Natural Language Processing (2025), 1–21
doi:10.1017/nlp.2025.10009

ARTICLE

Controllable abstractive summarization with arbitrary textual context

Tatiana Passali and Grigorios Tsoumakas

School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
Corresponding author: Tatiana Passali; Email: scpassali@csd.auth.gr

(Received 27 January 2024; revised 3 September 2024; accepted 24 July 2025)

Abstract

Controllable summarization models are typically limited only to a short text, such as a topic mention, a keyword, or an entity, to control the output summary. At the same time, existing models for controllable summarization are prone to generate artificial content, resulting in unreliable summaries. In this work, we propose a method for controllable abstractive summarization that can exploit arbitrary textual context from a short text to a collection of documents to direct the focus of the generated summary. The proposed method incorporates a sentence BERT model to extract an embedding-based representation of the given context, which is then used to tag the most representative words of the input document towards this context. In addition, we propose an unsupervised metric to evaluate the faithfulness of the topic-oriented sentences of the generated summaries with respect to the input document. Experimental results under different zero-shot setups demonstrate that the proposed method surpasses both state-of-the-art large language models (LLMs) and controllable summarization methods. The generated summaries are both reliable and relevant with respect to the input document.

Keywords: Text summarization; controllable summarization; abstractive summarization; evaluation; summarization and generation

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 a large 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; Song *et al.* 2019; Liu and Lapata 2020).

Early summarization methods, mainly focusing on *extractive summarization* (Fang *et al.* 2017; Mao *et al.* 2019), had limited success. The advent of deep learning led to much more powerful neural *abstractive summarization* methods (See, Liu, and Manning, 2017; Song *et al.* Song *et al.* 2019; Dong *et al.* 2019; Lewis *et al.* 2020; Zhang *et al.* 2020a), going beyond extracting unaltered sentences from the input to generating the summary using novel words and phrases that are not necessarily part of the input text.

The need to tailor the generated summary of a particular input document to the diverse interests and preferences of different users has fueled interest in controllable summarization methods. *Topic*-controllable summarization methods (Krishna and Srinivasan 2018; Bahrainian *et al.* 2021) influence the summary generation towards a given input topic. *Entity*-based summarization methods (Fan, Grangier, and Auli 2018a; He *et al.* 2022) influence the summary towards user-specified entities. Other methods focus on specific writing *style* guidelines (Fan, Grangier, and Auli, 2018a) or other non-semantic attributes of the text, such as length (Kikuchi *et al.* 2016; Liu, Luo, and Zhu, 2018; Takase and Okazaki 2019; Saito *et al.* 2020).

State-of-the-art approaches employ *prepending* for controlling the output of the summary (Fan, Grangier, and Auli, 2018a; He *et al.* 2022; Zhang *et al.* 2023; Yang *et al.* 2023). Prepending is versatile, as it can be used for controlling the summary towards arbitrary short texts, including entities, keywords, and even topics (Passali and Tsoumakas 2024), while also being able to control length (Fan, Grangier, and Auli, 2018a). However, it cannot scale to large textual contexts, like a document or a collection of documents. The latter is important in applications like news personalization, where we would like to influence the summary of an article towards the prior reading history of a user.

To overcome this limitation, we propose a flexible method for controlling the summary of a document towards arbitrary textual *context*, from a short text to a collection of documents, called *contextual abstractive summarization*. During inference, our method employs a pre-trained embedding model in order to obtain representations of both the context and each word of the input document. The words that are closer to the context in this embedding space are prepended with a special tag (see Figure 1). The same process is used for training any summarization model using any existing summarization dataset by considering the ground truth summary of each document as the context. This way models learn to influence the summary towards the tagged words.

Another limitation of prepending is that models tend to copy verbatim the given short text in the generated summary. While for entities this makes sense, for topics, it can lead to the introduction of misleading information that is not part of the input document (see Table 1). However, existing evaluation metrics for this task are mainly based on the presence of the given short text in the generated summary (Fan, Grangier, and Auli, 2018a; He *et al.* 2022) and can therefore be misleading in the case of topics. This limitation extends to the broader field of text generation, where there is a lack of a comprehensive and reliable evaluation metric to address these phenomena (Zhang *et al.* 2003). In this paper, we propose a new evaluation metric that assesses the similarity of the input document with the sentence of the generated summary that is closest to the context.

The contributions of this paper can be summarized as follows:

- We propose a flexible approach for controlling the summary of a document towards any arbitrary textual context, from a short text to a collection of documents. Our approach can be tied to any model's architecture and can be easily combined with any of the already existing generative summarization approaches.
- We propose an appropriate metric to evaluate the relevance and the reliability of the generated summary to ensure that it does not contain any artificially generated content. A human evaluation study confirms the reliability of the proposed metric.
- We provide an extensive empirical evaluation of the proposed approach, demonstrating both its generality and its effectiveness.
- We demonstrate under different zero-shot setups that the proposed method can influence the summary towards different contexts while preserving the reliability of the generated summary.

The rest of the paper is structured as follows. Section 2 reviews related work on controllable summarization. Section 3 introduces the proposed contextual summarization approach, while Section 4 presents the reliability metric. Section 5 discusses the experimental results. Section 6 draws conclusions from this work and discusses interesting future research directions.

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 1. Examples of summaries generated by CTRLsum and the proposed $BART_{tag}$ for different topics of the same document. Blue and violet demonstrate indicative tagged words for the topics "Science & and Health" and "Neuroscience," respectively. Orange indicates common tagged words for both topics. Bold red indicates the artificially generated content.

Original Document: (CNN) Everybody loves a **good comeback story** – **especially** one that's dino-sized. After its name was **booted** from **science books** for more than a century, a new study suggests that the Brontosaurus belongs to its own genera, and therefore deserves its own name. O.C. Marsh first named the Brontosaurus in 1879, after he received 25 crates of **bones** discovered at Como Bluff, Wyoming, **according** to the **Yale** Peabody **Museum** of **Natural** History. Similar to, though not as large as the Apatosaurus discovered a couple of years prior, Marsh named the dinosaur, "Brontosaurus," or "thunder lizard." Apatosaurus had three sacral **vertebrae** in its **hip** region and Brontosaurus had five, **according** to the museum's website, so Marsh gave the dinosaurs two different names. Later it was discovered that the number of sacral **vertebrae** is **related** to age: as the **animal** gets older, two of the **vertebrae** fuse to the sacrum. **Paleontologist** Elmer Riggs concluded in 1903 that the Brontosaurus was really a young Apatosaurus, and therefore must go by that name, according tothe museum. Emanuel Tschopp, a **paleontologist** at the Nova **University** of Lisbon, Portugal, led this latest study, which took five years and included **visits** to 20 **museums** in Europe and the United States to collect data. By **examining** "500 **anatomical** traits," **Tschopp** said he was able to "reconstruct the family tree" of Diplodocids – the family Apatosaurus and Brontosaurus belong to. **Tschopp** "looked at **every** single specimen," **according** to Jacques Gauthier, **curator** of **vertebrate paleontology** at **Yale** Peabody **Museum** and "noted their characteristics," **plugging data** into an **algorithm** that revealed "five **species** in the group of Diplodocids." Two oft hem are Brontosaurus-like and three are Apatosaurus-like, **Gauthier** said. While his **findings** have earned the excitement of Brontosaurus fans everywhere, **Tschopp** said he is "expecting a **debate** about the case of Brontosaurus." "We think we have **good** evidence," he said, "But I know there are other **researchgroups** that do not completely agree." For Gauthier, Tschopp's **discovery** means changes are **coming** to the original Brontosaurus **skeleton** that has lived at the **Yale** Peabody **Museum** of **Natural History** since 1936. "We're definitely **going**to have to change the label," he said.

CTRLsum

Science & Health: A new study suggests the Brontosaurus belongs to its own genera, and therefore deserves its own name. The study was conducted by a paleontologist at the Nova University of Lisbon, Portugal. **It was published in the journal Science & Health.**

Neuroscience: New study suggests Brontosaurus belongs to its own genera. **"Neuroscience is the future of dinosaurs," says paleontologist Emanuel Tschopp.**

$BART_{tag}$

Science & Health: A new study suggests that the Brontosaurus belongs to its own genera. The study took five years and included visits to 20 museums in Europe and the United States.

Neuroscience: A new study suggests that the Brontosaurus belongs to its own genera. O.C. Marsh first named the dinosaur in 1879, after he received 25 crates of bones.

2. Related work

We first present related methods and models for controllable summarization. Then, we discuss two related fields of study: personalized summarization and query-focused summarization. Finally, we provide an overview of the contributions and differences of the proposed method compared to previous works.

2.1 Controllable summarization

Prior work on controllable summarization focused on influencing the generated summaries according to different aspects. These aspects can be related either to content, such as a *thematic category* (Krishna and Srinivasan 2018; Ailem *et al.* 2019; Wang *et al.* 2020; Bahrainian *et al.* 2021; Passali and Tsoumakas 2024; Bahrainian, Feucht, and Eickhoff, 2022; Lu *et al.* 2024), an *entity* (Fan, Grangier, and Auli, 2018a; He *et al.* 2022; Chan, Wang, and King, 2021; Dou *et al.* 2021), and a *narrative style* (Fan, Grangier, and Auli, 2018b), or to form-related aspects such as the *length* (Kikuchi *et al.* 2016; Fan, Grangier, and Auli, 2018b; Takase and Okazaki 2019; Bian *et al.* 2019; Liu, Luo, and Zhu, 2020; Saito *et al.* 2020) of the output summary. This work focuses on content-related aspects.

Early approaches for controllable summarization were based on recurrent neural networks and required adaptations to the architecture of existing models (Krishna and Srinivasan 2018;

Frermann and Klementiev 2019; Bahrainian *et al.* 2021). Krishna and Srinivasan (2018) integrate topical embeddings with a generic pointer generator network (See *et al.* 2017) for topic-controllable summarization. The topical embeddings are extracted from the Vox Dataset (Vox Media 2017), a topical news dataset that contains more than 180 different thematic categories. Recently, Passali and Tsoumakas (2024) scaled this method to Transformers by summing topic embeddings along with token and positional embeddings to guide the summary generation towards a user-requested topic. Bahrainian *et al.* (2021) adapt the attention mechanism of a pointer generator network to work with topical information derived from an LDA model. Even though this model was trained with the topical attention mechanism, no topical information is used during inference.

Recent approaches have shown the effectiveness of special tokens and prompts for controlling the output of Transformer language models (Fan, Grangier, and Auli, 2018b; Keskar *et al.* 2019; He *et al.* 2022; Passali and Tsoumakas 2024; Bahrainian *et al.* 2022; Zhang *et al.* 2023; Yang *et al.* 2023). Fan *et al.* (2018b) propose a controllable model to generate summaries of a specific writing style or based on a requested entity from the input document. To influence the summary generation towards the desired entity, they prepend special entity markers in the source text. Style control is achieved in a similar way by prepending special style markers (He *et al.* 2022). Similarly, He *et al.* (2022) perform entity-based generation using control tokens in the form of keywords or prompts. Passali and Tsoumakas (2024) introduced different Transformer methods for topic-controllable summarization using special tokens in prepending or tagging the most representative words based on tf-idf weights for each topic.

With the advent of large language models (LLMs) such as ChatGPT (OpenAI, 2022b), LLaMA (Touvron *et al.* 2023a, b; Dubey *et al.* 2024), and Mistral (Jiang *et al.* 2023), prompting has achieved remarkable performance in the field of controllable summarization. For example, ChatGPT has been used for topic-controllable summarization (Yang *et al.* 2023) to direct the summarization output towards a desired topic. In addition, Zhang *et al.* (2023) propose a method for controlling multiple attributes of a document, such as topic, length of summary, extractiveness, and specificity, using prompt and prefix-tuning strategies.

2.2 Personalized summarization

An interesting application of controllable summarization is the personalization of a summary according to the interests of different users. Most of the existing methods for personalized summarization are more than a decade old and extractive. These methods use either textual user annotations, such as keywords (Zhang *et al.* 2003; Móro *et al.* 2012), or leverage information from more interactive features, such as user clicks (Yan, Nie, and Li, 2011) and gaze-based eye tracking (Dubey *et al.* 2020). Zhang *et al.* (2003) use a user's annotations, i.e., any user's word of interest, to generate personalized summaries. Yan *et al.* (2011) perform multi-document personalized summarization through interactive user clicks, while Dubey *et al.* (2020) exploit users' reading patterns with gaze-based eye tracking during a reading session. Díaz and Gervás (2007) combine a short-term and a long-term model based on different user-defined parameters, such as domain, categories, keywords, and feedback terms. Yang *et al.* (2012) use a relevance-based model along with a user model to retrieve the preferences of mobile users and select higher-ranked candidate sentences for the generated summary. Some limited steps have been made towards personalized abstractive review summarization using user characteristics and user-specific word-using habits from online reviews (Li, Li, and Zong, 2019) and headline generation for new articles (Ao *et al.* 2023).

2.3 Query-focused summarization

Another line of research that is close to controllable summarization is query-focused summarization, which refers to the task of generating a summary with respect to a given query (a question,

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 2. Examples of different tagging schemes according to different topics for a document from CNN/DailyMail (Hermann *et al.* 2015).

Climate Change: Peer-reviewed [TAG]**Environmental** Protection Agency studies say that the Clean Air Act and subsequent amendments have reduced early deaths associated with exposure to ambient fine particle [TAG]**pollution** and [TAG]**ozone**, and reduced [TAG]**illnesses** such as chronic bronchitis and acute myocardial infarction. The EPA estimates that, between 1970 and 2010, the act and its amendments prevented 365,000 early deaths from particulate matter alone. "No challengeposes more of a public threat than [TAG]**climate** [TAG]**change**," the President told me.
Science and Health: Peer-reviewed Environmental Protection Agency [TAG]**studies** say that the Clean Air Act and subsequent amendments have reduced early deaths associated with exposure to ambient fine particle pollution and ozone, and reduced [TAG]**illnesses** such as chronic [TAG]**bronchitis** and acute[TAG]**myocardial** [TAG]**infarction**. The EPA estimates that, between 1970 and 2010, the act and its amendments prevented 365,000 early [TAG]**deaths** from particulate matter alone. "No challenge poses more of a public threat than climate change," the President told me.

a word, or a short title). The majority of methods for query-based summarization are extractive (Fisher and Roark, 2006; Daumé III and Marcu 2006; Feigenblat *et al.* 2017; Xu and Lapata 2020), retrieving sentences from the input document that are closest to the given query. However, these methods typically lack coherence. Recent abstractive methods (Nema *et al.* 2017; Xu and Lapata 2021; Su, Yu, and Fung, 2021) achieve much more coherent and fluent results than extractive methods. Nema *et al.* (2017) incorporate query attention into an encoder-decoder RNN model, while other works (Xu and Lapata 2021; Su *et al.* 2021) are based on the Transformer paradigm to generate even higher quality summaries.

2.4 Differences with previous methods

Our work differs from all the aforementioned approaches in several ways. First, the proposed method is model-agnostic, unlike earlier approaches that required modifications to the model's architecture (Krishna and Srinivasan 2018; Frermann and Klementiev 2019; Bahrainian *et al.* 2021), enabling it to be applied effortlessly across any model. Second, it can be used to guide the summary generation towards any form of textual context, whether it is part of or external to the source document. While existing methods depend on short input prompts (Fan, Grangier, and Auli, 2018a; He *et al.* 2022; Zhang *et al.* 2023; Yang *et al.* 2023) to direct the summary, the proposed method is not limited to such constraints since it can incorporate arbitrary textual information to steer the focus of the output summary. This can also be extended to broader contexts, including collections of documents, as we experimentally demonstrate (see subsection 5.2.2). To the best of our knowledge, this is the first work that can be used for *abstractive personalized document summarization* with such diverse textual contexts, including individual documents or entire collections such as a user's reading history.

3. Contextual abstractive summarization

To influence the summary of an input document towards a textual context, our approach prepends during inference the special token [TAG] to the words of the input document that are semantically close to the context. Table 2 demonstrates the result of this process by considering an article from CNN/DailyMail (Hermann *et al.* 2015) as the input document and two different topics ("climate change" and "science and health") as the context.

To achieve this tagging, our method employs SBERT (Reimers and Gurevych 2019), a pre-trained BERT-based sentence embedding model, in order to obtain representations of both the context and each word of the input document (see subsection 3.1). It is important to note that our method is versatile, and any embedding model could be used to obtain these representations at this stage. Then, it computes the cosine similarity between the representations of each word of the input document and the context and tags a word if this similarity is higher than a threshold (see subsection 3.2). An illustration of this process is shown in Figure 1. To train a model to

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

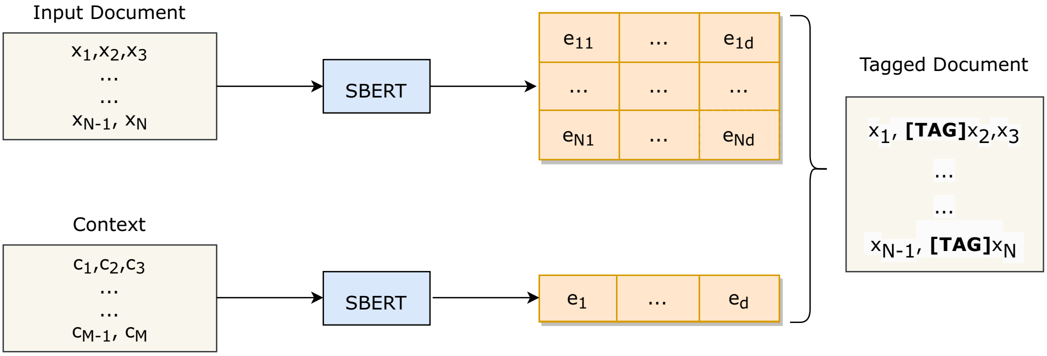


Figure 1. We encode each word of the input document into the same space with the given context using SBERT. Words of the input document that are semantically close to the context are prepended with the special token [TAG].

influence a summary towards the tagged words, our approach relies on existing summarization datasets, treating the ground truth summaries as context (see subsection 3.3).

3.1 Context and input word encoding

One of the key advantages of our approach is that the given context can be any textual information, beyond a short text, such as a document or even a collection of documents. For instance, the collection of documents could be the history of articles that a user has read on a website, in order for summaries of other articles to be aligned with his reading interests. The document could be the Wikipedia page of a topic (e.g., *climate change*) or entity (e.g., *Barack Obama*) a user is interested in, providing a richer source of information for influencing the summary. As another example, it could be a scientific paper that a researcher is drafting; in order to obtain summaries of related papers, she is studying. Our approach can, of course, also accept short text, such as one or more entities, topics, or arbitrary keywords (e.g., *battery life*, *screen*) that are of interest to a user.

The encoding, $e(c)$, of a context, c, differs among the different types of context. When it is a short text, we directly give it as input to SBERT, as the number of its tokens is not expected to exceed the input size of the model:

$$e(c) = \mathrm{SBERT}(c). \tag{1}$$

When the context is a single document, consisting of n sentences $\{s_1, \ldots, s_n\}$, we average all the sentence representations of the document into a final representation:

$$e(c) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{SBERT}(s_i). \tag{2}$$

In the case of a collection of n documents $\{d_1, \ldots, d_n\}$, for each document we follow the same process as in the case of a single document, and then we average all the document representations into a final representation:

$$e(c) = \frac{1}{n}\sum_{i=1}^{n} e(d_i). \tag{3}$$

The encoding of each word $\{x_1, x_2, \ldots, x_n\}$ of the input document x is also computed via SBERT so that it belongs to the same space as the encoding of the context:

$$e(x_i) = \mathrm{SBERT}(x_i). \tag{4}$$

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Note that since Transformers use sub-word tokenizers, words that comprise more than one token are represented by the average of the embeddings of their tokens.

The length of the context can affect the tagging process and, as a result, the quality of the generated summaries. We expect that using a larger context, such as a full document or a collection of documents, will provide a richer and more informative representation compared to a short input text, which may lack contextual depth. Our experimental results show that even with a large context, like a collection of documents, the performance remains high and is not negatively affected by the longer context length, demonstrating the versatility of our approach. At the same time, our method also performs well with shorter input contexts. This flexibility allows our approach to adapt effectively to different types and lengths of contexts.

3.2 Tagging

Given an input document x consisting of n words $\{x_1, x_2, \ldots, x_n\}$ and a context c, we first compute the cosine similarity between the encoding of each word $e(x_i)$ and the encoding of the context $e(c)$. We will tag those words whose similarity to the context is higher than a threshold t.

To compute this threshold, we learn a Gaussian mixture model with two components from these similarities, under the assumption that one corresponds to the similarities of words that are relevant to the context and one to those that are irrelevant. A Gaussian mixture model GMM for each similarity sim_{x_i} between a word x_i and the context c can be defined as follows:

$$GMM(sim_{x_i}) = \pi_1 N(sim_{x_i}|\mu_1, \Sigma_1) + \pi_2 N(sim_{x_i}|\mu_2, \Sigma_2), \tag{5}$$

where π_1 and π_2 indicate the weight coefficients of each component and $N(x|\mu_k, \Sigma_k)$ represents the probability density function with μ_1, μ_2 and Σ_{12}, Σ_2 to be the mean and the variance of each component.

We then take t the average of the means of the two Gaussian distributions as follows:

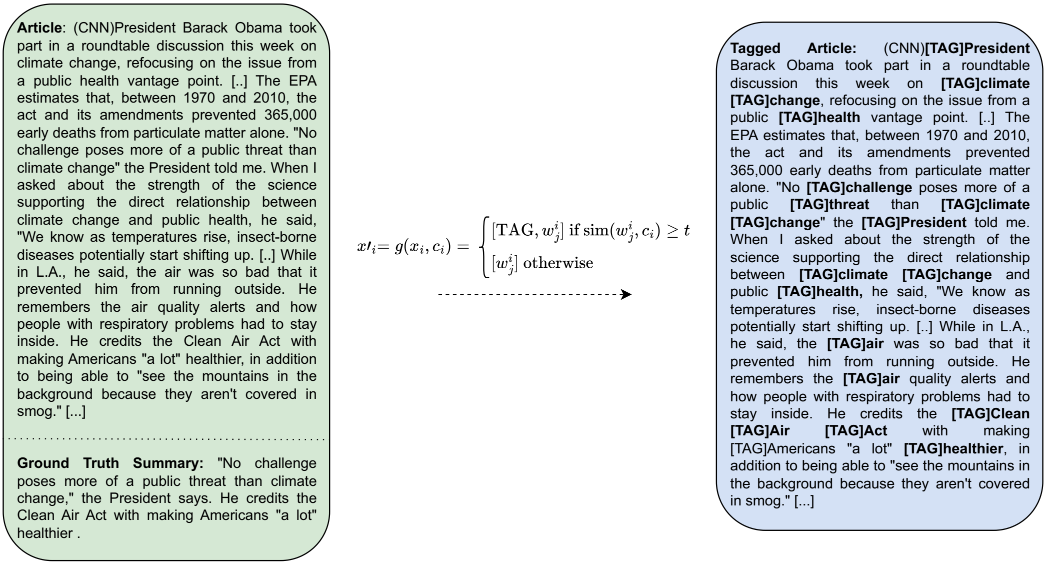
$$t = \frac{\mu_1 + \mu_2}{2}. \tag{6}$$

3.3 Training dataset

Large language models like Transformers typically require a large training dataset. Even though these datasets exist for general-purpose summarization, there is a lack of specific-task datasets with controllable attributes or additional information about different contexts. At the same time, existing controllable datasets with human annotations are very small (Bahrainian *et al.* 2022; Zhang *et al.* 2023), limiting their usefulness for training such models. Prior work for topic-controllable summarization relied mostly on synthetic datasets Krishna and Srinivasan (2018); Passali and Tsoumakas (2024). However, this raises concerns regarding the reliability of such datasets, which may potentially negatively impact the models' performance on real-world datasets. For example, Krishna and Srinivasan (2018) generate a dataset for topic-controllable summarization by combining sentences from two distinct documents on different topics paired with a single summary derived from only one of them. While this approach can serve as a good baseline for evaluation, it may yield unreliable results or even be too simple for models to accurately distinguish between different topics. In practice, documents, such as news articles, will typically discuss relevant topics, making this setup unrealistic.

The proposed method overcomes these limitations by exploiting the ground truth summaries of existing summarization datasets like CNN/DailyMail (Hermann *et al.* 2015), MultiNews (Fabbri *et al.* 2019), and XSum (Narayan, Cohen, and Lapata, 2020) to tailor the summary generation towards this context. Following this process, the proposed method incorporates the inherent structure of any summarization dataset without the need for additional labeled data.

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press



Article: (CNN)President Barack Obama took part in a roundtable discussion this week on climate change, refocusing on the issue from a public health vantage point. [..] The EPA estimates that, between 1970 and 2010, the act and its amendments prevented 365,000 early deaths from particulate matter alone. "No challenge poses more of a public threat than climate change" the President told me. When I asked about the strength of the science supporting the direct relationship between climate change and public health, he said, "We know as temperatures rise, insect-borne diseases potentially start shifting up. [..] While in L.A., he said, the air was so bad that it prevented him from running outside. He remembers the air quality alerts and how people with respiratory problems had to stay inside. He credits the Clean Air Act with making Americans "a lot" healthier, in addition to being able to "see the mountains in the background because they aren't covered in smog." [...]

Ground Truth Summary: "No challenge poses more of a public threat than climate change," the President says. He credits the Clean Air Act with making Americans "a lot" healthier .

$$x\prime_i = g(x_i, c_i) = \begin{cases} [\text{TAG}, w_j^i] \text{ if } \text{sim}(w_j^i, c_i) \geq t \\ [w_j^i] \text{ otherwise} \end{cases}$$

Tagged Article: (CNN)**[TAG]President** Barack Obama took part in a roundtable discussion this week on **[TAG]climate [TAG]change**, refocusing on the issue from a public **[TAG]health** vantage point. [..] The EPA estimates that, between 1970 and 2010, the act and its amendments prevented 365,000 early deaths from particulate matter alone. "No **[TAG]challenge** poses more of a public **[TAG]threat** than **[TAG]climate [TAG]change**" the **[TAG]President** told me. When I asked about the strength of the science supporting the direct relationship between **[TAG]climate [TAG]change** and public **[TAG]health,** he said, "We know as temperatures rise, insect-borne diseases potentially start shifting up. [..] While in L.A., he said, the **[TAG]air** was so bad that it prevented him from running outside. He remembers the **[TAG]air** quality alerts and how people with respiratory problems had to stay inside. He credits the **[TAG]Clean [TAG]Air [TAG]Act** with making [TAG]Americans "a lot" **[TAG]healthier**, in addition to being able to "see the mountains in the background because they aren't covered in smog." [...]

Figure 2. Training Dataset Creation. We exploit the ground truth summaries of existing large-scale summarization datasets to tailor the summary generation towards this context.

More specifically, given a summarization dataset that consists of documents accompanied by their respective target summaries, the proposed method uses these ground truth summaries as contexts and extracts their representation. The extracted context representation from the target summary is used to tag the most representative words of the input document. More specifically, given a document x_i that consists of n words $\{w_1^i, w_2^i, \ldots, w_{n_i}^i]$ and its corresponding context c_i as extracted from the target summary, we define a similarity function $\text{sim}(w_j^i, c_i)$ which measures the similarity between the word w_j^i and the context c_i. This process is described by the function g, where:

$$x_i' = g(x_i, c_i) = \begin{cases} [\text{TAG}, w_j^i] & \text{if } \text{sim}(w_j^i, c_i) \geq t \\ [w_j^i] & \text{otherwise} \end{cases} \tag{7}$$

Here, x_i' represents the version of the document x_i with specific words tagged with a special token [TAG], based on their similarity to the context exceeding the threshold t. The process of compiling the training dataset is shown in Figure 2.

During training, the model learns to intuitively guide the summary generation towards the target summary by back-propagating the cross-entropy loss between the predicted and the ground truth summary. As a result, the model learns to intuitively give more attention to the words that are prepended with the special tag token. During inference, the model can direct the summary towards any textual representation by adjusting the position of control tokens.

4. Relevance measure

Existing controllable summarization models are prone to generating hallucinated content in order to ensure the presence of the user-requested topic in the summary. For example, CTRLsum might force the generation of the requested topic in the summary regardless of its relevance to the input document, leading to unfaithful summaries with misleading information, as shown in Table 1.

The evaluation of such models typically includes metrics that simply count how many times the requested topic appears in the summary (He *et al.* 2022) or compute the similarity between the

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

generated summary and the given topic (Passali and Tsoumakas 2024). In this way, summaries that include the requested topic might get a high relevance score without being reliable and relevant to the input document.

To overcome this limitation, we propose an unsupervised relevance measure (REL) to evaluate how faithful the generated summaries are to the input document. REL builds on the assumption that if a summary is relevant to the input document, we expect that the sentence of the generated summary that is closest to the requested topic should be close to at least one sentence of the original document. In addition, it requires only the generated summary and the input document, so it can be easily used without the need for ground truth summaries.

REL consists of the following steps: First, given a generated summary S, we extract the sentence from the summary that is closest to the requested topic. Then, REL is computed as the maximum of all the similarities between the selected sentence representation and each of the sentence representations of the original document.

More specifically, REL is computed as follows:

$$REL(S, D) = \max_{i \in D}\{sim(e(s_i), e(s_r)\}, \tag{8}$$

where D represents the set of sentences in document d and $sim(e(s_j), e(s_r)$ denotes the cosine similarity between the representation of the sentence j of document d and the sentence r from summary s. The sentence r is defined as

$$r = \arg\max_{j \in S} \text{sim}\ (e(s_j), e(c)), \tag{9}$$

where S represents the set of sentences in a summary s and $sim(e(s_j), e(c)$ denotes the cosine similarity between the representation of the sentence j of summary s and the context c.

A high REL score indicates that the generated summary is semantically close to the document, while a lower REL score is a strong indicator that the generated summary might contain content that is not reliable. In cases where more than one sentence is semantically close to the context, REL considers the most relevant one and still can provide an indicator of the reliability of the summary. However, note that the proposed metric is not intended to replace existing metrics for summarization, such as ROUGE score, but rather to serve as an additional indicator of the reliability of a summary in relation to the provided context. Exploring alternative variations of the metric (e.g., incorporating multiple relevant sentences) is left for future research.

5. Empirical evaluation

This section presents the results of the empirical evaluation of the proposed method. First, we provide details about the experimental setup, and then we present and discuss the experimental and human evaluation results.

5.1 Experimental setup

In this subsection, we present the datasets that were used for the evaluation of the proposed methods. Also, we provide details about the models and the training as well as discuss the evaluation metrics that we used.

5.1.1 Datasets

We use the six following abstractive summarization datasets for our experiments:

- **CNN/DailyMail (Hermann *et al.* 2015)**: a news summarization dataset that consists of more than 300K articles and summaries from CNN and DailyMail news sources. We use the non-anonymized version 3.0.0 of the dataset.

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 3. Dataset statistics. Size is measured in articles for train, validation, and test set while the average length for documents and summaries is measured in tokens.

	Size (articles)			Length (tokens)	
Dataset	Train	Validation	Test	Document	Summary
CNN/DailyMail	287,113	13,368	11,490	781	56
MultiNews	44,972	5,622	5,622	2,103	264
XSum	204,045	11,332	11,334	431	23
NEWTS	4,800	–	1200	568	70
MacDoc	4,278	554	547	835	54
Debatepedia	12,000	719	1,000	72	10

- **MultiNews (Fabbri *et al.* 2019)**: a summarization dataset with more than 56K news articles accompanied by human-written summaries from more than 1,500 websites.
- **XSum (Narayan *et al.* 2020)**: a news summarization dataset with more than 200K articles and human-written summaries collected from the BBC website.
- **NEWTS (Bahrainian *et al.* 2022)**: a human-annotated topic-controllable summarization dataset based on CNN/Dailymail. It consists of 6,000 article-summary pairs annotated with topics.
- **MacDoc (Zhang *et al.* 2023)**: a human-annotated controllable summarization dataset from the MacSum dataset. It is based on the CNN/DailyMail dataset and contains over 5,000 articles and human-written summary pairs with five different control attributes, including topic, among others.
- **Debatepedia (Nema *et al.* 2017)**: a dataset that contains debate topics, including pros and cons arguments for 663 topic areas, in the form of triplets with a document, query, and summary. Each debate topic can include several queries that result in a dataset of 12,695 triplets.

Some statistics regarding the size and length of the datasets are shown in Table 3.

5.1.2 Context sources

We evaluate the proposed method under three different context scenarios: a) a topic mention, b) a document, and c) a collection of documents. To simulate these context representations, we employ Vox (Vox Media 2017), a topical news corpus with 23,024 articles divided into 185 thematic categories. We remove categories that are assigned to less than 30 documents as well as general categories (e.g., *"The Latest," "On Instagram," "Vox Sentences," "Podcasts," "Episode of the Week," "Reviews, " "2016ish," "First Person," "Identities,"* and *"The Big Idea"*). After preprocessing, we result in 61 out of 185 categories. We then use the filtered corpus to extract the different context representations. More specifically, in the case of the short text, we use the category itself. For a single document, we randomly select an article from the corpus, while for a collection of documents, we use all the documents that are assigned to the same category.

We use the test sets of the CNN/DailyMail, XSum, and MultiNews datasets to influence the summary according to the different context scenarios as extracted from the Vox Dataset. For each document in the dataset, we use the top-3 closest context representations for each scenario (short text, single document, document collection) to generate the tagging scheme for each document. For each document in the test set, we generate three new documents with different tagging

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

schemes according to the different contexts. Note that since articles do not typically include highly diverse topics, we expect that the top-3 topics will be close to each other. For the Debatepedia dataset, we use the query as the context.

5.1.3 Models and training

The proposed method is built on top of two state-of-the-art summarization models, BART-large and PEGASUS-large. BART-large is a Transformer model with 12 encoder/decoder layers, a bidirectional encoder, and an autoregressive decoder. It consists of approximately 406 M parameters. PEGASUS-large is another Transformer model with 16 encoder/decoder layers and consists of more than 560 M parameters. To compare the performance of the proposed method against state-of-the-art models, we also employ a variety of LLMs, including GPT-3.5, GPT-4, LLaMA-3 8B, Mistral 7B, and Claude v2. For all the LLMs, we set the temperature to 0 for reproducibility of the experiments. Following prior work on topic-controllable summarization by Yang *et al.* (2023), we use the following prompt: ”Summarize this article with respect to Aspect [Aspect]. Write directly the summary.”

To extract all the embedding-based representations, we use the multi-qa-distilbert-cos-v1, a lightweight DistillBERT model as introduced by Sentence Transformers Reimers and Gurevych (2019). We use the Hugging Face (Wolf *et al.* 2020) implementation for all the models.

We fine-tune both summarization models with a batch size set to 6 and a learning rate set to 3e-05, following He *et al.* (2022). Both BART-large and PEGASUS-large were trained for 150,000 steps with early stopping on the validation set, as they typically converged before reaching this limit. Note that further fine-tuning or extensive hyperparameter searching could lead to better performance. All training experiments were conducted on an NVIDIA T4 Tensor 16 GB GPU using Google Colab. For the inference of LLMs, we used the AWS Amazon Bedrock service, while the GPT models were accessed via the OpenAI API.

The evaluation results include the following models:

- **BART** (Lewis *et al.* 2020), the vanilla BART model, based on the BART-large architecture, without controllable attributes.
- **PEGASUS** (Zhang *et al.* 2020a), the vanilla PEGASUS model, based on PEGASUS-large architecture, without controllable attributes.
- $BART_{tag}$ model, which is based on the BART-large architecture.
- $PEGASUS_{tag}$ model, which is based on the PEGASUS-large architecture.
- **CTRLsum** (He *et al.* 2022), a controllable summarization model fine-tuned on the CNN/DailyMail dataset that works by prepending the requested entity to the input document.
- **BART-FT** (Su *et al.* 2021), a BART model for query-focused summarization that works by concatenating the query with the document using a special [SEP] token.
- **GPT-3.5** (OpenAI, 2022a), an LLM, developed by OpenAI, based on the GPT-3 architecture with 175 billion parameters.
- **GPT-4** (OpenAI, 2023), an improved version of GPT-3.5, also developed by OpenAI, with 1.5 trillion parameters.
- **LLaMA 3** (Dubey *et al.* 2024), a family of LLMs developed by Meta AI, with multiple versions of different parameters. In this setup, we use the 8 billion version.
- **Mistral** (Jiang *et al.* 2023), an LLM developed by Mistral AI, with 7 billion parameters.
- **Claude** (Anthropic, 2024), an LLM, developed by Anthropic AI, with 130 billion parameters.

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 4. Experimental results on CNN/DailyMail dataset using different input context for a short text (topic), document (doc.), and collection of documents (col.). F-1 scores for ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) are reported.

	R1	R2	RL	BERTScore	REL	cos	sim_{sum}
CTRLsum (topic)	37.33	16.66	34.55	87.06	0.67	0.27	0.80
$PEGASUS_{tag}$ (topic)	41.97	19.85	38.74	87.82	0.90	0.19	0.93
$BART_{tag}$ (topic)	43.59	20.85	40.66	88.29	0.86	0.19	0.92
$PEGASUS_{tag}$ (doc.)	41.96	19.86	38.72	87.82	0.90	0.19	0.95
$BART_{tag}$ (doc.)	43.52	20.78	40.58	88.28	0.86	0.19	0.92
$PEGASUS_{tag}$ (col.)	41.89	19.76	38.64	87.81	0.90	0.20	0.95
$BART_{tag}$ (col.)	43.63	20.87	40.70	88.30	0.86	0.25	0.93

5.1.4 Evaluation metrics

We use the family of ROUGE (Lin, 2004) metrics and BERTScore (Zhang *et al.* 2020b) for evaluating the quality of the generated summaries as well as the proposed REL metric to assess the reliability of the summary with respect to the input document. Since we do not have a ground truth summary for the different contexts, we calculate ROUGE with the target summary of the corresponding document. Even though ROUGE and BERTscore metrics are based on the target summary, they can provide an overview of whether the generated summaries are still in line with the general meaning of the input document. In addition, we report the cosine similarity ("cos") between the generated summary and the requested topic to evaluate how close the generated summary is to the given context. The higher the cosine similarity, the closer the summary is to the given context. Also, we compute the variance ("sim_{sum}") between the different summaries generated towards different contexts for the same input document. For calculating the variance, we compute the average cosine similarity between all the different pairs of the different generated summaries. A lower score indicates that the summaries have high variance, while a higher score indicates low variance with similar summaries.

5.2 Results

This subsection provides the results of the experimental evaluation on different datasets. First, we discuss the results for the different textual contexts, including short (single word or phrase) and arbitrary textual contexts (document or collection of documents). Then, we proceed with a comparison with existing controllable summarization models such as CTRLsum (He *et al.* 2022), state-of-the-art LLMs, and query-focused models such as BART-FT (Su *et al.* 2021).

5.2.1 Short textual context

The results of the CNN/DailyMail test set for the different context scenarios are shown in Table 4. We notice that the average cosine similarity (cos) between the generated summaries and the requested short text for a single topic is similar for both the $PEGASUS_{tag}$ and $BART_{tag}$ models (~ 0.19). Furthermore, the average similarity between the different summaries for the same document for both models is around 0.9, indicating that the models do generate slightly different summaries in different contexts. Note that since the top 3 topics that are nearest to the input document are typically similar to each other, we do not expect major changes between the different summaries.

In addition, the evaluation results in Table 4 demonstrate a significantly higher cosine similarity for the CTRLsum model (0.29 compared to 0.19), along with a higher variance between the

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 5. Examples of summaries generated by CTRLsum and the proposed $\mathit{PEGASUS}_{tag}$ for different topics of the same document.

Original Document: **Amanda** **Knox** may have finally cleared her name, but eight years of **legal battles** have left the **Seattle** native penniless, exhausted and **traumatized** from stress, claims her biographer. “In prison, she was **threatened** with **rape** from a **male** guard, it was a really **terrible** experience,” says **Trial** By **Jury** author, **Douglas** Preston. “I think it has really **affected** her, it’s really hard to **lead** a normal life after that.” Scroll down for **video** . Strain and relief: **Amanda** **Knox** spoke outside her parents’ home on March 27 flanked by her fiancé **Colin Sutherland** after she was **exonerated** for the **murder** of Meredith Kercher in **November** 2007 . Three weeks after an Italian **court overturned** her **conviction** for the 2007 **murder** of British student, Meredith Kercher, Prestontold RadarOnline that **Knox** is living out a **bittersweet** victory. “She went to Italy as a normal 20-year-old, West Coast girl, a soccer player, rock climber, very naïve,” says Preston about Knox’sill-fated adventure to Perugia. “So many **terrible** **things** happened to her in Italy, so many lies were said, and the online **savagery** directed at her, I’ve never seen **anything** like it.” So **affected** has **Knox** been by her **struggle** to be **exonerated** by the Italian courts, that Preston hinted to RadarOnline that she is **suffering PTSD** and seeking **professional** help. “Anyone who has been through whatshe’s been through is going to have issues. It’s like a **soldier** come back from Iraq, having children **shot** in **front** of him, you’ve been through a **horrific** experience. If it doesn’t affect you physically,it affects you emotionally,” he says to Radar. Family struggle: **Amanda** **Knox** stands on the **front** porch with her fiancé, her mother Edda **Mellas** and stepfather, Chris Mellas. According to herbiographer, **Knox** and her family have spent **millions** **fighting** her **conviction** . On top of her **fight** to be **declared** innocent, Preston claims that **Knox** has been **ravaged financially** by **legal** fees. This despite **receiving** a reported 4 million **advance** for her 2013 book, Waiting To Be Heard. “The **ordeal** has been hugely emotional, but the **financial impact** has been catastrophic,” says Preston.“Amanda received a 4 million **advance** for her book, that **money** has all gone, immediately out oft he door to **pay** **legal** **fees** and there is nothing left,”Preston explains. “She really wants to **establish** that she’s **innocent** in Italy, they’ve paid a lot of **fees** to Italian lawyers.”
CTRLsum
Criminal Justice: The Seattle native was cleared oft he 2007 murder of Meredith Kercher on March 27 by the Italian Court of **Criminal Justice.**
Gun Violence: Amanda Knox’s biographer, Douglas Preston, tells RadarOnline that she is suffering from PTSD and is seeking professional help. **Preston is the author oft he book, “Gun Violence:A Biography of Amanda Knox”.**
$\mathit{\textbf{PEGASUS}}_{\textbf{tag}}$
Criminal Justice: Amanda Knox and her family have spent millions fighting her conviction for the 2007 murder of British student Meredith Kercher. Her biographer, Douglas Preston, claims thatKnox has been leftpenniless, exhausted and traumatized from stress.
Gun Violence: Trial By Jury author Douglas Preston tells RadarOnline that Amanda Knox is living out a bittersweet victory. “She went to Italy as a normal 20-year-old, West Coast girl, a soccerplayer, rock climber, very nave,” says Preston about Knox’s ill-fated adventure to Perugia. “So many terrible things happened to her in Italy, so many lies were said, and the online savagery directedat her, I’ve never seen anything like it.” So affected has Knox been by her struggle to be exonerated by the Italian courts, that Preston hinted to RadarOnline that she is suffering PTSD and seekingprofessional help.

generated summaries. In addition, CTRLsum has a lower performance in terms of ROUGE and BERTscore metrics compared to the proposed models, which indicates that the generated summaries for the different topics might differ. On the other hand, these summaries might diverge from the general meaning of the input document. Despite the fact that the lower scores show a greater variance between the different summaries, the REL metric reveals a significant weakness in CTRLsum’s ability to generate relevant summaries with respect to the input document, in contrast to the proposed methods, $\mathit{PEGASUS}_{tag}$ and BART_{tag}, which show significantly higher reliability. More specifically, a qualitative evaluation of the generated summaries shows that CTRLsum succeeds in achieving higher performance by forcing the generation of the requested topic in the output summary. Some indicative examples of this behavior are shown in Table 1 and 5.

More specifically, we can see that CTRLsum forces the generation of the requested topic without ensuring the content validity of the summary. Thus, it can generate artificial content without preserving the document’s original meaning, resulting in unreliable summaries. For example, in Table 1 for both requested topics, *Science & Health* and *Neuroscience*, CTRLsum tends to generate inaccurate information, as shown in bold red text, that is not stated in the original document. The proposed BART_{tag} does not suffer from the same limitation. Similar conclusions can be drawn for

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 6. Experimental results on the MultiNews dataset using different input context for a short text (topic), document (doc.), and collection of documents (col.). F-1 scores for ROUGE-1 (R1), ROUGE-2(R2), and ROUGE-L (RL) are reported.

	R1	R2	RL	BertScore	REL	cos	sim_{sum}
$PEGASUS_{tag}$(topic)	43.05	16.10	37.97	86.37	0.84	0.22	0.86
$BART_{tag}$ (topic)	42.97	15.05	37.69	86.40	0.77	0.24	0.82
$PEGASUS_{tag}$ (doc.)	42.67	15.89	37.63	86.31	0.83	0.22	0.85
$BART_{tag}$ (doc.)	41.70	14.03	36.49	86.06	0.74	0.22	0.83
$PEGASUS_{tag}$ (col.)	42.93	16.02	37.86	86.35	0.84	0.33	0.85
$BART_{tag}$ (col.)	42.34	14.60	37.09	86.23	0.76	0.34	0.83

Table 7. Experimental results on the XSum dataset using different input context for a short text (topic), document (doc.), and collection of documents (col.). F-1 scores for ROUGE-1 (R1), ROUGE-2(R2), and ROUGE-L (RL) are reported.

	R1	R2	RL	BertScore	REL	cos	sim_{sum}
$PEGASUS_{tag}$(topic)	45.43	22.70	37.31	92.11	0.66	0.20	0.83
$BART_{tag}$ (topic)	41.08	18.33	32.99	91.36	0.64	0.23	0.82
$PEGASUS_{tag}$ (doc.)	45.00	22.31	36.85	92.04	0.65	0.20	0.82
$BART_{tag}$ (doc.)	39.61	17.32	31.80	91.11	0.63	0.19	0.81
$PEGASUS_{tag}$ (col.)	45.30	22.57	37.17	92.08	0.65	0.28	0.82
$BART_{tag}$ (col.)	40.87	18.16	32.80	91.32	0.64	0.30	0.82

the $PEGASUS_{tag}$ model, as shown in Table 5. Again, CTRLsum generates inaccurate information by imposing the requesting topic in the summarization output for both *criminal justice* and *gun violence* topics. In contrast, the proposed $PEGASUS_{tag}$ model generates summaries that are both topic-aware and provide reliable and accurate information according to the input document.

5.2.2 Arbitrary textual context

In contrast to existing controllable models like CTRLsum, the proposed method is not restricted to a specific word or entity for controlling the output summary. This means that the proposed method can also work effectively with a whole document or a collection of documents for guiding the summary generation, as shown in the second and third sections of Table 4. Note that we do not report results for CTRLsum in these sections since it cannot readily work with this type of information. More specifically, CTRLsum receives the input by prepending the requested topic in the original document. Thus, it is limited to single words or phrases since it is not possible to fit an arbitrary textual context into the input of the model. In Table 4, we notice that $BART_{tag}$ achieves a high cosine similarity (0.25) when more information is available (collection of documents) compared to a single document (0.13) or a single topic (0.19), while the REL metric remains high for both cases (0.86). Both the proposed models ($BART_{tag}$ and $PEGASUS_{tag}$) achieve a high REL score, but BART seems to outperform $PEGASUS_{tag}$ in terms of sim_{sum} and cos metrics.

The same conclusions can be drawn when evaluating different datasets, as shown in Table 6 and 7, where the results on the MultiNews and XSum datasets are reported, respectively.

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 8. Experimental results on the MacDoc dataset. F-1 scores for ROUGE-1 (R1), ROUGE-2(R2), and ROUGE-L (RL) are reported.

	R1	R2	RL	BertScore	REL	cos
BART	30.36	10.49	20.41	87.13	–	–
PEGASUS	27.51	9.10	19.10	86.29	–	–
GPT-3.5	26.17	8.45	16.80	87.00	0.77	0.42
GPT-4	26.93	8.55	16.86	87.00	0.76	0.46
Claude	25.42	7.77	16.03	85.60	0.74	0.52
LLaMA	25.68	8.32	16.56	85.78	0.74	0.44
Mistral	27.09	8.68	17.18	86.54	0.77	0.39
CTRLSum	25.75	9.77	19.64	87.57	0.82	0.41
$BART_{tag}$ (Ours)	29.84	10.50	20.79	86.98	0.85	0.34

More specifically, for the MultiNews dataset, we observe that the $BART_{tag}$ achieves again the higher cosine similarity (0.35) in the collection of documents setting compared to the single document (0.22) and single topic (0.24) settings. In addition, for the XSum dataset, the higher cosine similarity (0.30) is achieved with the $BART_{tag}$ in the collection of documents setting.

The higher variance that is observed between the different summaries for both the XSum and MultiNews datasets compared to the CNN/DailyMail dataset (average similarity ~ 0.82 compared to ~ 0.92) confirms the effectiveness of the proposed method to generate diverse summaries according to the different topics.

5.2.3 Topic-controllable summarization

We evaluate the proposed method ($BART_{tag}$) on two topic-controllable summarization datasets, MacDoc (Zhang *et al.* 2023) and NEWTS (Bahrainian *et al.* 2022), and compare its performance with state-of-the-art models. Unlike the previous datasets (CNN/DailyMail, XSum, and MultiNews), where ground truth summaries were used to simulate context, both MacDoc and NEWTS contain human-annotated summaries explicitly annotated for different topics. In this experiment, ROUGE scores can serve as a strong indicator of the topical focus of the summary since they can measure the quality of the summaries in relation to the topic-oriented reference summaries. Therefore, higher ROUGE scores indicate better performance in generating accurate and relevant summaries that align with the requested topic.

The results on the MacDoc dataset are presented in Table 8. We report the performance of both topic-controllable methods and baseline models (BART and PEGASUS) that do not incorporate topic control. It is important to note that the baseline PEGASUS and BART models, without topic control, generate the same summary for each topic, as they are not designed to adapt to different topics within the same document. Overall, we observe that the proposed $BART_{tag}$ model outperforms all other methods in terms of ROUGE scores (R1: 29.84, R2: 10.50, RL: 20.79), demonstrating its effectiveness in generating summaries that not only align with the requested topic but also maintain high overall quality. This indicates that the proposed method can successfully shift the topic of the generated summary. At the same time, the proposed method achieves the highest REL score (0.85), showing that the summaries are trustworthy and accurately reflect the input document's content.

Although the cosine similarity with the topic for $BART_{tag}$ (0.34) is lower compared to other models, such as Claude (0.52), this is paired with a significantly higher REL metric and ROUGE

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 9. Experimental results on the NEWTS dataset. F-1 scores for ROUGE-1 (R1), ROUGE-2(R2), and ROUGE-L (RL) are reported.

	R1	R2	RL	BertScore	REL	cos
BART	32.89	11.26	21.27	86.33	–	–
PEGASUS	30.93	10.39	20.22	85.48	–	–
GPT-3.5	33.11	10.69	20.46	86.90	0.77	0.23
GPT-4	32.53	10.31	19.94	86.58	0.71	0.26
Claude	32.55	9.92	19.12	85.38	0.73	0.29
LLaMA	30.20	9.20	18.32	84.56	0.58	0.40
Mistral	33.55	10.75	20.42	86.41	0.76	0.22
CTRLSum	19.49	6.70	14.86	86.15	0.87	0.18
$BART_{tag}$ (Ours)	33.27	11.73	21.87	85.56	0.87	0.20

scores. This suggests that while the proposed method may not align as closely with the topic in terms of cosine similarity, it produces more reliable and accurate summaries, as shown by the higher ROUGE and REL scores. Additionally, we observe that Claude's higher cosine similarity (0.52) is accompanied by a lower reliability score (0.74 compared to 0.85) and lower overall ROUGE scores (25.42 compared to 29.84 R1). This indicates that although Claude's summaries may be semantically closer to the requested topic, they are less aligned with the target summaries. Overall, there appears to be a correlation between ROUGE scores and the reliability metric, suggesting that the proposed REL metric is a trustworthy indicator of the topical focus and accuracy of the summaries.

Similar conclusions can be drawn from the results on the NEWTS dataset (see Table 9). The proposed $BART_{tag}$ model consistently achieves the highest ROUGE scores across all the models. While Mistral achieves a similar ROUGE-1 score to $BART_{tag}$ (33.55 vs. 33.27), our model outperforms Mistral in terms of ROUGE-2 and ROUGE-L scores (11.73 R2 and 21.87 RL), indicating a higher overall quality of the summaries. Moreover, $BART_{tag}$ also achieves the highest REL score (0.87), confirming the trustworthiness and accuracy of the generated summaries. In contrast, while CTRLSum achieves the same high REL score (0.87), it significantly underperforms in terms of ROUGE-1 scores (19.49 compared to 33.27 R1).

Finally, we observe that models like LLaMA, which show a higher cosine similarity with the topic (0.40), demonstrate low REL and ROUGE scores. This pattern suggests that while these models generate summaries that appear closely aligned with the requested topic, those might not be reliable or well-aligned with the target summaries.

5.2.4 Query-focused summarization

We also compare the proposed method with a query-focused summarization model fine-tuned on the Debatepedia dataset, as shown in Table 10. Similar performance is observed for both models ($BART_{tag}$ and BART-FT), with the latter slightly outperforming the $BART_{tag}$ in terms of ROUGE and BERTScore. More specifically, both $BART_{tag}$ and BART-FT achieve ~ 66 ROUGE-1 score, with the BART-FT slightly outperforming $BART_{tag}$ in ROUGE-2 (54.71 compared to 53.79) and ROUGE-L (65.39 compared to 64.68). Also, both REL and cosine similarity metrics yield the same results, indicating a close distance between the generated summary and the given query. More specifically, both models achieve a ~ 0.63 REL score and 0.51 cosine similarity. These results demonstrate that the proposed method is more effective when more information is available, for

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Table 10. Experimental results on the Debatepedia dataset with the given query used as the available context. F-1 scores for ROUGE-1 (R1), ROUGE-2(R2), and ROUGE-L (RL) are reported.

	R1	R2	RL	BertScore	REL	cos
BART-FT	66.84	54.71	65.39	93.59	0.63	0.51
$BART_{tag}$	66.05	53.79	64.68	93.50	0.62	0.51

Table 11. Correlation Between REL Metric and Human Ratings.

Metric	Correlation	p-value
Pearson	0.830	1.15×10^{-14}
Spearman	0.809	1.36×10^{-13}

example when a collection of documents is available for extracting the context representation. However, it is worth noting that query-focused summarization is a slightly different task from topic-controllable summarization. While queries can represent topics in some cases, there are scenarios, such as dealing with an arbitrary textual context, that is a collection of articles from a user's history, where the topic cannot be easily expressed as a query. In these situations, the challenge is to control the summarization output according to a broader and more complex input; thus, query-focused summarization approaches cannot readily be applied.

5.3 Human evaluation

To validate the proposed REL metric, we conducted a human evaluation study. In this study, each participant was presented with the original document and its corresponding summary and asked to rate the summary's reliability on a scale from 1 (not reliable) to 10 (highly reliable). Each participant evaluated three different summaries, resulting in a total of 33 summaries rated by 29 undergraduate students. Each summary was annotated by an average of 1.64 participants. To ensure the reliability of the data, we excluded annotators where ratings differed by more than 4 points from the mean. Inter-annotator agreement, measured with Krippendorff's alpha coefficient (Krippendorff, 2018) with ordinal weights, was 0.852, indicating a high level of consistency among annotators.

The primary goal of this evaluation was to determine how well the REL metric aligns with human judgments. We measured the correlation between the evaluation results and the REL scores using both Spearman's and Pearson's correlation coefficients. The results reveal a strong correlation with Spearman's correlation coefficient of 0.83 and Pearson's correlation coefficient of 0.809, as shown in Table 11. The correlation coefficients being very close to 1 support the REL metric's effectiveness in capturing the reliability of summaries as perceived by human annotators. In addition, both very low p-values ($p \ll 0.01$) indicate a statistically significant correlation between the REL metric and human annotations.

6. Conclusions and future work

In this work, we proposed a controllable summarization method for guiding the summary generation towards arbitrary textual context from a short text, like a topic or an entity, to a document or a collection of documents. The proposed method works by first extracting a BERT-based representation of the given context that is then used to tag the most representative words of the

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

input document. The main advantage of our method is that it can exploit all the different types of textual information beyond a short text, such as a document or a collection of documents, in order to direct the focus of the summary generation.

In addition, our findings revealed that existing controllable summarization methods are prone to generating artificial content in order to ensure the presence of the requested topic in the generated summary. To detect this behavior, we proposed an appropriate evaluation metric to measure the reliability of the topic-oriented sentences of the summary with respect to the input document. We also conducted a human evaluation study to confirm the validity of the proposed metric. The experimental results demonstrated that our method can effectively shift the generation towards the given context under different zero-shot scenarios while surpassing state-of-the-art LLMs in preserving the quality and reliability of the generated summaries.

Even though this work is focused on news summarization, it can also be easily applied in other domains. An interesting future research direction would be to employ the proposed method for scientific article summarization to obtain personalized summaries based on the related papers that a researcher is interested in. In addition, variations of the proposed metric could be explored by incorporating multiple sentences or using an average to compute the relevance of the summary.

Acknowledgements. AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.

Competing interests. The authors report no conflict of interest.

References

Ailem M., **Zhang B.** and **Sha F.** (2019). Topic augmented generator for abstractive summarization, arXiv preprint arXiv: 1908.07026.

Anthropic (2024). Meet claude. Accessed on 29.08.2024. *Available at:* https://www.anthropic.com/claude.

Ao X., **Luo L.**, **Wang X.**, **Yang Z.**, **Chen J.-H.**, **Qiao Y.**, **He Q.** and **Xie X.** (2023). Put your voice on stage: Personalized headline generation for news articles. *ACM Transactions on Knowledge Discovery from Data* **18**(3), 1–20.

Bahrainian S. A., **Feucht S.** and **Eickhoff C.** (2022). NEWTS: A corpus for news topic-focused summarization. In **Muresan S.**, **Nakov P.** and **Villavicencio A.** (eds), *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, pp. 493–503.

Bahrainian S. A., **Zerveas G.**, **Crestani F.** and **Eickhoff C.** (2021). Cats: Customizable abstractive topic-based summarization. *ACM Transactions on Information Systems* **40**(1), 1–24.

Bian J., **Lin B.**, **Zhang K.**, **Yan Z.**, **Tang H.** and **Zhang Y.** (2019). Controllable length control neural encoder-decoder via reinforcement learning, arXiv preprint arXiv: 1909.09492

Celikyilmaz A., **Bosselut A.**, **He X.** and **Choi Y.** (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*, Vol. **1** (Long Papers), pp. 1662–1675.

Chan H. P., **Wang L.** and **King I.** (2021). Controllable summarization with constrained Markov decision process. *Transactions of the Association for Computational Linguistics* **9**, 1213–1232.

Daumé H. III and **Marcu D.** (2006). Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia: Association for Computational Linguistics, pp. 305–312.

Díaz A. and **Gervás P.** (2007). User-model based personalized summarization. *Information Processing & Management* **43**(6), 1715–1734.

Dong L., **Yang N.**, **Wang W.**, **Wei F.**, **Liu X.**, **Wang Y.**, **Gao J.**, **Zhou M.** and **Hon H.-W.** (2019). Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pp. 13042–13054.

Dou Z.-Y., **Liu P.**, **Hayashi H.**, **Jiang Z.** and **Neubig G.** (2021). GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4830–4842, Online. Association for Computational Linguistics.

Dubey A., **Jauhri A.**, **Pandey A.**, **Kadian A.**, **Al-Dahle A.**, **Letman A.**, **Mathur A.**, **Schelten A.**, **Yang A.**, **Fan A.** and others (2024). The LLaMA 3 herd of models. arXiv preprint arXiv: 2407.21783

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Dubey N., **Setia S.**, **Verma A. A.** and **Iyengar S.** (2020). Wikigaze: Gaze-based personalized summarization of wikipedia reading session. In *Proceedings of the 3rd Workshop on Human Factors in Hypertext*, pp. 1–9.

Fabbri A., **Li I.**, **She T.**, **Li S.** and **Radev D.** (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, pp. 1074–1084.

Fan A., **Grangier D.** and **Auli M.** (2018a). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 45–54.

Fan A., **Grangier D.** and **Auli M.** (2018b). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 45–54.

Fang C., **Mu D.**, **Deng Z.** and **Wu Z.** (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications* **72**, 189–195.

Feigenblat G., **Roitman H.**, **Boni O.** and **Konopnicki D.** (2017). Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*, New York, NY, USA: Association for Computing Machinery, 961–964.

Fisher S. and **Roark B.** (2006). Query-focused summarization by supervised sentence ranking and skewed word distributions. In *Proceedings of the Document Understanding Conference, DUC-2006*, New York, USA: Citeseer.

Frermann L. and **Klementiev A.** (2019). Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, pp. 6263–6273.

He J., **Kryscinski W.**, **McCann B.**, **Rajani N.** and **Xiong C.** (2022). CTRLsum: Towards Generic Controllable Text Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5879–5915.

Hermann K. M., **Kocisky T.**, **Grefenstette E.**, **Espeholt L.**, **Kay W.**, **Suleyman M.** and **Blunsom P.** (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, Vol. 28, pp. 1693–1701.

Jiang A. Q., **Sablayrolles A.**, **Mensch A.**, **Bamford C.**, **Chaplot D. S.**, **Casas D.d l**, **Bressand F.**, **Lengyel G.**, **Lample G.** and **Saulnier L.** (2023). and others Mistral 7b. arXiv preprint arXiv: 2310.06825.

Keskar N. S., **McCann B.**, **Varshney L. R.**, **Xiong C.** and **Socher R.** (2019). Ctrl: conditional transformer language model for controllable generation, arXiv preprint arXiv: 1909.05858.

Kikuchi Y., **Neubig G.**, **Sasano R.**, **Takamura H.** and **Okumura M.** (2016). Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1328–1338.

Krippendorff K. (2018). *Content Analysis: An Introduction to Its Methodology*. Sage publications.

Krishna K. and **Srinivasan B. V.** (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*, Vol. **1** (Long Papers), pp. 1697–1705.

Lewis M., **Liu Y.**, **Goyal N.**, **Ghazvininejad M.**, **Mohamed A.**, **Levy O.**, **Stoyanov V.** and **Zettlemoyer L.** (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*, pp. 7871–7880.

Li J., **Li H.** and **Zong C.** (2019). Towards personalized review summarization via user-aware sequence network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. **33**, pp. 6690–6697.

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

Liu Y. and **Lapata M.** (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*.

Liu Y., **Luo Z.** and **Zhu K.** (2018). Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, pp. 4110–4119.

Liu Y., **Luo Z.** and **Zhu K. Q.** (2020). 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*.

Lu R., **Chen B.**, **Guo D.**, **Wang D.** and **Zhou M.** (2024). Hierarchical topic-aware contextualized transformers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **32**, 841–852.

Mao X., **Yang H.**, **Huang S.**, **Liu Y.** and **Li R.** (2019). Extractive summarization using supervised and unsupervised learning. *Expert Systems with Applications* **133**, 173–181.

Móro R. and **Bielikova M.** (2012). Personalized text summarization based on important terms identification. In *2012 23rd International Workshop on Database and Expert Systems Applications*, IEEE, pp. 131–135.

Nallapati R., **Zhou B.**, **dos Santos C.**, **Gulcehre C.** and **Xiang B.** (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 2016 SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290.

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Narayan S., **Cohen S. B.** and **Lapata M.** (2020). Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Nema P., **Khapra M. M.**, **Laha A.** and **Ravindran B.** (2017). Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, pp. 1063–1072.

OpenAI (2022a). Gpt-3.5 turbo. Accessed on 29.08.2024. *Available at:* https://platform.openai.com/docs/models/gpt-3-5-turbo.

OpenAI (2022b). OpenAI: Introducing chatGPT. Accessed on 29.08.2024. *Available at:* https://openai.com/blog/chatgpt.

OpenAI (2023). Gpt-4 turbo and gpt-4. Accessed on 29.08.2024. *Available at:* https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.

Passali T. and **Tsoumakas G.** (2024). Topic-aware evaluation and transformer methods for topic-controllable summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16282–16292.

Reimers N. and **Gurevych I.** (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992.

Saito I., **Nishida K.**, **Nishida K.**, **Otsuka A.**, **Asano H.**, **Tomita J.**, **Shindo H.** and **Matsumoto Y.** (2020). Length-controllable abstractive summarization by guiding with summary prototype, arXiv preprint arXiv: 2001.07331.

See A., **Liu P. J.** and **Manning C. D.** (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics*, pp. 1073–1083.

Song S., **Huang H.** and **Ruan T.** (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications* **78**, 857–875.

Su D., **Yu T.** and **Fung P.** (2021). Improve query focused abstractive summarization by incorporating answer relevance. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online. Association for Computational Linguistics, pp. 3124–3131,

Takase S. and **Okazaki N.** (2019). Positional encoding to control output sequence length. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota: Association for Computational Linguistics, Vol. **1** (Long and Short Papers), pp. 3999–4004.

Touvron H., **Lavril T.**, **Izacard G.**, **Martinet X.**, **Lachaux M.-A.**, **Lacroix T.**, **Rozière B.**, **Goyal N.**, **Hambro E.**, **Azhar F.** and et al. (2023a). LLaMA: Open and efficient foundation language models, arXiv preprint arXiv: 2302.13971.

Touvron H., **Martin L.**, **Stone K.**, **Albert P.**, **Almahairi A.**, **Babaei Y.**, **Bashlykov N.**, **Batra S.**, **Bhargava P.**, **Bhosale S.** and et al. (2023b). LLaMA 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv: 2307.09288.

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

Wang Z., **Duan Z.**, **Zhang H.**, **Wang C.**, **Tian L.**, **Chen B.** and **Zhou M.** (2020). Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 485–497.

Wolf T., **Debut L.**, **Sanh V.**, **Chaumond J.**, **Delangue C.**, **Moi A.**, **Cistac P.**, **Rault T.**, **Louf R.**, **Funtowicz M.**, **Davison J.**, **Shleifer S.**, **von Platen P.**, **Ma C.**, **Jernite Y.**, **Plu J.**, **Xu C.**, **Le Scao T.**, **Gugger S.**, **Drame M.**, **Lhoest Q.** and **Rush A.** (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45.

Xu Y. and **Lapata M.** (2020). Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 3632–3645.

Xu Y. and **Lapata M.** (2021). Generating query focused summaries from query-free resources. 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)*, Online. Association for Computational Linguistics, pp. 6096–6109.

Yan R., **Nie J.-Y.** and **Li X.** (2011). Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 1342–1351.

Yang G., **Wen D.**, **Chen N.-S.**, **Sutinen E.** and others (2012). Personalized text content summarizer for mobile learning: An automatic text summarization system with relevance based language model. In *2012 IEEE Fourth International Conference on Technology for Education*, IEEE, pp. 90–97.

Yang X., **Li Y.**, **Zhang X.**, **Chen H.** and **Cheng W.** (2023). Exploring the limits of chatGPT for query or aspect-based text summarization, arXiv preprint arXiv: 2302.08081.

https://doi.org/10.1017/nlp.2025.10009 Published online by Cambridge University Press

Zhang H., **Chen Z.**, **Ma W.-y.** and **Cai Q.** (2003). A study for document summarization based on personal annotation. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pp. 41–48.

Zhang J., **Zhao Y.**, **Saleh M.** and **Liu P.** (2020a). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119, pp. 11328–11339.

Zhang T., **Kishore V.**, **Wu F.**, **Weinberger K. Q.** and **Artzi Y.** (2020b). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia: OpenReview.net. April 26-30, 2020

Zhang Y., **Liu Y.**, **Yang Z.**, **Fang Y.**, **Chen Y.**, **Radev D.**, **Zhu C.**, **Zeng M.** and **Zhang R.** (2023). MACSum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics* **11**, 787–803.

Cite this article: Passali T. and Tsoumakas G. Controllable abstractive summarization with arbitrary textual context. *Natural Language Processing* https://doi.org/10.1017/nlp.2025.10009