"Gute Arbeit": Topic Exploration and Analysis Challenges for Corpora of German Qualitative Studies

Nam Khanh Tran*, Sergej Zerr*, Kerstin Bischoff*, Claudia Niederée*, Ralf Krestel**
*L3S Research Center, Hannover, Germany
{ntran,zerr,bischoff,niederee}@L3S.de
**Bren School of Information and Computer Sciences, University of California, Irvine
krestel@uci.edu

ABSTRACT

Given their long-standing research traditions, a tremendous body of data has been collected in the social sciences by observing or interviewing people regarding their behavior, attitudes, beliefs, etc. The Sociological Research Institute (SOFI) in Göttingen (Germany) carried out a number of studies observing working situation in German automobile and shipyard industry after the rapid economic growth in post-World War II Germany - the so-called German "economic miracle". Qualitative data in form of worker interviews was collected during the period of over the last 40 years, starting from early 60's (i.e Volkswagen and German dockyard studies) and findings of these studies made a significant impact on the working situation in German industry. Intelligent access to this heritage of qualitative data would turn such data collection into a valuable source for a secondary research, e.g., for longitudinal (meta)analysis or historical investigations. By using modern information technologies the project "Gute Arbeit" aims at providing intelligent access to qualitative social science data on the subject of "good work". Topic modeling has gained a lot of popularity as a means for identifying and describing the topical structure of textual documents and whole corpora. However, when applied to the corpora directly, topic modelling leads to poor quality topic models due to the limited number of sociological surveys in our dataset.

In our previous work we proposed *topic cropping* a fully automated process for selecting and incorporating additional domain-specific documents with similar topical content which can expand a dataset and significantly improve the quality of inferred topic models. We tested our approach on thematically close English and German document corpora and investigated that the produced results for German corpora slightly outperformed those of the English dataset.

Keywords

digital humanities, qualitative data, topic modeling

1. INTRODUCTION

For social sciences, sharing of 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 (see e.g. studies in UK Data Archive). 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 topic exploration in German 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 it is, for example, expressed in the theme of the Digital Humanities 2013 conference: "Freedom to Explore". Figure 1 visualizes the main building blocks for intelligent support of secondary qualitative analysis envisaged in this project: Contextualization, Information Extraction, Opinion Mining/Sentiment Analysis, and Anonymization.

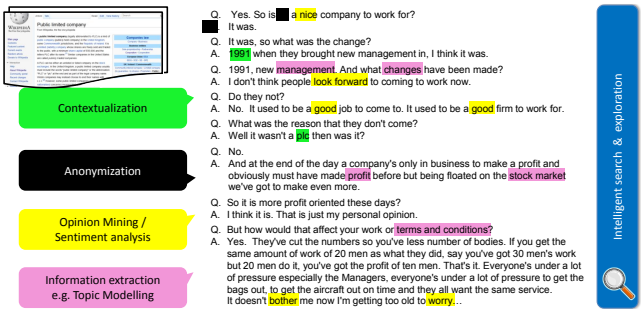

Figure 1: Modules supporting secondary analysis

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 for a quick overview. For example, topics extracted for a collection of studies, cases, or samples show the commonalities of themes while comparing topics of individual studies, cases 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 primary documents are often very long, and thus it is hard to easily grasp their thematic coverage – let alone to manually analyze and code them.

Due to the enormous resources required for conducting qualitative research by means of interviews (holding the interview, 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. [19]). In our recent work [23] we presented 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. Our previous experiments were conducted on a dataset of workers interviews in English language.
In this paper we present first topic cropping outcomes for the original German studies. Our results show a slight improvement in quality metrics for the German documents due to, as we believe, some properties of the German language such as wide usage compounds. We plan to evaluate the latter hypothesis in our future work

2. "GUTE ARBEIT": IT TOOLS

Ever since the period of rapid economic growth in post-World War II Germany the working environment has fundamentally changed. The rise of the service sector in industrialized countries, in particular, stimulated discussion about "subjectivization" of work, i.e. changes within post-tayloristic management strategies as well as altered work-ethics and attitudes. By re-analyzing data collected during more than four decades with the help of modern information technologies, Gute Arbeit studies how conceptions of "good work" evolved over time.

However, sharing and re-using data, e.g., for longitudinal (meta)analysis is not common practice so far. Gute Arbeit will enable intelligent access to such qualitative data gathered within diverse sociological studies regarding the workplace. For this, we will adapt and advance computational approaches from the fields of Information Retrieval and Data Mining, thus promoting the area of digital humanities. Gute Arbeit will contribute to the area of digital humanities by, amongst others, providing best practices and guidelines, tools and methodologies on how to facilitate re-use of sensitive primary data as well as on the exploitation of intelligent computational techniques for exploring them.

Mining additional knowledge from such valuable data, re-using it with new research questions in mind or sharing it with other researchers involves many challenges though. Besides warranting participants' anonymity, keeping and visualizing context is crucial to correctly interpret utterances of interviewees not surveyed by oneself. Since "good work" is a subjective concept, a major task will be the automatic extraction of topics and people's opinions regarding these topics. Here, the usefulness of popular text mining strategies like Topic Modeling and Sentiment Analysis has to be proven for the kind of material at hand: qualitative data from structured and unstructured interviews.

In secondary analysis, contextualization is crucial, thus it has taken the center stage in the debate so far. There are different kinds and levels of context of the interview, e.g., conversational, situational, regarding the research project, or institutional/cultural (see [1]). Here, we will focus on using external knowledge bases, e.g. Wikipedia or news corpora, to enrich primary data with background information on the socio-cultural context present at the time of data collection. Information extraction – here of Topics and Named Entities – will be beneficial for advanced exploration and navigation support, to get a quick overview over collection contents, to filter via corresponding facets, or to retrieve similar documents. Exploiting opinion mining for secondary analysis is especially interesting for our project as (1) the sociological research focuses on subjective conceptualizations of work and (2) we assume that opinionated passages with negative or critical statements about certain names or topics may receive special attention with respect to anonymization. In contrast to traditional software for qualitative text analysis, DigDeeper will offer various intelligent tools for searching and exploring (parts of) documents, samples, and collections.
Figure 2 gives an overview of the system architecture.

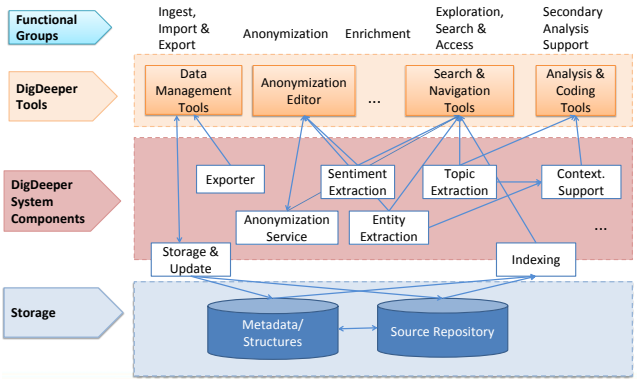

Figure 2: DigDeeper Architecture

Of course, the DigDeeper tool to be developed within the project will also support standard features for qualitative data analysis like coding, annotating, linking, and highlighting.

3. RELATED WORK

Tools for (Secondary) Analysis of Qualitative Data: When it comes to 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, Nvivo are well developed software products enabling the manual coding, annotation and linking of data in a variety of formats. Other features are simple search procedures, the definition of variables, automatic coding of specified text strings, and sometimes also visualization of co-occurrences are features or word frequency 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 build 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 [12].

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. Via lexicons, patterns and rules information extraction, e.g. of sentiment, can be achieved. To name just a few – mostly commercial – tools that (claim to) provide additional text mining capabilities: Catpac, SAS Text Miner, SPSS TextSmart, WordStat.

In [8], the usage of unsupervised learning methods is discussed, here a self-organizing map (SOM) build upon manually selected terms from interviews, for qualitative data analysis. They argue that such text mining procedures can aid both data-driven, inductive research by finding emergent categories/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) [2] is a further development of probabilistic latent semantic analysis (PLSA) [5] modeling documents using latent topics. LDA was developed in the context of large document collections, such as scientific articles, news collections, etc. with the goal of getting a quick topical overview. 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 [10].

There is also some work applying topic modeling to transcribed text. In [22], 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. [7] 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 [21] improves significantly over learning and applying it on short texts only. Learning it on both [24] and applying it on short texts improves further. Jin et al. [9] 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, 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 is described in [26] addresses this problem using manual selection of training corpora(s). A topic model is trained using auxiliary textbook chapters and 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 in an automated process. Applicability of topic modelling for multilingual IR were identified in [11] the authors attempt to construct accurate and comparable relevance models in the source and target language, and use that models to rank the documents in the target collection. The advantage of this approach is that it does not rely on a word-by-word translation of the query and the relevance of the target collection can be estimated more accurately. In [17] the authors proposed polylingual topic modelling using the Wikipedia interlinked pages. In this work we show that a language could be an important factor for topic modelling and topic cropping quality.

4. EXPERIMENTS

In our experiments we compared German and English document corpora. Both corpora comparably consist of qualitative sociological data, specifically surveys and interviews on topics related to working environment within different industrial areas.

4.1 English Dataset

This corpus consists of qualitative data shared for research purposes via the ESDS Qualidata / the UK Data Archive, which is currently moving to 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" [14]. Each case has verbatim transcriptions or summaries of in-depth Face-to-face interviews conducted in England and Scotland between 1999 and 2002. The study surveyed employees from inter-organizational networks as new organizational forms, analyzing how they operate in practice and focusing on the aspect of employment relationship.

- *Airport case*: four airlines, engineering department, airport security, baggage handling, full handling, cleaning company, fire service (30 files online)
- *Ceramics case*: five ceramics manufacturers (32 files)
- *Chemicals case*: a pigment manufacturing plant, two Suppliers, two Transportation specialists, two Business Service Contractors (28 files)
- *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[1]. For example, questions asked for how and why changes in organizational form arose and how much progress has been done on implementation. Regarding learning, interviewees were asked on knowledge and skills required for the jobs, on how and by whom training and learning is organized, or how customer/production pressures are handled. Subjective attitudes and experiences of work were captured via questions on changing patterns in and changing perceptions of team work, working time, pay, contracting, etc. For performance, definition of criteria at different levels, measurement and monitoring as well as source of performance pressure were talked about. In particular, the focus was on links between changing organizational forms and the four broader topics.

4.2 German Dataset

This corpus consist of qualitative data obtained during the time period 2001 - 2009 from the employee interviews of the vehicle manufacturing company "Auto 5000" which was set up inside the Volkswagen complex in Wolfsburg, Germany. This lower cost model company was set up aiming of keeping manufacturing jobs in Germany instead of moving production to other areas of Europe. The stuff was mainly composed of formerly unemployed people and those looking to have more flexible working hours.

The dataset is composed of three parts

- 19 individual interviews with skilled workers (2002)
- 14 individual interviews with production engineers (2003)
- 8 group discussions (2005)

Interviews include the employment history of the former unemployed workers and engineers, shift work and relations between the Volkswagen and "Auto 5000" employees. The average number of the pages per document is about 40.

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

4.3 Experimental Settings

For tailoring the Cropping corpus, we used top (selected by MI) representative terms identified in the Working corpus analysis phase. The terms were used individually to search for relevant Wikipedia pages using the Bing Search engine. 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, which is a parameter of the topic model learning process, the topics get ever more fine grained. The challenge here is to find a number, which results good topic coverage for the study (all relevant topics are in) and in sufficiently fine grained topics to help in 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 targeted application. We decided to take the diversity of the topics assigned to the study based on the topics learned from the Cropping corpus as a measure for a sufficient number of topics. The intuition behind this is that we need a sufficiently large topic model to cover all aspects of the study. As long as this is not yet reached the diversity still increases with the number of topics. 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.

5. A GENERAL APPROACH FOR TOPIC CROPPING

The goal of our approach described in [23] 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. To obtain this target, in recent work [23] we proposed the topic cropping workflow which is a four step process (see also Figure 3):

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

Analyzing Working Corpus Coverage: The goal of this step is 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. For overcoming this, we relied on a method for determining the most relevant terms by using a counter corpus and used the metric of Mutual Information (MI) [13], which measures how much the joint distribution of terms deviates from a hypothetical distribution in which features and categories are independent of each other. The measure ranks higher terms which are frequent in the working corpus but not in general.

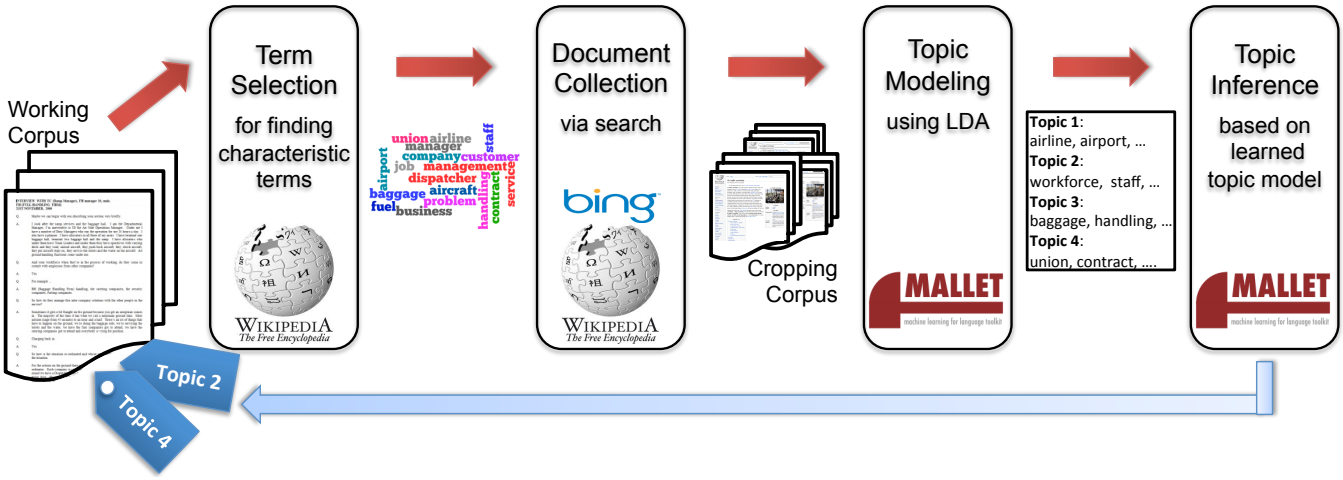

Figure 3: Workflow for Topic Cropping

Tailoring a Cropping Corpus: The top-ranked subset of those terms is used for tailoring the Cropping corpus. 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: 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 [18]) together with SparseLDA sampling scheme and data structure from [25]. LDA is based on a generative probabilistic model that models documents as mixtures over an underlying set of topic distributions.

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

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.

Applying the Topic Model: Using learned models from the previous step, we determine the topics for working corpus using topic inference as offered by the Mallet toolkit (cc.mallet.topics.TopicInferencer). It assigns to each of the topics in the topic model a probability of it being relevant for a study document. As stated in [23], 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. We analyze this issue further in Section 6.

6. EVALUATION

We judge the quality of the automatically detected topics exploiting both, internal (intrinsic) and external (extrinsic) evaluation [13, 20]. 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.

6.1 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 [4] 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 [13].

Figure 4 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 already in the beginning of the curve, indicating that our approach covers more aspects of the data even for smaller number of topics.

6.2 Topic Coherence

We tackle the task of topic coherence evaluation by rating coherence or interpretability based on an adaptation of the Google similarity distance (NGD), which performs effectively in measuring similarity between words [3]. The more similar, i.e less distant, the representative words within a topic, the higher or easier is its interpretability (see details in [23]).

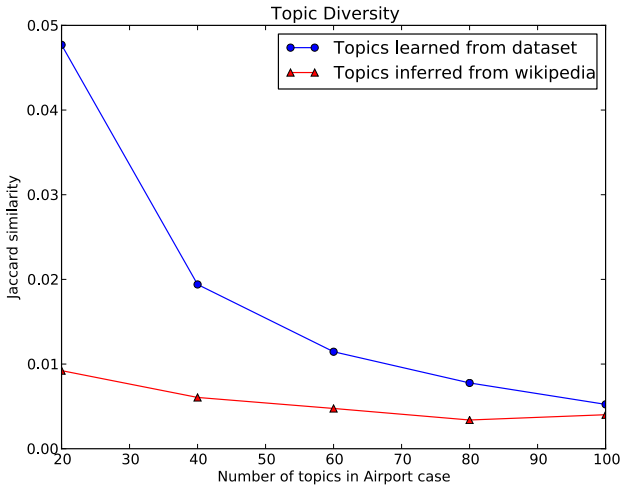

Figure 4: Topic diversity, measured via Jaccard similarity, and its variance for different numbers of topics learned during topic modeling.

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

	Topics	**NGD**
English		
W	bag day company baggage	0.44
W	airline service issue baggage	0.38
C	workers labor work employment	0.19
C	employee employees tax employer pay	0.19
German		
W	hörensagen standard schweißerpass block	0.60
W	antwort endeffekt wolfsburg gmbh	0.43
C	arbeitnehmer arbeitgeber gewerkschaften arbeit	0.19
C	unternehmen management ergebnisse mitarbeiter	0.32

For example, for the topic T_i that is presented by a list of words {airline, service, issue, baggage}, its NGD is determined by the average of the scores of all possible word pairs {(airline, service), (airline, issue), (airline, baggage), ...} (see also Table 1).

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 English cases as reported in [23] and for the one German case (Auto5000). For both corpora and all cases evaluated, we obtain consistent improvement. This indicates that the topics inferred from the German and English Cropping corpus are also significantly more coherent than those only learned directly from the working corporas (measured significance of a t-test $p < 0.001$).

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
Auto 5000	0.38	0.09	0.29	0.08

6.3 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 and 8 documents from English and German corpora respectively to be manually annotated by five users. Each document was split into smaller units – typically question and answer pairs – resulting in about 60 units per document. Thus, a total of 1500 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, for English dataset we made use of word synonyms from WordNet [16] to extend sets of topic words before matching. Due to the lack of German WordNet and the language property, we compute these scores for German corpus without any synonym extensions. 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 $\mathbf{t} = (t_1, ..., t_r)$

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

Figure 5 compares topics learned from the documents in the English 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). Similarly, Figure 6 presents the relevance results for the German corpus. 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.

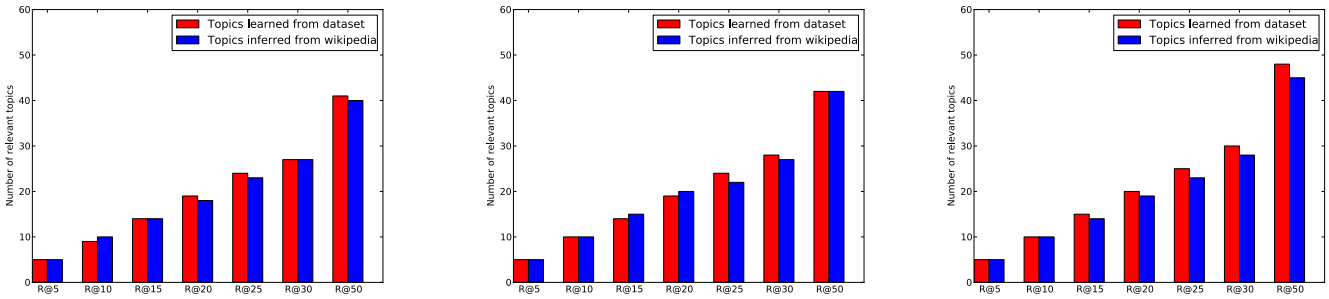

Figure 5: Topic relevance as the number of relevant topics at rank k, for the documents in English dataset

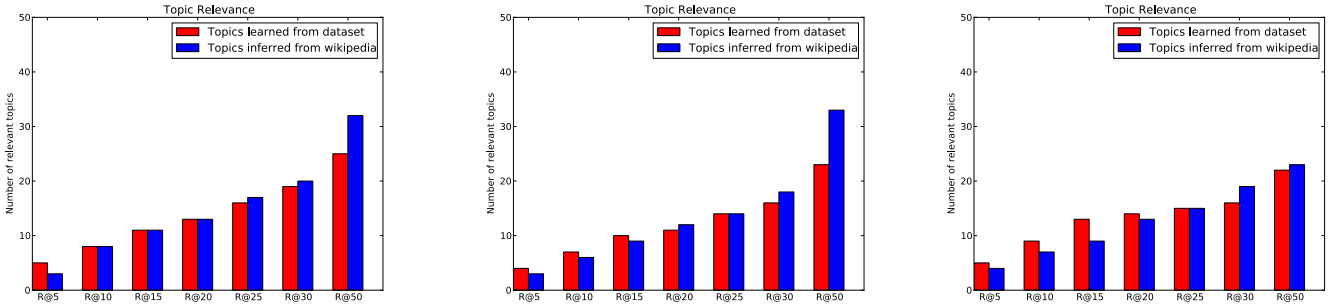

Figure 6: Topic relevance as the number of relevant topics at rank k, for the documents in German dataset

Table 3: Example topics inferred from the German working corpus (W) and the Cropping corpus (C).

Corpus	Topics
W	magdeburg erwartungshaltung april fliessbandarbeit rad blechteile ruck garderlingen bildungsträgern
W	art endeffekt umfeld niveau stendal nummer automobilbauer not bahn
W	antwort fachtalent autos umschulung mal jungs band hammer mechaniker
W	test aufgaben gestaltungswerkzeuge kinderbetreuung leistung maß wissen maschine leiter
W	monat ahnung bereich lack betriebsingenieur brief band fachwissen halle
C	dresden chemnitz dresdner zwickau sachsen radebeul clear style div
C	französischen paris frankreich französische saint jean louis dreyfus les
C	münchen deutscher geboren hans karl friedrich archäologe august verstorben
C	film filme films rolle regisseur filmen schauspieler regie
C	formula verfahren test unternehmen methoden management ergebnisse mitarbeiter methode

6.4 Results for the German Corpus

We conducted the Topic Cropping Procedure for the German corpus and obtained comparable results. Also in this case both, the diversity between topics and the coherence within each topic were increased. Additionally we noticed a slight increase in the relevance, meaning that inferred topic slightly better reflected the users annotations of German compared to the English dataset. In this paper we report the results and let the further investigations about the reason for the increase to the future work. A hypothesis is that the improvement could be due to a particular property of the German language the use of compounds. Compound is a word which consists of more than one word. English examples of compounds are words like: “smalltalk”, “makeup”, “notebook” and so on. In German language it is usual to use compounds and create them “on-the-fly”, if necessary. The English phrase “car body pressing” turns into a single word in German - “Karosseriebau”. Considering topic cropping strategy, the German word is more concise, as a query resulting in documents well focused on the particular topic. Compared to English case, the terms “car”, “body”, “pressing”, each would give a large number of noise documents when searching in Wikipedia. This hypothesis is supported by the fact that German queries in our experiments generally led to less Wikipedia pages, however the relevance of the pages was obviously higher compared to English dataset using our relevance annotation measure described in 6.3.

The Table 3 provides some examples for topics obtained from the working corpus only("W") and the Cropping corpus ("C"). In the future work we plan to evaluate the outcomes in more details, however already on the first glance, for German speaking person it will be more difficult to identify and label the topics obtained using straight forward modelling on original dataset compared our cropping approach. For example, the labels for "C" could be (top-down) "East Germany", "France", "archaeology", "movie", "company management", whether the labels for "W" are hard to extract.

7. CONCLUSION AND FUTURE WORK

In our recent work [23] we proposed a method for a *fully automated* adaptable process of tailoring a domain-specific sub-corpus from a general corpus (e.g. Wikipedia).

In this paper we present first results of an application of our Topic Cropping approach within the German national BMBF Project "Gute Arbeit" and a large scale qualitative interviews in German language. Our experiments show slight improvements of the results for German dataset and it seems it is due to some specific language properties.

We believe that wide usage of compounds in German language can lead selecting more concise representative query terms for the related document search in Topic Cropping and plan to investigate these details in our future work.

8. ACKNOWLEDGMENTS

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) (01UG1249C) and by the European project ForgetIT (GA600826). Responsibility for the contents lies with the authors.

9. REFERENCES

[1] L. Bishop. A proposal for archiving context for secondary analysis. *Methodological Innovations Online*, 1(2):10–20, 2006.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, 2007.

[4] F. Deng, S. Siersdorfer, and S. Zerr. Efficient jaccard-based diversity analysis of large document collections. In *Proceedings CIKM*, pages 1402–1411, 2012.

[5] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings UAI*, pages 289–296, 1999.

[6] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings 1st Workshop on Social Media Analytics*, SOMA, pages 80–88, 2010.

[7] C. Howes, M. Purver, and R. McCabe. Investigating topic modelling for therapy dialogue analysis. In *Proceedings IWCS Workshop on Computational Semantics in Clinical Text (CSCT)*, pages 7–16, 2013.

[8] N. Janasik, T. Honkela, and H. Bruun. Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods*, 12(3):436–460, 2009.

[9] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings CIKM*, pages 775–784, 2011.

[10] R. Krestel, P. Fankhauser, and W. Nejdl. Latent Dirichlet Allocation for Tag Recommendation. In *Proceedings RecSys*, pages 61–68, 2009.

[11] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proceedings of SIGIR '02*, New York, NY, USA, 2002.

[12] K. H. Leetaru. *Data Mining Methods for the Content Analyst: An Introdution to the Computational Analysis of Content*. Routledge, New York, USA, 2012.

[13] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[14] M. Marchington, J. Rubery, and H. Willmott. Changing organizational forms and the re-shaping of work : Case study interviews, 1999-2002 [computer file], 2004.

[15] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[16] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[17] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. Mccallum. Polylingual topic models. In *In EMNLP*, 2009.

[18] D. Newman, A. U. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

[19] D. Newman, E. V. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In *Proceedings NIPS*, pages 496–504, 2011.

[20] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Proceedings Human Language Technologies*, HLT, pages 100–108, 2010.

[21] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings WWW*, pages 91–100, 2008.

[22] M. Purver, K. P. Körding, T. L. Griffiths, and J. B. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings ACL*, pages 17–24, 2006.

[23] N. K. Tran, S. Zerr, K. Bischoff, C. Niederée, and R. Krestel. Topic cropping: Leveraging latent topics for the analysis of small corpora. In *Proceedings of TPDL 2013*, LNCS. Springer: to appear., 2013.

[24] G.-R. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged plsa for cross-domain text classification. In *Proceedings SIGIR*, pages 627–634, 2008.

[25] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings KDD*, pages 937–946, 2009.

[26] X. Zhu, X. He, C. Munteanu, and G. Penn. Using latent dirichlet allocation to incorporate domain knowledge for topic transition detection. In *Proceedings INTERSPEECH*, pages 2443–2445, 2008.