Topic modeling on Tatar news with An Attention Based Aspect Extraction model

Danis Sayfullin

December 2020

Abstract

This document is a report for project on topic modeling.
`https://github.com/dantatartes/tatar_topic_modeling`

1 Introduction

The Tatar language experience both lack of implemented natural language processing techniques and ready to use data, but demand for them is relatively big. Existing approaches to common NLP problems work well in widespread languages.

The goal of this project is to implement a full pipeline for aspect extraction with An Unsupervised Neural Attention Model on the dataset we have collected ourselves.

1.1 Team

Danis Sayfullin made this project.

2 Related Work

Initially, methods were mainly rule-based: [Hu and Liu, 2004] proposed to extract different product features through finding frequent nouns and noun phrases. They also extracted opinion terms by finding the synonyms and antonyms of opinion seed words through WordNet.

But most topic models build on latent Dirichlet Allocation [Blei et al., 2003]. LDA is a three-level Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. The majority of existing works [Brody and Elhadad, 2010, Zhao et al., 2010, Mukkerjee and Liu, 2012, Chen et al., 2014] are based on variants and extensions of LDA.

In contrast to LDA-based models, Attention Based Aspect Extraction model [He et al., 2017], which will be more detailed described below, explicitly encodes word-occurrence statistics into word embeddings, uses dimension reduction to

extract the most important aspects in the corpus, and uses an attention mechanism to remove irrelevant words to further improve coherence of the aspects.

The authors [Luo et al., 2019] presented a hierarchical model AE-CSA which is similar to ABAE. In addition to word vectors and aspect vectors, this model also considers sense and sememe vectors in computing the attention distribution. AE-CSA needs a dictionary of senses and sememes, which might not be available for all languages or domains.

3 Model Description

Each input sample is a list of indexes for words in a news article s. Each word w in vocabulary maps with a vector $e_w \in R^d$ from the pretrained 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. $\mathbf{W}$ is the weighted matrix parameter and $\mathbf{b}$ 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 $\mathbf{T}$. An aspect embedding matrix is $\mathbf{T} \in \mathbb{R}^{K \times d}$, where K - the number of aspects.

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

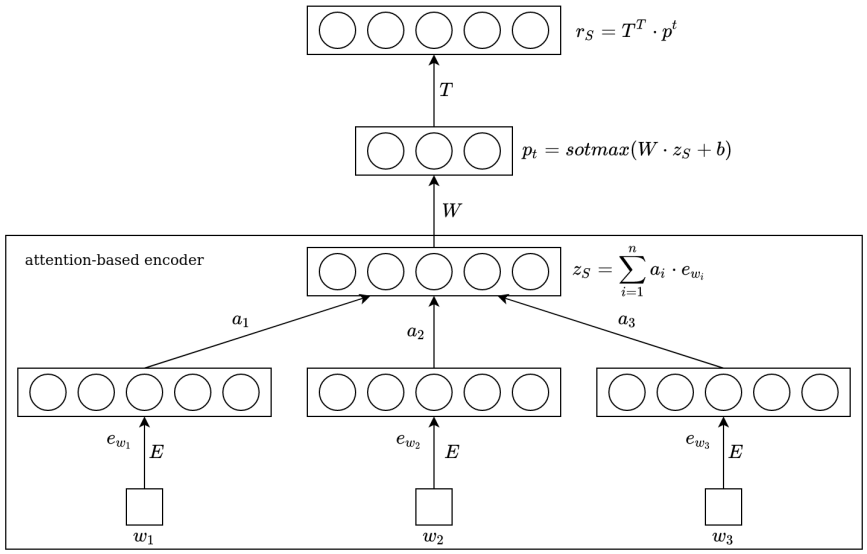

Figure 1: Structure of the ABAE model.

4 Dataset

Dataset under this project contains more than 62000 news from `https://tatar-inform.tatar/` website related to the period from 25 February 2020 to 24 February 2017.

We selected news with only five labels: agriculture, culture, incident, religion, and sport because they are less intersecting each other and show more reliable annotation.

	Train	Test
Articles	18449	4613

Table 1: Statistics of the selected part of the dataset.

The news was parsed using Python and Selenium library, stored in CSV format. Parser's code and data are available and put in the project's repository. Each row in the dataset includes a headline, topic, date, and full text.

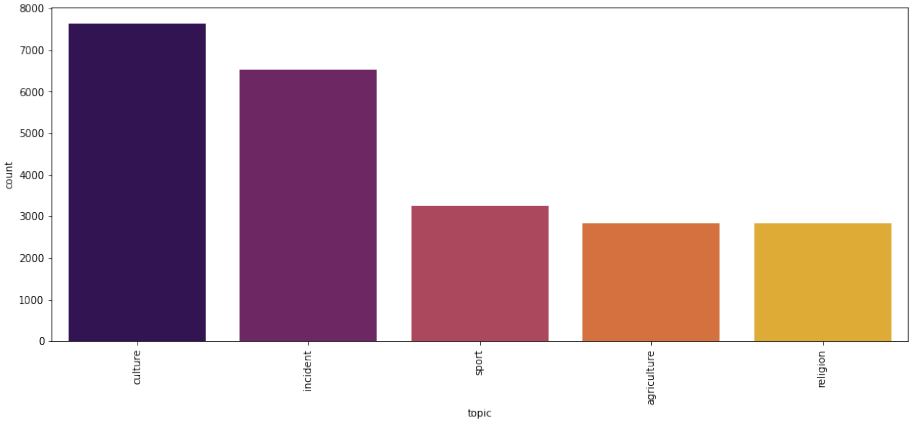

Figure 2: Number of samples on each selected topic.

	title	topic	date	text
0	Чаллыда яшәүче мөмкинлекләре чикле кешеләр күр...	society	25-11-2020	(Казан, 25 ноябрь, «Татар-информ», «Шәһри Чалл...
1	Актанышта бушлай дарулар бирә башладылар	health	25-11-2020	(Казан, 25 ноябрь, «Татар-информ», «Актаныш»)....
2	Шәле балалар бакчасында татарча курчак спектак...	science	25-11-2020	(Казан, 25 ноябрь, «Татар-информ», Гөлнар Гари...
3	Татарстанның җиде зур шәһәрендә бер атна эченд...	society	25-11-2020	(Казан, 25 ноябрь,«Татар-информ»). Ноябрь дәва...
4	Татарстанда болытлы һава торышы, кар һәм бозла...	ecology	25-11-2020	(Казан, 25 ноябрь, «Татар-информ»). 26 ноябрьд...
...	...	...	...	...

Figure 3: Dataset examples.

At the preprocessing stage, all first sentences with (location, date, news agency) pattern were deleted. All fully duplicated rows were dropped.

5 Experiments

5.1 Metrics

The model was evaluated by precision, recall, and F1 scores and their micro and weighted averages. Classes of aspects were tagged manually.

5.2 Experiment Setup

For the dataset, lowering, punctuations stopwords removal, and lemmatization were applied.

For LDA, we set Dirichlet priors $\alpha = 0.05$ and $\beta = 0.1$, run 1000 iterations.

We trained word2vec on 16 epochs with embedding size 200, window size 5, and negative sample size 5 with vocabulary size restricted to 20000.

The aspect embedding matrix initialized with the centroids of clusters results resulting from running k-means on word embeddings. Other parameters are

initialized randomly.

During the training process, the word embedding matrix was fixed, other parameters were optimized using Adam with a learning rate 0,001 for 16 epochs and batch size of 50. The number of negative samples was set to 15.

The number of aspects for the corpus was set to 15.

5.3 Baselines

LDA model implementation from Gensim library was used as a baseline.

6 Results

On average, ABAE model shows better results in Recall and F_1 metrics but loses Precision. ABAE model has better performance on most (culture) and least (agriculture and religion) presented labels.

Some words corresponding to aspect labels are listed in Tab. 4.

Aspect	Method	Precision	Recall	F_1
culture	LDA	0.654	**0.975**	0.783
	ABAE	**0.872**	0.893	**0.883**
incident	LDA	**0.982**	0.925	**0.953**
	ABAE	0.706	**0.986**	0.823
sport	LDA	**0.971**	**0.518**	**0.676**
	ABAE	0.927	0.486	0.638
agriculture	LDA	0.769	**0.857**	0.811
	ABAE	**0.856**	0.825	**0.840**
religion	LDA	**0.957**	0.238	0.381
	ABAE	0.872	**0.522**	**0.653**

Table 2: Aspect identification results per each label.

Average	Method	Precision	Recall	F_1
macro	LDA	**0.867**	0.703	0.721
	ABAE	0.847	**0.742**	**0.767**
weighted	LDA	**0.844**	0.792	0.771
	ABAE	0.830	**0.808**	**0.798**

Table 3: Average aspect identification results.

culture	спектакль, әкият, театр, режиссёр performance, fairy tale, theatre, producer
incident	пычак, җинаять, полиция, наркотик knife, criminal, police, drug
sport	ярыш, көндәш, азеведо, хоккей competition, competitor, azevedo, hockey
agriculture	ферма, сөт, гектар, терлек farm, milk, hectare, cattle
religion	мөселман, намаз, аллаһ, мәчет muslim, namaz, god, mosque

Table 4: Aspect labels and some coherence words to them with translation.

7 Conclusion

By performing this project we collected a dataset of Tatar news, trained the Attention Based Aspect Extraction model on it, and achieved better results on the majority of metrics compared to LDA.

References

[Blei et al., 2003] Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. volume 3, pages 993–1022.

[Brody and Elhadad, 2010] Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews.

[Chen et al., 2014] Chen, Z., Mukherjee, A., and Liu, B. (2014). Aspect extraction with automated prior knowledge learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

[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.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Dicovery and Data Mining*.

[Luo et al., 2019] Luo, L., Ao, X., Song, Y., Li, J., Yang, X., He, Q., and Yu, D. (2019). Unsupervised neural aspect extraction with sememes. pages 5123–5129.

[Mukkerjee and Liu, 2012] Mukkerjee, A. and Liu, B. (2012). Aspect extraction through semi-supervised modelling. In *Proceedings of the 50th Annual Meetings of the Association for Computational Linguistics.*

[Zhao et al., 2010] Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010). Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Emperical Methods in Natural Language Processing.*