Topic-driven Ensemble for Online Advertising Generation

Egor Nevezhin
ITMO University
egornevezhin@itmo.ru

Nikolay Butakov
ITMO University
alipoov.nb@gmail.com

Maria Khodorchenko
ITMO University
mariyaxod@yandex.ru

Maxim Petrov
ITMO University
djvipmax@gmail.com

Denis Nasonov
ITMO University
denis.nasonov@gmail.com

Abstract

Online advertising is one of the most widespread ways to reach and increase a target audience for those selling products. Usually having a form of a banner, advertising engages users into visiting a corresponding webpage. Professional generation of banners requires creative and writing skills and a basic understanding of target products. The great variety of goods presented in the online market enforce professionals to spend more and more time creating new advertisements different from existing ones. In this paper, we propose a neural network-based approach for the automatic generation of online advertising using texts from given webpages as sources. The important part of the approach is training on open data available online, which allows avoiding costly procedures of manual labeling. Collected open data consist of multiple subdomains with high data heterogeneity. The subdomains belong to different topics and vary in used vocabularies, phrases, styles that lead to reduced quality in adverts generation. We try to solve the problem of identifying existed subdomains and proposing a new ensemble approach based on exploiting multiple instances of a seq2seq model. Our experimental study on a dataset in the Russian language shows that our approach can significantly improve the quality of adverts generation.

1 Introduction

The task of abstractive summarization assumes getting short but still a semantically relevant representation of some text. The summarization helps advertisement to catch user attention, get him or her interested, and finally to engage in the full text reading. This is especially true for online advertising when it flashes in front of eyes on most website in the Internet.

Crafting such short texts for online advertising requires a level of mastering and some degree of subject understanding. Humans doing this job have to read at least the base text, which describes products or services and then to spend additional time thinking of suitable describing phrase. That may be very time-consuming and thus a good option for automation.

Modern natural language processing methods and techniques allow solving the task up to some degree of quality while still leaving space for improvement. The most advanced methods are based on neural network seq2seq models (Gehrmann et al., 2018) and consume large manually labeled datasets for supervised learning.

Due to being so popular and widespread, online advertising gives us an opportunity to obtain labeled data (in this particular case, a pair of the source text and banner text) using only data crawling techniques significantly decreasing human efforts needed to build a sufficient dataset for the task.

At the same time, ollected open data may consist of multiple subdomains with high data heterogeneity. Moreover, what data will come from the crawling depends on advertising providers and their strategies adopted to the end-user. The subdomains belong to different topics and may vary substantially in used vocabularies, phrases, text styles. This heterogeneity may be amplified by overrepresentation of some domains while underrepresentation of others. As it will be demonstrated in our experimental study, all

Proceedings of the 28th International Conference on Computational Linguistics, pages 2273–2283
Barcelona, Spain (Online), December 8-13, 2020

mentioned above may negatively affect the quality of adverts generation especially for text belonging to some topics leading semantic inconsistencies or wrong endings.

In this work, we try to smooth the consequences of such peculiarities in data by addressing arising issues with direct topic identification in texts and extracting groups of semantically similar texts in the dataset. It allows multiple models ensemble to be built with fine-tuned specifics to extract topic-driven groups. Eventually, this approach demonstrates improvement in the overall quality of generation.

Contributions of the paper are as follows: (1) a case study of online advertising generation as an abstractive summarization problem; (2) a new topic-driven ensemble approach; (3) an experimental study on Russian dataset of the new approach performed on online advertising generation task that demonstrates improved quality of the approach.

2 Related work

There are multiple efforts in scientific community directed to automate building of advertising and presenting it in a right way (Rosenkrans, 2010; Liu et al., 2018; Kim and Moon, 2020; Aguilar and Garcia, 2018). Some works analyze impact of banners' visual appearance to user behavior and psychological reactions (Rosenkrans, 2010; Liu et al., 2018) thus answering what banner should contain. Others try to automate time-consuming job of banners generation (Kim and Moon, 2020; Aguilar and Garcia, 2018) including smart banners concept that creates personalized banners. These works assume different approaches such as constructing a banner using already prepared parts or generating several banners by substituting words in a template or initial banner with synonyms. Our approach is different in two things: first, it works directly with the text from product page instead of altering some base banner or template text and thus it doesn't require additional efforts to prepare these materials; second, we use banner generation methods that differ from the ones from existing works – based on a modern neural architecture called Transformer.

In the field of natural language generation, one can find several major approaches based on encoder-decoder architectures: LSTM-based solutions with or without attention layers (Nema et al., 2017; Song et al., 2019), pointer-generator model (See et al., 2017) and transformer-based solutions gained significant popularity recently (Gehrmann et al., 2018; Gatt and Krahmer, 2018; Gavrilov et al., 2019). While pointer-generator specifically oriented on solving problems with the tendency of recurrent models to reproduce details incorrectly, caused by an unbalanced dictionary (a rare word is replaced by a more common one), Transformer may give better quality and easily parallelize. Work (Gavrilov et al., 2019) is dedicated specifically for generating short news headlines from articles and has similar conditions of text generating and methods being applied. However, online advertising, in contrary to news headlines, has a different style (related to getting the user to be engaged to sell him/her something) that should be reflected in the generated text. Moreover, our approach goes further dealing with heterogeneity in data and utilize ensemble technique.

Concerning approaches working with dataset peculiarities and exploiting its structure, several directions may be found. Transfer learning methods (usually in the form of pretraining and fine-tuning models) show promising results in many applications. However, they deal with various differences in data for datasets in various domains and not in the same dataset (Pan and Yang, 2009; Ramachandran et al., 2016; Peters et al., 2018; Devlin et al., 2019). Within the dataset a hard negative mining (Felzenszwalb et al., 2009; Zhang et al., 2016; Jin et al., 2018) tries to identify the most informative negative samples and to reduce the amount of other ("easy") negative samples in training data to bring more focus on hard ones. In (Braden Hancock and Ré, 2019 accessed May 25 2020) authors discuss data slicing approach, which assumes the presence of subsets in data that have something in common, and that may influence the performance of the model. They train encoder using different heads (e.g. top layers that solve classification task) with only a corresponding subset of data. Our approach is different in two points: we use topic modelling to identify semantically related subsets of data, we use pretraining/fine-tuning routine, which is more widespread in domain adaptation tasks.

To the best of our knowledge, there is no neural network-based generation method proposed for the problem of online advertising (banners) creation using natural language at the present moment. More-

over, no method considers dataset heterogeneity for this task.

3 The topic-driven ensemble approach

We will start by describing the collected dataset of online advertising followed by its cleaning procedure and will end with the proposed transformer-based approach.

3.1 Banner-page dataset

To obtain the required data for our research, we used a crawler software described in (Butakov et al., 2018). Initially, we collected the dataset consisting of banner texts and URL references to their respective product web-pages. To collect the dataset we used HTML tags selectors from several popular adblocks to obtain text and references of banners. The data was collected from several popular websites (to name a few: avito.ru, auto.ru, youla.ru). Then we collected the content of web-pages referenced by banners. Most sites are dynamic and use a lot of AJAX requests to get content and further apply JavaScript to render it. Therefore, we developed an extension of the crawler with Puppeteer[1] which allows controlling Google Chrome browser and multiple asynchronous requests that alongside need JavaScript code.

The following steps were taken for cleaning in the crawling phase: excluding pages with 404 errors; removing headers and footers; filtering classes by content (removing tags like "home", "delivery", "to cart", "log in"); filtering invisible elements (checks for visibility were performed); recursive DOM traversal and text extraction.

We left only pairs unique by text in the dataset as well as removed all pairs from the dataset with length less then 15 words. Finally, we got a dataset of 363596 records, containing textual $\langle banner, page \rangle$ pairs.

3.2 Abstractive summarization for online advertising generation

Online advertising generation as an abstractive summarization problem can be formulated as follows: having set of page-banner pairs $D = \{(x, y)|x \in X, y \in Y\}^N$ one needs to build a model that estimates a conditional probability $P(y_t|\{y_1, ..., y_{t-1}\}, x, \theta)$, where: N – size of the dataset; X – a set of webpages; Y – a set of online advertising (banners); $x = x_1, ..., x_m$ – a text of webpage consisting of tokens that belong to some vocabulary (m – length of a source text in tokens); $y = y_1, ..., y_k$ – a text of a banner (k – length of a banner in tokens belonging to the same vocabulary); t – token index in a banner's text; θ – a set of model parameters being used for modelling this conditional probability. Thus finding the best parameters is formalized as:

$$\theta_{MLE} = argmax_\theta \sum_{i=1}^{N} \log P(y_t|\{y_1, ..., y_{t-1}\}, x_i, \theta) \tag{1}$$

where i iterates through the training dataset, x_i is a web-page text from the dataset.

The task assumes construction of a language model conditioned on text that can generate banners in a token-by-token autoregressive manner. This approach is also known as sequence-to-sequence modeling and has gained great popularity in recent time.

However, the performance of this approach can decrease due to high data heterogeneity which appears in the diversity of the analyzed texts related to different products and different topics written by using dissimilar words and phrases. On the other hand, using synonyms and various stylistic techniques to describe the same products or services leads to the same effect. These differential peculiarities in the source web-pages either lexical or semantic may influence advertising texts by itself.

Hereafter, we will call sets of banner-page pairs having similar topics as subdomains. Thus subdomains should have similarity in their page or banner texts, while data heterogeneity in the datasets may be expressed as the presence of several subdomains with different marginal distributions of text features. In this case, X may consist of several subdomains, e.g. $X = \{X_1, X_2, ..., X_m|X_i \neq X_j\}$ and $P(X_i) \neq P(X_j)$, where X_i - samples from the dataset belonging to i-th subdomain, $P(X_i)$ - marginal distribution of i-th subdomain. In other words, different texts that discuss different topics may also affect

[1] https://github.com/puppeteer/puppeteer

how generated banners advertise their products and what style they use. Furthermore, the presence of such subdomains in data may add contributions to data sparsity by reducing co-occurences of words and phrases.

The differences in distributions of the features require a generator model to implicitly identify a subdomain and make appropriate adjustments to the generated advertising. Additional information (in the form of features or objectives) may help to do it more efficiently and thus improve quality of generation.

The situation may get worse from the imbalance of subdomains' sizes. It may lead to a higher degree of data sparsity and may negatively affect the performance comparing to the case with equal size of subdomains.

To provide such information, we propose an approach that takes into account peculiarities of different subdomains in the initial data. The approach consists of 3 main steps: (1) topic modelling to characterize each data sample by the topic mixture representing semantics (will be discussed in topic modelling subsection); (2) extraction of semantically-related groups of samples that may be viewed as subdomains and building of a simple classification routine for new samples; (3) training an ensemble of seq2seq models to handle online advertising with improved quality.

3.2.1 Topic modelling.

In the first step, we use all available data to build a topic model using unsupervised training. As it was mentioned above, we do not have labels to classify the dataset into subdomains, but we can use topic modelling to build semantically meaningful vectors in some k-dimensional space consisted of latent variables. Value of each variable represents the relevance of a text to a topic corresponding. These vectors are mixture of topics.

For topic modelling we use additive regularization approach (Vorontsov et al., 2015) (or ARTM) as one of the most advanced tools at the current moment. It is based on PLSA-oriented (Hofmann, 1999) matrix factorization and produces two matrices that approximate latent distribution over topics for documents $p(t|d)$ (matrix ϕ_{wt}) and distribution over words for topics $p(w|t)$ (matrix θ_{td}):

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, d \in Dt, w \in W \tag{2}$$

where Dt is a collection of documents, W is a finite set of vocabulary words and T is a set of topics.

The main idea of additive regularization is based on the maximization of the log-likelihood with the addition of weighted sum of regularizers to produce a unique solution for matrix factorization task. These regularizers allow achieving fine-grained control over the process of topic model creation (for instance, types of distributions for individual topics) which eventually leads to improved quality in comparison to other methods.

Such an approach allows bypassing costly manual labeling though it requires choosing a suitable set of parameters to obtain meaningful topics and topic mixtures for texts. Because there is still no reliable metric well-correlated with human assessment of quality (Khodorchenko and Butakov, 2018), topic modelling requires human intervention, albeit significantly reduced.

3.2.2 Subdomains identification

Acquired text embeddings (in the form of topics mixture for individual text) can be used further for the identification of subdomains in the dataset. To extract semantically related groups, we perform the following steps: (1) leave only top components containing most of the probability mass in a topic mixture while zeroing the rest components for each text; (2) perform k-Means clustering to assign a specific label to a subset of data in the dataset thus forming subdomains (which are clusters in terms of K-means clustering).

Step (1) allows discarding low probability tails which don't carry useful information about texts and act like noise letting clustering model to have better perplexity on a multitude of very small subdomains. Human assessment on these subdomains shows that some can be joined thus resulting in fewer number of bigger subdomains. Discarding tails give us a chance to obtain bigger subdomains (documents in them still semantically relevant) which is better for fine-tuning in the ensemble described later in this section.

Mixture components to be left are determined by a threshold (a component should be higher than the threshold) identified empirically.

The classification of new incoming text happens by simply finding the closest subdomain by measuring cosine similarity between centroids and the topics mixture of the text (using only top components).

3.2.3 Seq2seq ensemble for online advertising generation

Having extracted subdomain groups, we can now build an ensemble of subdomain fine-tuned models to improve the quality of generation.

Our ensemble approach assumes the following steps: (1) all data in train split are used to train a base seq2seq model; (2) for each extracted subdomain we create a separate model by fine-tuning the base seq2seq model for few epochs using only data from this subdomain group; (3) for each subdomain we assess quality of the fine-tuned and the base models on validation data, if a fine-tuned model has better metrics it is included into the ensemble otherwise we leave the base model for the subdomain; (4) we combine steps of topic mixture evaluating, classification into subdomain and the ensemble of fine-tuned and base models into a single inference pipeline for advertising text generation.

Step (1) can be implemented with any seq2seq model. In Section 4 we compare several state-of-the-art models and choose the best one to be used in the ensemble. Initially using all the data for training helps to achieve better performance than training on subdomains from the very beginning. On step (2) we perform fine-tuning using now only a subset of data respective for a subdomain. Thus, trying to better adapt the model to this particular subdomain and improve the quality of generation using this "specialized" model. On step (3) we verify the quality of the resulting fine-tuned models and may decide which one (fine-tuned or base model) to use for the subdomain. It allows us to ensure that the resulting quality is better or at least equal to the base model.

The scheme of the overall approach is presented in Fig 1. It depicts training and inference pipelines. The first 3 steps are implemented in the training pipeline while the 4-th represents the whole inference pipeline.

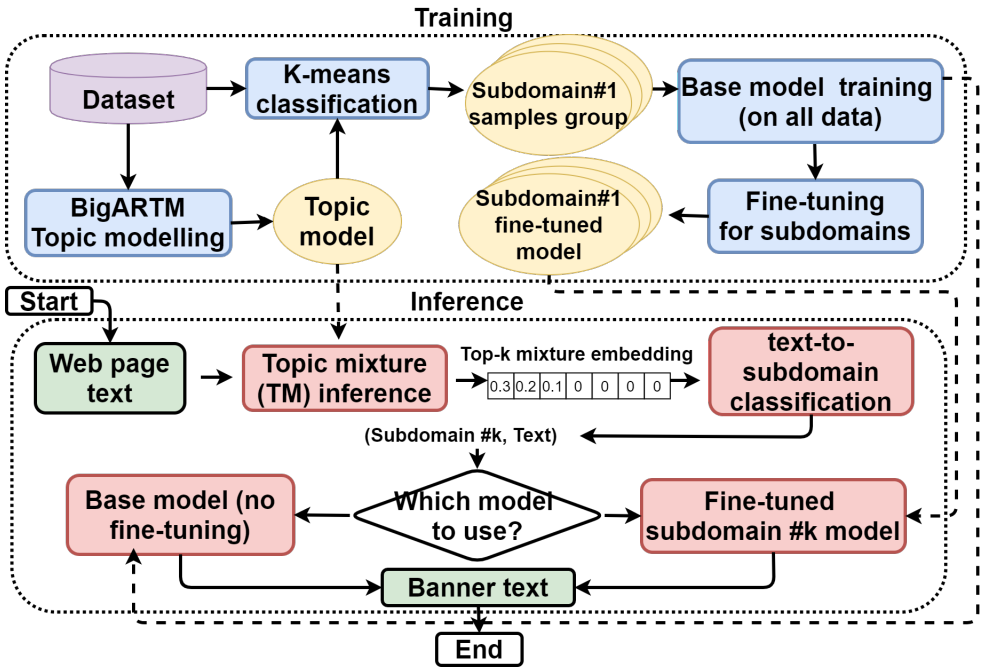

Figure 1: The scheme of the proposed ensemble approach: training pipeline (top) and inference pipeline (bottom) for online advertising generation.

It should be noted that while multiple fine-tuned models require additional time for their learning (though, performing a few more epoch with reduced dataset is not a full-scale learning), it still uses only one model instance for generation and has no other significant overheads.

4 Experimental study

In this section, we present the results of our experimental investigation for the proposed ensemble approach. We start by describing the results of topic modeling and subdomains extraction than proceeding to the comparison of the performance of the proposed ensemble method with existing alternatives. We end with discussing of the ensemble method performance for individual subdomains and its impact on overall performance. For our experiments including topic modeling, we used a random sample of data (roughly half of it) to significant reduce training time. The final size of the dataset used for training and testing was 181798 pairs.

4.1 Topic modeling and subdomains

There were multiple variants of topic models with different parameters, most notably with different variants of dictionary and count of main and background topics. Though the texts are initially in Russian, some words from other languages, mainly English, also appear from time to time. Thus, we experimented with the full dictionary variant and only Russian variant when all non-Russian words were removed. The regularization was done in accordance with strategies provided in (Nikolenko et al., 2017). As a result, we prepared the following models for human assessment: 1 – model with 200 topics (180 main and 20 background) with Russian vocabulary; 2 – model with 200 topics (180 main and 20 background) with full vocabulary; 3 – model with 300 topics (280 main and 20 background) with Russian vocabulary; 4 – model with 400 topics (360 main and 40 background) with full vocabulary; 5 – model with 400 topics (350 main and 50 background) with Russian vocabulary.

In order to select the best topic model, we randomly sampled 100 texts and inferred topics with a probability higher than 0.1 threshold, where each topic is a list of 20 most probable words. Each of the 8 assessors were given several examples where each text was supplied with topics from one of the models (the names of the models were hidden). The quality categories are the following: **Good** – all topics are related to the text, the main subject is clearly identified along with accompanying topics; **Interpretable** – some topics do not describe the content or there is no subject topic; **Non-interpretable** – there are no topics above the 0.1 threshold; **Bad** – all the found topics have no relation to the document. The threshold's value was identified empirically by the assessors based on matching between mixtures and semantic of the texts. This value was used for subdomain identification step.

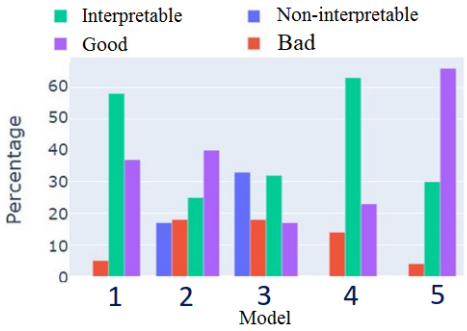

Figure 2: Scoring of different variants of topic models by human judgment. The numbers correspond to models described in the section. Percentage is a part of topics belonging to a category.

The results of comparison in Figure 2 clearly indicate that model 5 performed the best. Using this model we identified several semantically related subdomains in data that we used in further experiments. We give them semantically-related abbreviations and report their sizes in round squares: Goods (38059); Medicine (15362); Realty (15703); Business (21184); Work (11044); Travelling (8532); Services (11494); Sales (4435); Hi-tech (4891); Different (72538).

4.2 Comparative study of generators

To show the improvement that gives our model, we compare modern seq2seq methods (Pointer-network generator, Transformer and BERT-based abstract summarization model) on the task of online advertising

generation with our approach. Almost 80% of the dataset was used for training and the rest for testing. The subdomains mentioned above were extracted for train and test. The resulting sentences were obtained with a greedy decoding approach that is widely used in the field of natural language generation (Gatt and Krahmer, 2018) and finds the most probable token at each step.

To estimate the quality of the generated banners we use classical metrics in the field of abstractive summarization: rouge-1, rouge-2, and rouge-L (hereafter, R-1, R-2, R-L, correspondingly) (Lin, 2004). There are also repeated methods designated with 'macro'. In this case, we applied a different strategy that is named macro-averaging to calculate quality metrics: we calculate the average of individual subdomains' averages. Macro-averaging is often used for multi-class classification. This macro metric allows us to compare the performance of methods taking care of subdomains independently of their sizes. We do that only for Transformer-based and BERT-based methods as their performance is the best among alternatives. We chose the following parameters for Transformer: 6 encoder layers, 6 decoder layers, embedding size 512. This setting was similar to (Vaswani et al., 2017) and demonstrated good results in a number of applications. Our dataset contains almost 300 000 unique words most of which with only a few occurrences leading to issues with data sparsity and slowing down calculations. Thus, words from pages and banners were converted to tokens using byte-pair-encoding (Gavrilov et al., 2019). Tokens are frequently occurring words or parts of words such as roots, prefixes, suffixes and even letters in some cases. Such preprocessing allows reducing the size of the dictionary to 70 000 tokens for pages and banners. It solves mentioned above issues and allows us to save words almost entirely including their prefixes and endings. Pointer-generator network was used with the following settings: 256-dimensional hidden states and 128-dimensional word embeddings. We use one of the best models for abstractive summarization proposed in (Liu and Lapata, 2019): pre-trained Bert model as an encoder and 6 layers of seq2seq decoder (standard transformer decoder) with randomly initialized weights. The authors of this work demonstrate the ability of the Bert-based model to achieve high values of Rouge metrics and to generate more novel n-gram (hereafter we will refer to this model as the BERT model). Because of BERT, this model has many more parameters in comparison with other Transformer-based models we try.

To train the models, we used the following parameters: batch size was 96 for all models, for Pointer-generator, Transformer, Transformer-balanced 30 epochs was used for training; for topic-driven ensemble 20 epochs were used to train the base model and 10 epochs were used for fine-tuning. For training we used 2 Nvidia Tesla V100 32gb. For Transformers, the epochs number was fixed at 30, because the error function reached a plateau and practically stopped decreasing. For Pointer-generator we chose 30 epochs as this number is similar to one found in the original paper (See et al., 2017). For the BERT model, we trained it for 30 epoch. At this point, the score was not getting any better with subsequent iterations. We then trained BERT-based subdomains models for up to 5 epoch more.

For all resulting tables, we use the best found metrics values of all iterations. We conducted a series of experiments using methods mentioned above including the approach we propose (named as topic-driven ensemble) and a separate transformer instance with balanced by subdomains sizes training dataset. For the topic-driven ensemble, we state the metrics number which includes the whole pipeline. To train balanced instance of Transformer 10000 banner-page pairs were sampled from each subdomain. If a subdomain contained less than 10000 we took all available observations. The resulting training dataset consisted of 90945. The test set was the same as for other methods. According to results of our experiments (Table 1), Transformer showed itself as a better model to be served as a base model for our approach. Thus topic-driven ensemble method is also Transformer-based.We also included a simple heuristic based on selecting a title from the page as the text for banners. The comparative study results are in Table 1. The results with more details for individual subdomains are provided in subsection 4.3. It can be seen from Table 1, that Transformer method performs better than other base alternatives and topic-driven ensemble shows the best performance among all methods outperforming even transformer for 2 – 2.3%. And if we compare these methods using macro-averaging taking into account the found subdomains, we may see that the performance gains are even higher up to 5%. The variant with the balanced training dataset (Transformer-balaned) showed itself worse than both topic-driven ensemble

and transformer. Despite the balancing, the model has fewer examples to train on that significantly influences its performance. The BERT-based model which has many more layers shows better results than based on a plain transformer. Still, even for the BERT-based model application of the topic-driven ensemble allows gaining an additional increase in quality metrics for 1 - 2% for both macro and micro averaging. For both types of models, there are significant improvements in individual subdomains.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Titles	9.30	2.13	6.55
Pointer-generator	17.93	4.70	14.09
Transformer	37.21	23.38	34.59
Transformer (macro)	37.11	23.26	34.43
Transformer-balanced	28.61	13.57	26.18
Transformer-balanced (macro)	28.27	13.30	25.89
topic-driven ensemble	37.97	23.94	35.33
topic-driven ensemble (macro)	39.00	24.66	36.27
BERT	41.58	26.95	40.70
BERT + topic-driven ensemble	**42.21**	**27.50**	**41.27**
BERT (macro)	41.42	27.07	40.54
BERT + topic-driven ensemble (macro)	**41.98**	**27.59**	**41.08**

Table 1: ROUGE scores for different generator models (bold values are the best for macro and micro strategy of assessment)

4.3 Results for individual subdomains

In Table 2 we provide metrics for individual subdomains obtained by the ensemble of fine-tuned Transformer-based generators and base Transformer model. As it is shown in Table 2, fine-tuning boosts the results for different subdomains for 1%–7% on ROUGE-1, for 2%–10% on ROUGE-2 and for 2%–7% on ROUGE-L. Table 3 presents the results of the same kind for the BERT model and its ensemble-based modification. For individual domains the ensemble shows the boost for 1%–10% on ROUGE-L and ROUGE-1 while for ROUGE-2 the boost is up to 15%.

.

Subdomain	Topic-driven ensemble			Transformer			Size
	R-1	R-2	R-L	R-1	R-2	R-L	
Medicine	**34.87**	**21.40**	**32.31**	33.60	20.37	31.18	2048
Different	32.77	19.81	30.45	**35.76**	**22.50**	**33.22**	9216
Services	**38.92**	**24.40**	**36.27**	36.10	22.40	33.57	1280
Hi-Tech	**37.95**	**24.43**	**35.34**	36.54	23.58	33.99	512
Travelling	**38.17**	**24.76**	**36.01**	37.63	24.28	35.25	1280
Goods	34.06	20.06	31.73	**37.94**	**23.58**	**35.29**	5120
Business	**40.25**	**25.37**	**37.38**	38.41	22.84	35.35	2560
Sales	**40.44**	**24.24**	**36.85**	38.72	22.85	35.35	256
Work	**41.41**	27.50	**38.89**	39.96	**29.57**	37.44	1280
Realty	**44.26**	**28.39**	**41.18**	42.18	27.06	39.29	2304

Table 2: ROUGE scores of base Transformer model and the proposed topic-driven ensemble (Size is a size of a subdomain in the test split). The better values for each subdomain are in bold.

For some subdomains, mainly the ones which have less samples (and thus weaker representation) in the dataset, there is a significant rise of metrics, while for big subdomains the situation is different. Con-

Subdomain	BERT + topic-driven ensemble			BERT			Size
	R-1	R-2	R-L	R-1	R-2	R-L	
Medicine	39.4	25.8	38.4	**40.3**	**26.0**	**39.5**	2048
Different	**45.4**	**30.4**	**44.5**	44.9	29.5	43.9	9216
Services	39.9	25.3	38.9	**40.4**	**25.7**	**39.4**	1280
Hi-Tech	**51.3**	**37.3**	**50.3**	50.3	36.2	49.5	512
Travelling	39.0	24.6	38.2	**39.1**	**24.7**	**38.3**	1280
Goods	**49.2**	**35.6**	**48.7**	48.5	34.7	48.0	5120
Business	**36.5**	**20.2**	**35.4**	33	18.1	32.5	2560
Sales	**33.1**	**19.8**	**32.1**	32.0	18.6	31.1	256
Work	**46.8**	**31.7**	**45.2**	42.4	27.3	40.6	1280
Realty	47.3	31.8	46.5	**49.0**	**32.9**	**47.7**	2304

Table 3: ROUGE scores of the BERT model and BERT + topic-driven ensemble (Size is a size of a subdomain in the test split). The better values for each subdomain are in bold.

structing the ensemble from specialized fine-tuned models and the general (base) one allows achieving better performance. We should also note that training the base model and then fine-tuning on specific subsets represented by subdomains is a better option than either training with a balanced dataset or training separate models for each subdomain from the very beginning. In Table 4 we present examples of original and generated banners

Original Banner	**Generated text**
eco slim will help you with this have time to buy at a reduced price	losing weight without exhausting diets and stress read more
free blood and health analysis at the international hematology clinic in Moscow	have time to check the state of the body with a free blood test without queues

Table 4: Examples of generated advertising (intentionally translated in English)

4.4 Experimental study on MS MARCO dataset

To verify the proposed approach, we have conducted an experimental study on a different dataset called MS MARCO (Nguyen et al., 2016). This dataset is the biggest of its kind and was compiled for the task of question answering over text (QAT for short). This task assumes a question and a text (which may contain information useful for answering) as input into the model and an answer for the input question as an output. While the task of the dataset is different, applied for solving seq2seq neural architectures are not. Used by us in previous experiments BERT-based model may be applied without changes (except for concatenating of question and text into a single input text). We applied our approach to the BERT model trained on this dataset for QAT task and compare it with the base BERT model. For the base BERT model, we used the following parameters: training for 4 epochs (There was not any improvement in quality metrics on the test part of the dataset after 4 epochs. The dataset itself is huge.), Adam optimizer with b1 = 0.9 and b2 = 0.999, the learning rate of the encoder 0.002 and warmup steps 20000, the learning rate of decoder 0.2 and warmup steps 10000. We trained subdomains models for up to 20 epochs. To apply the approach we built a topic model and divided the dataset (in train and test parts) into 8 subdomains according to the model.

The results are presented in Table 5. In comparison with the base BERT model, the ensemble was able to improve results up to 3.5% and for 1.5% for macro- and micro- averaging. MS MARCO has much more data for training and thus all found subdomains have better representation in the dataset what

reduces the effect from the proposed approach. Nevertheless, there is still an opportunity to obtain at least minor improvements for large datasets and more of them on small and medium datasets with the proposed approach.

Subdomain	BERT + topic-driven ensemble			BERT			Size
	R-1	R-2	R-L	R-1	R-2	R-L	
Health Diseases	**62.82**	**51.61**	**63.96**	60.78	51.56	61.94	1000
Biographies	**57.89**	43.72	**58.39**	56.78	**44.01**	57.23	1000
Cooking	**42.79**	24.29	**45.12**	42.22	**24.39**	44.36	1000
Chemical Staff	**56.38**	**43.97**	**57.34**	55.89	43.72	56.84	1000
Social Services	**62.21**	52.07	**63.46**	61.81	**53.59**	63.03	1000
Weather	52.39	**40.99**	**54.66**	**52.41**	40.86	54.65	1000
Cities Info	59.44	44.72	59.98	59.59	**45.01**	**60.22**	1000
Salary and jobs	44.68	34.21	45.49	**45.03**	**34.46**	**45.83**	1000

Table 5: ROUGE scores of the BERT model and BERT + topic-driven ensemble (Size is a size of a subdomain in the test split). The better values for each subdomain are in bold.

5 Conclusion and Future Work

In this work, we presented an approach to generate banners for online advertising from texts of web pages. A high degree of data heterogeneity that may be found in a dataset and represented by various topic-based subdomains reduces the quality of generation. The proposed topic-driven ensemble approach helps to solve this issue and shows an increase in efficiency for up to 2% – 5% on average for all subdomains and up to 10% – 15% for individual subdomains. Probably, the approach may be further improved by introducing a topic mixture directly into the generator model. In this case, it would play a role of extended context and thus provide additional features the generator can rely on. Such an extension may improve quality even further due to eliminating of intermediate subdomains classifier and prevent errors that can happen on this level. Also, there is a potential to integrate knowledge about user profiles into the generator to obtain varying banners according to user behavior to better reach the target audience. Further efforts should be directed to the investigation of how different topic models and other schemes of a dataset dividing into separate domains influence the overall quality of generators.

Acknowledgements

This research is financially supported by The Russian Science Foundation, Agreement 20-11-20270.

References

J. Aguilar and G. Garcia. 2018. An adaptive intelligent management system of advertising for social networks: A case study of facebook. *IEEE Transactions on Computational Social Systems*, 5(1):20–32.

Ines Chami Vincent S. Chen Sen Wu Jared Dunnmon Paroma Varma Max Lam Braden Hancock, Clara McCreery and Chris Ré, 2019 (accessed May 25, 2020). *Massive Multi-Task Learning with Snorkel MeTaL: Bringing More Supervision to Bear*.

Nikolay Butakov, Maxim Petrov, Ksenia Mukhina, Denis Nasonov, and Sergey Kovalchuk. 2018. Unified domain-specific language for collecting and processing data of social media. *Journal of Intelligent Information Systems*, 51(2):389–414.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL 2019*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170, January.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In *ECIR*.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on UAI*, UAI'99, page 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller. 2018. Unsupervised hard example mining from videos for improved object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 307–324.

Maria Khodorchenko and Nikolay Butakov. 2018. Developing an approach for lifestyle identification based on explicit and implicit features from social media. *Procedia Computer Science*, 136:236 – 245. 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece.

Gwang Kim and Ilkyeong Moon. 2020. Online banner advertisement scheduling for advertising effectiveness. *Computers Industrial Engineering*, 140:106226.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders.

Chih-Wei Liu, Shao-Kang Lo, Ai-Yun Hsieh, and Yujong Hwang. 2018. Effects of banner ad shape and the schema creating process on consumer internet browsing behavior. *Computers in Human Behavior*, 86:9–17, 09.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. *CoRR*, abs/1704.08300.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Sergey I. Nikolenko, Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*.

Ginger Rosenkrans. 2010. Maximizing user interactivity through banner ad design. *Journal of Promotion Management*, 16(3):265–287.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Shengli Song, Haitao Huang, and Tongxiao Ruan. 2019. Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools Appl.*, 78(1):857–875, January.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. 2015. Bigartm: Open source library for regularized multimodal topic modeling of large collections. pages 370–381, 04.

Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. 2016. Is faster r-cnn doing well for pedestrian detection? In *European conference on computer vision*, pages 443–457. Springer.