NLP Course. Topic modeling and classification news on Hebrew with Neural Text Summarizer model

Vladimir Gurevich

December 2020

Abstract

This document is a report for the final project of the NLP course. Code repository: https://github.com/imvladikon/huawei-nlpcourse-project.

1 Introduction

Topic modeling methods are powerful smart techniques that widely applied in natural language processing to topic discovery, semantic mining from unordered documents. It's a probabilistic model that can help to find hidden semantic structure through massive amounts of raw text and cluster similar groups of documents together in an unsupervised way. The main topic modeling technique is LDA and it has applications in many NLP tasks such as classification. In this work, we tend to use LDA output for a news classification with different contextual embeddings in order to improve results and apply a neural text summarizer based on a variational autoencoder to decrease the dimension of the embeddings of the big new articles.

1.1 Team

Vladimir Gurevich prepared this document and the author of the project.

2 Related Work

As was mentioned, primarily topic models based on the Latent Dirichlet allocation (LDA) [Blei et al., 2003], where LDA is a generative probabilistic model of a corpus that was introduced by Blei, Ng, and Jordan in 2003 [1]. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. The development of this idea were multi-grain topic models [Titov and McDonald, 2008] and lda2vec [Moody, 2016] combining LDA and Word2vec [Mikolov et al., 2013] approach. Unsupervised aspect extraction has previously been presented with topic model hybrids [García-Pablos et al., 2018] - a topic modeling approach that biases word-aspect associations by computing the similarity from a word to a set of aspect terms. Using attention for topic modeling also was represented in the paper - ABAE - [He et al., 2017], where the main approach based on the attention mechanism [Bahdanau et al., 2015] for aspect extraction. Recently the Contrastive Attention was introduced [Tulkens and van Cranenburgh,] as a single-head attention mechanism based on an RBF kernel for aspect identification in sentiment analysis. In addition in this method, each document labeled based on the cosine similarity between the weighted document vector and the label vector. Interesting approach was proposed recently in [Hoyle et al., 2020], - using knowledge distillation to combine the best attributes of probabilistic topic models and pretrained transformers with the implementation of different Dirichlet Variational Autoencoder Topic Models. Another one usage of the transformers was proposed also in Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence [Bianchi et al., 2020], and it's a similar approach to current work and purpose of it, - the solutions are often not coherent enough, and thus harder to interpret. Coherence can be improved by adding more contextual knowledge to the model [Roee Aharoni, 2020]. Recently, neural topic models have become available, while BERT-based representations have further pushed the state of the art of neural models in general. And the combination of the pre-trained representations and topic models could improve it. But in the cited paper used BERT sentence embeddings pre-trained model, that have a limitation on large text (Bert model architecture limitation, is 512 tokens, although at nowadays there are some models, such as longformer [Beltagy et al., 2020] that solve this problem)

3 Model Description

In terms of the LDA modeling, we have a preprocessed corpus of texts that is fed to the model, where input sample is a list of indexes for words in a review sentence s. Each word w in vocabulary maps with a vector $e_w \in R^d$ from the embedding matrix $\mathbf{E} \in \mathbb{R}^{V \times d}$, where V - the vocabulary size. For each word in the sentence, a positive weight a_i is computed by an attention model, which is conditioned on the global context of the sentence:

$$a_i = \frac{exp(d_i)}{\sum_{j=1}^{n} exp(d_j)}$$

$$d_i = \mathbf{e_{w_i}^T} \cdot \mathbf{M} \cdot \mathbf{y_s}$$

$$\mathbf{y_s} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{e_{w_i}}$$

where $\mathbf{y_s}$ is the average of the word embeddings, which supposed to capture the global context of the sentence. $M \in \mathbb{R}^{d \times d}$ is a matrix which is learned as in the training process, $i = 1, ..., n$ are the word indexes in the sentence. A vector representation $\mathbf{z_s}$ constructs as:

$$\mathbf{z_s} = \sum_{i=1}^{n} a_i \mathbf{e_{w_i}}$$

Sentence embeddings of the filtered sentences $\mathbf{z_s}$ transforms into their reconstructions $\mathbf{r_s}$. To obtain the weight vector over K aspect embeddings $\mathbf{p_t}$, $\mathbf{z_s}$ is reduced from d to K dimensions. Applying a softmax non-linearity makes normalized non-negative weights:

$$\mathbf{p_t} = softmax(\mathbf{W} \cdot \mathbf{z_s} + \mathbf{b})$$

where each weight represents the probability that the input sentence belongs to the related aspect. **W** is the weighted matrix parameter and **b** is the bias vector, they are learned in the training process.

Now, the reconstructed vector representation of sentence is obtained as a linear combination of aspect embeddings from **T**. An aspect embedding matrix is $\mathbf{T} \in \mathbb{R}^{K \times d}$, where K - the number of aspects.

$$\mathbf{r_s} = \mathbf{T^T} \cdot \mathbf{p_t}$$

And then we create the represatations of the news text article. The embedding for the text sequence of an article is obtained by using a pre-trained Bert model initialized from HeBERT - pre-trained Hebrew language model [Avichay Chriqui, 2020] followed by a bidirectional GRU as a text summarizer[figure 2].

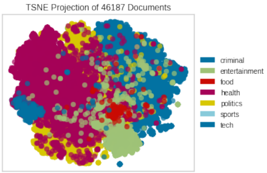

Figure 1: tSNE on the summarizer's embeddings.

This is done, as the pre-trained BERT model has a fixed maximum sequence length (512) of n limiting the sequential scope of the text. The news articles length could be a very long (max length of the article in the dataset is 63773 tokens). In order to obtain one text representation, the aggregate articles are batched into a set of sentences $(u_1, u_2, , u_m)$, where $u_i \in Z_n$ and m is the maximum occurring length in the corpus after batching each news text into sentences of n words. The resulting set of sentences for one record is embedded using the BERT model which results in a matrix $U \in R^{md_{BERT}}$

$$h_{1:m} = GRU_{enc}(U_m, h_0)$$

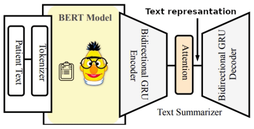

Figure 2: The text representation module takes as input article text and preprocesses them into tokens, which are embedded using Bert. Subsequently, it is fed to a text summarizer autoencoder network. In this work, a LSTM-AE is used to learn the text representation.

$$E_U = softmax(\frac{h_{1:m}h_{1:m}^T}{\sqrt{d^{text}}}h_{1:m}$$

$$\hat{U} = GRU_{dec}(U_m, h_m)$$

$$L(U_{1:m}, \hat{U}_{1:m}) = \sum_{i}^{m}(u_i - \hat{u}_i)^T(u_i - \hat{u}_i)$$

Main idea was proposed by the paper [Darabi et al., 2019]"TAPER: Time-Aware Patient EHR Representation"

Subsequently, the set of sentence are summarized into a single article text representation using a neural text summarizer module. This module follows an auto-encoder architecture with GRU's as the building block. The input is a set of sentence representations obtained by the BERT module applied on the aggregate text followed by a self-attention head on the hidden representations, where the decoder is tasked to output a sentence representations following the bottleneck. The objective of the neural text summarizer is reducing the MSE loss objective between the input of the text embeddings and the models predicted representation at the correspondingly. Finally, the text representation is obtained by summarizing the set of sentence representations $U_1, U_2, ..., U_m$, where the output of the attention head applied on the encoders hidden representations is used as the representation. For comparing results were used also other embeddings: Hebrew fasttext (mean pooled on text and on title), multilangual Bert model, language agnostic Bert model.

4 Dataset

Dataset was built from scratch. Top 10 news, economy, and technology-related sites were scraped, such as walla.co.il, calcalist.co.il, kikar.co.il, globes.co.il, maariv.co.il, ynet.co.il, geektime.co.il etc. There are 164606 records in the dataset and 46147 was annotated (reduced) according to related aspect label, such as Sports, Tech, Politics, Economics, Health, Criminal, Entertainment, Food, using keywords, tags (SEO fields and metadata gathered from html page) and category that were also scraped from the sites. Datasets fields are article body, title, or headline of the news, keywords, description and category. The repository contains both datasets, the original one with raws categories and processed. As post-processed stage was performed comparison inference from classification by several Bert models (DistillBERT, multilingual Bert) on Hebrew title field and title translated (machine translation) to English, with the annotated label in terms of fixing, incorrect or ambiguous labels, when each model had other predictions, part of such record were fixed, part were removed. Hebrew Bert model wasn't used in order to prevent any leakages in evaluation part where was used this model. In spite of this, need to note that the dataset is not reliably annotated, but probably it's the first publicly available dataset on Hebrew with news classification that contains around 50k annotated records and 160k raws records. The Link to the dataset there is in the repository

	Train	Valid	Test
Articles	36949	4619	4619
Tokens	13,277,942	1,630,720	1,646,100

Table 1: Statistics of the dataset.

5 Experiments

5.1 Metrics

We seek to discover a latent space of topics that is meaningful and useful to people (Chang et al., 2009). Accordingly, we evaluate topic coherence, which is significantly correlated with human judgments of topic quality (Aletras and Stevenson, 2013; Lau et al., 2014) and widely used to evaluate topic model. It is calculated as follows:

$$C(z,S^z) = \sum_{n=2}^{N}\sum_{l=1}^{n-1} \log \frac{D_2(w_n^z, w_l^z)+1}{D_1(w_l^z)}$$

for aspect z and its set of N top words $S^z = \{w_1^z, ..., w_N^z\}$, where $D_1(W)$ is the document frequency of word w and $D_2(w_1, w_2)$ is the co-document frequency of words w_1 and w_2. A higher coherence score indicates a better aspect semantic coherence.

For a classification task metrics are: precision, pecall and f1-score for each news topic and classification report from sklearn.

5.2 Experiment Setup

Firstly text, titles were prepossessed, were removed punctuation symbols, diacritical signs (except geresh '), and stop words that was defined in the project based on spacy stop words defined for Hebrew. For LDA was used Gensim implementation with tuning on coherence-score with number of the topics - 8,20,25, and Dirichlet priors - 0.05, 0.08, 0.1 with 1000 iterations.

We initialize the BERT model with the pre-trained weights on Hebrew language as presented in [Avichay Chriqui, 2020], and was created vector embeddings of the articles with dim [20x768]. To train the text summarizer we use a 2-layer bidirectional GRU autoencoder with the intermediate representation set to denc = 128. A teacher-forcing ratio of 0.5 is used with a step learning rate schedule decay of 0.1 every 30 epochs with initial lr set to 10^3. And was obtained final embeddings with dim [512]. also was created different embeddings - fasstext, bert embeddings on title and language agnostic embeddings.

5.3 Baselines

- LDA model from Gensim
- Random forest classifier on pre-trained embeddings

6 Results

The results are not the best, we need to pay attention to the unreliable way of the annotation datasets. But even so, adding summarization embeddings improved the model according to F1-score. For classification task was performed simple Random forest classifier without tuning and hyperparameter optimization, the main goal was to show the impact of the contextual embeddings and embeddings that was gotten from neural feature extractor (summarizer). And finally, as features were used concatenation of the output of LDA on test corpus (distribution of the probabilities topics) and pre-trained embeddings.

Model	alpha|topics	8	20	25
LDA	0.05	0.483	0.56	0.584
	0.08	0.492	0.558	0.579
	0.1	0.484	0.553	0.592

Table 2: Coherence scores, LDA model

7 Conclusion

Was done work by collecting news dataset with articles, descriptions, and categories. Performed a default LDA model using Gensim implementation to establish the baseline coherence score and reviewed practical ways to optimize the LDA hyperparameters. Then LDA topics distributions, contextual embeddings, neural text summarizer features were used for classification tasks and compare with each other. The best result was done by Bert Hebrew embeddings model extracted from its neural features by VAE model.

Model	Labels	Precision	Recall	F1-score
RF+Neural article features	sports	0.90	0.48	0.63
	tech	0.92	0.95	0.93
	politics	0.87	0.84	0.85
	entertainment	0.94	0.82	0.88
RF+LABSE(title)	sports	0.86	0.49	0.62
	tech	0.93	0.93	0.93
	politics	0.85	0.82	0.84
	entertainment	0.94	0.82	0.88
RF+fasttext(mean)	sports	0.77	0.52	0.62
	tech	0.93.	0.93	0.93
	politics	0.81	0.83	0.82
	entertainment	0.89	0.79	0.84

Table 3: Results of models

business	"Experience, Life, Develop, Advantage, Jobs"
economics	"billion, most, this year, times, per, year,company"
politics	"Netanyahu, Government, Head, White, Ganz, Benjamin, Law, Knesset"

Table 4: Some words corresponding to the news labels(here machine-translated representation of the words from Hebrew to English)

References

[Avichay Chriqui, 2020] Avichay Chriqui, Inbal yahav, T. C. S. L. A. L. (2020). Hebrew bert pretrained model.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

[Beltagy et al., 2020] Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer.

[Bianchi et al., 2020] Bianchi, F., Terragni, S., and Hovy, D. (2020). Pretraining is a hot topic: Contextualized document embeddings improve topic coherence.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

[Darabi et al., 2019] Darabi, S., Kachuee, M., Fazeli, S., and Sarrafzadeh, M. (2019). TAPER: time-aware patient EHR representation. volume abs/1908.03971.

[García-Pablos et al., 2018] García-Pablos, A., Cuadros, M., and Rigau, G. (2018). W2vlda: Almost unsupervised system for aspect based sentiment analysis. Expert Syst. Appl., 91:127–137.

[He et al., 2017] He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2017). An unsupervised neural attention model for aspect extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada. Association for Computational Linguistics.

[Hoyle et al., 2020] Hoyle, A. M., Goel, P., and Resnik, P. (2020). Improving Neural Topic Models using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1752–1771, Online. Association for Computational Linguistics.

[Mikolov et al., 2013] Mikolov, T., tau Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In HLT-NAACL.

[Moody, 2016] Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. ArXiv, abs/1605.02019.

[Roee Aharoni, 2020] Roee Aharoni, Y. G. (2020). Unsupervised domain clusters in pretrained language models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, page 7747–7763.

[Titov and McDonald, 2008] Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08.

[Tulkens and van Cranenburgh,] Tulkens, S. and van Cranenburgh, A. Embarrassingly simple unsupervised aspect extraction. arXiv:2004.13580 [cs.CL].