BART-TL: Weakly-Supervised Topic Label Generation

Cristian Popa Traian Rebedea
University Politehnica of Bucharest

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.

Introduction

Topic modeling is a popular unsupervised method for exploring large corpora of documents. Topics are represented as distributions over words, while documents as mixtures of topics. This NLP task is typically solved using Blei's LDA, which we used ourselves throughout all experiments.

While 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. Historically, this was achieved by establishing a pool of labels and ranking them using certain scoring functions.

The main reference topic labeling method we used was the one described in the paper "Automatic Labelling of Topics with Neural Embeddings" (Bhatia et al., 2016), which we will refer to as "NETL" from here on. The authors extract topics using LDA from different sources (blogs, books, news, PubMed) and use titles of Wikipedia articles as candidate labels. They gather annotator feedback to obtain gold-standard labels and showcase the performance of supervised and unsupervised rankers.

Methodology and Performance

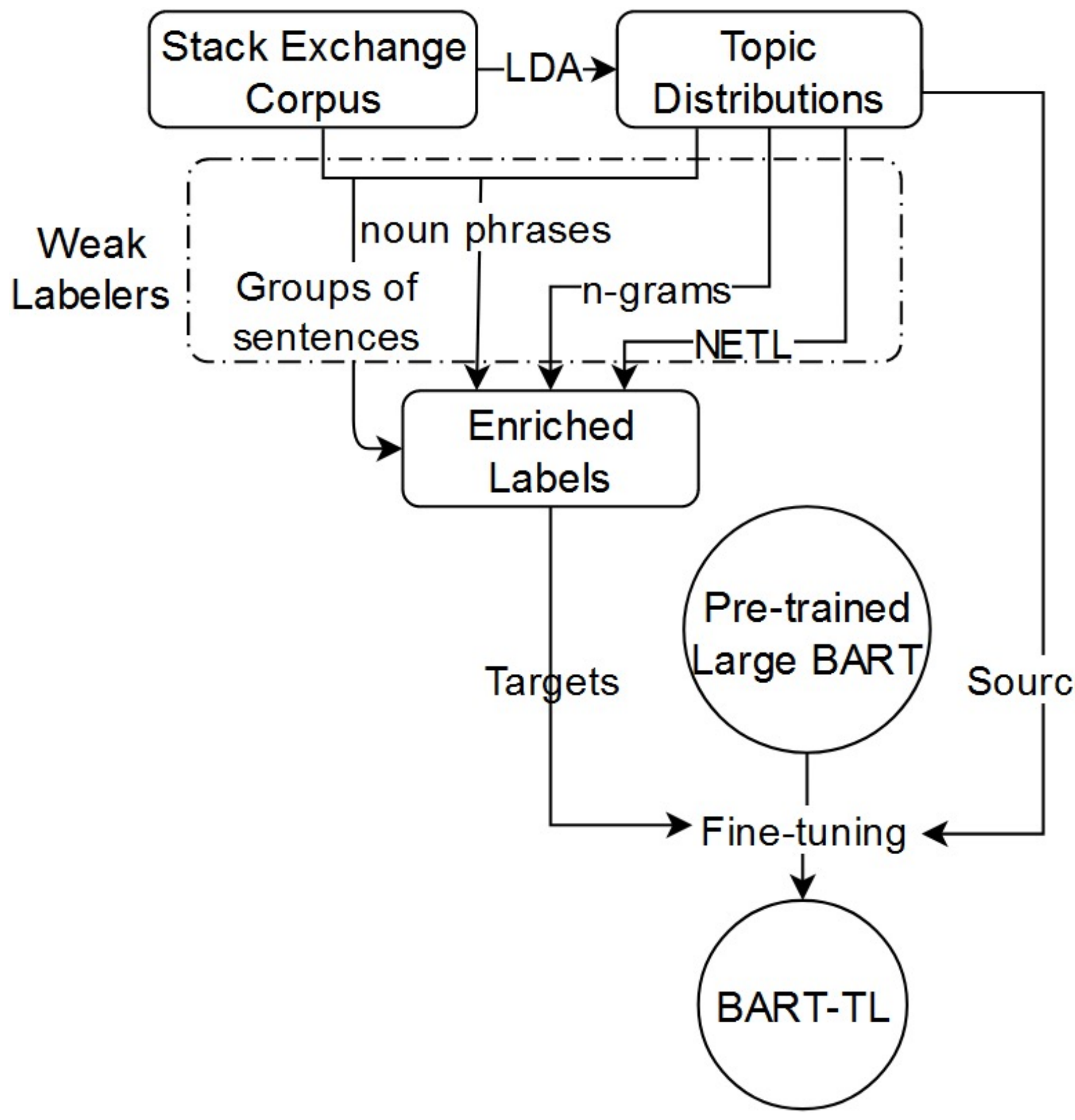
We aim to reframe the task of topic labeling as a sequence-to-sequence task, rather than a ranking one. Our method utilizes a pre-trained BART (Lewis et al., 2019) transformer model, with a denoising autoencoder architecture, hence the name *BART-TL*.

To finetune the BART model, we first construct a weakly-supervised dataset. We apply LDA (Blei et al. 2003) on posts crawled from the Stack Exchange forums on 5 different subjects: English, Biology, Economics, Law, and Photography. These are thoroughly pre-processed beforehand by removing, among others, XML artifacts and stop-words.

We label the sequence-to-sequence dataset by starting from the NETL labeler, extracting initial candidate labels for each topic. Because the number of topics is low, with an upper limit of 100 topics for each of the 5 subjects, we decided to enhance the training data by having the dataset be a one-to-many mapping from topics, represented as the top-20 terms in the topic distribution, to labels. These labels are the NETL-extracted labels, along with our own enrichments:

- Random words sampled from the topic distribution (n-grams)
- Short paragraphs relevant to the topic (Gourru et al., 2018)
- Noun phrases found relevant to the topic

We finetuned models using the combination of NETL labels and n-grams (BART-TL-ng), as well as all of them (BART-TL-all). The finetuning of the pre-trained BART model uses the recommended hyper-parameters from the RoBERTa paper (Liu et al., 2019).



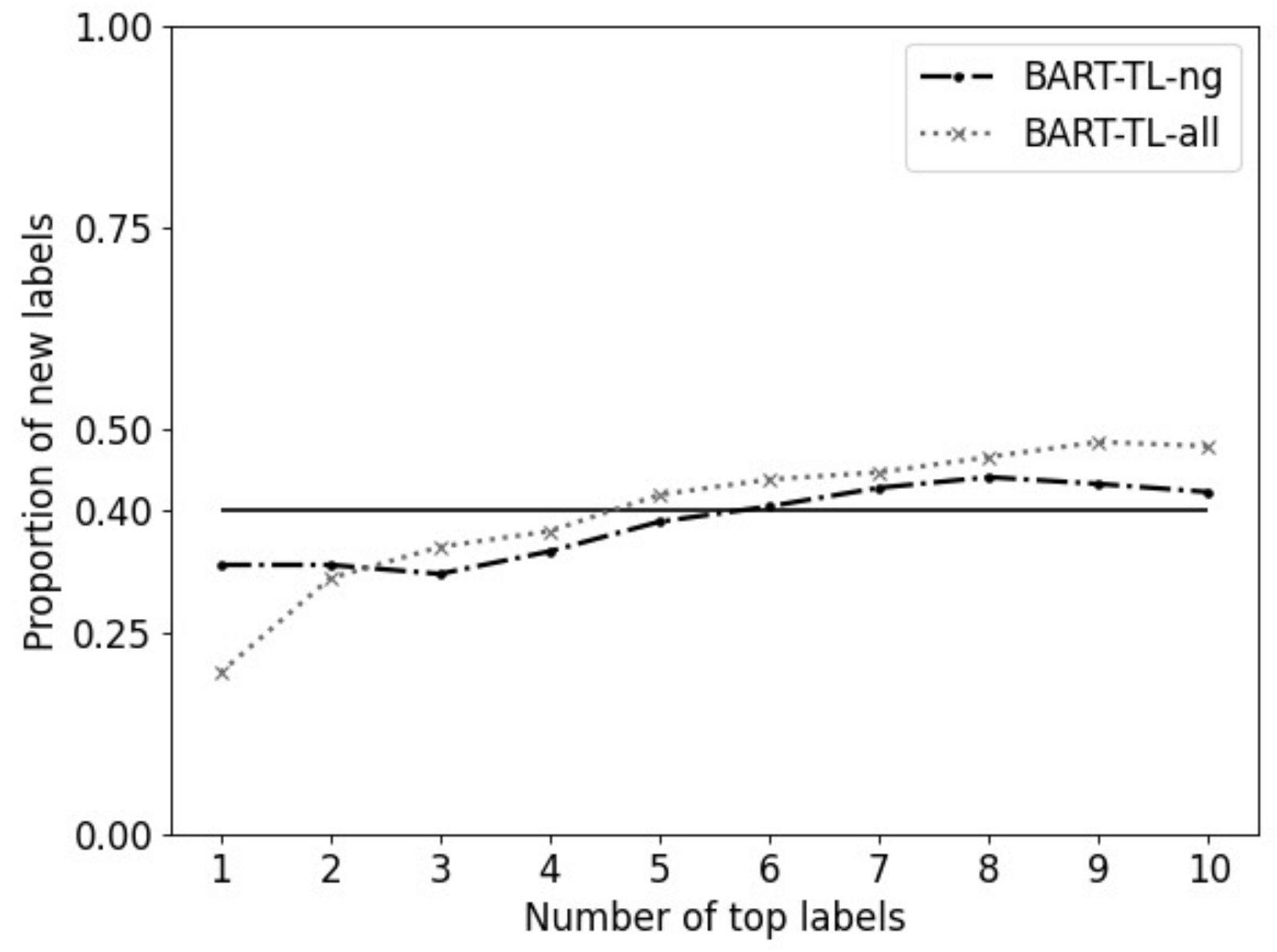
To assess the performance of our method, we make a comparison against the NETL supervised and unsupervised rankers. The top labels generated by the BART-TL models, as well as the ones ranked highly by the NETL supervised and unsupervised methods are sent to human annotators for feedback in the form of surveys. The responses are filtered using control labels that have no relevance (e.g. stop-words).

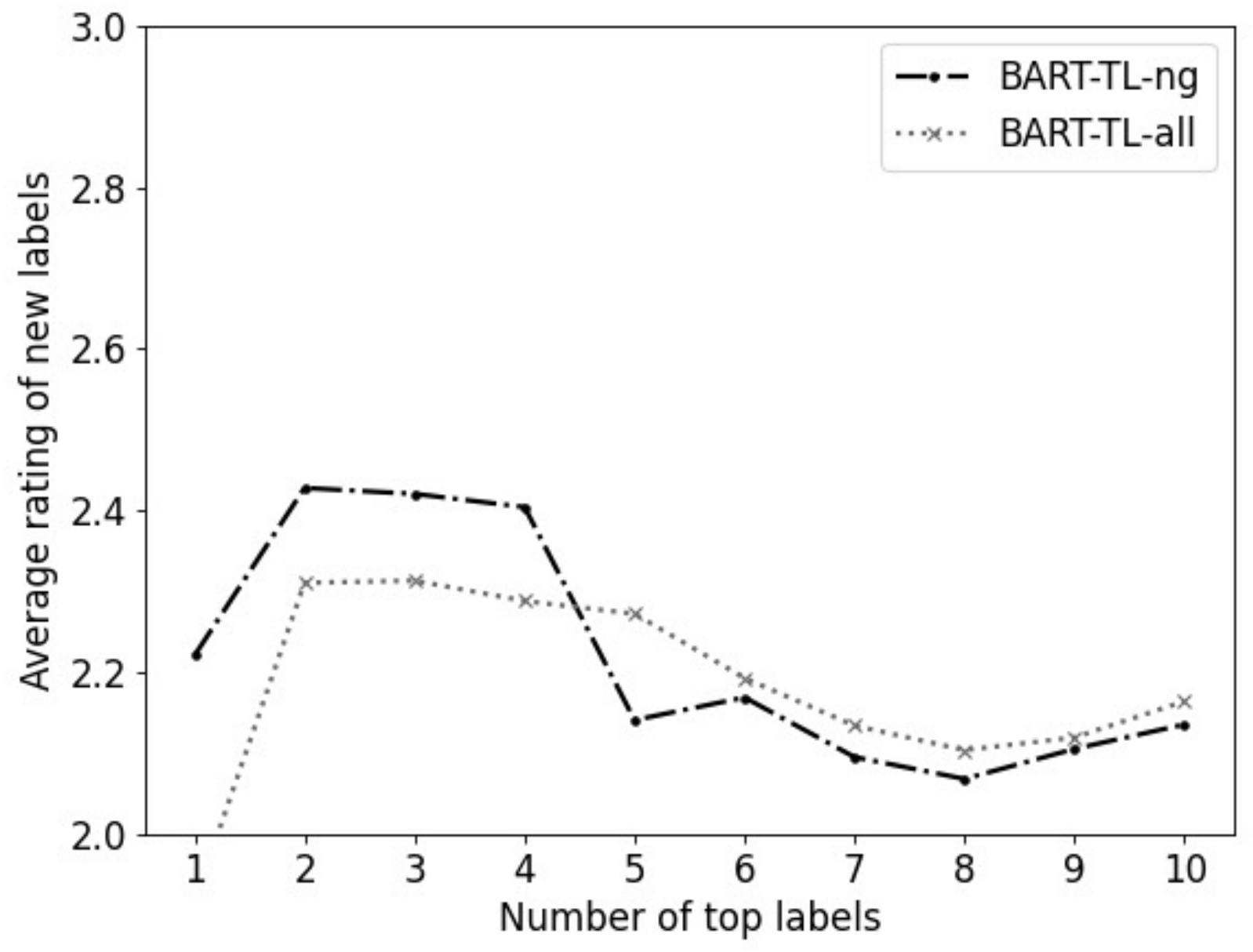
The main results are presented below. We are interested in the both the overall quality of the labels, measured using the top-k average annotator rating, as well as the order of the best labels, measured using nDCG-k.

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

Additional Results

A great advantage of generative methods is that the models can create original labels for the topics. We are interested in this model novelty, how frequent is it and whether these novel labels are good. The results show that ~40% of labels are completely new, never seen in the finetuning data, and that their average rating does not stray far from regular labels:





Conclusion

We introduced the *BART-TL* model that builds upon previous topic labeling solutions by adopting a generative deep learning strategy. Large transformer models are fine-tuned in a weakly-supervised manner using unsupervised labelers to obtain meaningful labels. While current results have varying quality compared to NETL, BART-TL is able to generate novel labels of similar quality.