Deep neural networks modeling for natural language question answering on semi-structured datasets / Omri Suissa

Abstract

Applying information technologies in cultural heritage domains is an ongoing effort to make cultural heritage resources digitally available, searchable, and analyzable. In recent years, deep neural networks (DNNs) have dominated the field of automatic text analysis and natural language processing (NLP), presenting a super-human performance in some cases. DNNs are state-of-the-art machine learning algorithms solving many NLP tasks relevant to Digital Humanities (DH) research, such as spell checking, language detection, entity extraction, author detection, question answering, and other tasks. These supervised (and self-supervised) algorithms learn patterns from a large number of "right" and "wrong" examples and apply them to new examples. However, using DNNs for analyzing text resources in DH research presents challenges, such as the (un)availability of training data and a need for domain adaptation. This collection of research articles explores these challenges and how to overcome them by (1) designing a practical decision model for DH experts for when and how to choose the appropriate machine learning or deep learning approaches for their research based on an analysis of multiple use-cases from DH studies in recent literature, (2) designing and empirically validating an end-to-end pipeline and a new novel domain adaptation methodology for the factual question-answering task (i.e., close reading) using DNNs in one DH field - genealogical knowledge graphs combined with unstructured texts, and (3) designing and empirically validating an end-to-end pipeline and a new novel domain adaptation methodology for a natural language numerical aggregation question-answering (i.e., distant reading) using DNNs in the same DH field.

The primary genealogical dataset for these articles is 3,140 family trees containing 1,847,224 different individuals from the corpus of the Douglas E. Goldman Jewish Genealogy Center in the

Anu Museum[1]. The Douglas E. Goldman Jewish Genealogy Center contains over 5 million individuals and over 30 million family tree connections (edges) to families, places, and multimedia items. To comply with the Israeli privacy regulation[2] and the European general data protection regulation[3] (GDPR), only family trees for which the Douglas E. Goldman Jewish Genealogy Center in Anu Museum has been granted consent or rights to publish online were used in the dataset generation. Moreover, as far as possible, all records containing living individuals have been removed from the dataset. Furthermore, all personal information and any information that can identify a specific person in these articles' examples, including the examples in the figures, have been altered to protect the individuals' privacy.

As described in the following chapters, due to genealogical datasets' unique structure and characteristics, this domain requires specific research and models. As described in the research articles, other open datasets were used for pre-training of the DNN models.

These articles present the main two challenges almost every DH/LIS (Digital Humanities and Library and Information Science) researcher can expect to encounter using DNN models in her research. The first article, "Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science" (published in the Journal of the Association for Information Science and Technology), presents the main challenges of using DNNs in DH research, a decision model for handling these challenges, and the potential adoption of DNN methods. Moreover, this article argues that DH/LIS researchers should expand their arsenal of computational skills and methods. While these challenges (i.e., domain adaptation and unavailable training data) are not unique to the DH domain, the DH domain mainly deals with non-modern language and data. Therefore, these challenges become considerably harder to solve (i.e., harder to generate synthetic data, harder to transfer learning from one domain to another).

[1] https://dbs.anumuseum.org.il/skn/en/c6/e18493701

[2] https://www.gov.il/BlobFolder/legalinfo/data_security_regulation/en/PROTECTION%20OF%20PRIVACY%20REGULATIONS.pdf

[3] https://gdpr-info.eu/

The second article, "Question answering with deep neural networks for semi-structured heterogeneous genealogical knowledge graphs" (published in the Semantic Web journal), presents a method that tackles several challenges for question-answering in the genealogical domain, such as representing a genealogical graph as a knowledge graph, adapting graph traversal algorithms to the genealogical interpretation of the relationships, automatically generating datasets from genealogical format (GEnealogical Data COMmunication; see The GEDCOM Genealogical Data Standard for further details), and defining genealogical question types in the genealogical domain. The fine-tuned model trained on the genealogical dataset with second-degree relationships, Uncle-BERT$_2$, yielded an F1 score of 81.45 and outperformed the baseline BERT model (only 60.12). Moreover, this article examined the effect of the type of question on the accuracy of the neural network models in question-answering in the genealogy domain.

The third article, "Around the GLOBE: Numerical Aggregation Question-Answering on Heterogeneous Genealogical Knowledge Graphs with Deep Neural Networks" (published in the Journal on Computing and Cultural Heritage), outlines and implements an end-to-end, multi-phase methodology for DNN-based answering aggregative numerical natural questions in the genealogical domain. The results show that the GLOBE model outperformed the state-of-the-art generic model (TaPas) on genealogical data. Moreover, this article examined the effect of the dataset design on the accuracy of the question-answering model and shows that the complexity of the genealogical domain requires a more complex pipeline that can split and reconstruct the dataset tables based on the question. Furthermore, this article shows the impact of the mathematical operation on the ability of the DNN model to predict answers close to the correct answer.

In summary, these articles' contributions are (1) a decision model for handling DH\LIS challenges when using DNNs, (2) a genealogical knowledge graph representation of the GEDCOM standard, (3) a dedicated graph traversal algorithm for genealogical data (Gen-BFS), (4) an automatically generated SQuAD-style genealogical training dataset (Gen-SQuAD), (5) a fine-tuned question-answering BERT-based model for the genealogical domain (Uncle-BERT), (6) a genealogical dataset representation and generation of GEDCOM knowledge graph structure to a structured dataset for numerical aggregation question-answering (GenAgg), (7) evaluation of the optimal dataset design for numerical aggregation question-answering for the genealogical domain, (8) an

end-to-end numerical aggregation question-answering pipeline for the genealogical domain, and (9) customized and fine-tuned numerical question-answering BERT-based model for the genealogical domain (GLOBE).

The proposed methodologies can be applied to other downstream NLP tasks in the genealogical domain (and potentially other DH fields), such as entity extraction, text classification, summarization, etc. Researchers can utilize these articles' results to reduce time, cost, and complexity and improve accuracy in the genealogical domain NLP research.

MMS: 9926875312505776