Preprint from: https://www.gipp.com/pub/

A. Zhukova, F. Hamborg, and B. Gipp, "Interpretable and Comparative Textual Dataset Exploration Using Near-Identity Mention Relations," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2020.

Interpretable and Comparative Textual Dataset Exploration Using Near-Identity Mention Relations

Anastasia Zhukova[1], Felix Hamborg[2], Bela Gipp[1]

[1] University of Wuppertal, Germany
{lastname}@uni-wuppertal.de

[2] University of Konstanz, Germany
felix.hamborg@uni-konstanz.de

ABSTRACT

Dataset exploration is a set of techniques crucial in many research and data science projects. For textual datasets, commonly used techniques include topic modeling, document summarization, and methods related to dimension reduction. Despite their robustness, these techniques suffer from at least one of the following drawbacks: document summarization does not explicitly set documents in relation, the others yield summaries or topics that often are difficult to interpret and yield poor results for topics that consist of context-dependent terms. We propose a method for dataset exploration that employs cross-document near-identity resolution of mentions of semantic concepts, such as persons, other named entity types, events, actions. The method not only sets documents in relation and thus allows for comparative dataset exploration, but also yields well interpretable document representations. Additionally, due to the underlying approach for cross-document resolution of concept mentions, the method is able to set documents in relation as to their near-identity terms, e.g., synonyms that are not universally valid but only in the given dataset.

ACM Reference format:

Anastasia Zhukova, Felix Hamborg, and Bela Gipp. 2020. Interpretable and Comparative Textual Dataset Exploration Using Near-Identity Mention Relations. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL'20), Virtual Event, China, August 2020,* 2. Pages.

1 INTRODUCTION AND RELATED WORK

The exploration of textual datasets is a commonly required task, especially today with an ever-increasing amount of electronically available text. Commonly used techniques for text dataset exploration include methods for *topic modeling, document summarization,* and *dimension reduction.* In this paper, we focus on comparative exploration of textual datasets. While the summaries of state-of-the-art methods for document summarization are well-interpretable, they do not allow for efficient comparison of documents. Topic modeling produces topics and distributions over documents so that users can efficiently understand which documents share which topics. By looking at highly weighted terms of a topic or deriving a summary using them, users can comprehend the topic's content. Dimension reduction yields high-level vector representations of documents, e.g., by decomposing the document-term frequency matrix using LSA, NMF, or PCA, or by extracting highly weighted tokens.

While methods for both topic modeling and dimension reduction enable comparative dataset exploration, they suffer from two main drawbacks: interpretation difficulty and incapability to resolve vaguealy defined topics.

Topics derived by topic modeling are often not easy to interpret or not meaningful to humans, despite being statistically valid [2]. One reason is that topics not only represent semantic relations but also latent relations, which may be difficult if not impossible for humans to make sense of. Statistical topics lack of hyperonymous relations between comprosed words, but represent more a collection of frequently coocurring concepts.

Further, topic modeling and dimension reduction approaches require that documents share words or at least use commonly valid synonyms. Both techniques will fall short if the analyzed dataset contains many semantic concepts with synonyms, different wording of same concepts, e.g., "Cuba" and "the Caribbean Pearl", near-identity mentions, e.g., "US government" or "President Trump" [8], or indirectly mentioned concepts, e.g., Iran's nuclear restart mentioned as "Tehran's nuclear ambitions" and "resuming its nuclear program" [3,4]. Even recent neural models (cf. [1]) that are trained on large datasets to find mentions across related documents are incapable of resolving near-identity or highly context-dependent mentions.

To enable dataset exploration using interpretable and comparative document representations, we propose an approach that employs 1) topic modeling to find related documents based on their *high-level representations* and leverages 2) near-identity cross-document coreference resolution (NI-CDCR) to provide *detailed representations* of the documents' concept compositions.

2 METHODOLOGY

The proposed approach performs three phases: deriving 1) *high-level* and 2) *detailed* document representations, and 3) *visualizing* the results for dataset exploration by users. First, for high-level representations, we use LDA and further calculate for each document-pair a similarity using cosine similarity.

JCDL '20, August 1–5, 2020, Virtual Event, China
 978-1-4503-7585-6/20/06...$15.00
https://doi.org/10.1145/3383583.3398562

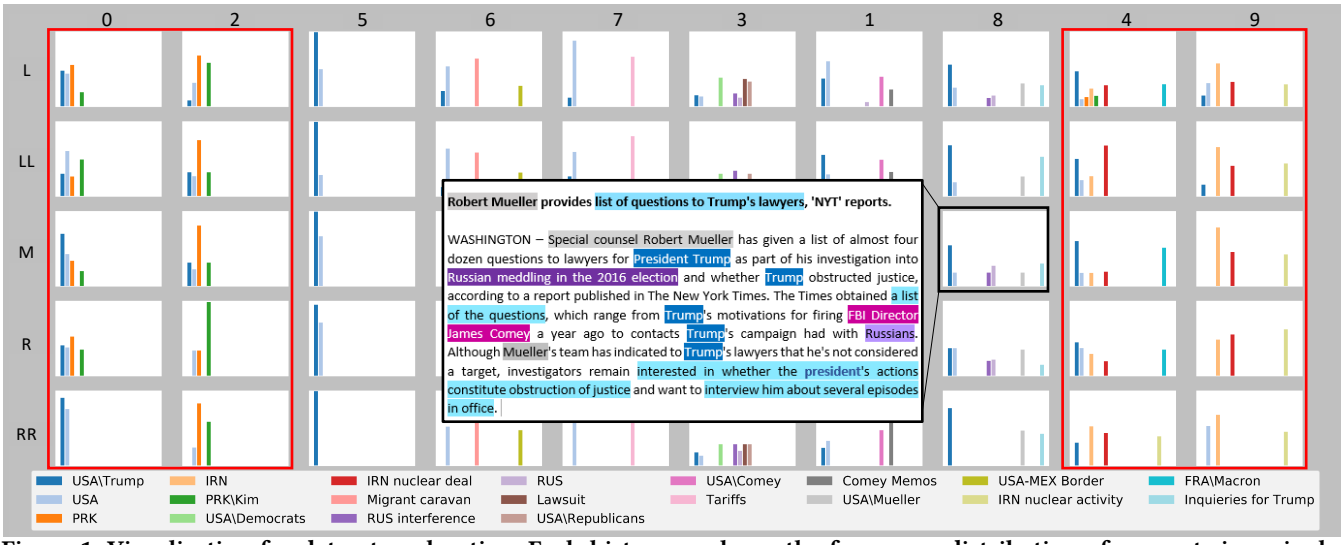

Figure 1: Visualization for dataset exploration. Each histogram shows the frequency distribution of concepts in a single document. In this example, documents are organized visually in columns and rows according to metadata dimensions.

Second, for detailed document representations, the approach extracts mentions of semantic concepts from each document and links them across subsets of documents determined to be similar in the first task (cosine similarity $s \geq 0.7$). This way, similar to topics in topic modeling, the identified concepts 1) *summarize* the documents as well as allow for 2) comparison of the *documents*. The second phase consists of two tasks: mention extraction and mention linking. First, we extract noun phrases (NPs) and verb phrases (VPs) as candidate mentions of named entities (NEs), events, actions, and other concepts covered in the subset of related documents. The mention linking consists of four types of CDCR sieves [6], where each sieve analyzes specific properties, such as core meaning or context, to determine whether to mentions refer to the same semantic concept (cf. [3]).

Third, the system visualizes the results for dataset exploration. The prototype shows a *histogram-matrix* as depicted in Figure 1, showing one histogram for each document. Users can organize documents in rows and columns according to metadata or use an automated ordering, which uses document-to-document similarities from the first task to place similar document nearby. In each histogram, each bar represents one concept and its height the frequency of that concept in the histogram's document [7]. Users select the number of concepts for exploration, e.g., Figure 1 depicts 20 bars representing 20 most frequently occurring concepts.

Figure 1 shows the concept distribution of 50 news articles of the NewsWCL50 dataset [3] depicted in a 10x5 matrix layout, where columns are events and rows are stance. We find that "Trump" is the globally the most frequently mentioned concept (leftmost, blue bar). When comparing its frequency across the documents, we find that the most similar pairs of events identified by high-level document representations, i.e., events 0 and 2 as well as 4 and 9 (red boxes). The histograms of each event pair show that although the events were identified as similar, content-wise we see difference based on the identified concepts. For example, in events 4 and 9 we see that event 4 focuses more on the country leaders Trump and Macron, whereas event 9 reports more about Iran, it's nuclear activities, and its nuclear deal.

3 CONCLUSION AND FUTURE WORK

We present an approach for dataset exploration that is, in contrast to the state-of-the-art, capable of providing comparable and easy-to-interpret document representations and summaries. By using topic modeling and near-identity cross document coreference resolution (NI-CDCR), the approach is able to narrow down the search space for CDCR, which in turn produces interpretable and comparable summaries. Further, it resolves context- and event-specific mentions, e.g., in contrast to the state-of-the-art, the approach can resolve topics that contain terms that are normally contradictory but may be valid in the dataset's context, such as "freedom fighters" and "terrorists" in media coverage on the Ukrainian crisis. In the future, we plan to add more interactive functions to the visualization, such as a semantic zoom to reveal details on documents and topics (cf. [5]).

REFERENCES

[1] Shany Barhom et al. 2019. Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution. 2, (2019), 4179–4189. DOI:https://doi.org/10.18653/v1/p19-1409

[2] Jonathan Chang et al. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Journal of Chemical Information and Modeling* 53, (2009), 1689–1699. DOI:https://doi.org/10.1017/CBO9781107415324.004

[3] Felix Hamborg et al. 2019. Automated Identification of Media Bias by Word Choice and Labeling in News Articles. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2019), 196–205. DOI:https://doi.org/10.1109/JCDL.2019.00036

[4] Felix Hamborg et al. 2019. Illegal Aliens or Undocumented Immigrants? Towards the Automated Identification of Bias by Word Choice and Labeling. *in Proceedings of the iConference 2019* (2019). DOI:https://doi.org/10.13140/RG.2.2.10120.06402

[5] Felix Hamborg et al. 2019. NewsDeps: Visualizing the Origin of Information in News Articles. *arXiv:1909:10266*, 2019, 1–7.

[6] Jing Lu and Vincent Ng. 2016. Event coreference resolution with multi-pass sieves. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016* (2016), 3996–4003.

[7] David Newman et al. 2010. Evaluating topic models for digital libraries. January (2010), 215. DOI:https://doi.org/10.1145/1816123.1816156

[8] Marta Recasens et al. 2010. A typology of near-identity relations for coreference (NIDENT). *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010* May 2014 (2010), 149–156.