popa_2021_bart_tl_weakly_supervised_topic_label_generation

Year

2021

Author(s)

Popa, Cristian and Rebedea, Traian

Title

BART-TL: Weakly-Supervised Topic Label Generation

Venue

EACL

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel

Underlying technique

Transformer-based (with denoising autoencoder architecture)

Topic labeling parameters

- Fine-tuning epochs: 2 epochs
- Optimizer: Adam
  - β1 = 0.9

    - $\beta 2 = 0.999$
    - $\varepsilon = 10-8$
    - 0.1 weight decay
    - 0.1 dropout
    - 0.1 attention dropout
    - 0.1 la- bel smoothing
    - 6% warmup steps
    - learning rate = 3e-5.
- The final labels are generated using beam search with a beam size of 25

Label generation

Method of performing a weakly-supervised fine-tuning on transformer models pre-trained on English data in order to obtain human-comprehensible and meaningful topic labels.

The proposed method utilizes a pre-trained BART Lewis et al., 2020 transformer model, with a denoising autoencoder architecture, as a sequence- to-sequence approach is adopted for the task of topic labeling.
The model is fine-tuned on the baseline dataset.

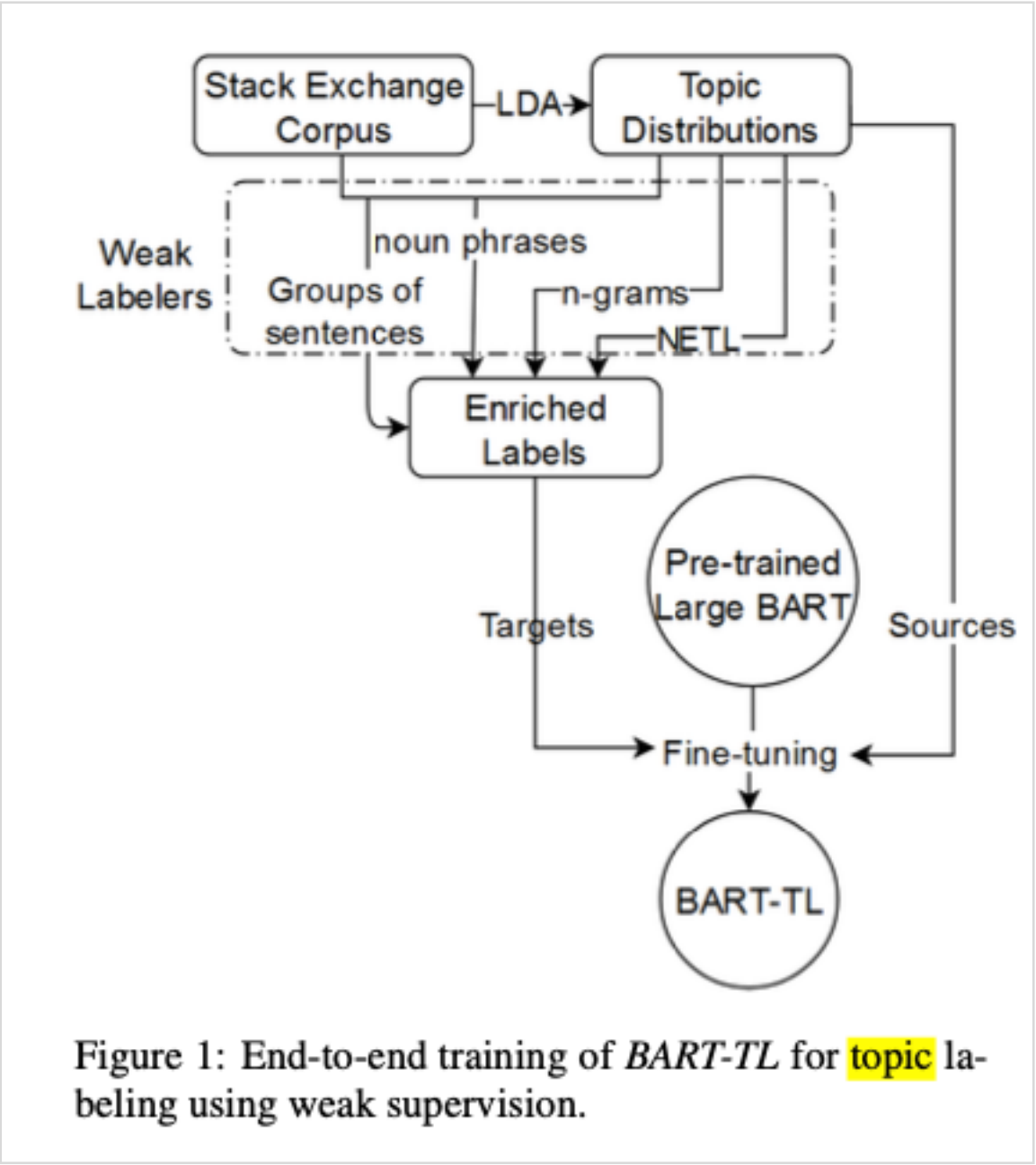

Figure 1: End-to-end training of *BART-TL* for topic labeling using weak supervision.

Table 2: Samples of good and bad quality *new* labels generated by *BART-TL* models.

Top-10 topic terms	Good new labels	Bad new labels
crime center institution chain prison facility prisoner transformation jail custody	criminal justice system administrative court	guarantee principle
plate vehicle state license motor shall registration law apostille issued	driver's license license plate law	no matter what vehiclelicense *(no space)*
rate interest price inflation bond increase real money supply nominal	investment rate discount rate	rate interest rate principle

Motivation

Although the resulting distributions of topic models are useful for computational purposes, such as measuring the similarity of two documents, these may prove difficult to interpret by humans. Topic labeling aims to solve this issue by computing labels for each

topic.

Topic modeling

LDA

Topic modeling parameters

Nr of topics (K): 100

Nr. of topics

100 per subject (those with CV score under 0.3 are removed)
303 topics after pre-processing

Label

Single or multi-word label generated by the fine-tuned Transformer model.

Label selection

\

Label quality evaluation

We focus on both the overall quality of the labels through top-k average rating,
as well as how well the labels are ordered through normalized discounted cumulative gain

We gather annotations in the form of surveys with 7 questions, one per topic, on the quality of topic labels on a scale from 0 to 3.
For each of the 5 subjects in the corpus, we select 6 coherent topics for evaluation.
The labels are taken from the unsupervised and supervised versions of the original NETL method, along with BART-TL-ngram, and BART-TL-all.
For each method, only top-10 labels are considered for evaluation.

An extra stopword label is introduced as a distractor, removing answers from annotators with over 25% of these scores ≥ 1.

A topic is presented using its top-10 terms, along with 2 relevant short paragraphs, to offer additional context when the topic is unclear.
Each survey has balanced topics based on the 5 subjects and each question contains 9 balanced labels based on the models.

Comparison models

BART-TL-ng and BART-TL- all

After having generated the 100 labels per topic in the baseline dataset…

For the weak labelers, we choose to extract 5 n- grams with a n varying between 2 and 4, 5 groups of sentences with a character threshold of 120 and 10 noun phrases with a length of 2 to 4 words that have at least 25 occurrences. We experimented with each strategy individually but provided results for a model employing only the n-grams enrichment BART-TL-ng, and one using all of them, BART-TL- all.

NETL (Bhatia et al., 2016).

This method uses names of Wikipedia articles as candidate labels and trains word2vec and doc2vec models on Wikipedia dumps.

Table 1: Qualitative comparison of labels between *NETL* and *BART-TL* models.

Models	**All**						**English**					
	Top-k Avg.			**nDCG-k**			**Top-k Avg.**			**nDCG-k**		
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5
NETL (U)	2.66	**2.59**	**2.50**	0.83	**0.85**	0.87	2.19	2.46	**2.38**	0.57	0.78	0.84
NETL (S)	**2.74**	2.57	2.49	**0.88**	**0.85**	**0.88**	2.63	2.47	2.28	0.84	0.86	0.86
BART-TL-all (U)	2.64	2.52	2.43	0.83	0.84	0.87	2.58	2.33	2.20	0.81	0.83	0.89
BART-TL-all (S)	2.64	2.55	2.42	0.81	0.84	0.87	2.58	2.36	2.15	0.81	0.86	0.89
BART-TL-ng (U)	2.62	2.50	2.33	0.82	0.84	0.85	2.58	**2.49**	2.26	0.81	**0.91**	**0.93**
BART-TL-ng (S)	2.73	2.46	2.25	0.87	0.83	0.83	**2.75**	2.40	2.21	**0.91**	0.88	0.91

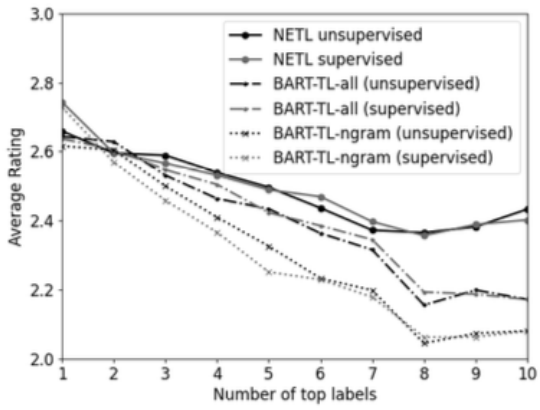

Figure 2: Evolution of average rating considering top-k labels.

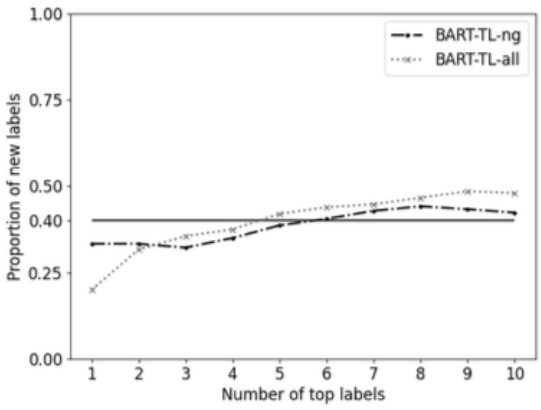

Figure 3: Average proportion of new labels in top-k.

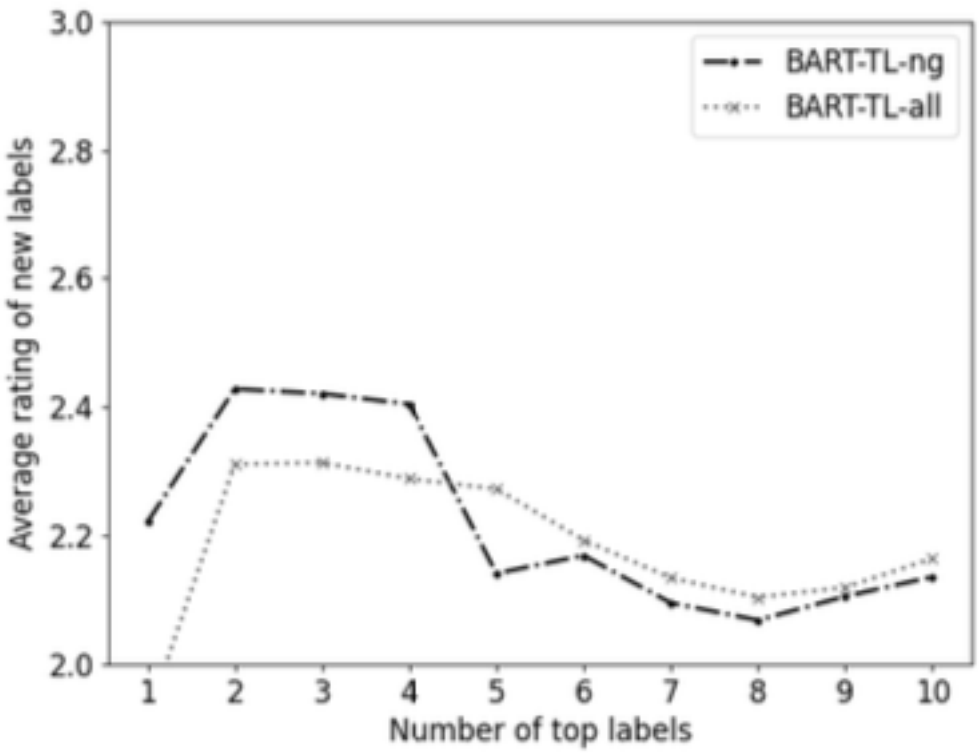

Figure 4: Average rating of new labels in top-k.

Assessors

The annotators have varying backgrounds, including computer science, medicine, law, and economics.

We gathered a total of 35 survey responses and filtered out the labels that had only a single annotation.

This annotation was performed pro-bono and we estimate that the average time per annotated survey was 10 minutes.

Domain

Domain (paper): Topic labeling
Domain (corpus): English, Biology, Economics, Law, and Photography

Problem statement

Proposing a novel solution for assigning labels to topic models by using multiple weak labelers by leveraging generative transformers to learn accurate representations of the most important topic terms and candidate labels.
Achieving this by fine-tuning pre-trained BART models on a large number of potential labels generated by state of the art non-neural models for topic labeling, enriched with different techniques.

Corpus

Fine-tuning dataset (baseline dataset)
Origin: Stack Exchange (stackexchange directory listing)
Nr. of documents: 419,189
Details:

- Corpora crawled from Stack Exchange on 5 different subjects: English, Biology, Economics, Law, and Photography

Document

A single Stack Exchange thread

Pre-processing

Fine-tuning dataset (baseline dataset)
Pre-processing

- Removal of XML artifacts, stop- words, and individual numbers.
- Documents with fewer than 20 words are removed from the corpus, along with words that occur less than 10 and more than 50,000 times.

Topic modeling

- Topics are filtered based on coherence, removing the ones with a CV score under 0.30

Labeling

- Dataset built for fine-tuning BART starting from the NETL labeler (Bhatia et al., 2016).
- Top 100 terms for each topic are considered and used to generate 100 labels per topic
- Initial candidate labels extracted for each topic after the embeddings similarity filtering but modify this process by assigning a greater weight in the scoring based on the importance of the word in the topic distribution.
- To avoid overfitting the most important word, the weights of the top-5 terms are equalised.
- The labels that consist only of stopwords are removed.
- Finally, a one-to-many sequence mapping from topics, represented as a concatenation of the top-20 terms separated by spaces, to the corresponding labels is constructed.
- Other enrichment approaches for this dataset, using other weak labelers as follows.:
  - Entries consisting of space-separated n-grams sampled from the most important words in the topic. The sampling is weighted by the underlying probability distribution and these do not have to be consecutive.
  - Inspired by the work of Gourru et al. (2018), groups of sentences are added as targets using a variant of the COS10 technique for sentence extraction. The best sentences are joined one-by-one into a short paragraph until a minimum character threshold is met.
  - One last idea for improving the baseline dataset is including popular noun phrases from the corpus. They are ranked based on the relevance to the topic and must appear at least a certain number of times in the corpus.

```
@inproceedings{popa_2021_bart_tl_weakly_supervised_topic_label_generation,
    title = "{BART}-{TL}: Weakly-Supervised Topic Label Generation",
    author = "Popa, Cristian  and
      Rebedea, Traian",
    booktitle = "Proceedings of the 16th Conference of the European Chapter of
the Association for Computational Linguistics: Main Volume",
    month = apr,
    year = "2021",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2021.eacl-main.121",
    doi = "10.18653/v1/2021.eacl-main.121",
```

```
    pages = "1418--1425",
    abstract = "We propose a novel solution for assigning labels to topic models by using multiple weak labelers. The method leverages generative transformers to learn accurate representations of the most important topic terms and candidate labels. This is achieved by fine-tuning pre-trained BART models on a large number of potential labels generated by state of the art non-neural models for topic labeling, enriched with different techniques. The proposed BART-TL model is able to generate valuable and novel labels in a weakly-supervised manner and can be improved by adding other weak labelers or distant supervision on similar tasks.",
}
```

#Thesis/Papers/Initial