapters, and sections of each ICC book.

The obtained results, which are reported in [1], have shown the effectiveness of LamBERTa w.r.t. all book-specific query sets, and its superiority against widely used deep-learning text classifiers, namely *BiLSTM* [22], a bidirectional LSTM model as sequence encoder, *TextCNN* [23], a convolutional-neural-network-based model with multiple filter widths for text encoding and classification, *TextRCNN* [24], a bidirectional LSTM with a pooling layer on the last sequence output, *Seq2Seq-A* [25, 26], a Seq2Seq model with attention mechanism, and the *Transformer* model for text classification, which is adapted from the model originally proposed in [27] for machine translation. Also, we considered a few-shot learner conceived for an attribute-aware prediction task [28] that we have originally adapted based on the availability of ICC metadata.

On both global and local learning scenarios, our LamBERTa models outperform all the above mentioned competing methods, which has confirmed our initial expectation on the superiority of LamBERTa in learning classification models from few per-class labeled examples under a tough multi-class classification scenario.

Explainability. Explainability is one crucial aspect that typically arises in deep/machine learning models, and is clearly of high interest also in artificial intelligence and law (e.g., [29, 30]). In this regard, in [1] we investigate explainability of our LamBERTa models focusing on (i) understanding of how they form complex relationships between the textual tokens, and (ii) providing insights into the patterns generated by LamBERTa models through a visual exploratory analysis of the learned representation embeddings.

3. Conclusions

Our work falls into the corpus of recently developed studies that aim to show how artificial intelligence tools can be helpful not only to legal experts to reduce their workload, but also to citizens who can benefit from such tools to serve their search and consultation needs.

In this respect, we have presented LamBERTa, a BERT-based language understanding framework for law article retrieval as a prediction task. One key feature of LamBERTa is its ability to deal with a challenging learning scenario, where the multi-class classification setting is characterized by hundreds or thousands of classes and very few, per-class training instances that are generated in an unsupervised fashion.

It should be emphasized that, while focusing on the Italian Civil Code in its current version, the LamBERTa architecture can easily be generalized to learn from other law code systems.

For all technical and experimental details on our research study and the LamBERTa framework, the interested reader is referred to [1].

References

[1] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code, Artif. Intell. Law (2021) 1–57. doi:10.1007/s10506-021-09301-8.

[2] F. Dadgostari, M. Guim, P. Beling, M. A. Livermore, D. Rockmore, Modeling law search as prediction, Artif. Intell. Law 29 (2021) 3–34.

[3] B. Luo, Y. Feng, J. Xu, X. Zhang, D. Zhao, Learning to predict charges for criminal cases with legal basis, in: Proc. EMNLP, 2017, pp. 2727–2736.

[4] H. Ye, X. Jiang, Z. Luo, W. Chao, Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions, in: Proc. NAACL-HLT, 2018, pp. 1854–1864.

[5] J. O'Neill, P. Buitelaar, C. Robin, L. O'Brien, Classifying sentential modality in legal language: a use case in financial regulations, acts and directives, in: Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL), 2017, pp. 159–168.

[6] I. Chalkidis, I. Androutsopoulos, A. Michos, Obligation and Prohibition Extraction Using Hierarchical RNNs, in: Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2018, pp. 254–259.

[7] P. Do, H. Nguyen, C. Tran, M. Nguyen, M. Nguyen, Legal question answering using ranking SVM and deep convolutional neural network, CoRR abs/1703.05320 (2017).

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. NAACL-HLT, 2019, pp. 4171–4186.

[9] J. Rabelo, M. Kim, R. Goebel, Combining similarity and transformer methods for case law entailment, in: Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL), 2019, pp. 290–296.

[10] I. Chalkidis, I. Androutsopoulos, N. Aletras, Neural legal judgment prediction in english, in: Proc. ACL, Association for Computational Linguistics, 2019, pp. 4317–4323.

[11] L. Sanchez, J. He, J. Manotumruksa, D. Albakour, M. Martinez, A. Lipani, Easing legal news monitoring with learning to rank and BERT, in: Proc. ECIR, volume 12036 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 336–343.

[12] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, S. Ma, BERT-PLI: modeling paragraph-level interactions for legal case retrieval, in: Proc. IJCAI, 2020, pp. 3501–3507.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: the muppets straight out of law school, CoRR abs/2010.02559 (2020).

[14] H. Nguyen, P. M. Nguyen, T. Vuong, Q. M. Bui, C. M. Nguyen, T. B. Dang, V. Tran, M. L. Nguyen, K. Satoh, JNLP team: Deep learning approaches for legal processing tasks in COLIEE 2021, CoRR abs/2106.13405 (2021). URL: https://arxiv.org/abs/2106.13405. arXiv:2106.13405.

[15] M. Yoshioka, Y. Aoki, Y. Suzuki, BERT-based ensemble methods with data augmentation for legal textual entailment in COLIEE statute law task, in: Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL), ACM, 2021, pp. 278–284.

[16] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. NAACL-HLT, 2018, pp. 2227–2237.

[17] A. Radford, I. Sutskever, Improving language understanding by generative pre-training,

in: arxiv, 2018.
[18] S. Bengio, K. Dembczynski, T. Joachims, M. Kloft, M. Varma, Extreme Classification, Technical Report, Report from Dagstuhl Seminar 18291, 2019. doi:10.4230/DagRep.8.7.62.
[19] F. Tamburini, How "BERTology" Changed the State-of-the-Art also for Italian NLP, in: Proc. 7th Italian Conf. on Computational Linguistics (CLiC-it), volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.
[20] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proc. 6th Italian Conf. on Computational Linguistics (CLiC-it), volume 2481 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.
[21] D. Puccinelli, S. Demartini, R. E. D'Aoust, Fixing comma splices in italian with BERT, in: Proc. 6th Italian Conf. on Computational Linguistics (CLiC-it), volume 2481 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.
[22] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, in: Proc. IJCAI, 2016, pp. 2873–2879.
[23] Y. Kim, Convolutional neural networks for sentence classification, in: Proc. EMNLP, 2014, p. 1746–1751.
[24] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: Proc. AAAI, 2015, p. 2267–2273.
[25] C. Du, L. Huang, Text classification research with attention-based recurrent neural networks, Int. J. Comput. Commun. Control 13 (2018) 50–61.
[26] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proc. ICLR, 2015.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. NIPS, 2017, pp. 5998–6008.
[28] Z. Hu, X. Li, C. Tu, Z. Liu, M. Sun, Few-shot charge prediction with discriminative legal attributes, in: Proc. COLING, Association for Computational Linguistics, 2018, pp. 487–498.
[29] K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, A. S. Yeh, Semi-supervised methods for explainable legal prediction, in: Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL), 2019, pp. 22–31.
[30] P. Hacker, R. Krestel, S. Grundmann, F. Naumann, Explainable AI under contract and tort law: legal incentives and technical challenges, Artif. Intell. Law 28 (2020) 415–439.