

rosati_2022_moving_beyond_word_lists_toward s_abstractive_topic_labels_for_human_like_topi cs_of_scientific_documents

Year

2022

Author(s)

Rosati, Domenic

Title

Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents

Venue

WIESP

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Topic labeling parameters

Label generation

Topic modeling can be reframed as a set of two tasks: (1) finding meaningful clusters for documents and (2) performing MDS on those individual clusters to find meaningful topic labels.

After obtaining document clusters, MDS models such as (Lu et al., 2020) can be used to produce natural language summaries that synthesize common themes from documents.

In this work, we treat topic labels as a "citation intent" label

Topic 0
NPMI and Topic Coherence are measures used to measure the semantic coherence of topics .
Topic 4
Topic model quality and interpretability are two different metrics used to measure the semantic interpretability of a topic .
Topic 2
Evaluation metrics: Log predictive probability (LPP) and topic interpretability

Table 3: **Topic** representations produced by multi-lexsum-tiny. Compared to word lists they are much more readable and closer to everyday notions of **topics**.

Motivation

Finding approach to topic labeling that is closer to a natural language description of a topic than a word list.

Topic modeling

Contextualized topic models (CTM)

Topic modeling parameters

SPECTER for constructing embeddings

Nr of topics: 3 to 50

Nr. of topics

10

Label

we propose that the *topic representation be a sentence or paragraph*

Label selection

\

Label quality evaluation

Table 2 shows the Rouge-1 (R-1) and BERTScore (average F1 across topics) for each of the models selected for generating topic representations using MDS as well as the topic lists generated by CTM.

Model	R-1	BERTScore
multi-lexsum-long	38	85
multi-lexsum-tiny	3	81
ms2	3	81
multixscience	15	80
topic lists	1	76

Table 2: Rouge-1 (R-1) and BERTScore (F1) results for each models topic representations measured against.

Topic word lists have the worst R-1 and BERTScore. The MDS models do a little bit better with multi-lexsum-long having the best overall score. multixscience also does well with regards to R-1. Since multixscience and multi-lexsum-long are long form summaries, it appears that R-1 is potentially biased towards longer summaries and may not be a good measure across representations, in particular it may be uninformative for evaluating the performance of topic lists. ms2 and multi-lexsum-

tiny are smaller and have better BERTScore than multixscience indicating they might provide more semantically similar representations.

Assessors

\

Domain

Paper: Topic labeling

Dataset: Topic labeling

Problem statement

This work asks whether an alternative approach to topic labeling can be developed that is closer to a natural language description of a topic than a word list.

To this end, we present an approach to generating human-like topic labels using abstractive multi-document summarization (MDS).

We investigate our approach with an exploratory case study.

We model topics in citation sentences in order to understand what further research needs to be done to fully operationalize MDS for topic labeling.

Corpus

Origin: (citation statements) to the paper (Lau et al., 2014)

Nr. of documents: 183

Details:

- citation statements

Document

in-text passages which contain citations (citation statements) to a given paper paper

Pre-processing

\

```
@inproceedings{rosati_2022_moving_beyond_word_lists_towards_abstractive_topic_labels_for_human_like_topics_of_scientific_documents,
  title = "Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents",
  author = "Rosati, Domenic",
  booktitle = "Proceedings of the first Workshop on Information Extraction from Scientific Publications",
  month = nov,
  year = "2022",
  address = "Online",
  publisher = "Association for Computational Linguistics",
  url = "https://aclanthology.org/2022.wiesp-1.11",
  pages = "91--99",
  abstract = "Topic models represent groups of documents as a list of words (the topic labels). This work asks whether an alternative approach to topic labeling can be developed that is closer to a natural language description of a topic than a word list. To this end, we present an approach to generating human-like topic labels using abstractive multi-document summarization (MDS). We investigate our approach with an exploratory case study. We model topics in citation sentences in order to understand what further research needs to be done to fully operationalize MDS for topic labeling. Our case study shows that in addition to more human-like topics there are additional advantages to evaluation by using clustering and summarization measures instead of topic model measures. However, we find that there are several developments needed before we can design a well-powered study to evaluate MDS for topic modeling fully. Namely, improving cluster cohesion, improving the factuality and faithfulness of MDS, and increasing the number of documents that might be supported by MDS. We present a number of ideas on how these can be tackled and conclude with some thoughts on how topic modeling can also be used to improve MDS in general.",
}
```