Topic-controllable Abstractive Summarization

Anonymous ACL submission

Abstract

Existing approaches for topic-controllable summarization either incorporate topic embeddings or modify the attention mechanism. The incorporation of such approaches in a particular summarization model requires the adaptation of its codebase, a process that can be complex and time-consuming. Instead, we propose a model-agnostic topic-controllable summarization method employing a simple tagging-based formulation that can effortlessly work with any summarization model. In addition, we propose a new topic-oriented evaluation measure to quantitatively evaluate the generated summaries based on the topic affinity between the generated summary and the desired topic. Experimental results show that the proposed tagging-based formulation can achieve similar or even better performance compared to the embedding-based approach, while being at the same time significantly faster.

1 Introduction

The exponential rise in the volume of textual data available through various sources, ranging from social media to financial reports, makes it virtually impossible for humans to digest all the important information for their needs, without spending an enormous amount of effort. Automatic summarization methods can mitigate this problem, by shortening texts to a more concise form (Nallapati et al., 2016; Celikyilmaz et al., 2018; Liu and Lapata, 2020; Song et al., 2019).

Even though early methods had limited success on this task, mainly focusing on *extractive summarization* (Fang et al., 2017; Mao et al., 2019), the advent of deep learning led to much more powerful neural *abstractive summarization* (See et al., 2017; Song et al., 2019; Dong et al., 2019; Lewis et al., 2020; Zhang et al., 2020) methods. These methods go beyond extracting unaltered sentences from the input, allowing for generating the summary using novel words and phrases that are not necessarily part of the input text.

Despite the success of deep learning models, there is often the need to go beyond delivering a generic summary of the document, and instead produce a summary that focuses on a specific topic that pertains to the user's interests. For example, a newswire article may discuss two topics, such as sports and politics, yet the user may be interested only in the sports aspect. Existing topic-controllable summarization models address this need either by incorporating topic embeddings into the model's architecture (Krishna and Srinivasan, 2018) or by modifying the attention mechanism (Bahrainian et al., 2021). However, they are restricted to very specific neural architectures and it is not straightforward to use them with any summarization model.

At the same time, there is no clear way to evaluate such approaches, since there is no evaluation measure designed specifically for topic-controllable summarization. Indeed, existing methods just use the typical ROUGE score (Lin, 2004) for measuring the summarization accuracy and then employ user studies to qualitatively evaluate whether the topic of the generated summaries indeed matches the users' needs (Krishna and Srinivasan, 2018; Bahrainian et al., 2021).

Based on the aforementioned observations, we propose a model-agnostic topic-controllable summarization method that can be effortlessly combined with any neural architecture. Given a topic labeled collection, the proposed method works by first extracting keywords that are semantically related to the topic the user requested and employing special tokens to tag them before feeding the document to the summarization model. Experimental results show that this can be an effective and efficient way to influence summarization models towards the users' needs.

Furthermore, we propose a topic-aware evaluation measure for quantitatively evaluating topic-controllable summarization methods in an objective way without involving expensive and time-consuming user studies. In particular, we propose calculating prototype term weighting representations, namely tf-idf, of different topics, and then calculating the cosine similarity between the generated summaries and the prototype topic vectors.

The contributions of this paper can be summarized as follows:

- We propose a simple, yet effective and efficient model-agnostic way to perform topic-controllable summarization.
- We adapt an existing topic-controllable method to work with Transformer-based architectures, scaling up from existing RNN-based formulations, establishing a strong, yet computationally demanding baseline for topic-oriented summarization.
- We propose a topic-oriented measure to quantitatively evaluate the generated summaries without the need for resorting to human studies.
- We provide an extensive empirical evaluation as well as a zero-shot experimental evaluation, demonstrating both the generality of the proposed method, as well as its effectiveness.

The rest of the paper is organized as follows. In Section 2 we review the existing topic-oriented summarization related literature. In Section 3 we introduce the proposed methods while in Section 4 we provide the experimental results. Finally, conclusions are drawn and interesting feature research directions are discussed in Section 5.

2 Topic-oriented Summarization

Methods for topic-oriented summarization belong to two broader categories: a) methods that employ topical information to enhance the quality of the generated summaries and b) topic-controllable methods that use topical information to control the output of the generated summaries.

2.1 Improving summarization using topical information

The integration of topic modeling into summarization models has been initially used in the literature to improve the quality of existing state-of-the-art models (Ailem et al., 2019; Wang et al., 2020; Liu and Yang, 2021). Statistical topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or Poisson Factor Analysis (PFA) (Zhou et al., 2012) are used to supply summarization models with global topic semantics, allowing the generation of more coherent and consistent summaries.

Ailem et al. (2019) use LDA to influence the model to generate summaries based on both the input text and the underlying document topics and as a result to improve the quality of the generated summary. To achieve this, the decoder of a pointer generator network is enhanced with the information of the latent topics that are derived from an LDA model. Thus, the integration of topic modeling can capture hidden semantic structures based on word co-occurrences, allowing the model to generate better summaries conditioned on a more global context. Similar methods have been applied by Wang et al. (2020) using PFA with a plug-and-play architecture that can be adapted to any Transfomer-based model. This architecture consists of 3 independent modules: Semantic-informed attention (SIA), Topic Embedding with Masked Attention (TEMA), and Document-related modulation (DRM). SIA is embedded as an additional head into the multi-head attention mechanism. This added head is extracted from a fixed semantic-similarity attention matrix for each topic. TEMA uses topic embeddings as an additional decoder input based on the top-n topics from the input document. Since a topic can be represented as a distribution over all the tokens from the vocabulary, topic embeddings can be derived from a mixture of all the corresponding token embeddings. Finally, DRM is used to modulate a hidden layer for each decoder adding a topic feature bias vector.

Liu and Yang (2021) propose to enhance summarization models using an Extreme Multi-Label Text Classification (XMTC) model to improve the consistency between the underlying topics of the input document and the summary, leading to summaries of higher quality.

Even though Wang et al. (2020) refers to the potential of controlling the output conditioned on a specific topic using GPT-2 (Radford et al., 2019) with TEMA, all the aforementioned approaches are focused on improving the accuracy of existing summarization models.

2.2 Topic-controllable summarization methods

Some steps towards controlling the output of a summarization model conditioned on a thematic category have been made by Krishna and Srinivasan (2018) proposing a controllable summarization setting that builds upon the pointer generator network (See et al., 2017). The topical information is integrated into the model as a topic vector, which is then concatenated with each of the word embeddings of the input text. Each topic vector is computed as a Bag of Words (BoW) representation that is derived from Vox Dataset (Vox Media, 2017), a news dataset that contains articles from 185 different news topics.

Krishna and Srinivasan (2018) created a topic-oriented training dataset that builds upon CNN/DailyMail as follows. First, the dot-product between the BoW representation of the summary and all the BoW topic representations is computed. The topic with the highest similarity is assigned to the corresponding article while articles with more than one dominant topic are discarded. All the topic-assigned articles are used to compile a temporary intermediate dataset. To create the topic-oriented dataset, two articles a_1 and a_2 with different topics, are randomly selected from the intermediate dataset. A new article a' is created by sequentially selecting sentences from both articles. The new article a' is assigned with the summary from one out of two selected articles and the same process is repeated to create a new article a'' which is now assigned with the remaining summary. Then, the initially selected articles a_1 and a_2 are discarded from the intermediate dataset. This process is continued until there are no articles in the intermediate dataset or all the remaining articles belong to the same topic. Finally, the new topic-oriented dataset consists of super-articles that discuss two distinct topics but are assigned each time with one of the corresponding summaries so the model learns to distinguish the most important sentences for the corresponding topic during training.

Recently, Bahrainian et al. (2021) propose to incorporate the topical information from each document to modify the attention mechanism of the pointer generator network (See et al., 2017). The modification of the attention mechanism is introduced as topical attention generated by an LDA model. More specifically, each word is represented as a topic vector that is derived from LDA and then is combined with the original attention weights of the model to compute the final attention weights. It is important to note that even though the model is trained with the topical attention mechanism during training, no topical information is used during inference. Thus, the aforementioned method allows for controlling the topic of the generated summary only from the perspective of the restriction of unwanted topics during training, contrary to the proposed method, which allows for guiding the generation towards a topic, during inference.

3 Contributions

In this section, we present the main contributions of this paper. More specifically, we introduce two different topic-controllable methods to guide the summary generation towards a specific topic: a) tagging-based formulation and b) embedding-based formulation. We also present the proposed topic-oriented similarity measure which is used for evaluating the topic affinity between the desired topics and the generated summaries.

3.1 Tagging-based formulation

The proposed tagging-based method employs a trivial, yet effective mechanism to shift the summary generation towards the desired topic, assuming the existence of a set of representative terms for each thematic category. More specifically, after lemmatization, the most representative words for the desired topic are tagged with special tag tokens before feeding to the summarization model. As demonstrated in Section 4, this can be an effective way to intuitively guide the model towards the tagging words during both training and inference.

To apply this mechanism, a topic-oriented training set is required. However, this is not a straightforward process due to the lack of appropriate topic-oriented summarization datasets. Indeed, there are no existing datasets for summarization that contain multiple summaries for each input document, according to the different topical aspects of the text (Krishna and Srinivasan, 2018). Thus, we adopt the same approach with (Krishna and Srinivasan, 2018) to create a topic-oriented dataset that builds upon the CNN/DailyMail (Hermann et al., 2015). We apply the tagging mechanism to each document of the topic-oriented dataset according to the assigned topic of the corresponding summary. More specifically, for each document of the compiled dataset, we tag the terms that belong to the intersection of words between the lemmatized document and the top-N most representative terms for the corresponding topic.

The most representative words can be extracted either by simple prototype term weighting representations such as BoW or tf-idf, statistical topic modeling algorithms such as Labeled LDA (Ramage et al., 2009) or even more sophisticated keyword extraction models (Ding and Luo, 2021; Liang et al., 2021). In this work, we use tf-idf to demonstrate the efficacy of our method, even when a simple mechanism is employed.

More specifically, we use tf-idf to extract document representations and then calculate the topical vectors. Given a corpus $\mathcal{D}$, we can represent a document d as a vector $\mathbf{x}_d$ which contains the tf-idf scores for each term of the document. The tf-idf score for each term t of a document d, belonging to a corpus $\mathcal{D}$, is computed as:

$$x_{dt} = tf(t, d, \mathcal{D}) \times idf(t, \mathcal{D}), \tag{1}$$

where $tf(t, d, \mathcal{D})$ indicates the number of times that term t appears in document d, while $idf(t, \mathcal{D})$ indicates the inverse document frequency of term t in corpus $\mathcal{D}$ which is computed as follows:

$$idf(t, \mathcal{D}) = \log \frac{|\mathcal{D}| + 1}{df(t, \mathcal{D}) + 1} + 1, \tag{2}$$

where $df(t, d)$ is the frequency of term t in D. Note that the length of each tf-idf vector is equal to the size of the vocabulary V of the corpus $\mathcal{D}$, i.e., $\mathbf{x}_d \in \mathbb{R}^{|\mathcal{V}|}$, where $|\mathcal{V}|$ denote the cardinality of the vocabulary $\mathcal{V}$. Finally, we normalized the extracted vectors to have unit length as:

$$\mathbf{x}_d^{(n)} = \frac{\mathbf{x}_d}{||\mathbf{x}_d||_2}, \tag{3}$$

where $||\mathbf{x}||_2$ is the l_2 norm of the vector $\mathbf{x}_d$.

Then, given a topic-assigned collection of documents $\mathcal{X}$, we can follow the aforementioned procedure to extract a topical vector representation $\mathbf{y}_c$ for each topic c, by grouping together documents of the same topic and averaging their tf-idf representation as follows:

$$\mathbf{y_c} = \frac{1}{|\mathcal{X}_c|} \sum_{\mathbf{x} \in \mathcal{X}_c} \mathbf{x} \tag{4}$$

The topical vector extraction is summarized in Figure 1.

Table 1: Representative terms for topics from 2017 KDD Data Science+Journalism Workshop (Vox Media, 2017)

Topic	Terms
Politics	policy, president, state, political, vote, law, country, election
Sports	game, sport, team, football, fifa, nfl, player, play, soccer, league
Health Care	patient, uninsured, insurer, plan, coverage, care, insurance, health
Education	student, college, school, education, test, score, loan, teacher
Movies	film, season, episode, show, movie, character, series, story
Space	earth, asteroid, mars, comet, nasa, space, mission, planet, astronaut

Finally, we extract the top-N most important terms for each topic according to the top tf-idf scores of each topical vector. An example of some indicative representative words for a number of topics in a topical corpus is shown in Table 1.

Given the set of representative words for each topic, a document, and the desired topic, the tagging mechanism works as follows:

1. All the words of the input document are lemmatized to their roots.

2. We identify the common words between the existing lemmatized tokens and the representative words for the desired topic.

3. Finally, we tag each token of the input document with a special token, i.e., [TAG], only if the lemmatized form of this token is contained in the set of the most representative words for the corresponding topic.

For example, suppose that we pre-process the sentence below, as a part of an input document, from which we aim to guide the generation towards the topic "*Business & Finance*".

> *"By one estimate, American individuals and **businesses** together spend 6.1 **billion** hours complying with the **tax** code every year."*

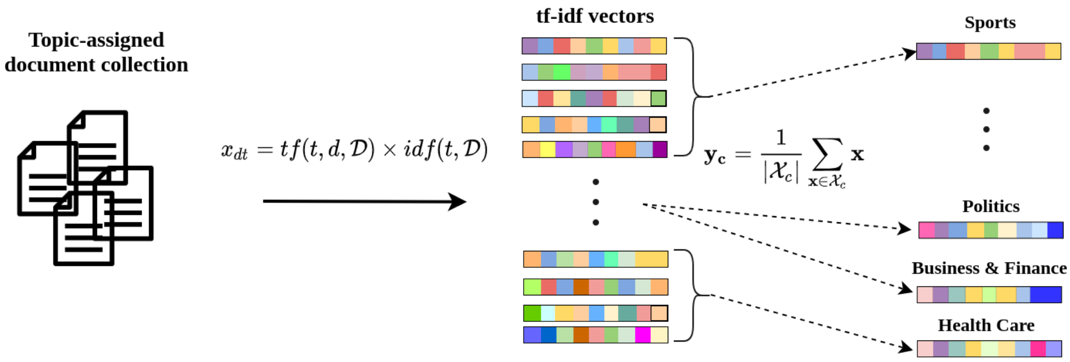

Figure 1: Topical vector extraction using tf-idf scores, given a topic-assigned document collection. First, we calculate tf-idf scores for each document. Then, documents of the same topic are grouped and their tf-idf representation is averaged.

Following the aforementioned procedure, we will enclose with special tokens, the words "*businesses*", "*billion*" and "*tax*" since they belong to the set of the most representative words for the desired topic.

During training, the model learns to intuitively give more "attention" to the tagged words and as a result shift the generation towards the desired topic. The tagging mechanism can be used during inference to guide the summary generation towards the user-requested topic provided by any set of representative terms. Also, since this method does not affect the architecture of the summarization model, it can easily be applied to any model's architecture.

3.2 Embedding-based formulation

To establish a strong baseline for comparing the tagging-based method with existing methods in the literature, we adapted the method proposed in Krishna and Srinivasan (2018) to work with Transformer-based architectures. As described in Section 2, Krishna and Srinivasan (2018) use a pointer generator network (See et al., 2017) to concatenate topic embeddings with token embeddings allowing for generating topic-oriented summaries. The topic embeddings are represented as one-hot encoding vectors with a size equal to the number of the total topics. During training, the model takes as inputs the corresponding topic embedding along with the input document.

However, this method cannot be directly applied to pre-trained Transformer-based models due to the different shapes of initialized weights of the word and position embeddings. Unlike RNNs, Transformer-based models are typically trained for general tasks and then fine-tuned with less data for more specific tasks like Summarization. Thus, the architecture of a pre-trained model is already defined and cannot be altered easily to initialize the pre-trained model's weights with the exact same shape of the concatenated word and topic embeddings. Another option would be to initialize the model from scratch with random weights with the appropriate shape of the concatenated word and topic embeddings but this would be very computationally demanding as it would require a large amount of data and time for training.

To this end, instead of concatenation, we propose to sum the topic embeddings following the same concept with positional encoding where token embeddings are summed with positional encoding representations to create an input representation that contains the position information. Instead of one-hot encoding embeddings, we use trainable embeddings allowing the model for optimizing them accordingly during training. The topic embeddings have the same dimensionality as the token embeddings.

To sum the trainable topic embeddings with token and positional embeddings, we modify the input representation as follows:

$$z_i = WE(x_i) + PE(i) + TE(i), \qquad (5)$$

where WE, PE and TE are the word embeddings, positional encoding and topic embeddings respectively, for token x_i in position i.

Then, we use the same created topic-oriented dataset from Krishna and Srinivasan (2018) to fine-tune the summarization model for topic-oriented summarization allowing for establishing a strong comparison between the proposed tagging-based method and the more powerful embedding-based one.

3.3 Topic-focused evaluation measure

As explained in Section 1, there is currently no structured way to evaluate the performance of topic-oriented summarization methodologies. To this end, we propose a new topic-oriented measure, Summarization Topic Affinity Score (STAS), to evaluate the generated summaries according to the semantic similarity between the vector representation of the desired topic and the generated summary. More specifically, we compute the similarity between the vector representation of the summary and the vector representation of the desired topic, divided by the maximum value of all the similarities between the vector representation of the summary and all the topic vector representations. Given the vector of the target topic $\mathbf{x}_t$ and the vector representation of the predicted summary $\mathbf{x}_s$, STAS is computed as follows:

$$STAS(\mathbf{x}_s, \mathbf{x}_t) = \frac{s(\mathbf{x}_s, \mathbf{x}_t)}{\max\{s(\mathbf{x}_s, \mathbf{x}_{ti}) : i = 1 \ldots N_t\}}, \tag{6}$$

where N_t is the number of topic and $s(\mathbf{x}_s, \mathbf{x}_t)$ indicates the cosine similarity between the two vectors $\mathbf{x}_s$ and $\mathbf{x}_t$ which is computed as follows:

$$s(\mathbf{x}_t, \mathbf{x}_s) = \frac{\mathbf{x}_t \mathbf{x}_s}{\|\mathbf{x}_t\| \|\mathbf{x}_s\|}. \tag{7}$$

Thus, summaries that are similar to the requested topic are rewarded while summaries that are dissimilar are penalized.

4 Experimental Evaluation

In this section, we present the experimental results of the proposed method. First, we introduce the experimental setup used for the evaluation, including the dataset generation procedure, the evaluation metrics, and employed deep learning architectures. Then, we proceed by presenting and discussing the experimental evaluation using both the proposed tagging-based method, as well as the embedding-based method, appropriately adapted to work on Transformers.

4.1 Experimental setup

Datasets and Evaluation Metrics In order to create the topic-oriented dataset as described in Section 3, we use the Vox Dataset (Vox Media, 2017), which consists of 23,024 news articles of 185 different topical categories. We discarded topics with relatively low frequency, i.e. lower than 20 articles, as well as articles assigned to general categories that do not discuss explicitly a topic, i.e. "*The Latest*", "*Vox Articles*", "*On Instagram*" and "*On Snapchat*".

In the experiments, we investigate two different setups: a) fine-tuning without pre-processing the Vox dataset, keeping also noisy categories that do not discuss a particular topic, and b) fine-tuning after pre-processing the Vox dataset as described. All summaries of the created dataset are assigned with a topic according to the similarity between the derived topical vector representations and the vectorized summary. Thus, keeping noisy topics might lead to false topic assignments to the training summaries.

After pre-processing, we end up with 14,312 articles from 70 categories out of the 185 initial topical categories. Then, following the same procedure as Krishna and Srinivasan (2018), we create the topic-oriented dataset combining sentences from article-pairs from the CNN/DailyMail (Hermann et al., 2015). We use the anonymized version of CNN/Dailymail similar to See et al. (2017). The final topic-oriented dataset consists of 132,766, 5,248, and 6,242 articles for training, validation, and test, respectively. The average document and summary length of the created dataset is 1,544 and 56 tokens, respectively.

All the tags for the tagging-based method were applied to the dataset after lemmatization using NLTK (Bird, 2006) based on the top-N=100 most representative terms for each topic. We also use the Vox Dataset (Vox Media, 2017) to extract the tf-idf vector representations for each document in the corpus. To this end, we employed the tf-idf vectorizer provided by the Scikit-learn library (Pedregosa et al., 2011).

All methods were evaluated using both the well-known ROUGE (Lin, 2004) score, to measure the quality of the generated summary, as well as the proposed STAS measure.

Models and Training For all the conducted experiments we have employed a BART-large (Lewis et al., 2020) architecture, which is a transformer-based model with a bidirectional encoder and an auto-regressive decoder. BART-large consists of 12 layers for both encoder and decoder and 406M parameters. We used the implementation provided by Hugging Face for the BART-large architecture (Wolf et al., 2020).

We fine-tune all the models for 100,000 steps with a learning rate of 0.00003 and batch size 4 with early stopping on the validation set. We use the established parameters for BART-large architecture using label smoothed cross-entropy loss (Pereyra et al., 2017) with the label smoothing factor set to 0.1.

For all the experiments, we use PyTorch version 1.10 and Hugging Face version 4.11.0. All the models were trained using available GPUs in Google Colab[1], with approximate average training runtime 9.5 and 18 hours for the tagging-based and embedding-based method, respectively. Both data and code will be publicly available.

4.2 Results

The evaluation results on the generated dataset are shown in Table 2. We report results using five different methods. First, we employ both the generic Pointer Generation method ("PG") (See et al., 2017), as well as the topic-oriented PG ("Topic-Oriented PG") (Krishna and Srinivasan, 2018). We also use the generic BART (Lewis et al., 2020) model ("BART") fine-tuned on the regular CNN/DailyMail dataset for summarization, as well as both the adapted embedding-based formulation ("BART_{emb}") and the tagging-based formulation ("BART_{tag}").

The experimental results reported in Table 2 for the two different pre-processing setups indicate that topic-oriented methods indeed perform significantly better compared to the baseline methods that do not take into account the topic requested by the user. Furthermore, the proposed BART-based formulation significantly outperforms the generic PG approach, regardless of the applied topic mechanism (BART_{emb} or BART_{tag}). Also, the proposed tag-based mechanism seems to be more robust to noise, leading to slightly better results when no pre-processing is applied. On the other hand, when the data are pre-processed, both the embedding and the topic tagging approach lead to quite similar results. However, as we further demonstrate later, the proposed tagging method is significantly faster than embedding-based approaches, leading to the overall best trade-off between accuracy and speed.

The results of the inference time for both methods are shown in Table 3. The inference time of the proposed method is significantly smaller, improving the performance of the model by almost one order of magnitude. Indeed, the proposed method can perform inference on 100 articles in less than 40 seconds, while the embedding-based formulation requires more than 300 seconds for the same task.

[1] https://research.google.com/colaboratory/

Table 2: Experimental results on the created topic-oriented dataset based on CNN/DailyMail dataset. We report f-1 scores for ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L).

	R-1	R-2	R-L
PG (See et al., 2017)	26.8	9.2	24.5
BART (Lewis et al., 2020)	30.46	11.92	20.57
Topic-Oriented PG (Krishna and Srinivasan, 2018)	34.1	13.6	31.2
Proposed BART_{emb} (all topics)	37.64	16.94	26.20
Proposed $\text{BART}_{\text{tags}}$ (all topics)	**37.94**	**17.21**	**26.49**
Proposed $\text{BART}_{\text{tags}}$ (pre-processed topics)	39.30	18.06	27.49
Proposed BART_{emb} (pre-processed topics)	40.15	18.53	28.06

Table 3: Inference time for 100 articles. All numbers are reported in seconds.

	Tagging	Inference	Total time
BART_{emb}	-	303.0	303.0
$\text{BART}_{\text{tags}}$	**7.1**	**32.0**	**39.1**

In Table 4, we also provide an experimental evaluation using the proposed Summarization Topic Affinity Score (STAS) measure. The effectiveness of using topic-oriented approaches is further highlighted using the proposed method since the improvements acquired when applying the proposed method are much higher compared to the ROUGE score. Also, both the embedding and the tagging method lead to similar results ($\sim$68.5%) using STAS measure, even though the tagging approach is significantly faster and easier to apply. Note that when no pre-processing is used, the tagging-based approach is more robust to noise, leading to a better STAS score (49.65%) compared to the embedding-based approach (46.70%).

4.3 Zero-shot experimental evaluation

The tagging mechanism allows the model to intuitively guide the summary generation according to the tagged words of the desired topic which can also be an effective way to generalize to unseen topics. To demonstrate the efficacy of the tagging-based model on unseen topics, we fine-tune the BART model on the same training set of the created topic-oriented dataset but removing 5% of the

Table 4: Evaluation based on the proposed Summarization Topic Affinity Score (STAS).

	STAS (%)
BART (Lewis et al., 2020) (all topics)	33.99
Proposed BART_{emb} (all topics)	46.70
Proposed $\text{BART}_{\text{tags}}$ (all topics)	49.65
BART (Lewis et al., 2020) (pre-processed topics)	51.86
Proposed $\text{BART}_{\text{tags}}$ (pre-processed topics)	68.42
Proposed BART_{emb} (pre-processed topics)	68.50

topics. More specifically, we randomly remove 3 topics out of the 70 topics (i.e., "*Movies*", "*Transportation*" and "*Podcasts*") of the training set and evaluate the model both on the test set of seen topics and on the zero-shot test, which consists of 264 articles of unseen topics, as shown in Table 5.

Table 5: Experimental results on both test set with seen topics and zero-shot test set with unseen topics. We report STAS measure scores and f-1 scores for R-1, R-2 and R-L.

	R-1	**R-2**	**R-L**	**STAS (%)**
BART_{tag} (seen topics)	38.31	17.27	26.48	68.21
BART_{tag} (unseen topics)	37.52	16.99	26.71	74.80

Even though the model has not seen the zero-shot topics during training, it can successfully generate topic-oriented summaries for these topics achieving similar results in terms of ROUGE-1 score (∼38% for both test sets) and even better results in terms of STAS measure on the zero-shot test (∼68%) compared to the test set with the seen topics (∼74%). This finding confirms the capability of the tagging-based method to generalize successfully to unseen topics, provided that a set of representative terms is given.

4.4 Examples of generated summaries

We present some examples generated by the tagging-based model on the created dataset for different topics as shown in Table 6. Indeed, the proposed model can shift the generation towards the desired topic of the super-article which contains different topics. Furthermore, the generation of the summary according to the corresponding topic is not affected by the presence of the other topic which is also discussed in the input article.

Table 6: Generated summaries of our proposed tagging-based model according to the two different topics of the super-article containing articles of these topics. Part of summaries is truncated due to size limitations.

Sports: Jenson Button and Fernando Alonso failed to finish the Malaysian Grand Prix ... Button lasted double the amount of time as his teammate. ***Gun Violence:*** Adam Lanza killed his mother, Nancy, inside the home before killing 20 first-graders and six members of staff at Sandy Hook Elementary School in 2012. ...
Transportation: Ford unveiled two prototype electric bikes at Mobile World Congress in Barcelona. ... The bikes are part of an experiment by Ford called Angle on Mobility. ***Neuroscience:*** Researchers from Bristol University measured biosonar bat calls to calculate what members of group perceived as they foraged for food ...

5 Conclusions and Future Work

We proposed a model-agnostic topic-controllable method that can work with any summarization model to influence the summary generation towards the desired topic. The proposed method works by employing special tokens to tag semantically-related words for each topic and then guide the generation towards this topic. To establish a strong baseline, we also adapt an existing topic-controllable embedding-based method to a more powerful Transformer-based model, scaling up from traditional RNNs. We also proposed STAS, a structured way to evaluate the generated summaries according to the affinity of the requested topic with the topic of the generated summary. Experimental results under two different pre-processing setups demonstrate that the proposed method can achieve similar or even better performance than the adapted embedding-based mechanism, while being significantly faster and easier to apply.

Future research could examine other controllable aspects, such as style (Fan et al., 2018) or entities (He et al., 2020). In addition, it would be very interesting to extend the proposed method towards working with any arbitrary topic, bypassing the requirement of having a labeled document collection of a topic to be able to guide the summary towards this topic.

References

Melissa Ailem, Bowen Zhang, and Fei Sha. 2019. Topic augmented generator for abstractive summarization. *arXiv preprint arXiv:1908.07026*.

Seyed Ali Bahrainian, George Zerveas, Fabio Crestani, and Carsten Eickhoff. 2021. Cats: Customizable abstractive topic-based summarization. *ACM Trans. Inf. Syst.*, 40(1).

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Haoran Ding and Xiao Luo. 2021. AttentionRank: Unsupervised keyphrase extraction using self and cross attentions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1928, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Changjian Fang, Dejun Mu, Zhenghong Deng, and Zhiang Wu. 2017. Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, 72:189–195.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *ArXiv*, abs/2012.04281.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating topic-oriented summaries using neural attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1697–1705.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Unsupervised keyphrase extraction by jointly modeling local and global context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 155–164, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (WAS 2004)*.

Jingzhou Liu and Yiming Yang. 2021. Enhancing summarization with text classification via topic consistency. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 661–676. Springer.

Yang Liu and Mirella Lapata. 2020. Text summarization with pretrained encoders. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Xiangke Mao, Hui Yang, Shaobin Huang, Ye Liu, and Rongsheng Li. 2019. Extractive summarization using supervised and unsupervised learning. *Expert systems with applications*, 133:173–181.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 2016 SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Stroudsburg, PA, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

Shengli Song, Haitao Huang, and Tongxiao Ruan. 2019. Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications*, 78.

Vox Media. 2017. Vox Dataset (DS+J Workshop).

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. 2012. Beta-negative binomial process and poisson factor analysis. In *Artificial Intelligence and Statistics*, pages 1462–1471. PMLR.