Machine Learning Architectures for Scalable and Reliable Subject Indexing

Fusion, Knowledge Transfer, and Confidence

Martin Toepfer(✉)

ZBW – Leibniz Information Centre for Economics,
Düsternbrooker Weg 120, 24105 Kiel, Germany
m.toepfer@zbw.eu

Abstract. Digital libraries desire automatic subject indexing as a scalable provider of high-quality semantic document representations. The task is, however, complex and challenging, thus many issues are still unsolved. For instance, certain concepts are not detected accurately, and confidence estimates are often unreliable. Accurate quality estimates are, however, crucial in practice, for example, to filter results and ensure highest standards before subsequent use. The proposed thesis studies applications of machine learning for automatic subject indexing, which faces considerable challenges like class imbalance, concept drift, and zero-shot learning. Special attention will be paid to architecture design and automatic quality estimation, with experiments on scholarly publications in economics and business studies. First results indicate the importance of knowledge transfer between concepts and point out the value of so-called fusion approaches that carefully combine lexical and associative subsystems. This extended abstract summarizes the main topic and status of the thesis and provides an outlook on future directions.

Keywords: Automatic subject indexing · Machine learning · Quality control

1 Introduction

By subject indexing, libraries create concise yet comprehensive descriptions of documents with terms of controlled vocabularies like MeSH[1], LCSH[2], GND[3], or STW[4]. These structured semantic document representations are highly valuable for digital libraries since they support services like semantic browsing, multilingual information retrieval and recommendation, or trend detection. According to the TPDL-2017 theme “Part of the machine: turning complex into scalable”[5],

[1] www.nlm.nih.gov/mesh/.

[2] id.loc.gov/authorities/subjects.html.

[3] www.dnb.de/gnd.

[4] www.zbw.eu/stw.

[5] www.tpdl.eu/tpdl2017/.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 644–647, 2017.
DOI: 10.1007/978-3-319-67008-9_61

digital libraries demand for solutions that ease access to large amounts of heterogeneous data, thus making accurate automatic subject indexing a key technology for their infrastructure. The challenges that need to be solved are, however, considerable: subject indexing is a complex cognitive task which involves several aspects of the human mind, such as natural language processing and semantic reasoning under uncertainty. Its automation thus requires artificial intelligence and machine learning, having many options to model the task and encode it in terms of input, output, features, dependencies, and objectives. In the past, such architectures have been realized in different ways.

For instance, several researchers regarded subject indexing as a multi-label classification task [2,9], which differs from standard classification in that multiple possibly interrelated classes have to be assigned per document. Since the number of different classes is remarkably large, often exceeding several thousands of concepts, system engineers have to be very careful. Complex models with many variables can quickly suffer from too few training examples to estimate parameters reliably, and in order to recognize previously unseen concepts, zero-shot learning [8] and exploitation of external knowledge sources have to be regarded.

For these reasons, researchers have made several assumptions and simplifications. For instance, Medelyan and Witten [7] made a specific *invariance assumption* that allows to learn parameters shared by all concepts, *transferring knowledge* between them. This reduces the minimum amount of documents necessary for training and enables zero-shot learning, however, it requires extensive linguistic knowledge from thesauri. Other approaches make "naive" independence assumptions for the sake of efficiency. Very common are binary relevance approaches like the work of Wilbur and Kim [11], using independent classifiers for each subject heading, thus learning associations between terms and concepts. More complex architectures have been constructed by *combining different approaches* using meta-learning [4] or learning to rank [3]. Yet, research on the design of machine learning architectures with heterogeneous modules for automatic subject indexing is limited. Their composition and setup has numerous parameters, configuration options, and design choices.

Interestingly, although in practice accurate *quality estimates* are fundamental to ensure highest standards of generated meta-data, there have been negative experiences with confidence values provided by systems[6], hence, it becomes an attractive research topic. Finally, note that evaluation in automatic indexing is non-trivial and may require more detailed analysis than standard metrics like precision and recall.

2 Contributions and Research Questions

The proposed thesis aims to make contributions in the fields of digital libraries and computer science, gaining insights into applications of machine learning

[6] For instance, due to experiments at the ZBW and correspondence with the German National Library at a recent workshop on "Computer-assisted Subject Cataloguing", 2017 in Stuttgart, Germany.

techniques, and exploring different approaches, their effectiveness and efficiency. Related to Manning's thoughts on computational linguistics and deep learning [5], the thesis strives for thorough problem analysis and meaningful composition of machine learning architectures, which may provide more general insights.

In this regard, the thesis will be directed by the following questions:

1. Architectures:
   (a) How do aspects of *architecture design* (modelling: encoding of input, features, output, dependencies, objectives) and requirements of the environment (properties of thesauri, characteristics and availability of training data, dynamics of the domain) relate to each other? In particular, which effects does architecture design have according to concept drift and zero-shot learning?
   (b) What is the role of *invariance* assumptions, which enable to share parameters in learning? Where do they apply? How are they modeled and integrated into systems?
   (c) Can different approaches be "meaningfully" combined, that is, can we leverage individual advantages effectively?
2. How can (reliable) confidence estimates be computed for automatic subject indexing? What are relevant aspects of "confidence" in subject indexing? How do the terms "quality" and "confidence" relate to each other in the field of automatic subject indexing? How do confidence and quality estimation fit into encompassing architectures?

Finally, many research activities focused on the medical domain where subject-specific solutions are available and can be incorporated. This thesis will investigate a less-studied domain, namely scholarly publications related to economics, where comparable solutions do not exist. Differing challenges may emerge, with the prospect of novel insights.

3 Approach

As part of the thesis, different automatic subject indexing architectures and confidence estimation approaches are analysed, designed, implemented, and evaluated.

The project looks at approaches like dictionary matching, ranking and associative methods, and determines how they fit into encompassing architectures, especially with respect to concept drift.

Regarding confidence estimation, techniques similar to the DeepQA architecture of IBM Watson [1] will be considered. The project will first collect ideas for implementation: confidence estimation features for automatic indexing must be developed, and meaningfully grouped into evidence profiles. Finally, experimental evaluation will be performed.

Theoretical analysis will be taken into account, however, it may be limited due to the fuzzy nature of the tasks [9]. Therefore, experiments will be conducted to test systems and justify hypothesis empirically. Commonly used metrics are precision, recall and F_1 [2,9], although they may be too shallow to assess quality reasonably. Semantic relations between concepts [6] and graded ratings may therefore be taken into account. Confidence estimation methods may be evaluated using ranking metrics. The data that will mainly be used, has the following properties: scholarly publications related to economics, written in English, only descriptive metadata (short texts), indexed with descriptors of the STW.

4 Status Summary

A major contribution has already been accomplished by analysis of architectures, development of specific fusion systems, and experiments on short texts (titles and author keywords) [10]. This work will be supplemented by certain extensions, their analysis, and experiments. For instance, different approaches to apply learning components for fusion will be explored. Work on confidence estimation is at an early stage and thus may profit from exchange with the digital libraries community.

References

1. Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J.M., Schlaefer, N., Welty, C.A.: Building watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
2. Gibaja, E., Ventura, S.: A tutorial on multilabel learning. ACM Comput. Surv. **47**(3), 52: 1–52: 38 (2015)
3. Huang, M., Névéol, A., Lu, Z.: Recommending MeSH terms for annotating biomedical articles. JAMIA **18**(5), 660–667 (2011)
4. Jimeno-Yepes, A., Mork, J.G., Demner-Fushman, D., Aronson, A.R.: A one-size-fits-all indexing method does not exist: automatic selection based on meta-learning. JCSE **6**(2), 151–160 (2012)
5. Manning, C.D.: Computational linguistics and deep learning. Comput. Linguist. **41**(4), 701–707 (2015)
6. Medelyan, O., Witten, I.H.: Measuring inter-indexer consistency using a thesaurus. In: Proceedings of Joint Conference on Digital Libraries, pp. 274–275. ACM (2006)
7. Medelyan, O., Witten, I.H.: Domain-independent automatic keyphrase indexing with small training sets. J. Am. Soc. Inf. Sci. Technol. **59**(7), 1026–1040 (2008). http://dx.doi.org/10.1002/asi.20790
8. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22, pp. 1410–1418. Curran Associates, Inc. (2009). http://papers.nips.cc/paper/3650-zero-shot-learning-with-semantic-output-codes.pdf
9. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)
10. Toepfer, M., Seifert, C.: Descriptor-invariant fusion architectures for automatic subject indexing. In: Proceedings of Joint Conference on Digital Libraries (2017). Accepted
11. Wilbur, W.J., Kim, W.: Stochastic gradient descent and the prediction of MeSH for PubMed records. Proc. AMIA Ann. Symp. **2014**, 1198–1207 (2014)