korencic_2018_document_based_topic_coherence_measures_for_news_media_text

Year

2018

Author(s)

Damir Korencic and Strahil Ristov and Jan Snajder

Title

Document-based topic coherence measures for news media text

Venue

Expert Systems With Applications

Topic labeling

Partially automated

Focus

Primary

Type of contribution

Novel

Underlying technique

Manual labeling assisted by associated documents

Topic labeling parameters

Nr of inspected topic words: 20
Document-topic probability used for document inspection: 10%

Label generation

The topics were then manually labeled by two annotators, with the annotation procedure set up as follows.

The annotators were instructed to inspect a model topic represented as a list of article titles and words and to infer, if possible, the corresponding semantic topics. A semantic topic was described as either an abstract concept or a concept corresponding to an entity, event, or a story.
After inspecting a model topic, the annotators consulted a shared list of semantic topics discovered so far and either updated the list with new topics or re-used the existing ones.

The model topic was then labeled with semantic topics and, if the topic contained unrelated random documents or words, with an additional "noise" label.

The top-ranked documents were presented to the annotators as a list of titles sorted by document-topic probabilities, with the full texts of articles available for inspection.
The annotators were instructed to inspect the documents in sorted order and to stop when the document-topic probability fell below 10%.

Topic-related words were presented as a list of top 20 topic words.

For labeling, the annotators relied primarily on documents, as the words proved either seemingly unrelated, vague, or too abstract whereas lists of well-formed document titles provided more accurate and specific information. Consequently, the decisions on the semantic topics and their correspondence to model topics, as well as the decision on existence of noise, were made based on topic-related documents, while the words at best served to confirm this decision.

In summary, 52% of the model topics were labeled with one semantic topic, 15% were labeled with one semantic topic and the noise label, 17% were labeled with two semantic topics, 4% were labeled with two semantic topics plus noise, while 12% were labeled as noise.

Motivation

"One can obtain the coherence judgments directly, be asking the annotators to judge the coherence of each topic, or indirectly, by first asking the annotators to semantically interpret and label each topic, and then use these labels to derive the coherence judgments. We chose the latter approach as it allowed us to reuse existing datasets with manually labeled topics.”

“Using the above-described labeled dataset as input [see `label generation`], we defined as coherent all topics that have been annotated with a single semantic topic, possibly with the addition of noise, and as incoherent otherwise.
In other words, we consider a topic as coherent as long as a human annotator can recognize that the topic corresponds to a single semantic topic […], whereas an incoherent topic corresponds either to noise or to a mixture of two or more semantic topics. This resulted in 235 topics (67%) being labeled as coherent and 115 top- ics (33%) labeled as incoherent.”

In other words, topic labels are used as a proxy for coherence.

Topic modeling

LDA (on the original dataset, which is made up of LDA topics)

Topic modeling parameters

Dataset 1 - Nr of topics (k): 50 (for three models), 100 (for two models)
Dataset 2 - Nr of topics (k): 50 (for three models), 100 (for one models)

Nr. of topics

Dataset 1 - 350 (topics from the five models were pooled together)
Dataset 2 - 250 (topics from the four models were pooled together)

Label

Either one semantic topic (described as either an abstract concept or a concept corresponding to an entity, event, or a story), one semantic topic and the noise label, two semantic topics, two semantic topics plus noise or simply noise.

Label selection

No label selection, since topics were divided among the two annotators, but:

- “For calibration, annotators labeled a shared set of 50 topics and updated the labeling conventions.”

- “On a sample of 50 topics annotated by both annotators, the annotators agreed on 88% of the topics, while the chance-corrected kappa agreement coefficient is 0.674”

Label quality evaluation

\

Assessors

\

Domain

Paper: Topic coherence
Dataset: News

Problem statement

In this paper, we introduce the notion of document-based topic coherence and propose novel topic coherence measures that estimate topic coherence based on topic documents rather than topic words.
We evaluate the proposed measures on two datasets containing topics manually labeled for document-based coherence, on which the proposed measures outperform a strong baseline as well as word-based coherence measures.
We also demonstrate the usefulness of document-based coherence measures for automated topic discovery from news media texts.

Corpus

Dataset 1
Origin: Various US news outlets
Nr. of documents: 24.000
Details:

- political news articles from mainstream US news outlets

Dataset 2
Origin: Various Croatian news outlets
Nr. of documents:

Details:

- Topics derived from a corpus of news texts in Croatian language
- Serving as an additional test set for assessing the robustness of the results

Document

LDA topic extracted from the set of news articles

Pre-processing

\

```
@article{korencic_2018_document_based_topic_coherence_measures_for_news_media_t
ext,
  abstract = {There is a rising need for automated analysis of news text, and
topic models have proven to be useful tools for this task. However, as the
quality of the topics induced by topic models greatly varies, much research
effort has been devoted to their automated evaluation. Recent research has
focused on topic coherence as a measure of a topic's quality. Existing topic
coherence measures work by considering the semantic similarity of topic words.
This makes them unfit to detect the coherence of transient topics with
semantically unrelated topic words, which abound in news media texts. In this
paper, we introduce the notion of document-based topic coherence and propose
novel topic coherence measures that estimate topic coherence based on topic
documents rather than topic words. We evaluate the proposed measures on two
datasets containing topics manually labeled for document-based coherence, on
which the proposed measures outperform a strong baseline as well as word-based
coherence measures. We also demonstrate the usefulness of document-based
coherence measures for automated topic discovery from news media texts.},
  author = {Damir Koren{\v c}i{\'c} and Strahil Ristov and Jan {\v S}najder},
  date-added = {2023-03-19 19:01:07 +0100},
  date-modified = {2023-03-19 19:01:07 +0100},
  doi = {https://doi.org/10.1016/j.eswa.2018.07.063},
  issn = {0957-4174},
  journal = {Expert Systems with Applications},
  keywords = {Topic models, Topic coherence, Topic model evaluation, Text
analysis, News text, Exploratory analysis},
  pages = {357–373},
  title = {Document-based topic coherence measures for news media text},
  url = {https://www.sciencedirect.com/science/article/pii/S0957417418304883},
  volume = {114},
  year = {2018}}
```

#Thesis/Papers/Initial