The Challenges of German Archival Document Categorization on Insufficient Labeled Data

Fabian Hoppe[1,2⋆], Tabea Tietz[1,2], Danilo Dessì[1,2], Nils Meyer[3,4], Mirjam Sprau[5], Mehwish Alam[1,2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany
[3] Baden-Württemberg State Archives, Germany
[4] Deutsche Digitale Bibliothek, Germany
[5] German Federal Archives, Germany

Abstract. Document exploration in archives is often challenging due to the lack of organization in topic-based categories. Moreover, archival records only provide short text which is often insufficient for capturing the semantic. This paper proposes and explores a dataless categorization approach that utilizes word embeddings and TF-IDF to categorize archival documents. Additionally, it introduces a visual approach built on top of the word embeddings to enhance the exploration of data. Preliminary results suggest that current vector representations alone do not provide enough external knowledge to solve this task.[6]

Keywords: Dataless Categorization · Text Categorization · Document Exploration · Cultural Heritage

1 Introduction

Exploring cultural heritage data is inevitably connected to searching through historical records in archives. This task is complicated because of the huge amount of documents stored in hierarchical file systems. In fact, the retrieval of relevant information usually requires significant human effort. Hence, the demand for topic-based categorization and more diverse exploration methods, like visual exploration, increases with the growing number of electronically available archival records. Despite the ongoing trend of prominent digitization campaigns the majority of archival objects so far are neither available in digital form nor transcribed. Thereby, their content often is inaccessible and only descriptive metadata can be used to categorize and organize them. For most of the electronically available archival documents merely a title has been registered probably together with identifier and archive file system information. As a consequence, numerical vector representations required by modern categorization systems face several challenges if applied to archival data:

⋆ Corresponding author's email: fabian.hoppe@fiz-karlsruhe.de

- Document titles provide only short texts which are insufficient to capture semantics due to the amount of data and natural language ambiguity issues.
- Archival objects are organized in a hierarchical file system which is ignored by current representations.
- Document understanding requires extensive world knowledge, like historical context information.
- The annotation of data is hindered by disagreements of experts about the detailed interpretation of data, and the fine-grained domain-specific information need. Consequently, only insufficient training data is available.

This paper addresses these shortcomings by exploring the use of word embeddings and TF-IDF as a way to introduce external knowledge, and interpret the semantics of document titles and category labels within a dataless categorization approach on German archive data. Furthermore, the paper introduces a visualization which provides graphical exploration of the archive when the categorization approach is unfeasible. The contribution of this paper is three-fold:

- We propose a dataless categorization approach for German archive data based on vector representations.
- We include the archive structure in our vector representations to improve categorization results.
- We provide a visual exploratory tool that can be used to retrieve documents and support their annotation.

2 Related Work

In the Digital Humanities community the use of semantic technologies has attracted a fair amount of interest in order to make the retrieval and exploration of digitized archives easier. Nevertheless, the use of semantic representations in this field has not been fully investigated yet. In fact, previous works were often based on supervised classification, e.g., [8] used various classifiers for supporting historians in enhancing their work, or topic modeling methods such as [3] where the authors categorized a collection of 24,787 archive documents with 100 topics. However, these methods usually rely on complete digitized documents where each document provides a large amount of text. In case of short texts these approaches fail, which makes them unfeasible to use for metadata of archival resources. The sparsity problem for short text categorization is addressed by current deep learning approaches. For example, to learn a context within short texts, neural network-based systems such as Convolutional Neural Networks (CNN) were proposed [10]. However, such kind of approaches usually need a lot of labeled data to train models, and often deal with a small number of fixed classes. Recent methods deal with insufficient training data by considering external knowledge about the categories, e.g., category labels are used to infer their semantics. In particular, these approaches exploit a vector space model where both texts and categories are represented and compared by similarity measurements [1,7]. One model, KBSTC [9] showed that this dataless categorization

approach provides reasonably good results for short texts by using entity embeddings. However, these methods are applied on English datasets, as a contrast, the archive documents considered in this study are in German. Inspired by this kind of methodologies, we investigate general-purpose embeddings for archival objects. This analysis considers the unique structure of German archival documents and a higher number of classes as compared to other datasets used for evaluating dataless approaches.

3 Dataset of Archival Holdings

Our dataset was collected by two digitization projects of the German Federal Archives and the Baden-Württemberg State Archives on the so-called Weimar Republic, the first German democracy [4]. The ongoing decade of anniversaries related to the Weimar Republic increases the current demand by historical researchers and the general public to find historical documents of the time. Over the last few years, the archives have selected a large number of relevant archival holdings from ministries, public institutions, corporate bodies and particular individuals from this period to be digitized and described, which cover aspects of politics, economy, society and everyday life in Germany from 1918 to 1933. The collection is composed of 21,042 documents and 799 categories defined by domain experts. Only 9% (2,011 documents) are manually annotated with 59% of all categories occurring at least once. The titles are on average 7 words long, which does not provide enough contextual information for capturing the semantics of a document. The documents are organized in a file system, e.g., the document *public welfare for the poor* is part of the document *welfare for war victims and survivors*, which is than again part of the document *supply affairs*, etc. On average one document is a part of 4 higher level documents.

4 Methodology

This section details our approach to support the categorization of archival data.

4.1 Word Embeddings Generation

The data consists of German text, therefore, new word embedding models were trained using Skip-gram Word2Vec [5] and FastText [2] on a dump of all German Wikipedia articles[7]. The references and link sections of Wikipedia articles were removed. The word embeddings were trained with 300 dimensions, a window of 5 words and 10 negative samples.

4.2 Dataless Categorization Approach

Our approach is illustrated in Fig. 1. The input is a document title d_j and a set of categories C. The approach is subdivided in the following four modules.

[7] https://dumps.wikimedia.org/dewiki/ retrieved on 20.11.2019

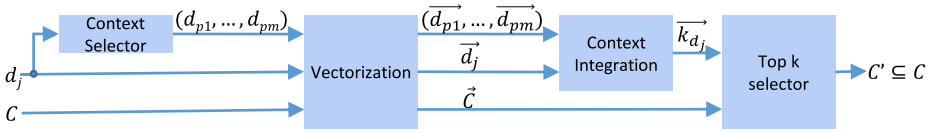

Fig. 1. Schema of the dataless approach.

Context Selector. This module extracts the context of d_j as a sequence of its ancestors, starting with the parent document d_p, and traversing up the archive file system by recursion, i.e., $k_{d_j} = (d_p, k_{d_p}) = (d_{p1}, \ldots, d_{pm})$.
Vectorization. This module transforms the input textual representations into vectors through the use of word embeddings or TF-IDF. For example, consider a document title d_j. First, it removes all stop words, yielding d'_j. Then, when it is set up to use word embeddings, it creates the vector representation $\overrightarrow{d_j}$ by applying (1), where n denotes the number of words of d'_j, and $\overrightarrow{w_{i,j}}$ denotes the word embedding of the i-th word. When the TF-IDF mode is set up, the module assign a numerical value for each word in d'_j building the TF-IDF vector $\overrightarrow{d_j}$.
Context integration. This module adapts the hierarchical information of the document context into the vector representation. More precisely, based on the context sequence k_{d_j}, an exponential weighting schema of the corresponding document vector representation $\overrightarrow{k_{d_j}}$ is calculated according to equation (2), where m is the length of the context sequence, $\overrightarrow{d_{pi}}$ denotes the vector representation of the i-th context object, and w denotes a hyperparameter to determine the importance of the context. The weights insure that document vectors are scaled based on their relative position in the hierarchy.
Top k selector. After the featurization process, both the document $\overrightarrow{k_{d_j}}$ and categories $\overrightarrow{C}$ are represented by vectors which can be used to detect their semantic similarity. For doing so, this module employs the cosine similarity, and yields the top k categories $C' \subseteq C$ as prediction.

$$\overrightarrow{d_j} = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{w_{i,j}} \quad (1) \qquad \overrightarrow{k_{d_j}} = \overrightarrow{d_j} + \sum_{i=1}^{m} \frac{1}{w^{-i}} \cdot \overrightarrow{d_{pi}} \quad (2)$$

5 Results and Discussion

This section reports our preliminary results of the proposed approach, and discusses the open challenges related to the archive document categorization.
Experimental setup. The weighing factor for the context embeddings is set to $w = 1.1$. Based on the average number of assigned categories per document in the gold standard of 3.16 the classification parameter k is set to 5.
Dataless results. The results of our dataless approach are reported in Table 1. The approach does not obtain relevant results in terms of precision, recall, and

Table 1. Comparison of vector representations within the dataless categorization.

Data	Representation	Precison	Recall	F-measure
Title	Word2Vec	**0.059**	**0.110**	**0.077**
	FastText	**0.059**	0.109	0.076
	TF-IDF	0.042	0.078	0.054
Title + Context	Word2Vec	0.075	0.140	0.098
	FastText	0.061	0.114	0.080
	TF-IDF	**0.123**	**0.229**	**0.160**

f-measure. The TF-IDF representation method obtains the best overall performance and outperforms the semantic representations when the title and context were considered. The fine grained classification task highlights small differences between specific words, which are neglected within a semantic space. For example *officer* and *navy officer* are different categories with a high cosine similarity for semantic embeddings, but a low cosine similarity for TF-IDF. Consequently, it is easier to differentiate between both categories within the TF-IDF space. In addition, the high dimensional space of TF-IDF is better suited to store the different aspects gathered by combining the title and context. Overall, Table 1 shows that 1) the context of documents plays an important role for achieving better performance 2) basic semantic representations are not sufficient to solve a fine-grained dataless classification task. However, considering the high number of possible categories (799) the results are encouraging , because a purely random assignment of the average number of categories would achieve a f-measure baseline of 0.004. Also, by manually revising the results, many suggested categories matter for the input document. For example, the input title "*public welfare for the poor*" is assigned to the categories *unemployment benefits* and *orphan welfare*. A considerable issue of this task is that archivists when classifying documents use their experience and expertise. It is a challenge to include this background knowledge in any automated process and most-likely this categorization task will always need humans in the loop to deliver satisfactory results.

6 Visual Exploration

In this section, we briefly introduce how the vector representations can be used to visualize the search space through the *Embedding Projector* tool [6]. This tool enables the interactive visualization of embeddings and utilizes PCA and t-SNE to perform dimensionality reduction and plots these representations as a point cloud. Additionally, it depicts the nearest neighbours of a selected embedding and provides functionalities to search based on metadata and restrict the plotted points to a specific subset. The visualization is available online[8]. It supports

[8] http://vocol-ise.fiz-karlsruhe.de/

further research on text categorization by presenting the document arrangement within a graphical space to domain experts. It can be used to evaluate the vector representations of documents and categories as well as to support the manual labeling process by enabling the search of similar categories, e.g. a domain expert finds the more specific category 'flu' by looking for neighbours of the generic category 'disease'. This provides a possibility to improve the accuracy of the gold standard by pointing to semantically similar, but less frequently used categories.

7 Conclusion and Future Work

In this paper preliminary results of a semantic-based approach on dataless categorization to support the activities of archivists in exploring and annotating archive documents is presented. Moreover, the first version of a visual exploratory tool for supporting the manually exploration and annotation tasks is introduced. Both proposed methods are based on semantics captured by training vector space models and can be used in many unsupervised settings. Future work will enhance the vector representations by exploiting taxonomic relations that occur between categories and integrate external resources such as DBpedia and authority files.

References

1. Chang, M.W., Ratinov, L.A., Roth, D., Srikumar, V.: Importance of semantic representation: Dataless classification. In: Aaai. vol. 2, pp. 830–835 (2008)
2. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893 (2018)
3. Hengchen, S., Coeckelbergs, M., Van Hooland, S., et al.: Exploring archives with probabilistic models: Topic modelling for the valorisation of digitised archives of the european commission. In: IEEE Int. Conf. on Big Data. pp. 3245–3249 (2016)
4. Herrmann, T., Zahnhausen, V.: Auf dem Weg zum Digitalen Lesesaal: Das Projekt 'Weimar – Die erste deutsche Demokratie'. In: Kulturelles Kapital und ökonomisches Potential. Zukunftskonzepte für Archive. 86. Deutscher Archivtag. Verband deutscher Archivarinnen und Archivare e.V. (2016)
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
6. Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F.B., Wattenberg, M.: Embedding projector: Interactive visualization and interpretation of embeddings. arXiv preprint arXiv:1611.05469 (2016)
7. Song, Y., Roth, D.: On dataless hierarchical text classification. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
8. Sprugnoli, R., Tonelli, S.: Novel event detection and classification for historical texts. Computational Linguistics **45**(2), 229–265 (2019)
9. Türker, R., Zhang, L., Koutraki, M., Sack, H.: Knowledge-based short text categorization using entity and category embedding. In: European Semantic Web Conference. pp. 346–362. Springer (2019)
10. Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.L., Hao, H.: Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing **174**, 806–814 (2016)