Topic Cropping: Leveraging Latent Topics for the Analysis of Small Corpora

Nam Khanh Tran[1], Sergej Zerr[1], Kerstin Bischoff[1], Claudia Niederée[1], and
Ralf Krestel[2]

[1] Leibniz Universität Hannover / Forschungszentrum L3S, Hannover, Germany
`NTran@L3S.de`, `zerr@L3S.de`, `bischoff@L3S.de`, `niederee@L3S.de`
[2] Bren School of Information and Computer Sciences, University of California, Irvine
`krestel@uci.edu`

Abstract. Topic modeling has gained a lot of popularity as a means for identifying and describing the topical structure of textual documents and whole corpora. There are, however, many document collections such as qualitative studies in the digital humanities that cannot easily benefit from this technology. The limited size of those corpora leads to poor quality topic models. Higher quality topic models can be learned by incorporating additional domain-specific documents with similar topical content. This, however, requires finding or even manually composing such corpora, requiring considerable effort. For solving this problem, we propose a fully automated adaptable process of *topic cropping*. For learning topics, this process automatically tailors a domain-specific Cropping corpus from a general corpus such as Wikipedia. The learned topic model is then mapped to the working corpus via topic inference. Evaluation with a real world data set shows that the learned topics are of higher quality than those learned from the working corpus alone. In detail, we analyzed the learned topics with respect to coherence, diversity, and relevance.

Keywords: digital humanities, qualitative data, topic modeling

1 Introduction

For social sciences, sharing qualitative primary data like interviews and re-using it for secondary analysis is very promising as data collection is very time consuming. Moreover, some qualitative data sources capture valuable information about attitudes, beliefs, etc. as people had them at other times – “realities” that cannot be captured anymore. Enabling secondary analysis of data not collected by oneself, analyzing it with new research questions in mind, imposes a lot of challenges though. In this paper, we focus on the aspect of advanced techniques for facilitating exploration of such data and for improving findability in digital data archives. Supporting intelligent access to and exploration of data shared for re-use is also a main goal within the digital humanities as expressed, for example, in the theme of the Digital Humanities 2013 conference: “Freedom to Explore”.

By exploiting information retrieval and topic modeling techniques we can mine additional knowledge about themes discussed in primary qualitative data.

This way, interview contents can be visualized by means of extracted topics to give a quick overview. For example, topics extracted from a collection of studies, or samples show the commonalities of themes while comparing topics of individual studies, or samples sheds light on the specifics. Interview topics as well aid an enhanced (automatic) content analysis and retrieval of similar documents. This is especially interesting as qualitative documents are often long, and thus it is hard to grasp their thematic coverage – let alone to manually analyze them.

Due to the enormous resources required for conducting qualitative research by means of interviews (holding the interview, transcription, document coding/analysis), the primary data resulting from such qualitative studies is usually limited to a small number of interviews per study case or sample. Topic models, however, are based on statistics and thus perform better on big data sets (see, e.g. [1]). Here, we present a generalizable framework for using topic modeling given such corpora restrictions as they occur in qualitative social science research. Our fully automated adaptable process tailors a domain-specific Cropping corpus by collecting relevant documents from a general corpus or knowledge base, here Wikipedia. The topic model learned on this substitute corpus is then applied to the original collection. Hence, we exploit state-of-the-art IT-methods adapting and integrating them for usage as research tools for the digital humanities. In detail, the contributions of this paper are:

- We propose a process for *topic cropping* and proof its improved performance for small corpora by analyzing diversity, coherence, and relevance.
- By integrating the automatic evaluation of topic quality we take a first step towards a self-optimizing process of selecting parameters for topic cropping in different settings.

2 Related Work

Tools for (Secondary) Analysis of Qualitative Data: Regarding software tools and techniques for supporting the (re-)analysis of qualitative data usually three groups are differentiated. Qualitative data analysis (QDA) tools like ATLAS.ti, MaxQDA, or Nvivo are well developed products enabling the manual coding, annotation, and linking of data in a variety of formats. Other common features are simple search procedures, the definition of variables, automatic coding of specified text strings, and word frequency or co-occurrence counts.

More advanced are tools for (quantitative) content analysis, e.g., General Inquirer, Diction, LIWC, TextPack, WordStat. Software in this category usually builds upon large dictionaries to analyze vocabulary use also semantically. Besides word frequencies, category frequency analysis as well as statistics or filtering for keywords in contexts (KWIC / concordance) are typical features. Programs may offer co-occurrence or correlation analysis of categories or words, ideally accounting for synonyms via the built-in dictionaries. Related is cluster analysis and multidimensional scaling for visualizing word or category correlations. Dictionaries can also be used for normative comparison, i.e., to find specifics of vocabulary usage in a document or a collection [2].

Text mining and statistical analysis are advanced techniques exploited to automatically find themes and trends in qualitative data. Tasks are, for example, supervised document classification requiring human input for the label or variable value to be learned, unsupervised clustering of similar documents, or document summarization. Various algorithms as well as standard data preprocessing procedures (stemming, stop word removal, etc.) exist. Information extraction, e.g., of sentiment, can be achieved via lexicons, patterns, and rules. To name just a few – mostly commercial – tools that (claim to) provide additional text mining capabilities: Catpac, SAS Text Miner, SPSS TextSmart, WordStat.

In [3], the usage of unsupervised learning methods for qualitative data analysis is discussed, here a self-organizing map (SOM) build upon manually selected terms from interviews. The authors argue that such text mining procedures can aid both data-driven, inductive research by finding emergent concepts as well as theory-driven, deductive research by checking the adequacy and applicability of defined schemes. The next section reports in detail on work regarding the related goal of topic modeling for qualitative data – the focus of this paper.

Topic Modeling: Topic modeling is a generative process that introduces latent variables to explain co-occurrence of data points. Latent Dirichlet allocation (LDA) [4] is a further development of probabilistic latent semantic analysis (PLSA) [5]. LDA was developed in the context of large document collections, such as scientific articles, news collections, etc. The success of LDA led to the application in other domains, such as image processing, as well as other types of documents, e.g. tweets [6] or tags [7]. Some work applies topic modeling to transcribed text. In [8], the standard LDA model is extended to identify not only topics but also topic boundaries within longer meeting transcripts. The authors show that topic modeling can be used to detect segments in heterogeneous text. Howes et al. [9] investigate the use of topic models for therapy dialog analysis. More specifically, LDA is applied to 138 transcribed therapy sessions to then predict patient symptoms, satisfaction, and future adherence to treatment using latent topics detected vs. hand coded topics. The authors find only the manually assigned topics to be indicative. Human assessment of the interpretability of the automatically learned topics showed high variance of topic coherence.

Using topic models where there is only limited data, e.g., very short documents or very few documents, has been studied as well. Micro-blogging services, such as Twitter, limit single documents to 140 tokens. Hong and Davison [6] study different ways to overcome this limitation when training topic models by aggregating these short messages based on users or terms. The resulting longer documents yield better topic models compared to training on short, individual messages. Unfortunately, this method only works if the number of short texts is sufficiently large. Using additional long documents to improve topics used for classification was proposed in various approaches: Learning a topic model from long texts and then applying it to short text [10] improves significantly over learning and applying it on short texts only. Learning it on both [11] and applying it on short texts improves performance further. Jin et al. [12] present their Dual LDA model to model short texts and additional long text explicitly, which outperforms standard LDA on long and short texts for classification. Our focus is not on classification of short documents but we use topic modeling to analyze (long) individual documents and focus more on a careful selection of the corresponding training corpus. Incorporating domain knowledge for topic transition detection using LDA as described in [13] addresses this problem using manual selection of training corpora. A topic model is trained using auxiliary textbook chapters and is used to compare slide content and transcripts of lectures. Because of sparse text on slides and possible speech recognition errors in the transcripts training a topic model on long, related documents improves alignment of slides and transcript significantly. In contrast, our method does not rely on a manual selection of a training set as cropping is performed as an automated process.

3 A General Approach for Topic Cropping

The goal of our approach is to enable the exploitation of the advantages of topic models, e.g., with respect to capturing latent semantics, even if the considered corpus is too small for their direct application. Smaller corpora such as qualitative studies in the humanities result in topic models of restricted quality. The approach we are following in this work is to use another larger corpus (the Cropping corpus) for learning the topic model. Subsequently, the learned topic model is applied to the study under consideration via topic inference. Qualitative studies are often very focused, which makes finding a good Cropping corpus a difficult task. Since we are looking for an approach, which is applicable in different settings (i.e., for studies in different application domains), there are two requirements to be satisfied: (1) having a Cropping corpus that is specific enough to produce a good and useful coverage of the topics in the study under consideration (2) while avoiding the effort of searching for an adequate Cropping corpus whenever working with studies in a new application domain.

For this purpose, we decided to include into the automated process of topic cropping a phase for analyzing the working corpus coverage and a phase of automatic corpus tailoring. The tailoring phase creates a tailored domain-specific corpus from a large corpus with a very wide coverage such as Wikipedia. This implies a four step process for topic cropping (see also Figure 1):

1. Analyzing working corpus coverage by selecting characteristic terms
2. Tailoring a Cropping corpus by collecting relevant documents
3. Learning a topic model from the Cropping corpus
4. Applying topic inference to the working corpus

This process is embedded into a generalizable framework, which can be adapted to different settings via parameters. The final aim is to learn those parameters of the process steps in a self-optimizing loop.

Analyzing Working Corpus Coverage: For tailoring the Cropping corpus, we first have to understand the topical coverage of the corpus under consideration. At first glance, this might look like a hen-egg problem: we need to know the main topics of the corpus for building a corpus for learning those topics.

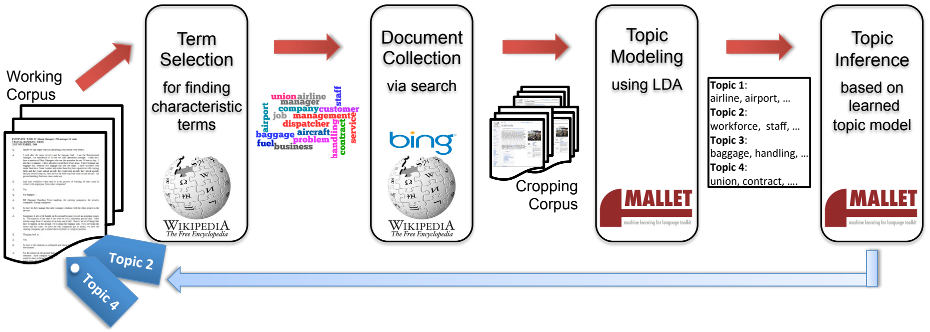

Fig. 1: Workflow for Topic Modeling on a Cropping corpus

For overcoming this, we relied on a method for determining the most relevant terms by using a counter corpus. Starting from a particular case in the study under consideration and a random subset of pages selected from Wikipedia, we used the metric of Mutual Information (MI) [14], which measures how much the joint distribution of terms deviates from a hypothetical distribution in which features and categories (working corpus and Wikipedia corpus in our case) are independent of each other. The measure ranks higher terms which are frequent in the working corpus but not in general. They are used as representative terms for corpus coverage.

Tailoring a Cropping Corpus: The top-ranked subset of those terms is used for tailoring the Cropping corpus. In our approach, we used a general Web search engine to identify the set of highest ranked Wikipedia pages for each of the terms. The Cropping corpus is created from the set union of all those pages. Wikipedia has been selected as the starting point for Cropping corpus creation because of its broad coverage providing information on seemingly every possible topic. Of course it is also possible to use large domain specific corpora or combinations of several corpora.

Learning the Topic Model: For learning the topic model, we made use of the Mallet topic modeling toolkit [15], namely the class ParallelTopicModel. This class offers a simple parallel threaded implementation of LDA (see [16]) together with SparseLDA sampling scheme and data structure from [17]. LDA models documents as probabilistic combinations of topics $P(z|d)$, with each topic described by terms following another probability distribution i.e. $P(w|z)$.

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

where $P(w_i)$ is the probability of the ith word for a given document and z_i is the latent topic. $P(w_i|z_i = j)$ is the probability of w_i within topic j. $P(z_i = j)$ is the probability of picking a word from topic j in the document. These probability distributions are specified by LDA using Dirichlet distributions. The number of latent topics T has to be defined in advance and allows to adjust the degree of specialization of the latent topics. For inference and parameter estimation, Gibbs sampling iterates multiple times over each word w_i in document d_i, and samples a new topic j for the word based on the probability $P(z_i = j|w_i, d_i, z_{-i})$ until the LDA model parameters converge.

Applying the Topic Model: In this step the topic model learned from the Cropping corpus is applied to the working corpus using topic inference as offered by the Mallet toolkit (cc.mallet.topics.TopicInferencer). It is not expected that the set of topics learned from the Cropping corpus is exactly the set of topics inherently included in the working corpus. Rather, the set of topics learned from the Cropping corpus is roughly a superset of the working corpus topics. Learned topics that are not available in the working corpus will however have no major impact on the topic inference process as long as the ”real” working corpus topics are also in the learned topic model. Topic inference will assign to each of the topics in the topic model a probability of it being relevant for a study document.

4 Experiments

4.1 Dataset

For our experiments, we re-used qualitative data shared via the ESDS Qualidata / the UK Data Service. We selected four out of the eight cases from the case study on “Changing Organizational Forms and the Re-shaping of Work” [18]. Each case has verbatim transcriptions or summaries of in-depth Face-to-face interviews conducted in England and Scotland between 1999 and 2002.

1. *Airport case*: four airlines, engineering department, airport security, baggage handling, full handling, cleaning company, fire service (30 files)
2. *Ceramics case*: five ceramics manufacturers (32 files)
3. *Chemicals case*: a pigment manufacturing plant, two Suppliers, two Transportation specialists, two Business Service Contractors (28 files)
4. *PFI case*: Hotel Services Company, Facilities Design Company, Special Purpose Vehicle, NHS Trust Monitoring Team (41 files)

Interviews were held in semi-structured form given guidelines for questions along the main research themes of managing, learning and knowledge development, experience of work, and performance – particularly investigating the links between these topics and changing organizational forms[3]. Participants were managers and employees at all levels, sometimes also union representatives. The number of pages per document varies between two and 32 for verbatim transcripts, summaries are usually of two to ten pages in length. These interview documents consist of transcribed spoken, natural language with answers being usually short, often elliptic, and requiring co-text and context for interpretation.

4.2 Experimental Settings

For tailoring the Cropping corpus we used the top 20 most representative terms as identified in the working corpus analysis phase. The Bing Search engine was

---

[3] For more details see: `http://discover.ukdataservice.ac.uk/catalogue?sn=5041`

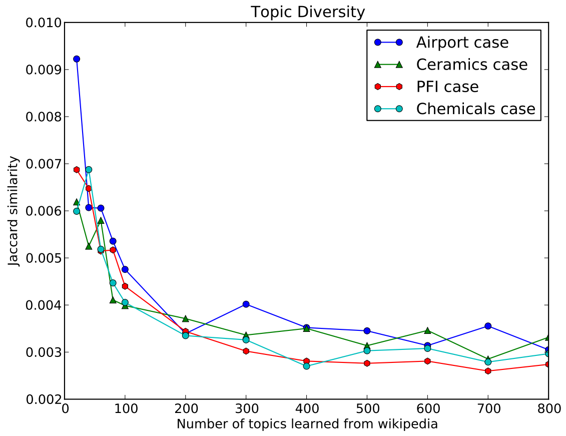

Fig. 2: Topic diversity, measured via Jaccard similarity for various number of topics learned from the Cropping corpus

queried for each of those terms individually to retrieve relevant Wikipedia pages. This resulted in a Cropping corpus of about 10,000 documents.

An important parameter in learning the topic model is the number of topics to be learned. With an increasing number of topics – a parameter of the topic model learning process – the topics get more fine-grained. The challenge here is to find a number, which results in good topic coverage for the study (all relevant topics are in) and in sufficiently fine-grained topics to help exploring unknown qualitative material while still being useful for human understanding and for spotting areas with similar topics. There is no general notion of a ”good” number of topics since this strongly depends on the corpus and the application. We decided to take topic diversity as a measure for an appropriate number of topics, more precisely the diversity of the topics assigned to the study based on the topics learned from the Cropping corpus. The intuition behind this is that we need a sufficiently large topic model to cover all aspects of the study. Once the diversity stops increasing substantially the newly added topics are either not relevant for the study or they just provide subtopics by splitting topics, which does not substantially add to the diversity. Figure 2 shows the increase in topic diversity for various numbers of topics learned from the Cropping corpus. For this topic inference we used a threshold of 0.01 to cut out “noisy” topics with very low probabilities. Figure 2 is discussed in more detail in the next section.

5 Evaluation

We judge the quality of the automatically detected topics exploiting both, internal (intrinsic) and external (extrinsic) evaluation [14, 19]. In topic analysis an internal evaluation prefers low similarity between topics whilst within a topic high similarity is favored. We adopt this idea by measuring *topic diversity* capturing variance between the different topics in a model and *topic coherence* within the single topics respectively. We additionally measure *topic relevance* externally by comparing with human annotators. In this section, we evaluate both the topics learned directly from the working corpus and those from the Cropping corpus with the same setting and analyze them with respect to these quality dimensions.

5.1 Topic Diversity

Topic diversity is an important criterion for judging the quality of a learned model. The more diverse, i.e. dissimilar, the resulting topics are, the higher will be the coverage regarding the various aspects talked about in our interview data. It has been shown in earlier work that the Jaccard Index is an adequate proxy for diversity [20] and its output value correlates with a number of clusters (topics in our case) within the dataset. Thus, to estimate the average similarity between produced clusters, we employ the popular Jaccard coefficient [14]. Given two topic models T_i and T_j, i.e. set of terms, their Jaccard similarity $JS(T_i, T_j)$ is defined as follows:

$$JS(T_i, T_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|}.$$

Given a collection of topic models $T_1, \ldots, T_n$, the refined (excluding self-similar pairs) average Jaccard similarity [20] is defined as follows ($1 \leq i < j \leq n$):

$$sim = \frac{2}{n(n-1)} \sum_{i<j} JS(T_i, T_j),$$

For alle available cases, Figure 2 plots topic diversity with respect to the number of inferred topics. We observe that similarity values sharply decrease until the number of topics reaches the range 80-100. They do not substantially change in the tail. This may be an indicator for a reasonable number of topics for our datasets. Similarly, Figure 3 shows the change of the average Jaccard similarity, comparing the diversity of topics learned from the working and the Cropping dataset. We observe that topics learned from the Cropping corpus are generally more diverse in the beginning of the curve, indicating that our approach covers more aspects of the data even for smaller number of topics.

5.2 Topic Coherence

We tackle the task of topic coherence evaluation by rating coherence or interpretability based on an adaptation of the Google similarity distance, which performs effectively in measuring similarity between words [21]. The more similar, i.e less distant, the representative words within a topic, the higher or easier is its interpretability. Cilibrasi and Vitanyi's *normalized Google distance* (**NGD**) function measures how close word x is to word y on a zero to infinity scale using the formula:

$$\mathbf{NGD}(x, y) = \frac{max\{logf(x), logf(y)\} - logf(x, y)}{logM - min\{logf(x), logf(y)\}}$$

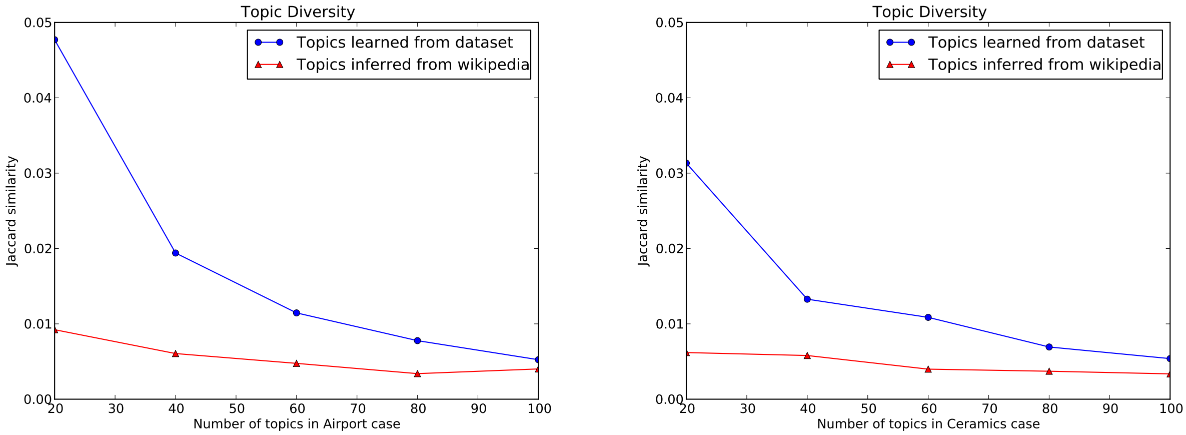

Fig. 3: Topic diversity, measured via Jaccard similarity, and its variance for different numbers of topics learned during topic modeling.

Table 1: Example topics with coherence measured via normalized Google distance (NGD), topics inferred from the working corpus (W) or the Cropping corpus (C).

Corpus	**Topics**	**NGD**
W	bag day company baggage ramp	0.44
W	airline service issue baggage handling	0.38
C	workers labor work employment workforce	0.19
C	employee employees tax employer pay	0.19

Table 2: Average (Avg) and standard deviation (SD) of topic coherence of three cases, measured via normalized Google distance (NGD). Topics are inferred from the working corpus (W) or the Cropping corpus (C).

Case	AvgNGD$_W$	SD$_W$	AvgNGD$_C$	SD$_C$
Airport	0.34	0.07	0.21	0.08
Ceramics	0.32	0.08	0.25	0.09
Pfi	0.35	0.1	0.22	0.08

where $f(x)$ and $f(y)$ are the number of hits of words x and y, respectively, $f(x,y)$ is the page-counts for the query $x\ AND\ y$ and M is the total number of web pages that Google indexes. A NGD of zero indicates that word x and word y are practically the same. They are independent when their distance reaches approximately one.

Given a topic T which is represented by its top-m words (we set m=5 in this experiment) denoted by $\mathbf{w} = (w_1, ..., w_m)$, its normalized Google distance is:

$$\mathbf{NGD}(T) = \frac{2}{m(m-1)} \sum_{w_i, w_j \in \mathbf{w}} \mathbf{NGD}(w_i, w_j)$$

To estimate overall topic coherence, we randomly choose a list of 30 learned topics per case ($T = (T_1, ..., T_n)$), compute NGD for each T_j, and then take the average of the list $\mathbf{AvgNGD}(T) = \frac{1}{n}\mathbf{NGD}(T_j)$.

Table 2 reports the average normalized Google distances and their deviations for topics inferred for three cases. For all cases evaluated, we obtain consistent improvement. Specifically, evaluating over the 90 topics of these three cases, we improve 32% in terms of normalized Google distance. This indicates that the topics inferred from the Cropping corpus are significantly more coherent than those learned directly from the working corpus (significance of a t-test $p < 0.001$).

5.3 Topic Relevance

While topic diversity and topic coherence can help to estimate the quality of the topics with respect to information-theoretic considerations, validity of our results, i.e., the usefulness of the derived topics for the working corpus, needs to be assessed by human evaluation of topic relevance. Here, we decided to compare our inferred topics with topics assigned by human annotators. For this evaluation, we randomly selected 16 documents from the study to be manually annotated by four users. Each document was split into smaller units – typically question and answer pairs – resulting in about 60 units per document. Thus, a total of 1000 units was annotated. We asked users to define topics discussed in each given unit. Each unit could have one or more topics and there were no restrictions on how topics are to be phrased. Typically the topics assigned were single words or short phrases.

Topic relevance is then assessed by automatically matching user defined topics with the learned ones. For this, the terms used by the user for a topic are matched with the top terms learned for a topic by the topic model. We consider it a match if the term used by the user appears in the top terms of the respective topic. By design, this evaluation gives preference to the topic model learned directly from the working corpus since the users tend to use terms that appear in the text. Similarly, the topic models learned directly on the working corpus use exactly those terms for their topics. In order to even out this terminology disadvantage, we made use of word synonyms from WordNet [22] to extend sets of topic words before matching. A learned topic T is considered to be relevant if its representative words and their synonyms $\mathbf{w} = (w_1, ..., w_k)$ share one or more terms with user defined topics $\mathbf{t} = (t_1, ..., t_r)$

$$\mathbf{Rel}(T) = \begin{cases} 1 & \text{if } \mid \mathbf{w} \cap \mathbf{t} \mid > 0 \\ 0 & \text{otherwise} \end{cases}$$

There are two reasons to use this type of evaluation in spite of its weakness: First, the alternative solution of showing the user the learned topic together with the text for relevance assessment puts a high burden on the user since it is not trivial to judge automatically learned topics. In addition, there is the risk that the user also unintentionally assesses topic quality in terms of coherence at the same time. Second, we are aiming for a self-optimizing loop, where parameters of the process are adapted iteratively through learning based on quality assessment. In this context, the evaluation of topic relevance chosen here only has to be done once and can be re-used in every iteration. The alternative manual evaluation

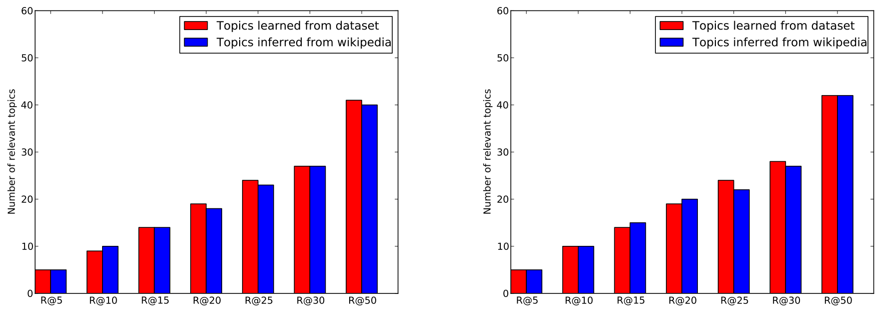

Fig. 4: Topic relevance as the number of relevant topics at rank k, for two documents

of the relevance of each learned topic as a whole would have to be repeated in every loop to assess the newly learned topics.

For two example documents, Figure 4 compares topics learned from the working and Cropping corpus with respect to the number of relevant topic at rank k $\mathbf{R}@k = \sum_{i=1}^{k} \mathbf{Rel}(T_i)$, where the rank is determined by the probability of the topic assignment (resulting from topic inference). We achieve similar results for other documents. On average, at rank 10 we obtain 9.8 relevant topics with a deviation of 0.35 for the working topics and 9.2 with a deviation of 1.0 for the Cropping topics. It can be seen from the results that the topics learned from Wikipedia reach a comparable level of relevance as those learned directly from the corpus, while being more coherent and diverse.

6 Conclusion and Future Work

In this paper we propose a method for a *fully automated* and adaptable process of tailoring a domain-specific sub-corpus from a general corpus such as Wikipedia and exploiting it to increase the topic model quality for limited size corpora such as studies in sociology and other qualitative material in the digital humanities. Our experiments show substantial improvements in diversity as well as in internal coherence of inferred topics compared to a naive approach using the limited size corpora exclusively. At the same time our method keeps the topic relevance high as confirmed by human annotators. We believe that our approach can be further improved by exploiting the automatic evaluation for adjusting the input parameters of the algorithm. In future work, we plan to modify the approach towards a self-optimizing automatic cycle. One important task, therefore, is to develop a more precise automatic evaluation of topic relevance through matching the user-annotated and the automatically inferred topics.

Acknowledgments. We are greatly thankful to M. Marchington, J. Rubery, and H. Willmott for sharing their primary data and to the UK Data Archive for making qualitative data available for (secondary) research. The work was supported by the project *"Gute Arbeit" nach dem Boom (Re-SozIT)* funded by the German Federal Ministry of Education and Research (BMBF) under mark 01UG1249C and by the European projects ForgetIT (GA600826) and ARCOMEM (GA270239). Responsibility for the contents lies with the authors.

References

1. Newman, D., Bonilla, E.V., Buntine, W.: Improving topic coherence with regularized topic models. In: Proceedings NIPS. (2011) 496–504
2. Leetaru, K.H.: Data Mining Methods for the Content Analyst: An Introdution to the Computational Analysis of Content. Routledge, New York, USA (2012)
3. Janasik, N., Honkela, T., Bruun, H.: Text mining in qualitative research: Application of an unsupervised learning method. Organizational Research Methods **12**(3) (2009) 436–460
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
5. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings UAI. (1999) 289–296
6. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings 1st Workshop on Social Media Analytics. SOMA (2010) 80–88
7. Krestel, R., Fankhauser, P., Nejdl, W.: Latent Dirichlet Allocation for Tag Recommendation. In: Proceedings RecSys. (2009) 61–68
8. Purver, M., Körding, K.P., Griffiths, T.L., Tenenbaum, J.B.: Unsupervised topic modelling for multi-party spoken discourse. In: Proceedings ACL. (2006) 17–24
9. Howes, C., Purver, M., McCabe, R.: Investigating topic modelling for therapy dialogue analysis. In: Proceedings IWCS Workshop on Computational Semantics in Clinical Text (CSCT). (2013) 7–16
10. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings WWW. (2008) 91–100
11. Xue, G.R., Dai, W., Yang, Q., Yu, Y.: Topic-bridged plsa for cross-domain text classification. In: Proceedings SIGIR. (2008) 627–634
12. Jin, O., Liu, N.N., Zhao, K., Yu, Y., Yang, Q.: Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings CIKM. (2011) 775–784
13. Zhu, X., He, X., Munteanu, C., Penn, G.: Using latent dirichlet allocation to incorporate domain knowledge for topic transition detection. In: Proceedings INTERSPEECH. (2008) 2443–2445
14. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
15. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
16. Newman, D., Asuncion, A.U., Smyth, P., Welling, M.: Distributed algorithms for topic models. Journal of Machine Learning Research **10** (2009) 1801–1828
17. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: Proceedings KDD. (2009) 937–946
18. Marchington, M., Rubery, J., Willmott, H.: Changing organizational forms and the re-shaping of work : Case study interviews, 1999-2002 [computer file] (2004)

19. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Proceedings Human Language Technologies. HLT (2010) 100–108
20. Deng, F., Siersdorfer, S., Zerr, S.: Efficient jaccard-based diversity analysis of large document collections. In: Proceedings CIKM. (2012) 1402–1411
21. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Trans. on Knowl. and Data Eng. **19**(3) (2007) 370–383
22. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11) (1995) 39–41