Discovering Interpretable Topics by Leveraging Common Sense Knowledge

Ismail Harrando
ismail.harrando@eurecom.fr
EURECOM
Sophia Antipolis, France

Raphaël Troncy
raphael.troncy@eurecom.fr
EURECOM
Sophia Antipolis, France

ABSTRACT

Traditional topic modeling approaches generally rely on document-term co-occurrence statistics to find latent topics in a collection of documents. However, relying only on such statistics can yield incoherent or hard to interpret results for the end-users in many applications where the interest lies in interpreting the resulting topics (e.g. labeling documents, comparing corpora, guiding content exploration, etc.). In this work, we propose to leverage external common sense knowledge, i.e. information from the real world beyond word co-occurrence, to find topics that are more coherent and more easily interpretable by humans. We introduce the *Common Sense Topic Model* (CSTM), a novel and efficient approach that augments clustering with knowledge extracted from the ConceptNet knowledge graph. We evaluate this approach on several datasets alongside commonly used models using both automatic and human evaluation, and we show how it shows superior affinity to human judgement. The code for the experiments as well as the training data and human evaluation are available at https://github.com/D2KLab/CSTM.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

Topic Modeling, common sense knowledge, Interpretable Topics

ACM Reference Format:
Ismail Harrando and Raphaël Troncy. 2021. Discovering Interpretable Topics by Leveraging Common Sense Knowledge. In *Proceedings of the 11th Knowledge Capture Conference (K-CAP '21), December 2–3, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3460210.3493586

1 INTRODUCTION

Topic modeling is a text mining task that is widely used for many applications, both for other NLP downstream tasks (e.g. text similarity, document retrieval, recommender systems), as well as a tool to explore, visualize and interpret the content of large collections of text. While the first application can be evaluated and improved by quantitatively measuring the performance on the downstream task, it is harder to capture the ability of a topic model to generate results that are understandable and useful for a human user. Several previous research efforts [4, 6, 13, 18, 27] have highlighted the discrepancy between most quantitative and automatic evaluation metrics (widely used in the literature) and human judgement, as these models tend to optimize for numerical objectives that rarely align or correlate well with what humans consider "topics".

Most topic modeling approaches focus on word co-occurrences statistics as the main signal to detect the latent semantic relations among them – an idea that goes all the way back to the 50s (*"You shall know a word by the company it keeps"*[8]). This makes them inherently incapable of capturing relations between words that are not explicitly present in the training data, which is bound to happen in any text collection with a large-enough vocabulary. A lot of work has been done to explore the possibility of injecting external knowledge (usually domain-specific) into the task of topic modeling (Section 2). Yet, to the best of our knowledge, no attempt to incorporate human general knowledge (or *common sense*) into the process of topic modeling has been proposed to bridge the gap between statistics-based optimization and human judgement.

In this paper, we try to answer the following research question: How to generate topics that humans can easily understand? To do so, we propose a method that combines the knowledge from a common sense knowledge graph [25] with a clustering algorithm to produce topics that are more correlated with the human judgement of coherence while scaling seamlessly to large datasets.

2 RELATED WORK

Our work touches on two aspects of the task of topic modeling: incorporating external knowledge into topic models, as well as the qualitative evaluation of topic models beyond automatic metrics.

Incorporating knowledge into topic modeling. Our work joins a growing pool of approaches aiming to incorporate external knowledge into the topic modeling training. [5] approached the problem of importing external "General knowledge" into the task of topic modeling by factoring lexical and semantic relations of words such as synonymy into the training of the topic model (LDA). They also proposed to leverage training (domain) data itself to correct some of the wrong knowledge that may have been injected into the process. [9] followed a similar approach, focusing mostly on synonymy to create "concepts" that replace words in the topic assignment phase of training LDA, and incorporate the external knowledge in the pre-processing step as well. [17] also proposed a modified LDA algorithm that uses synonyms sets from a Thesaurus in both word-topic assignment and document-topic assignment, conditioned on their co-occurrence. [24] leveraged a different source of external knowledge, by extracting and linking entities from the text, then using the embedding similarity for entities linked from the document

K-CAP '21, December 2–3, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8457-5/21/12...$15.00
https://doi.org/10.1145/3460210.3493586

as a constraint for training LDA. [29] introduced an efficient model based on a factor graph framework to integrate prior knowledge such as word correlation and document labels, by expressing the prior knowledge as sparse constraints on the hidden topic variables. Finally, several works [1, 26] explored using external knowledge for Topic Labeling, aiming to improve the overall interpretability of the generated labels.

Topic Modeling Interpretability and Evaluation. In [4], Chang et al. highlighted several shortcomings in the the use of automatic evaluation metrics such as Topic Coherence, as topic models can score high without creating "semantically meaningful" latent topics. They also proposed two human evaluation methods (*word intrusion* and *topic intrusion*) to examine the performance of 3 topic models, and found that the automatic coherence metric does not align well with human quality judgement. [7] found that using *Word Embedding Coherence*, i.e. using (external, pre-trained) word embedding similarity to score how coherent the top words of the generated topics are, and showed that it aligns better with human judgement. [6] reached a similar conclusion after presenting a thorough survey of the literature on topic interpretability and proposing a definition of it. They also proposed an experimental framework which tests both topic words quality and topic assignment, and studied how different models behave in it. [18] conducted an expert analysis of topic modeling results (based on LDA), and reported several results such as how *word intrusion* detection correlates well with human judgement of topic quality. They also devised a method to automatically identify some classes of bad topics.

common sense knowledge. This is a blossoming interest in modeling and reasoning using common sense knowledge, as demonstrated by the increasing numbers of common sense knowledge graphs [15, 22, 25] and models that use them [12, 19]. In this work, we only focus on ConceptNet [25], a widely used common sense knowledge graph, which models words of different languages and the lexical relations such as `Synonym` and `DerivedFrom`, but also semantic ones such as `LocatedAt` and `UsedFor`.

3 APPROACH

Similarly to previous works [23], we approach the task of topic modeling as a *document clustering problem*, i.e. we generate vector representations for all documents in the studied corpus that we call *common sense enriched bag of words* representation, and then we run a clustering algorithm to find N coherent clusters (N being the number of topics) which represent our topics. We refer to this combination henceforth as *CSTM* (Common-Sense Topic Model).

Common-sense Enriched Bag of Words (CS-BoW). Inspired by methods from the query expansion literature [2, 14], we propose to enrich the oft-used Bag of Words document representation with related terms from the ConceptNet Knowledge Graph. The advantage of using ConceptNet is that it is mostly populated by the common sense "Related To" relation, which implies a topical relatedness between terms. Concretely, for each word in the document, we query ConceptNet to retrieve all terms that are directly linked to it (one hop away on the graph), and we add them to the document, but only if they already appear in the corpus (to avoid increasing the the vocabulary size). For instance, a document that mentions the word "camera" would automatically be enriched with the words "photo", "lens", etc. The document representation is then constructed as the Bag of Words containing all the original words of the document, in addition to all words that are related to them in ConceptNet. We surmise that by appending all related terms to its words, each document becomes more representative of its topic.

We also use *ConceptNet Numberbatch* – pretrained graph embeddings for ConceptNet – to measure similarity between each word in the document and the words to be potentially added. We only keep the words above an empirically-defined threshold to avoid adding noisy terms to the document representation.

We note that because this process does not add any new vocabulary words to the vector representation, the performance of the clustering algorithm is constant, i.e. this operation comes at no cost except the preprocessing, which is done once and can be trivially parallelized. The filtering via embedding similarity can also be precomputed and cached so that the creation of the *CS-BoW* can be done with almost no extra overhead.

Clustering. There is a rich and diverse literature on the task of clustering. For the sake of simplicity and scalability, we choose *K-Means*, a commonly-used clustering algorithm that is fast and can handle bigger datasets using the highly optimized *FAISS*[1] implementation, and we run it on the CS-BoW representations of the corpus documents. Exploration of more advanced clustering methods is left for future work. To generate the topic top words, we consider the centroid vectors generated by K-Means and pick the N components (corresponding to words on the CS-BoW representation) with the largest coefficients to represent the topic.

4 EXPERIMENTS

In this section, we detail the experimental setup to test our model. We run CSTM alongside three baselines on 4 news datasets, all annotated with topical labels for each document. For each dataset, we consider the number of topics to be exactly the number of ground-truth labels, as we expect our topic models to be able to find the same ones automatically. For CSTM, we set the filtering threshold to 0, i.e. any term that has a negative cosine similarity with the original document term (through Numberbatch embeddings) is not added to the CS-BoW.

We then perform two evaluations: a quantitative analysis of the resulting topic assignment (computed by measuring the agreement between the resulting topic distribution among the corpus documents and the ground truth labels, using the *V-measure metric* [20]), and topic top words (via *Coherence*). We compute both the NPMI coherence (which is heavily corpus dependant) and the Word Embeddings coherence as defined by Fang et al. [7]. This measure has been shown to correlate better with human judgement because it relies on word similarity beyond a specific corpus (through the word embeddings). Both coherence metrics are computed over the top 10 words of each topic. We then perform a human evaluation to validate the claim that factoring common sense into topic models yield topics that are more easily interpretable by humans.

4.1 Baselines

We compare our model to two frequently used topic modeling algorithms: LDA [3] and NMF [28]. We also add K-Means on the

[1] https://github.com/facebookresearch/faiss/

traditional BoW representation to see how the common sense enrichment helps with the task. For LDA, we only slightly fine-tune the hyper-parameters, and we observed empirically that the default ones seem to provide the best results. We also note that the preprocessing of the dataset to remove the most and least frequent words is crucial to get decent results with LDA. Similarly with NMF, we vary the preprocessing and the generation of the BoW. For each model, we train using 5 different seeds and several hyperparameter configurations, and we keep only the results from the instance with the highest Word Embeddings coherence (which is positively correlated with the V-measure as well).

4.2 Datasets

For evaluation, we selected 4 news datasets with different characteristics in terms of number of documents, number of topical labels, vocabulary size, and writing style (editorial vs user-submitted). The topic labels are essential for evaluation as they give us an idea on what to expect our model to be able to find.

- **20 Newsgroups** [16]: a collection of 18000 user-generated forum posts arranged into 20 groups seen as topics such as *“Baseball", “Space", “Cryptography", and “Middle East"*.
- **AFP News** [21]: a dataset containing 70K English news articles issued by the French News Agency (*AFP*). The articles are tagged with one or more topics coming from IPTC NewsCode taxonomy[2]. We consider only documents with one label, and only the first level of this taxonomy such as *“Politics", “Art, Culture and Entertainment", “Environment"*. The label distribution is highly unbalanced.
- **AG News** [11]: a news dataset containing 127600 news articles from various sources, fairly distributed among 4 categories: *“World", “Sports", “Business" and “Sci/Tech"*.
- **BBC News** [10]: a news dataset from BBC containing 2225 English news articles classified in 5 categories: *“Politics", “Business", “Entertainment", “Sports" and “Tech"*.

5 RESULTS

5.1 Quantitative Analysis

We evaluate our model as well as the baselines on the 4 datasets and we report on the quantitative results on 3 metrics in Table 1. While our goal is to produce humanly understandable topics, we consider the two tasks of topic assignment (putting documents in clusters that are similar to what a human annotator would) and top words coherence (producing top words that are all semantically related) as proxies to such goal. We later explore the correlation between these metrics and human judgement.

On the automatically computed metrics, we see that CSTM generally performs the best or on par with the best on the V-measure and the Word Embedding coherence, suggesting that the addition of common sense knowledge indeed drives the resulting topics to be closer to human judgement. The low score on NPMI, which is solely based on word co-occurrences in the corpus, is justified by the fact that the top words generated by CSTM do not explicitly co-occur a lot in the corpus, but are rather semantically related through the external knowledge.

Dataset	Model	V-measure	WE_coherence	NPMI
BBC	CSTM	**0.789**	**0.382**	-0.139
	K-Means	0.662	0.346	0.105
	LDA	0.729	0.359	**0.122**
	NMF	0.172	0.371	0.0225
AG	CSTM	0.2506	**0.387**	-0.0539
	K-Means	0.171	0.225	**0.027**
	LDA	**0.542**	0.214	0.001
	NMF	0.095	0.306	-0.0017
20NG	CSTM	0.403	0.303	-0.055
	K-Means	**0.433**	0.246	**0.127**
	LDA	0.403	**0.353**	0.031
	NMF	0.274	0.281	0.092
AFP	CSTM	0.431	0.296	-0.0459
	K-Means	**0.444**	**0.329**	**0.159**
	LDA	0.397	0.322	0.075
	NMF	0.409	0.308	0.127

Table 1: Quantitative performance of CSTM and Baselines on 4 datasets. Best result on each dataset-metric pair is highlighted in bold

We also notice that K-means by itself is quite a good baseline for topic modeling, especially on topic assignment. Human evaluation, however, reveals that the topics found by K-Means are not easily interpretable by humans.

5.2 Human Evaluation

For human evaluation, we tasked 12 fluent English speakers (graduate students with limited to no knowledge of the task) to perform three assignments to evaluate the resulting topics. NMF, the worst performing model on all automatic metrics, was dropped from the comparison to make the experiment easier for the subjects.

1. **Word intrusion**: we follow the procedure as defined in [4]. To make the task tractable, we randomly choose one topic per dataset/model pair, resulting in 12 topic-words sets. Each set contains the top 5 words from a topic, with one top word from a different topic shuffled in the mix. We ask the evaluator to identify the odd word. The more the test is able to identify the odd word, the better we judge the model to be able to create coherent and understandable topics.
2. **Topic Labeling**: we give the evaluator a list of the ground-truth labels from each dataset (e.g. “Politics", “Technology"..), alongside the top words from one topic generated by each model. We then ask the evaluator to assign one of the labels to the topic, and give a score to how well they match (on a scale from 0 to 5, 5 corresponding to “all top words perfectly matching"). The more a model is able to generate topics that strongly match with the ground-truth labels, the higher its accumulative score will be.
3. **Topic Classification**: we give the evaluator a snippet (first 50 words) of a document picked at random from each dataset, as well as the top words from the topic that it was assigned to it by each model. The evaluators are then asked to choose which topic they prefer among them, and rate the matching. Each evaluator is asked to do so for 4 documents, one from each dataset.

[2] http://cv.iptc.org/newscodes/subjectcode/

To measure agreement, we divide the group into 6 pairs and we give identical questions to each pair. Given the randomized nature of the question, we expect the high correlation between answers from each pair to reflect a broader agreement over the compared topic models.

	Tasks		
Models	**Intrusion**	**Labeling**	**Classification**
CSTM	**83.3%**	**84.6%**	**27.5%**
K-Means	33.3%	81.7%	19.5%
LDA	29.2%	52.9%	13.3%

Table 2: Scores percentage (w.r.t the maximum obtainable) across datasets for CSTM, K-Means and LDA

In Table 2, we provide the results of our human evaluation. On all three tasks, *CSTM* outperforms the other two models, with a significant margin on two. On word intrusion specifically, CTSM seems to produce top topic words with clear semantic coherence: 83.3% of the word intrusions were correctly identified. On the task of labeling as well, evaluators were mostly able to identify labels in the original dataset that correspond to the topics created by the model and with high confidence. Finally, users mostly preferred the topic attribution from CSTM to the other topic models, showing how it can be used for automatic classification as well. The results of the human evaluation as well as the script used to generate the evaluation forms can be found at https://github.com/D2KLab/CSTM. It is worth noting that, although the sample size for the human experiment is relatively small, there was a high agreement among subjects (an average pair scores correlation of **0.78**), suggesting robust results.

6 CONCLUSION AND FUTURE WORK

We propose a simple yet effective approach to incorporating common sense knowledge into topic modeling to produce topics that are more readily interpretable by human assessors. On automatic and human evaluation, *CSTM* proves to be a promising method for generating topics that are fit for user-facing tasks such as guided corpus exploration or textual data analysis and visualization.

Based on this primary work, we can explore different directions of potential improvement: using TF-IDF variants to generate a more robust CS-enriched representations, experimenting with other clustering techniques and common sense knowledge graphs, combining the CS-enriched BoW with other topic modeling techniques, and studying the impact of all the hyperparameters (e.g. number of topics, filtering threshold) in improving the quality of the results. We also envision extending this work to the task of Topic Labeling, as human interpretability is a key requirement for good labels.

ACKNOWLEDGMENT

This work has been partially supported by the European Union's Horizon 2020 research and innovation program within the Odeuropa project (grant agreement No. 101004469), by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003) and by raisin.ai within the MyLittleEngine project.

REFERENCES

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Krys Kochut, and Hamid Reza Arabnia. [n. d.]. A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling. *IJACSA 2017* ([n. d.]).

[2] Dr. Hiteshwar Kumar Azad and A. Deepak. 2019. Query Expansion Techniques for Information Retrieval: a Survey. *Inf. Process. Manag.* 56 (2019), 1698–1735.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

[4] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models *(NIPS'09)*. Red Hook, NY, USA, 288–296.

[5] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Discovering Coherent Topics Using General Knowledge. In *CIKM '13* (San Francisco, California, USA). New York, NY, USA, 209–218.

[6] Caitlin Doogan and Wray Buntine. 2021. Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures. In *NAACL '21*.

[7] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data. In *SIGIR '16* (Pisa, Italy) *(SIGIR '16)*. New York, NY, USA.

[8] Adriana Ferrugento, Ana Alves, Hugo Gonçalo Oliveira, and Filipe Rodrigues. 1957. A synopsis of linguistic theory 1930-1955.

[9] Adriana Ferrugento, Ana Alves, Hugo Gonçalo Oliveira, and Filipe Rodrigues. 2015. Towards the Improvement of a Topic Model with Semantic Knowledge, Vol. 9273. Portuguese Conference on Artificial Intelligence, 759–770.

[10] Derek Greene and Pádraig Cunningham. 2006. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In *ICML 2006*.

[11] Antonio Gulli. 2005. *AG's corpus of news articles*.

[12] Ismail Harrando and Raphaël Troncy. 2021. Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph. In *LDK 2021*. Dagstuhl, Germany.

[13] Alexander Miserlis Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan L. Boyd-Graber, and P. Resnik. 2021. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. *ArXiv* abs/2107.02173 (2021).

[14] Ming-Hung Hsu, Ming-Feng Tsai, and Hsin-Hsi Chen. 2006. Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. In *Information Retrieval Technology*. Berlin, Heidelberg.

[15] Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. CSKG: The CommonSense Knowledge Graph. *Extended Semantic Web Conference (ESWC)* (2021).

[16] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *12th International Conference on Machine Learning (ICML)*. 331–339.

[17] Natalia Loukachevitch, Michael Nokel, and Kirill Ivanov. 2018. Combining Thesaurus Knowledge and Probabilistic Topic Models. In *Analysis of Images, Social Networks and Texts*.

[18] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *EMNLP '11* (Edinburgh, United Kingdom). Association for Computational Linguistics, USA.

[19] Janna Omeliyanenko, Albin Zehe, Lena Hettinger, and Andreas Hotho. [n. d.]. LM4KG: Improving Common Sense Knowledge Graphs with Language Models. In *ISWC 2020*. Cham.

[20] Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *EMNLP-CoNLL '07*. Association for Computational Linguistics, Prague, Czech Republic, 410–420.

[21] Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphaël Troncy, and Xavier Tannier. 2019. Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. In *5th Wiki Workshop*. 1232–1239.

[22] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. [n. d.]. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI 2019*.

[23] Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!. In *EMNLP '20*. Association for Computational Linguistics, Online, 1728–1736.

[24] Dandan Song, Jingwen Gao, Jinhui Pang, Lejian Liao, and Lifei Qin. 2020. Knowledge Base Enhanced Topic Modeling. In *ICKG 2020*. 380–387.

[25] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. 4444–4451.

[26] Ilaria Tiddi, Mathieu d'Aquin, and Enrico Motta. 2015. Using Linked Data Traversal to Label Academic Communities. In *WWW 2015* (Florence, Italy) *(WWW '15 Companion)*. New York, NY, USA.

[27] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation Methods for Topic Models. In *ICML '09* (Montreal, Quebec, Canada) *(ICML '09)*. New York, NY, USA, 1105–1112.

[28] Wei Xu, Xin Liu, and Yihong Gong. 2003. Document Clustering Based on Non-Negative Matrix Factorization *(SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 267–273.

[29] Yi Yang, Doug Downey, and Jordan Boyd-Graber. 2015. Efficient Methods for Incorporating Knowledge into Topic Models. In *EMNLP '15*. Association for Computational Linguistics, Lisbon, Portugal, 308–317.