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

A. Zhukova, F. Hamborg, and B. Gipp, "Interpretable Topic Modeling Using Near-Identity Cross-Document Coreference Resolution," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2020.

Interpretable Topic Modeling Using Near-Identity Cross-Document Coreference Resolution

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

Topic modeling is a technique used in a broad spectrum of use cases, such as data exploration, summarization, and classification. Despite being a crucial constituent of many use cases, established topic models, such as LDA, often produce statistically valid yet non-meaningful topics, i.e., that cannot easily be interpreted by humans. In turn, the usability of topic modeling approaches, e.g., in document summarization, is non-optimal. We propose a topic modeling approach that uses TCA, a method for also near-identity cross-document coreference resolution. TCA showed promising results when resolving mentions of not only persons and other named entities, but also broad, vague, or abstract concepts. In a preliminary evaluation on news articles, we compare the approach with state-of-the-art topic modeling. We find that (1) the four baselines produce statistically valid yet hollow topics or topics that only refer to events in the dataset but not the events' topical composition. (2) TCA is the only approach that extracts topics that distinctively describe meaningful parts of the dataset.

ACM Reference format:

Anastasia Zhukova, Felix Hamborg, and Bela Gipp. 2020. Interpretable Topic Modeling Using Cross-Document Concept Coreference Resolution. 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

Topic modeling methods extract groups of related words, so-called topics, from a set of texts. Topics may have a rather abstract nature since they consist not only of semantically related words but also latently related words that frequently cooccur. State-of-the-art topic modeling includes generative methods, e.g., LDA, dimensionality reduction methods, e.g., probabilistic Latent Semantic Analysis (pLSA), and approaches employing non-negative matrix factorization (NMF). Many variations of these exist, which fundamentally employ the same concepts as the core methods, e.g., hierarchical LDA uses a two-level topic modeling technique to produce topics and also sub-topics.

JCDL '20, August 1–5, 2020, Virtual Event, China
 978-1-4503-7585-6/20/06…$15.00
https://doi.org/10.1145/3383583.3398564

Topics resulting from state-of-the-art methods are often difficult to interpret by humans [1,7]. A commonly used evaluation measure is likelihood, which describes how well the extracted topics fit the documents [8]. Increasing the number of topics to be extracted often leads to a higher likelihood, since topics will become more fine-grained. While on the one hand many, fine-grained topics often lead to a good performance as to likelihood, such topics also tend to be less meaningful for humans. On the other hand, few and thus broad topics are difficult to interpret as well, since each topic will contain more semantic aspects [1].

Low topic interpretability remains a core issue in topic modeling research. Despite the recently increased use of evaluation measures aiming to optimize topic quality as to interpretability, e.g., coherence or word intrusion, often the resulting topics are still not meaningful to human assessors [2,6,8]. One reason is that there seems to be no reliable relation between automatic, intrinsic measures commonly used for evaluation of topic models and interpretability of topics as assessed by humans [1,2,6].

To increase interpretability, we propose a topic modeling method that employs near-identity cross-document coreference resolution (CDCR) [3,4]. In contrast to state-of-the-art CDCR, the underlying approach, named *target concept analysis* (TCA), has shown effective in resolving not only named entities (NEs), such as persons, but also abstract or broadly defined concepts. For such concepts different, in some cases even contradictory, terms refer to a single concept, e.g., the terms "freedom fighters" and "terrorists" in media coverage on the Ukrainian crisis.

2 CDCR-BASED TOPIC MODELING

We propose an approach that uses as topics the concepts that target concept analysis (TCA) [3] resolves across documents. Effectively, each concept, e.g., "immigrants" or "US military action," represents one topic and the mentions of each concept, e.g., "undocumented immigrants" and "illegal aliens," or "military presence" and "military threat," represent the topic's words. The frequencies of a topic's words in each document yield the weighted distribution of topics and documents (cf. [7]). Compared to the state-of-the-art, especially nominal coreference resolution and synonym resolution, TCA can resolve also near-identity coreferences, which are highly context-dependent and may refer to abstractly or contrarily mentioned semantic concepts [3].

TCA consists of three types of CDCR sieves [3], where each sieve analyzes specific properties of two candidates to determine

whether both belong to one semantic concept and thus should be merged. The first two sieves analyze the core meaning of phrases, e.g., heads of phrases. The next two sieves analyze core meaning modifiers, e.g., adjectives and compounds, and merge candidates if phrases bearing additional meaning in their modifiers are similar. The fifth and sixth sieves analyze frequent word patterns.

3 PRELIMINARY EVALUATION

We evaluate TCA-based topic modeling as to our objective of topic interpretability on NewsWCL50, a dataset consisting of 50 news articles (≈22k words) and over 5900 in-text annotations of concept mentions across articles [3]. Despite the lack of gold standard datasets for topic modeling, NewsWCL50 has two desired characteristics: first, its topical composition is of challenging, different complexity, e.g., we expect to find topic compositions that represent clearly defined concepts, such as actors involved in the events, as well as vaguely defined, implicitly mentioned, or latent concepts, such as Russian interference in the US elections. Second, its topical composition is known since it consists of well-defined news events and actors and actions involved in the events. The dataset contains ten events and for each event five articles, each from an online news outlet representing the political spectrum in the US from strongly left to strongly right.

We quantitively and qualitatively compare topic interpretability with four baselines. Three state-of-the-art topic modeling approaches, i.e., LDA, LSA, and NMF, and a baseline employing bag-of-concepts (BoC) [5]. We employ MALLET and gensim for LDA, and sklearn for LSA and NMF. BoC defines topics as clusters of words and phrases that are semantically related in the word2vec word embedding space. To find semantically related clusters in BoC, i.e., topics, we use affinity propagation.

Table 1 shows an excerpt of the results of our qualitative evaluation, reporting most representative terms (i.e., with high weights) for an exemplary topic of each method. TCA is the only approach that can extract topics that distinctively and uniquely describe meaningful parts of the reported events, such as the actors, actions, locations involved. Interestingly, despite its slightly lower C_v performance shown in Table 2, TCA's topics are strongly more meaningful and interpretable compared to the other topic modeling methods. This difference may be explained by the low correlation of purely intrinsic measures with human assessed interpretability. The topics by LDA and NMF are summaries of each of the events of NewsWCL50. BoC produces abstract concepts as well, e.g., we see a topic related to US intelligence.

Table 1: Excerpt of qualitative analysis

Method	Most representative terms of one topic
LSA	Mr. Trump, say, Iran, nuclear, Comey, deal
LDA	Comey, Trump, memo, write, president, Thursday
NMF	visit, London, Trump, Khan, protest, July, state
BoC	FBI, clandestine, Watergate, FBI, CIA, Comey, intel
TCA	Russia investigation, Mr. Mueller's investigation, an investigation into Russia's election meddling, a probe into Russian interference

Table 2: Quantitative evaluation using topic coherence

Method	Unit	C_v	# topics	Topic type
LSA	L, B	*0.47*	2	Abstract
LDA	L, B	0.62	10	Summary of events
NMF	L, B	0.68	10	Summary of events
BoC	L, B	0.52	534	Concepts
TCA	L, P	**0.52**	97	Concepts

4 CONCLUSION AND FUTURE WORK

We proposed an approach for well-interpretable topic modeling. Our method employs target concept analysis (TCA), which is a technique that capably resolves cross-document coreferences of also abstract and broadly defined semantic concepts. Such coreferences often occur in news texts, where different journalists tend to use different terms to refer to the same semantic concepts, such as individuals ("freedom fighters" vs. "terrorists") or actions ("cross the border" vs. "invade the country"). In an evaluation, we compare TCA with four state-of-the-art methods. In our quantitative experiment, we find that topic coherence of the TCA-based method C_v=0.52<0.68 achieved by NMF. However, when manually assessing the interpretability of produced topics, we find that TCA is the only method that extracts topics that distinctively and uniquely describe meaningful parts of the reported events. Other methods produce abstract topics or topics that only refer to events in the dataset but not the events' topical composition.

For future work, we plan to improve the quality of the topics produced by TCA. Specifically, we will devise a coreference resolution method capable of also resolving mentions of actions, i.e., a mix of noun and verb phrases [3], resolve local context-dependent coreferences, and disambiguate concepts, e.g., distinguish mentions of "USA-PRK meeting" and "USA-JAP meeting." While the current evaluation focuses on topic interpretability and news articles, being the main objective and target of our work, we will also evaluate the method using common measures, such as likelihood and perplexity, and further real-world datasets.

REFERENCES

[1] Loulwah Alsumait and Daniel Barbar. 2009. Topic Significance Ranking of LDA Generative Models. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I*, 67–82. DOI:https://doi.org/10.1007/978-3-642-04180-8_22

[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] Han Kyul Kim et al. 2017. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing* 266, (2017), 336–352. DOI:https://doi.org/10.1016/j.neucom.2017.05.046

[6] Jey Han Lau et al. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (2014), 530–539. DOI:https://doi.org/10.3115/v1/e14-1056

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

[8] Hanna M Wallach and David Mimno. 2009. Evaluation Methods for Topic Models. In *Proceedings of the 26th annual international conference on machine learning*, 1105–1112. DOI:https://doi.org/10.1145/1553374.1553515