Topic-Controllable Summarization:
Topic-Aware Evaluation and Transformer Methods

Tatiana Passali, Grigorios Tsoumakas

School of Informatics, Aristotle University of Thessaloniki, Greece
{scpassali, greg}@csd.auth.gr

Abstract

Topic-controllable summarization is an emerging research area with a wide range of potential applications. However, existing approaches suffer from significant limitations. For example, the majority of existing methods built upon recurrent architectures, which can significantly limit their performance compared to more recent Transformer-based architectures, while they also require modifications to the model's architecture for controlling the topic. At the same time, there is currently no established evaluation metric designed specifically for topic-controllable summarization. This work proposes a new topic-oriented evaluation measure to automatically evaluate the generated summaries based on the topic affinity between the generated summary and the desired topic. The reliability of the proposed measure is demonstrated through appropriately designed human evaluation. In addition, we adapt topic embeddings to work with powerful Transformer architectures and propose a novel and efficient approach for guiding the summary generation through control tokens. Experimental results reveal that control tokens can achieve better performance compared to more complicated embedding-based approaches while also being significantly faster.

Keywords: topic-controllable summarization, control tokens, evaluation metric

1. Introduction

Neural abstractive summarization models have matured enough to consistently produce high quality summaries (See et al., 2017; Dong et al., 2019; Zhang et al., 2020; Lewis et al., 2020). Building on top of this significant progress, an interesting challenge is to go beyond delivering a generic summary of a document, and instead produce a summary that focuses on specific aspects that pertain to the user's interests. For example, a financial journalist may need a summary of a news article to be focused on specific financial terms, like cryptocurrencies and inflation.

This challenge has been recently addressed by *topic-controllable summarization* techniques (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019; Bahrainian et al., 2022). However, the automatic evaluation of such techniques remains an open problem. Existing methods use the typical ROUGE score (Lin, 2004) for measuring summarization accuracy and then employ user studies to qualitatively evaluate whether the topic of the generated summaries matches the users' needs (Krishna and Srinivasan, 2018; Bahrainian et al., 2021). Yet, ROUGE can only be used for measuring the quality of the summarization output and cannot be readily used to capture the topical focus of the text. Some early steps have been made in this direction by (Bahrainian et al., 2022), employing a latent Dirichlet allocation (LDA) model to evaluate the topical focus of a summary. However, this serves as a simple indicator of the presence of the topic and cannot be directly used as an evaluation metric as it cannot be easily interpreted across different

documents. Therefore, there is no clear way to automatically evaluate such approaches, since there is no evaluation measure designed specifically for topic-controllable summarization.

At the same time, the majority of 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; Lu et al., 2024). Since these approaches are restricted to very specific neural architectures, it is not straightforward to use them with any summarization model. Even though control tokens have been shown to be effective and efficient for controlling the output of a model for entity-based summarization (Fan et al., 2018; He et al., 2020; Chan et al., 2021) without any modification to its codebase, there have been limited efforts for topic-controllable summarization (Bahrainian et al., 2021).

Motivated by these observations, 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 a summary representation of different topics and then calculating the cosine similarity between the generated summaries and the prototype topic vectors in relation with all the possible topics. The proposed measure assumes the existence of a pre-defined set of topics and thus can be easily adapted to any set of different topics. The effectiveness and reliability of the proposed measure are demonstrated through appropriately designed human evaluation.

In addition, we extend prior work on topic-controllable summarization by adapting topic embeddings (Krishna and Srinivasan, 2018) from the traditional RNN architectures to the more recent and powerful Transformer architecture (Vaswani et al., 2017). However, as shown in experimental evaluation, this approach suffers from significant limitations e.g., slow inference. To this end, we propose a novel control token approach via three different approaches: i) *prepending* the thematic category, ii) *tagging* the most representative tokens of the desired topic, and iii) using both prepending and tagging. For the tagging-based method, given a topic-labeled collection, we extract keywords that are semantically related to the topic that the user requested and then employ special tokens to tag them before feeding the document to the model. We also demonstrate that control tokens can successfully be applied to zero-shot topic-controllable summarization, while at the same time being significantly faster than the embedding-based formulation and can be effortlessly combined with any neural architecture.

Our contributions can be summarized as follows:

- We propose a topic-aware measure to quantitatively evaluate topic-oriented summaries, validated by a user study.

- We develop topic-controllable Transformers summarization methods.

- We provide an extensive empirical evaluation of the proposed methods, including an investigation of a zero-shot setting.

The rest of this paper is organized as follows. In Section 2 we review existing related literature on controllable summarization. In Section 3 we introduce the proposed topic-aware measure, while in Section 4 we present the proposed methods for topic-controllable summarization. In section 5 we discuss the experimental results. Finally, conclusions are drawn and interesting future research directions are given in Section 6.

2. Related Work

2.1. Controllable text-to-text generation

Controllable summarization belongs to the broader field of controllable text-to-text generation (Liu et al., 2021; Pascual et al., 2021). Several approaches for controlling the model's output exist, either using embedding-based approaches (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019), prepending information using special tokens (Fan et al., 2018; He et al., 2020) or using decoder-only architectures (Liu et al., 2021). Controllable

summarization methods can influence several aspects of a summary, including its topic (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019; Bahrainian et al., 2021), length (Liu et al., 2018; Chan et al., 2021), style (Fan et al., 2018), and the inclusion of named entities (Fan et al., 2018; He et al., 2020; Chan et al., 2021). Despite the importance of controllable summarization, there are still limited datasets for this task, including ASPECTNEWS (Ahuja et al., 2022) for extractive aspect-based summarization, EntSUM (Maddela et al., 2022) for entity-based summarization and NEWST (Bahrainian et al., 2022) for topic-aware summarization.

2.2. 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 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) refer to the potential of controlling the output conditioned on a specific topic, all the aforementioned approaches are focused on improving the accuracy of existing summarization models instead of influencing the summary generation towards a particular topic.

2.3. Topic-control in neural abstractive summarization

Some steps towards controlling the output of a summarization model conditioned on a thematic category have been made by Krishna and Srinivasan (2018); Frermann and Klementiev (2019), who proposed embedding-based controllable summarization models on top of the pointer generator network (See et al., 2017). Krishna and Srinivasan (2018) integrate the topical information into the model as a topic vector, which is then concatenated with each of the word embeddings of the input text. Bahrainian et al. (2021) propose to incorporate the topical information into the attention

16283

mechanism of the pointer generator network, using an LDA model.

With the advancements in Transformer architecture, Large Language Models (LLMs) such as GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023) can also be employed for this task. For example, Bahrainian et al. (2022) employ different prompting techniques to control the topic of the summary. Some steps towards integrating the Transformer architecture into topic-controllable summarization have been made by (Lu et al., 2024) who employ contextual embeddings within a constrained attention mechanism. In addition, Zesheng and Yucheng (2023) adopt a topic-aware graph network to generate representations of topic nodes that are then ingested into the decoder of the summarization model to generate a topic-oriented summary.

However, a significant limitation of the majority of these approaches is that they require modifications to the architecture of the summarization model. Our work adapts the embedding-based paradigm to Transformers (Vaswani et al., 2017), and employs control tokens, which can be applied effortlessly and efficiently to any model architecture as well as to the zero-shot setup.

3. Topic-aware Evaluation Measure

We propose a new topic-aware measure, called Summarization Topic Affinity Score (STAS), to evaluate the generated summaries according to their semantic similarity with the desired topic. STAS assumes the existence of a predefined set of topics $\mathcal{T}$, and that each topic, $t \in \mathcal{T}$, is defined via a set, $\mathcal{D}_t$, of relevant documents.

STAS is computed on top of vector representations of topics and summaries. Note that several options exist for extracting such representations, ranging from simple bag-of-words models to sophisticated language models like BERT (Devlin et al., 2019). However, for simplicity, this work uses tf-idf vector representations to demonstrate that even with a simple representation, STAS can be successfully used for evaluating topic-controllable summarization models. More specifically, we employ the tf-idf model, where idf is computed across all documents $\bigcup_{t \in \mathcal{T}} \mathcal{D}_t$. Note that the use of idf allows us to weigh down common words that typically do not contain any important information about the topic of interest.

For each topic t, we compute a topic representation $\mathbf{y}_t$ by averaging the vector representations, $\mathbf{x}_d$, of each document $d \in \mathcal{D}_t$ (see Fig. 1):

$$\mathbf{y}_t = \frac{1}{|\mathcal{D}_t|} \sum_{d \in \mathcal{D}_t} \mathbf{x}_d. \qquad (1)$$

The representation of the predicted summary, $\mathbf{y}_s$,

is computed using the same model too, so as to lie in the same vector space with the topic vectors.

To calculate STAS, we compute the cosine similarity between the representations of the summary, $\mathbf{y}_s$, and the desired topic, $\mathbf{y}_t$, divided by the maximum cosine similarity between the representations of the summary and all topics:

$$STAS(\mathbf{y}_s, \mathbf{y}_t) = \frac{s(\mathbf{y}_s, \mathbf{y}_t)}{\max_{z \in T}\{s(\mathbf{y}_s, \mathbf{y}_z)\}}, \qquad (2)$$

where $s(\mathbf{y}_s, \mathbf{y}_t)$ indicates the cosine similarity between the two vectors $\mathbf{y}_s$ and $\mathbf{y}_t$, computed as follows:

$$s(\mathbf{y}_t, \mathbf{y}_s) = \frac{\mathbf{y}_t \mathbf{y}_s}{\|\mathbf{y}_t\|\|\mathbf{y}_s\|}. \qquad (3)$$

Summaries that are similar to the requested topic receive high STAS values, while those that are dissimilar receive low STAS values. It is worth noting that cosine similarity values can differ significantly across topics, due to the varying overlap between the common words that appear in the summaries and the different topics. Dividing by the maximum value takes this phenomenon into account, leading to comparable STAS values across different topics and allows for normalizing the similarity over the dominant topic.

In addition, STAS can be an effective measure even when more than one dominant topic is discussed in the document, since it avoids distributing the similarity over topics that might appear to a smaller degree in a document, while allowing for normalizing the similarity over the dominant topics. For example, if the document contains two dominant topics, we expect that STAS will be near 1 for both dominant topics, while topics that appear to a smaller degree in a document will not affect the measure.

4. Topic Control with Transformers

In this section, we present the proposed topic-controllable summarization methods that fall into two different categories: a) incorporating topic embeddings into the Transformer architecture and b) employing control tokens before feeding the input to the model. Note that similar to the STAS measure, for all the proposed methods, we assume the existence of a predefined set of topics where each topic is represented from a set of relevant documents.

4.1. Topic Embeddings

Following other embedding-based methods for topic-controllable summarization (Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019), we adapt a topic-aware pointer generator to work with Transformer-based architectures. As described in
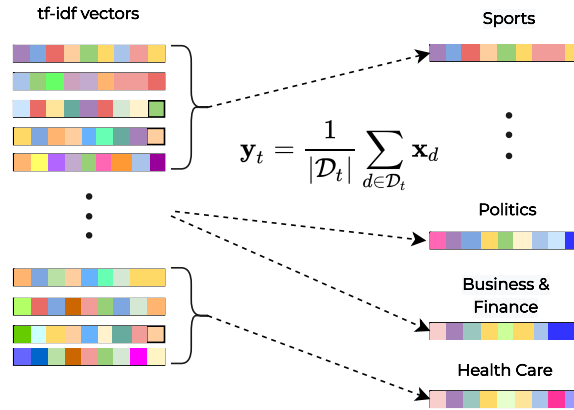
16284

Figure 1: Obtaining representative words, given a topic-assigned document collection. First, we calculate vector representations for each document. Then, documents of the same topic are grouped and their vector representations is averaged. Finally, we obtain the words with top N scores.

Section 2, Krishna and Srinivasan (2018) generate topic-oriented summaries by concatenating topic embeddings, which are represented as one-hot encoding vectors, to the token embeddings of a pointer generator network (See et al., 2017). 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 input 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 the 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 computationally demanding as it would require a large amount of data and time.

Instead of concatenation, we propose to sum the topic embeddings, following the rationale 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.

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

$$z_i = WE(x_i) + PE(i) + TE, \qquad (4)$$

where WE, PE and TE are the word embeddings, positional encoding and topic embeddings respectively, for token x_i in position i. During inference, the model generates the summary based on the trained topic embeddings, according to the desired topic.

4.2. Control Tokens

We propose three different approaches to control the generation of the output summaries using control tokens: a) *prepending* the thematic category as a special token to the document, b) *tagging* with special tokens the representative terms for each topic, and c) combination of both control tokens.

There exist several controllable approaches that prepend information to the input source to influence the different aspects of the text such as the style (Fan et al., 2018) or the presence of a particular entity (He et al., 2020; Chan et al., 2021). Even though this technique can be readily combined with topic controllable summarization, this direction has not been explored yet. We adapt these approaches to work with topic information by simply placing the desired thematic category at the beginning of the document. For example, prepending *"Sports"* to the beginning of a document represents that we want to generate a summary based on the topic *"Sports"*. During training, we prepend to the input the topic of the target summary, according to the training dataset. During inference, we also prepend the document with a special token according to the user's requested topic.

Going one step further, we propose another method for controlling the output of a model based on tagging the most representative terms for each

16285

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
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

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

thematic category using a special control token. The proposed method assumes the existence of a set of the most representative terms for each topic. Then, given a document and a requested topic, we employ a special token to tag the most representative terms before feeding the document to the summarization model, aiming to guide the summary generation towards this topic. We extract N representative terms for each topic by considering the terms with the top N scores in the corresponding topical vector, as extracted in Section 3. Before training with a neural abstractive model, we preprocess each input document by surrounding the representative terms of its summary's topic with the special token [TAG]. This way, during training the model learns to pay attention to tagged words, as they are aligned with the topic of the summary. To influence the summary generation towards a desired topic during inference, the terms of this topic that appear inside the input document are again tagged with the special token.

Given the set of representative words for each topic, a document, and the desired topic, the tagging mechanism works as follows. All the words of the input document are lemmatized to their roots. Then, we identify the common words between the existing lemmatized tokens and the representative words for the desired topic. 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. Some examples of representative terms for each topic are shown in Table 1.

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

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

4.3. Topical Training Dataset

All the aforementioned methods assume the existence of a training dataset, where each summary is associated with a particular topic. However, currently there are no existing large-scale training datasets for abstractive summarization that contain summaries according to the different topical aspects of the text. Thus, we adopt the approach of Krishna and Srinivasan (2018) to compile and release a topic-oriented dataset.

More specifically, Krishna and Srinivasan (2018) create a topic-oriented dataset which contains new super-articles by combining two different articles of the original dataset and keeping the summary of only one of them. First, they extract BoW vector representations for each topic from the Vox dataset (Media, 2017). Then, they 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 superarticles that discuss two distinct topics, but are assigned each time to one of the corresponding summaries. Therefore, the model learns to distinguish the most important sentences for the corresponding topic during training. Even though this procedure requires some additional effort, it allows us to effectively train our models on a topic-controllable setup.

This dataset is used to fine-tune all the aforementioned methods.

5. Empirical Evaluation

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

5.1. Experimental Setup

We use the following two datasets to evaluate the proposed models: (a) CNN/DailyMail and b) Topic-Oriented CNN/DailyMail. CNN/DailyMail is an abstractive summarization dataset with articles from CNN and DailyMail accompanied with human generated bullet summaries (Hermann et al., 2015). We use the non-anonymized version of the dataset similar to See et al. (2017). Topic-Oriented CNN/DailyMail is a synthetic version of CNN/Dailymail which contains super-articles of two different topics accompanied with the summary for the one topic.

To compile the topic-oriented CNN/DailyMail dataset any dataset that contains topic annotations can be used. Following Krishna and Srinivasan (2018), we also use the Vox Dataset (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*". After pre-processing, we end up with 14,312 articles from 70 categories out of the 185 initial topical categories.

The final synthetic topic-oriented CNN/DailyMail consists of 132,766, 5,248, and 6,242 articles for training, validation, and test, respectively while the original CNN/DailyMail consists of 287,113, 13,368 and 11,490 articles.

The Vox dataset is also used to extract the topic vector representations for the STAS measure. We use the tf-idf vectorizer provided by the Scikit-learn library (Pedregosa et al., 2011) to extract a vector representation for each document in the corpus. Then, all the representations of the same topic are averaged to extract the final vector representation for each topic.

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.

For all the conducted experiments we employ a BART-large architecture (Lewis et al., 2020), 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 use the established parameters for the BART-large architecture and the implementation provided by Hugging Face (Wolf et al., 2020). All the models are fine-tuned for 100,000 steps with a learning rate of 3×10^{-5} and batch size 4, with early stopping on the validation set. We use PyTorch version 1.10 and Hugging Face version 4.11.0. All the models were trained using GPUs available in Google Colab, and in specific the NVIDIA T4 Tensor 16 GB GPU. The code and the compiled dataset are publicly available[1].

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.

5.2. Results

The evaluation results on the compiled topic-oriented dataset are shown in Table 2. Our results include the following models:

1. **PG** (See et al., 2017) which is a generic pointer generator network, which is based on the RNN's architecture

2. **Topic-Oriented PG** (Krishna and Srinivasan, 2018) which is the topic-oriented pointer generator network also based on the RNN's architecture.

3. **BART** (Lewis et al., 2020) which is the generic BART model which is based on the Transformer-based architecture.

4. **BART_{emb}** which is the proposed topic-oriented embedding-based extension of BART.

5. **BART_{tag}** which is the proposed topic-oriented tagging-based extension of BART.

6. **BART_{pre}** which is the proposed topic-oriented prepending-based extension of BART.

7. **BART_{pre+tag}** which is the combination of the tagging and prepending extensions of BART.

The experimental results reported in Table 2 show that topic control methods perform significantly better compared to the corresponding baseline methods that do not take into account the topic

---

requested by the user. Furthermore, the proposed BART-based formulation significantly outperforms the topic-oriented PG approach, regardless of the applied method (BART$_{emb}$, BART$_{tag}$ or BART$_{pre}$). The best results are obtained when the tagging and prepending methods are combined. The effectiveness of using topic-oriented approaches is further highlighted using STAS, since the improvements acquired when applying the proposed methods are much higher compared to the improvements in ROUGE score. The embedding and tagging methods lead to similar results (around 68.5%), while the prepending method achieves better results (71.9%). Finally, when we combine the tagging and prepending methods, we observe additional gains, outperforming all the evaluated methods with a 72.36% STAS score.

In addition, the inference time of all the methods that use control tokens is significantly smaller, improving the performance of the model by almost one order of magnitude. Indeed, all the control tokens approaches 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.

A real example of the generated summaries with and without topic control is shown in Table 3 for a super-article that contains a mixture of the *transportation* and *neuroscience* topics. We notice that the summary of BART discusses only one of the two topics of the super-article, while the control tokens in BART$_{tag}$ can successfully shift the generation towards the desired topic of the super-article.

5.3. Zero-shot Experimental Evaluation

In contrast to the embedding-based models, all the methods that use control tokens can directly handle unknown topics. More specifically, for the prepending method we simply prepend the unknown topic to the document while for the tagging method we tag the most representative words for the unknown topic, assuming the existence of a representative set of documents for this topic. To demonstrate the efficacy of control tokens 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 (i.e., "*Movies*", "*Transportation*" and "*Podcasts*") out of the 70 topics of the training set and evaluate the models on the zero-shot test, which consists of 264 articles of unseen topics, as shown in Table 4. We also employ an LLM (GPT-3.5), prompting it to summarize articles on the requested topic given the prompt "'Summarize the following article for the topic [Topic]'".

Even though the models have not seen the zero-shot topics during training, they can successfully generate topic-oriented summaries for

these topics achieving similar results in terms of both ROUGE-1 score and STAS metric, with the BART$_{pre+tag}$ method outperforming all the other methods. In addition, the results indicate that all the proposed BART models outperform GPT-3.5, with BART$_{pre+tag}$ achieving 39.22 compared to 24.43 ROUGE-1. In addition, the proposed method achieves a significantly higher STAS score ($\sim$78%) compared to 58.16% of GPT-3.5. This finding further confirms the capability of methods that use control tokens to generalize successfully to unseen topics, paired with increased efficiency (406M parameters for BART-large vs 175B parameters of GPT-3.5).

5.4. Experimental Results on Original CNN/DailyMail

We evaluate on the original CNN/DailyMail test set all the proposed methods fine-tuned on the topic-oriented CNN/DailyMail training and validation sets, using both an oracle setup, where the topic information is extracted from the target summary according to the assigned topical dataset, and a non-oracle setup, where the topic information is extracted directly from the input document. More specifically, for the non-oracle setup, we extract the top-3 topics from the input article. For computational reasons, we sample 3,000 articles from the test set of the original CNN/DailyMail and we predict the summary for each of the three different topics. STAS is therefore computed on these 9,000 pairs of topics and articles.

The results are shown in Table 5. All models perform quite similarly in terms of ROUGE score in the oracle setup, while the best performance is achieved when tagging is combined with prepending, outperforming all the evaluated methods. We do not compute ROUGE scores in the non-oracle setup, as we lack a gold summary for each different topic in this case.

In terms of STAS, in both setups prepending leads to much better results compared to token embeddings and tagging, which have similar scores. The best results are again obtained when tagging is combined with prepending. The high STAS score of the combined BART$_{pre+tag}$ model in the non-oracle (70.09%) setup shows that this model can successfully shift the generation towards multiple different topics.

5.5. Human Evaluation

In order to validate the reliability of STAS, we conducted a human evaluation study. More specifically, we retrieved a set of 83 summary-topic pairs. Then, we asked 80 volunteer participants, including both graduate and undergraduate students, to participate in this evaluation study. Our main objective

	R-1	R-2	R-L	STAS (%)	Control Tokens	Inference	Total Time
PG	26.8	9.2	24.5	-	-	-	-
BART	30.46	11.92	20.57	51.86	-	30.8	30.8
Topic-Oriented PG	34.1	13.6	31.2	-	-	-	-
BART$_{tag}$ (Ours)	39.30	18.06	36.67	68.42	7.1	32.0	39.1
BART$_{emb}$ (Ours)	40.15	18.53	37.41	68.50	-	303.0	303.0
BART$_{pre}$ (Ours)	41.58	19.55	38.74	71.90	<0.1	30.9	30.9
BART$_{pre+tag}$ (Ours)	**41.66**	**19.57**	**38.83**	**72.36**	7.1	31.7	39.7

Table 2: Experimental results on the compiled 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) and inference time for 100 articles. Time is reported in seconds.

Generic summary: Ford unveiled two proto-type 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 Daubenton'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.

Table 3: Summaries generated by BART$_{tag}$ according to the two different topics of the super-article along with the generic summary generated by BART.

	R-1	R-2	R-L	STAS (%)
GPT-3.5	24.43	6.19	14.8	58.16
BART$_{tag}$	37.52	16.99	35.58	74.80
BART$_{pre}$	38.13	17.84	35.69	74.67
BART$_{pre+tag}$	**39.22**	**18.81**	**36.81**	**77.94**

Table 4: Results on the topic-oriented CNN/DailyMail test set with unseen topics.

was to gauge how well STAS is correlated with human judgments. We asked participants to evaluate how relevant is the generated summary with respect to the given topic. The human evaluation was presented as a multiple-choice question with 10 answers (1 to 10). Each time, we picked either a relevant or an irrelevant topic for the given summary. Each summary topic-pair was annotated by an average of 2.8 annotators. Inter-annotations with a more than 5-degree variance were subjected to manual evaluation and discarded. The inter-annotator agreement across the raters for each sample was 0.86 and measured using Krippendorff's alpha coefficient (Krippendorff, 2004), with ordinal weights (Hayes and Krippendorff, 2007). The high inter-annotator agreement is an indicator of consistency across different raters.

We also investigate the score ranges between human annotators and the STAS metric to better interpret the STAS results. The 10-point evaluation scale allows for better interpretation of the STAS scores, especially in cases where the topic might be relevant but not dominant in the given summary. At the same time, this also gives us an estimation of how to interpret STAS scores on a 1–10 scale based on the human annotations. Thus, we indicate a minimum threshold of the STAS metric for a summary to be strongly related to a topic. More specifically, if a human annotator evaluates the summary for a topic with more than 8, we suppose that this topic is also 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, i.e., 69.60%,to indicate the minimum threshold of the STAS metric for actively discussing a topic in the summary.

6. Conclusions and Future Work

We proposed STAS, a structured way to evaluate the generated summaries. In addition, we

	Oracle				Non-oracle
	R-1	R-2	R-L	STAS(%)	STAS(%)
BART$_{emb}$	42.93	20.27	40.14	71.14	66.76
BART$_{tag}$	42.54	20.11	39.80	71.51	67.09
BART$_{prepend}$	42.75	20.20	39.94	74.20	69.67
BART$_{prepend+tag}$	**43.35**	**20.66**	**40.53**	**74.23**	**70.09**

Table 5: Results on the original CNN/DailyMail test set, with oracle and non-oracle guidance.

Metric	Correlation	p-value
Pearson	0.94	7.8e-63
Spearman	0.87	8.4e-63

Table 6: Correlation between human evaluation and STAS measure.

conducted a user study to validate and interpret the STAS score ranges. We also proposed topic-controllable methods that employ either topic embeddings or control tokens demonstrating that the latter can successfully influence the summary generation towards the desired topic, even in a zero-shot setup. Our empirical evaluation further highlighted the effectiveness of control tokens achieving better performance than embedding-based methods, while being significantly faster and easier to apply.

Future research could examine other controllable aspects, such as style (Fan et al., 2018), entities (Chan et al., 2021) or length (Liu et al., 2018; Chan et al., 2021). In addition, the tagging-based method could be further extended to working with any arbitrary topic, bypassing the requirement of having a labeled document collection of a topic to guide the summary towards this topic. Finally, richer vector representations, such as contextual embeddings, could be explored to further improve the performance of the proposed methods.

7. Bibliographical References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECT-NEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506. Association for Computational Linguistics.

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

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

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

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained markov decision process. *Transactions of the Association for Computational Linguistics*.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 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*.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne,

Australia. Association for Computational Linguistics.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*.

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

K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage.

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,*.

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*. ACL.

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

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

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*. Springer.

Yizhu Liu, Zhiyi Luo, and Kenny Q. Zhu. 2018. Controlling length in abstractive summarization using

a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Ruiying Lu, Bo Chen, Dandan Guo, Dongsheng Wang, and Mingyuan Zhou. 2024. Hierarchical topic-aware contextualized transformers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:841–852.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. ACL.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

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*.

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*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *The 31st Conference on Neural Information Processing Systems (NIPS 2017)*.

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)*. 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*. Association for Computational Linguistics.

Shi Zesheng and Zhou Yucheng. 2023. Topic-selective graph network for topic-focused summarization. *arXiv preprint arXiv:2302.13106*.

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*, pages 11328–11339. PMLR.

Fangwei Zhu, Shangqing Tu, Jiaxin Shi, Juanzi Li, Lei Hou, and Tong Cui. 2021. TWAG: A topic-guided Wikipedia abstract generator. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

8.   Language Resource References

Karl Moritz Hermann and Tomas Kocisky and Edward Grefenstette and Lasse Espeholt and Will Kay and Mustafa Suleyman and Phil Blunsom. 2015. *Teaching machines to read and comprehend*. Advances in Neural Information Processing Systems.

Vox Media. 2017. *Vox Dataset (DS+J Workshop)*. Vox Media. PID https://sites.google.com/view/dsandj2017/datasets.

Abigail See and Peter J Liu and Christopher D Manning. 2017. *Get To The Point: Summarization with Pointer-Generator Networks*. Association for Computational Linguistics.