Tagging Words for Simple, Efficient, Effective and Zero-Shot Topic Control in Neural 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. Furthermore, we conducted a user study which validates the reliability of the proposed measure. Finally, the proposed method can be successfully used for zero-shot topic controllable summarization without any modification.

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; Zhang et al., 2020; Lewis 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 a document, and instead produce a summary that focuses on a specific aspect that pertains to the user’s interests. For example, a financial journalist might want to request summaries around a specific financial term or terms e.g., cryptocurrencies or/and inflation, etc. This need has fueled interest towards controllable summarization methods that can influence the summary according to different user’s requirements, such as guiding the summary’s topic (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019; Bahrainian et al., 2021), length (Takase and Okazaki, 2019; Liu et al., 2020; Chan et al., 2021), entity (Fan et al., 2018a; He et al., 2020; Liu and Chen, 2021; Dou et al., 2021) or style (Fan et al., 2018a).

This paper focuses on topic-controllable summarization. Existing models for topic-controllable summarization either incorporate topic embeddings into the model’s architecture (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019) or modify the attention mechanism (Bahrainian et al., 2021). However, it is not straightforward to use them with any summarization model, as they are restricted to very specific neural architectures. Furthermore, all existing methods for topic-controllable summarization can only work with predefined topics seen during training and cannot handle unknown topics during inference.

At the same time, there is no clear way to evaluate such approaches, since there is no eval-
The rest of this paper is organized as follows. In Section 2 we review 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 future research directions are discussed in Section 5.

2 Topic-oriented Summarization

Methods for topic-oriented summarization belong to two broader categories: a) those 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) enhance the decoder of a pointer-generator network using the information of the latent topics that are derived from LDA. Similar methods have been applied by Wang et al. (2020) using Poisson Factor Analysis (PFA) with a plug-and-play architecture that uses topic embeddings as an additional decoder input based on the most important topics from the input document. 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. Zhu et al. (2021) use a topic-guided abstractive summarization model for Wikipedia articles leveraging the topical information of Wikipedia categories. 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), 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), who proposed a controllable sum-
marization 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 the Vox Dataset (Vox Media, 2017), a news dataset that contains articles from 185 different news topics. Similarly, Freermann and Klementiev (2019) propose another embedding-based approach for topic-controllable summarization, which is based on embedding the desired topic into a pointer generator network using the same latent space with all the other items of the vocabulary.

Recently, Bahrainian et al. (2021) propose to incorporate the topical information from each document to modify the attention mechanism of the pointer generator network using an LDA model. Even though the model is trained with the topical attention mechanism during training, no topical information is used during inference. Thus, it 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.

All the aforementioned methods require modifying the architecture of existing models and can only work with topics seen during training, which limits their ability to generalize to unseen topics. In contrast, the proposed method is model agnostic and can be applied effortlessly and efficiently to a zero-setting setup.

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 a topic-oriented similarity measure for evaluating the topic affinity between the desired topics and the generated summaries.

3.1 Tagging-based formulation

The proposed tagging-based method assumes the existence of a training dataset where each summary is associated with a particular topic, as well as the existence of a set of representative terms for each of these topics. Before training with any neural abstractive model, we pre-process each input document by tagging the representative terms of its summary’s topic with a special token, i.e., [TAG]. This way, during training the model learns to pay attention to tagged words, as they are aligned with the topic of the summary. During inference, to influence summary generation towards a desired topic, the terms of this topic that appear inside the input document are again tagged with the special token. In the following subsections we discuss how we address the two assumptions of our simple, yet effective, method.

3.1.1 Representative terms for each topic

Representative terms for a set of topics can be obtained from a corpus of documents associated with these topics via simple prototype term weighting representations such as BoW (Krishna and Srinivasan, 2018), 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). This work uses tf-idf to extract representative words, as summarized in Figure 1, in order to demonstrate the efficacy of our method, even when a simple topic representation is employed.

Given a topic-assigned corpus \( \mathcal{D} \) and its vocabulary \( V \), we can represent a document \( d \in \mathcal{D} \) as a vector \( x_d \in \mathbb{R}^{|V|} \), containing the tf-idf score for each term \( t \) of the document:

\[
x_{dt} = tf(t, d, \mathcal{D}) \times idf(t, \mathcal{D}),
\]

where \( tf(t, d, \mathcal{D}) \) indicates the number of times that term \( t \) appears in document \( d \), while \( idf(t, \mathcal{D}) \)
Table 1: Representative terms for topics from 2017 KDD Data Science+Journalism Workshop (Vox Media, 2017)

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

indicates the inverse document frequency of term $t$ in corpus $D$:

$$iddf(t, D) = \log \frac{|D| + 1}{df(t, D) + 1} + 1,$$

(2)

where $df(t, D)$ is the frequency of term $t$ in $D$.

We normalize the extracted vectors to have unit length as follows:

$$x_d^{(n)} = \frac{x_d}{||x_d||_2},$$

(3)

where $||x||_2$ is the $l_2$ norm of $x_d$.

We obtain a topical vector representation $y_c$ for each topic $c$, by grouping together documents of the same topic and averaging their tf-idf representation as follows:

$$y_c = \frac{1}{|D_c|} \sum_{x_d \in D_c} x_d$$

(4)

where $D_c$ is the subset of $D$ belonging to topic $c$.

Finally, we extract $N$ representative terms for each topic by considering the terms with the top $N$ tf-idf scores in the corresponding topical vector. An example of some indicative representative words for a number of topics from the Vox dataset (Vox Media, 2017) is shown in Table 1.

3.1.2 Topical training dataset

To apply this mechanism, a topic-oriented training set is required. However, there are no existing datasets for summarization that contain summaries according to the different topical aspects of the text. Thus, we adopt the same approach with Krishna and Srinivasan (2018) to create a topic-oriented dataset that builds upon CNN/DailyMail (Hermann et al., 2015).

More specifically, we create the topic-oriented training dataset as follows. First, we extract BoW vector representations for each topic from the Vox dataset (Vox Media, 2017). Then, we compute the dot-product between the BoW representation of the summary and all the BoW topic representations. 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 final 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 of one of the two selected articles and the same process is repeated to create a new article $a'''$ that is assigned with the remaining summary. Then, the initially selected articles $a_1$ and $a_2$ are removed 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.

The final 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.

3.2 Embedding-based formulation

Following other embedding-based methods for topic-controllable summarization (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019), we establish a strong baseline for comparing the proposed tagging-based method adapting a topic-aware pointer generator 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 total number of 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. Concatenating the topic embeddings with the contextual word embeddings of a Transformer-based model would require retraining the whole summarization model from scratch with the appropriate dimension. However, this would be very computationally demanding as it would require a large amount of data and time.

To this end, instead of concatenation, we propose to sum the topic embeddings following the paradigm of 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 to optimize them 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:

\[ z_i = WE(x_i) + PE(i) + TE, \]

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

To fine-tune the embedding-based summarization model, we use the same topic-oriented dataset, as described in subsection 3.1.2.

3.3 Topic-focused evaluation measure

As explained in Section 1, there is currently no clear 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 representations of the summary and 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. The vector representation of the predicted summary is computed using tf-idf.

Given the vector of the target topic \( x_t \) and the vector representation of the predicted summary \( x_s \), STAS is computed as follows:

\[
STAS(x_s, x_t) = \frac{s(x_s, x_t)}{\max\{s(x_s, x_{ti}) : i = 1...N_t\}},
\]

where \( N_t \) is the number of topic and \( s(x_s, x_t) \) indicates the cosine similarity between the two vectors \( x_s \) and \( x_t \) which is computed as follows:

\[
s(x_s, x_t) = \frac{x_s^T x_t}{\|x_s\| \|x_t\|}.
\]

Thus, summaries that are similar to the requested topic are rewarded, while summaries that are dissimilar are penalized. It is worth noting that cosine similarity values might differ due to the varying overlap between the common words that may appear between the summaries and topics representations. This might lead to a significant variation between the final cosine similarity scores, even though the summary might belong to the same topic. Thus, dividing by the maximum value of the right topic takes into account this phenomenon, leading to a more interpretable metric.

4 Experimental Evaluation

In this section, we present and discuss the experimental evaluation results.

4.1 Experimental Setup

Datasets and evaluation metrics To create the topic-oriented dataset as described in Section 3, any dataset that contains topic annotations can be used. Following (Krishna and Srinivasan, 2018), we also use the Vox Dataset (Vox Media, 2017) which consists of 23,024 news articles of 185 different topical categories. Even though the Vox is a relatively small dataset, we demonstrate that it can achieve adequate results. 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”. Finally, we extract the tf-idf
topic 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).

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). The final topic-oriented dataset consists of 132,766, 5,248, and 6,242 articles for training, validation, and test, respectively. More details about the datasets can be found in the Appendix B.

For the tagging-based method, all the words of the input document are lemmatized to their roots using NLTK (Bird, 2006). Then, we tag the words between the existing lemmatized tokens and the representative words for the desired topic, based on the top-N=100 most representative terms for each topic.

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 (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. More details about the experiments can be found in Appendix C. 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 seven 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 proposed method can also be compared with other approaches that prepend the information to the input source using special tokens, which has been shown that can be effectively combined with Generative LM models like BART (He et al., 2020; Dou et al., 2021). Thus, we report results for BART_{prepend} after appropriately modifying entity-based approaches (Fan et al., 2018a; He et al., 2020; Chan et al., 2021; Dou et al., 2021) to work with topics and for BART_{tag+prepend} combining them with the proposed 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, when BART_{tag} is combined with the prepending mechanism, we obtain the best results of all the evaluated methods. Also, 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.

In Table 3, we also provide an experimental evaluation using the proposed Summarization Topic Affinity Score (STAS) measure. The ef-
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).

<table>
<thead>
<tr>
<th></th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>PG (See et al., 2017)</td>
<td>26.8</td>
<td>9.2</td>
<td>24.5</td>
</tr>
<tr>
<td>BART (Lewis et al., 2020)</td>
<td>30.66</td>
<td>11.92</td>
<td>20.57</td>
</tr>
<tr>
<td>Topic-Oriented PG (Krishna and Srinivasan, 2018)</td>
<td>34.1</td>
<td>13.6</td>
<td>31.2</td>
</tr>
<tr>
<td>BART (all topics)</td>
<td>37.64</td>
<td>16.94</td>
<td>35.20</td>
</tr>
<tr>
<td>BART (all topics)</td>
<td>37.94</td>
<td>17.21</td>
<td>35.37</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>39.30</td>
<td>18.06</td>
<td>36.67</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>40.15</td>
<td>18.53</td>
<td>37.41</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>41.58</td>
<td>19.55</td>
<td>38.74</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>41.66</td>
<td>19.57</td>
<td>38.83</td>
</tr>
</tbody>
</table>

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 (∼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%). Finally, when we combine the proposed method with the prepending mechanism, we observe additional gains, outperforming all the evaluated methods.

Table 3: Evaluation based on the proposed Summarization Topic Affinity Score (STAS).

<table>
<thead>
<tr>
<th></th>
<th>STAS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART (all topics)</td>
<td>33.99</td>
</tr>
<tr>
<td>BART (all topics)</td>
<td>46.70</td>
</tr>
<tr>
<td>BART (all topics)</td>
<td>49.65</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>51.86</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>68.42</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>68.50</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>71.90</td>
</tr>
<tr>
<td>BART (pre-processed topics)</td>
<td>72.36</td>
</tr>
</tbody>
</table>

The results of the inference time for both methods are shown in Table 4. 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.

Table 4: Inference time for 100 articles.

<table>
<thead>
<tr>
<th></th>
<th>Tagging</th>
<th>Inference (sec)</th>
<th>Total time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART (seen topics)</td>
<td>7.1</td>
<td>32.0</td>
<td>39.1</td>
</tr>
<tr>
<td>BART (unseen topics)</td>
<td>7.1</td>
<td>32.0</td>
<td>39.1</td>
</tr>
</tbody>
</table>

4.3 Zero-shot experimental evaluation

In contrary with all the evaluated models, the proposed method is the only one that can directly handle unknown topics. Since the tagging mechanism allows the model to intuitively guide the summary generation according to the tagged words of the desired topic, it 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 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. We report results only for the proposed method, since all the other evaluated methods cannot be directly applied to a zero-shot setting.

Table 5: Experimental results on test set with seen topics and zero-shot test set with unseen topics.

<table>
<thead>
<tr>
<th></th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
<th>STAS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART (seen topics)</td>
<td>38.31</td>
<td>17.27</td>
<td>26.48</td>
<td>68.21</td>
</tr>
<tr>
<td>BART (unseen topics)</td>
<td>37.52</td>
<td>16.99</td>
<td>26.71</td>
<td>74.80</td>
</tr>
</tbody>
</table>

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 Experimental results on real data

We evaluate the proposed methods on real data from CNN/DailyMail using both an oracle setup, where the topic information is extracted from the target summary, and a non-oracle setup, where the topic information is extracted directly from the input document. More specifically, for the non-oracle setup, we retrieve 3000 articles from the test set of the original CNN/DailyMail and then extract the top-3 topics from the input article, using tf-idf. We predict the summary for each of the three different topics.

The results are shown in Table 6. All the models perform quite similarly in terms of ROUGE score, while the best performance is achieved when the proposed method is combined with prepending, outperforming all the evaluated methods. The high STAS score (70.09%) of the combined BART\textsubscript{prepend+tag} model indicate that the proposed method can successfully shift the generation towards the desired topic, even when a more challenging real dataset is used, where the topics are much more related than the synthetic one.

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
<th>STAS(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART\textsubscript{emb}</td>
<td>42.93</td>
<td>20.27</td>
<td>40.14</td>
<td>66.76</td>
</tr>
<tr>
<td>BART\textsubscript{tag}</td>
<td>42.54</td>
<td>20.11</td>
<td>39.80</td>
<td>67.09</td>
</tr>
<tr>
<td>BART\textsubscript{prepend}</td>
<td>42.75</td>
<td>20.20</td>
<td>39.94</td>
<td>69.67</td>
</tr>
<tr>
<td>BART\textsubscript{prepend+tag}</td>
<td>43.35</td>
<td>20.66</td>
<td>40.53</td>
<td>70.09</td>
</tr>
</tbody>
</table>

Table 6: Experimental results on the test set of original CNN/DailyMail, with oracle and non-oracle guidance.

4.5 Human evaluation

In order to validate the reliability of STAS, we conduct a human evaluation study. More specifically, we retrieve 10 generated summaries using the proposed method and we ask 60 volunteer participants (i.e., graduate and undergraduate students) to evaluate how relevant is the generated summary with respect to the given topic in a scale of 1 to 10. Each time, we randomly pick a relevant or an irrelevant topic for the given summary. We also obtain the STAS scores for the same summary-topic pairs. To measure the correlation between human evaluation and the proposed metric, we use two statistical correlation measures: Spearman’s and Pearson’s correlation coefficient. As shown in Table 7, the correlation for both metrics is positive and very close to 1 (~0.8 for both metrics). Also, as indicated from both very low p-values, we can conclude with large confidence that STAS is strongly correlated with the results of human evaluation.

We also investigate the score ranges between human annotators and STAS to roughly indicate a minimum threshold of STAS metric for a summary to be strongly related to a topic. We suppose that if a human annotator evaluates the summary for a topic with more than 8, then the topic is strongly included in the summary. Thus, for all the human evaluations with a score equal to or higher than 8, we take the minimum value of STAS (~0.64) to indicate the minimum threshold, with an average STAS measure of 0.89.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Correlation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>0.80</td>
<td>1.1e-14</td>
</tr>
<tr>
<td>Spearman</td>
<td>0.77</td>
<td>4.2e-13</td>
</tr>
</tbody>
</table>

Table 7: Correlation between human evaluation and STAS measure.

5 Conclusions and Future Work

We proposed a tagging-based 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. We also proposed STAS, a structured way to evaluate the generated summaries. In addition, we conducted a user study to evaluate the validity of STAS. Experimental results demonstrate that the proposed method can achieve similar or even better performance than embedding-based methods, while being significantly faster and easier to apply. Also, the proposed method can be applied with no effort for zero-shot topic-controllable summarization.

Future research could examine other controllable aspects, such as style (Fan et al., 2018b) 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


A Example of generated Summaries

We present some examples of generated summaries using the tagging-based model on the created dataset for different topics along with the generic summary generated by the generic BART model as shown in Table 8 and 9. The proposed model can shift the generation towards the desired topic of the super-article which contains different topics in contrary with the generic BART model which discusses only one topic of the article.
Table 8: Generated summaries of our proposed tagging-based model according to the two different topics of the super-article containing articles of these topics along with the generic summary generated by generic BART model.

**Generic summary:** Sir Bradley Wiggins has been included in Team Sky’s line-up for Sunday’s Gent-Wevelgem sprint classic. Wiggins, the 2012 Tour de France champion, will be joined by fellow Great Britain Olympian Geraint Thomas for the 239-kilometre one-day event in Belgium. Bernhard Eisel, who won the event in 2010, is also included in an eight-man team along with Christian Knees, Ian Stannard, Andy Fenn, Luke Rowe and Elia Viviani.

**Gender-based Violence:** Staff at Venice High School in Los Angeles informed police of a suspected attack on a female student on Tuesday. However the LAPD found the sexual misconduct involved another student and had occurred frequently for more than a year. An investigation has since identified 14 suspects between the ages of 14 and 17 who have allegedly been involved in sexual activity at the school. Eight youngsters from the school were arrested on Friday while another was arrested at another location.

**Sports:** Sir Bradley Wiggins has been included in Team Sky’s line-up for Sunday’s Gent-Wevelgem sprint classic. Wiggins, the 2012 Tour de France champion, will be joined by fellow Great Britain Olympian Geraint Thomas for the 239-kilometre one-day event in Belgium. Giant-Alpecin’s John Degenkolb is the defending champion in Flanders.

### B Datasets

We provide some details and statistics about the datasets that used in the experimental evaluation. We use the anonymized version 3.0.0 of CNN/Dailymail similar to See et al. (2017), which was released under Apache-2.0 License. All the articles from CNN/DailyMail are in the English language. Some statistics of the original CNN/DailyMail along with the created dataset are shown in Table 10.

Table 9: Generated summaries similar to Table 8

**Generic summary:** Ford unveiled two prototype electric bikes at Mobile World Congress in Barcelona. MoDe:Me and Mo de:Pro are powered by 200-watt motors and fold to fit on a train or in the boot of a car. With pedal assist they help riders reach speeds of up to 15mph (25km/h) The bikes are part of an experiment by Ford called Handle on Mobility.

**Transportation:** Ford unveiled two prototype electric bikes at Mobile World Congress in Barcelona. The MoDe: Me and Mo de: Pro are powered by 200-watt motors. They fold to fit on a train or in the boot of a car. With pedal assist, riders reach speeds of up to 15mph (25km/h)

**Neuroscience:** Researchers from Bristol University measured biosonar bat calls to calculate what members of group perceived as they foraged for food. Pair of Daubentons’s bats foraged low over water for stranded insects at a site near the village of Barrow Gurney, in Somerset. It found the bats interact by swapping between leading and following, and they swap these roles by copying the route a nearby individual was using up to 500 milliseconds earlier.

### C Training details

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, with approximate average training runtime 9.5 and 18 hours for the tagging-based and embedding-based method, respectively.

Table 10: Basic Dataset Statistics. The size is measured in articles and length is measured in tokens.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN/DM (original)</td>
<td>287,113</td>
<td>13,368</td>
</tr>
<tr>
<td>CNN/DM (synthetic)</td>
<td>132,766</td>
<td>5,248</td>
</tr>
</tbody>
</table>

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