Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects - A Survey

Ashok Urlana[1] **Pruthwik Mishra**[2] **Tathagato Roy**[2] **Rahul Mishra**[2]
TCS Research, Hyderabad, India[1] IIIT Hyderabad[2]
ashok.urlana@tcs.com, pruthwik.mishra@research.iiit.ac.in,
tathagato.roy@research.iiit.ac.in, rahul.mishra@iiit.ac.in

Abstract

Generic text summarization approaches often fail to address the specific intent and needs of individual users. Recently, scholarly attention has turned to the development of summarization methods that are more closely tailored and controlled to align with specific objectives and user needs. Despite a growing corpus of controllable summarization research, there is no comprehensive survey available that thoroughly explores the diverse controllable attributes employed in this context, delves into the associated challenges, and investigates the existing solutions. In this survey, we formalize the Controllable Text Summarization (CTS) task, categorize controllable attributes according to their shared characteristics and objectives, and present a thorough examination of existing datasets and methods within each category. Moreover, based on our findings, we uncover limitations and research gaps, while also exploring potential solutions and future directions for CTS. We release our detailed analysis of CTS papers at https://github.com/ashokurlana/controllable_text_summarization_survey.

1 Introduction

Despite the significant advancements in automatic text summarization, its one-size-fits-all approach falls short in meeting the varied needs of different segments of users and application scenarios. For example, generic automatic summarization may struggle to produce easily understandable summaries of scientific documents for non-expert users or create extremely brief summaries of news stories for online feeds. Lately, a myriad of works have emerged aimed at generating more controlled (Fan et al., 2018a; Maddela et al., 2022; He et al., 2022; Zhang et al., 2023b; Pagnoni et al., 2023) and tailored text summaries that meet a wide range of user needs.

CTS task is centered around creating summaries of source documents that adhere to specific criteria.

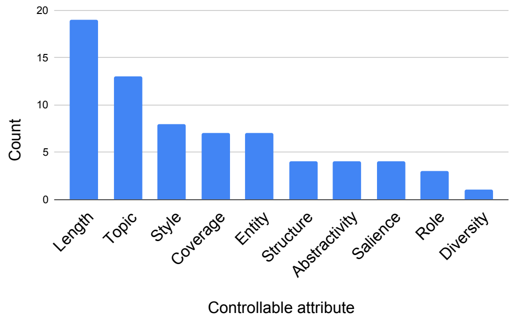

Figure 1: Number of controllable text summarization publications for various attributes.

These criteria are managed through various controllable attributes (CA) or aspects like summary length (Kwon et al., 2023), writing style (Goyal et al., 2022), coverage of key information (Li et al., 2018; Jin et al., 2020b), content diversity (Narayan et al., 2022), and more. These criteria vary depending on the task, user needs, and specific application context. For example, length-controlled summaries (Hitomi et al., 2019) are particularly useful in situations where brevity is crucial, like in social media posts, headlines, and abstracts. In areas such as marketing, academic writing, or professional communication, a style-controlled summary (Chawla et al., 2019) is essential to ensure that the information aligns with the intended tone and messaging strategy. Similarly, topic-controlled summaries (Bahrainian et al., 2021) are commonly used in research papers, reports, and content curation, providing an emphasis on a specific topic to enhance clarity and coherence in the presented information.

There is an uneven distribution of attention within the research community towards various CAs as depicted in Fig 1. The majority of CTS works concentrate on managing length, topic, and style. This could be attributed to two main factors. First, it is comparatively simpler to develop datasets for evaluating length, topic, and style compared to aspects like structure and diversity. Second, there is a plethora of application scenarios for

Source: (CNN)Novak **Djokovic** extended his current winning streak to 17 matches after beating Thomas Berdych 7-5, 4-6, 6-3 in the rain-interrupted **final** of the Monte Carlo Masters....After winning the Australian Open back in January, Djokovic has followed up with Masters' victories at Indian Wells and Miami. He then beat Rafa Nadal, arguably one of the greatest players on clay of all time...
Length:Long, **Coverage**:High, **Topic**:Djockovic,Final **Summary:** Djokovic wins 7-5, 4-6, 6-3 after a tight match with Berdych in the Monte Carlo Masters final. Djokovic also followed up with Masters' victories at Indian Wells and Miami.
Length:Normal, **Topic**:Djockovic, **Coverage**:Normal, **Summary:** It's been a sensational year for Djokovic after beating Berdych in the finals and also winning against clay expert Nadal.
Length:Short, **Coverage**:Normal, **Topic**: No Control **Summary:** Djokovic wins Monte Carlo Masters after beating Berdych 7-5, 4-6, 6-3 in the finals.

Table 1: Summaries obtained by varying Controllable Attributes from MACSUM (Zhang et al., 2023b)

length or topic-oriented summaries, such as generating concise news feeds or focused legal reports.

In this survey, we collect and analyze 61 research papers pertaining to various possible CAs. The filtration criteria for the selection of papers are described in Appenedix A. Subsequently, we classify these CAs into 10 categories, grouping similar ones based on shared characteristics and objectives. Moreover, we delve into the existing datasets, evaluating their creation methods and appropriateness for the respective task in each CA category. Furthermore, we scrutinize the current CTS methodologies for each CA category, drawing comparisons between their overarching frameworks and discussing relevant limitations. Subsequently, we discuss in detail the generic and specific evaluation strategies for CAs utilized by various works. Finally, we attempt to critique the current approaches and unravel potential future research trajectories. To the best of our knowledge, it is the first comprehensive survey on CTS.

2 Task Formulation

This section introduces the Controllable Text Summarization (CTS) task by outlining its definition and offering a categorized breakdown of different controllable attributes along with concise descriptions for each. Given a set of source documents $D = \{d_1, d_2, \ldots, d_k\}$. Each document, d_i, consists of a sequence of n tokens: $\{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$. S_i is the target summary of document d_i, which comprises of a sequence of m tokens: $\{s_{i,1}, s_{i,2}, \ldots, s_{i,m}\}$, where $m \ll n$. The user wants to control a set of controllable attributes C. The task can be framed as a conditional generative problem: $P(S|D, C) = \prod_i^k P(S_i|d_i, C)$

Attribute	Definition
Length	Controlling the length of the summary
Style	Controlling the readability levels, politeness, humor, and emotion
Coverage	Controlling the salient information in summary
Entity	Summary specific to pre-defined entities
Structure	Create summaries with predefined structure or order
Abstractivity	Controlling the novelty in sentence formation
Salience	Adjusting the presence of prominent information
Role	Providing role-specific summaries
Diversity	Generating semantically diverse summaries
Topic	Controlling topic-focused summary generation

Table 2: Controllable attributes definitions.

2.1 Controllable Attributes

A Controllable Attribute or Aspect (CA) refers to a user or application-driven trait of a summary designed to meet specific criteria or conditions, such as Length, Style, Role, etc. In the literature, it is evident that various authors use different terms to describe the same CAs, which exhibit similar characteristics and objectives (such as "Salience: Key information", and "Coverage: Granularity"). Additionally, numerous attributes can be encapsulated by a representative class; for instance, "Style" may serve as a class encompassing Tone, Readability, Humor, Romance, and similar aspects, facilitating their classification within the same category as shown in Table 3. Based on these observations, we group the CAs into 10 categories as listed in Table 2.

Class	**Attribute**
Style	Tone, Readability, Humor, Romance, Clickbait
Coverage	Coverage, Granularity
Entity	Entity, Keyword
Topic	Topic, Aspect, Decision of interest, Opinion based on user interest
Abstractivity	Abstractiveness, Extractiveness, Novelty
Salience	Salience, Key information

Table 3: Merging of attributes into representative classes.

3 Related Surveys

In the literature, a multitude of surveys center around conventional text summarization methods (El-Kassas et al., 2021; Nazari and Mahdavi, 2019; Allahyari et al., 2017; Gambhir and Gupta, 2017), including task-specific surveys such as multi-document summarization (Sekine and Nobata, 2003), cross-lingual summarization (Wang et al., 2022), and dialogue-based summarization (Tuggener et al., 2021). There are a few surveys that concentrate on text generation techniques (Zhang et al., 2023a; Prabhumoye et al., 2020) and the causal perspective (Wang et al., 2024; Hu and Li, 2021) on the same. On the contrary, this is the first survey, that focuses on controllable summarization by offering a thorough analysis of CTS methods, challenges, and prospects.

4 Datasets

This section provides a broad overview of the CTS datasets and corresponding creation/acquisition strategies. The CTS methods are evaluated in several ways: 1) by utilizing publicly available summarization datasets, 2) by datasets derived from generic datasets, and 3) by creating human-annotated datasets.

4.1 Generic Datasets

CTS research predominantly leverages widely used news summarization datasets. Notably, about 57% of CTS studies utilize either CNN-DailyMail (Nallapati et al., 2016) or DUC (Over and Yen, 2004; Dang, 2005). Other popular datasets, including Gigaword (Napoles et al., 2012), XSum (Narayan et al., 2018), NYTimes (Sandhaus, 2008), NEWSROOM (Narayan et al., 2018), and dialogue-based SAMSUM (Gliwa et al., 2019), along with opinion-based datasets (Angelidis and Lapata, 2018; Angelidis et al., 2021), are employed for controllable summarization. However, these generic datasets lack explicit annotations and nuances to evaluate the CA-specific summarization. CTS requires specialized datasets (as detailed in Table 4) to provide evaluation opportunities for specific aspects like length, topic, style, etc.

4.2 Derived Datasets

The derived datasets are obtained by applying the aspect-specific heuristics to the widely used generic datasets. In this section, we list out a few derived datasets and their creation strategies.

JAMUL. Hitomi et al. (2019) collect length-sensitive headlines for the Japanese language. Each article consists of three headlines with varying lengths of 10, 13, and 26 characters respectively. **TS and PLS.** In order to enhance the readability of biomedical documents, Luo et al. (2022) introduce two types of summaries. The *Technical Summary (TS)* is an abstract of a peer-reviewed bio-medical research paper and the *Plain Language Summary (PLS)* is the authors submitted summary as part of the journal submission process. **Wikiasp.** In order to construct the multi-domain aspect-based summarization corpus, Hayashi et al. (2021) utilize the Wikipedia articles from 20 domains. Further, the section titles and paragraph boundaries of each article are obtained as a proxy of aspect annotation. In another study, Ahuja et al. (2022) create the **ASPECTNEWS** dataset for aspect-oriented summarization. They achieve it by utilizing articles from the CNN/DailyMail dataset and identifying documents related to 'earthquakes' and 'fraud investigations' by using the universal sentence encoder (Cer et al., 2018). Further, Mukherjee et al. (2020) collect a CA-based opinion summarization dataset consisting of tourism reviews. These are obtained from the *TripAdvisor* website and identified the relevant aspects using the unsupervised attention-based aspect extraction technique (He et al., 2017).

4.3 Human annotated

This section provides the details of the human-annotated CTS datasets.

GranDUC. By re-annotating the DUC-2004 (Dang, 2005), Zhong et al. (2022) release a novel benchmark dataset for the granularity control. Annotators are instructed to create summaries of multiple documents with *coarse, medium,* and *fine* granularity levels. **Multi-LexSum.** Shen et al. (2022c) create a human-annotated corpus of 9,280 civil rights lawsuits and corresponding summaries with different degrees of granularity. The target summary length ranges from one-sentence to multi-paragraph level. **EntSUM.** (Maddela et al., 2022) is a human-annotated entity-specific controllable summarization dataset. It utilizes the articles from The New York Times Annotated Corpus (NYT) (Sandhaus, 2008) and includes annotated summaries for PERSON and ORGANISATION tags. The recent release of EntSUMV2 (Mehra et al., 2023) is the more abstractive version of EntSUM. **NEWTS.** Bahrainian et al. (2022) introduce the top-

Dataset	Controllable attribute(s)	Human-annotated	Size	Domain	Dataset URL
Multi-LexSum (Shen et al., 2022c)	Coverage	Yes	9280	Legal	https://tinyurl.com/22ksfase
GranDUC (Zhong et al., 2022)	Coverage	Yes	50	News	https://tinyurl.com/2x72ubrw
TS and PLS (Luo et al., 2022)	Style	No	28124	Biomedical	https://tinyurl.com/yck3v9px
MACSUM (Zhang et al., 2023b)	Length, Coverage, Topic	Yes	9686	News, meetings	https://tinyurl.com/3d2dsc7u
NEWTS (Bahrainian et al., 2022)	Topic	Yes	6000	News	https://tinyurl.com/36hzk3ew
WikiAsp (Hayashi et al., 2021)	Topic	No	320272	Encyclopedia	https://tinyurl.com/3u45hfbn
ASPECTNEWS (Ahuja et al., 2022)	Topic	No	2000	News	https://tinyurl.com/bdzxs8ej
Tourism ASPECTS (Mukherjee et al., 2020)	Topic	No	7000	Reviews	https://tinyurl.com/ypjhhrxv
EntSUM (Maddela et al., 2022)	Entity	Yes	2788	News	https://tinyurl.com/2pz9vzyw
JAMUL (Hitomi et al., 2019)	Length	No	1932398	News	https://tinyurl.com/3s3ecua9
CSDS (Lin et al., 2021)	Role	Yes	10700	Dialogues	https://tinyurl.com/adk7zc7u
MReD (Shen et al., 2022b)	Structure	Yes	7089	Meta reviews	https://tinyurl.com/4nn87fd6

Table 4: List of controllable summarization datasets.

ically focused summarization corpus by leveraging documents from CNN-DailyMail and employing crowd-sourcing to generate two distinct summaries with different thematic aspects for each document. **CSDS.** Lin et al. (2021) introduce the role-oriented Chinese Customer Service Dialogue Summarization (CSDS) dataset. It is meticulously annotated, segmenting the dialogues based on their topics and summarizing each segment as a QA pair. **MReD.** To tackle the task of structure-controllable summarization, Shen et al. (2022b) introduce the Meta-Review Dataset (MReD). It is created by gathering meta-reviews from the open review system and categorizing each sentence into one of nine pre-defined intent categories (abstract, strength, weakness, etc.,). **MACSUM.** Zhang et al. (2023b) develop a human-annotated corpus to control the mix of CAs (Topic, Speaker, Length, Extractiveness, and Specificity) together. MACSUM covers source articles from CNN/DailyMail and QMSUM (Zhong et al., 2021) datasets.

5 Approaches to Controlled Summarization

Various CAs have been investigated in controllable summary generation tasks, including style (politeness, humor, formality), content (length, entities, keywords), and structure. In this section, we describe various approaches to achieve CTS for the attributes mentioned in Table 2. Additionally, we list out the novel contributions and limitations for each paper in the Appendix C Table 9.

Length. Earlier methods lacked length control and only employed heuristics such as stopping the generation after a fixed number of tokens. To overcome this, four different approaches to integrate length as a learnable parameter are proposed.

Adding length in input: Fan et al. (2018b) propose a convolutional encoder-decoder-based summarization system, where it quantizes summary lengths into discrete bins of different size ranges. During training, the input data is prepended with the gold summary length represented by bin lengths. Due to a fixed number of length bins, the system ***fails*** to generate summaries of arbitrary lengths. CTRLSUM (He et al., 2022) presents a generic framework to generate controlled summaries using keywords specific to length. Instead of controlling a single attribute, Zhang et al. (2023b) allow different length attribute values (normal, short, long) to be used as inputs along with the source text for hard prompt tuning (Brown et al., 2020).

Adding length in encoder: Yu et al. (2021) propose a length context vector that is generated at each decoding step derived from the positional encodings. This vector is then concatenated with the decoder hidden state and encoder attention vectors. The ***limitation*** of the system is the generation of incomplete summaries for short desired lengths. Liu et al. (2022b) propose a length-aware attention model that adapts the source encodings based on the desired length by pretraining the model. Zhang et al. (2023b) add a hyperparameter for learning the prefix embeddings for different attributes at each layer of the encoder and decoder for soft prefix tuning (Li and Liang, 2021).

Adding length in decoder: Kikuchi et al. (2016) propose the first method to control length using a BiLSTM encoder-decoder architecture with attention (Luong et al., 2015) for sentence compression. In each step of the decoding process, an additional input for the remaining length is provided as an embedding. Instead of pre-defined length ranges, Liu et al. (2018) add a desired length parameter at the decoding step to each convolutional block of the initial layer of the convolutional encoder-decoder model. Févry and Phang (2018) design an unsupervised denoising auto-encoder for sentence compression, where the decoder has an additional input of the remaining summary length at each time step. While it produces grammatically correct summaries, but they are nonsensical or semantically different from the input. This leads to the generation of ***unfaithful*** summaries.

To handle the length constraint, Takase and Okazaki (2019) propose two modifications to the sinusoidal positional embeddings on the decoder side: length-difference positional encoding and length-ratio positional encoding. Sarkhel et al. (2020) present a multi-level summarizer that models a multi-headed attention mechanism using a series of interpretable semantic kernels to control lengths, reducing the trainable parameters significantly. The model does ***not encode*** the length attribute directly. Song et al. (2021) design a confidence-driven generator that is trained on a denoising objective with a decoder-only architecture, where the source and summary tokens are masked with position-aware beam search. Goyal et al. (2022) use a mixture-of-experts model with multiple transformer-based decoders for identifying different styles or features of summaries. Kwon et al. (2023) introduce the summary length prediction task on the encoder side and this predicted summary length is inserted with a length-fusion positional encoding layer.

Adding length in loss/reward function: Makino et al. (2019) propose a global minimum risk training optimization method under length constraint for the neural summarization tasks which is faster and generates five times fewer over-length summaries on an average than others. Chan et al. (2021) use an RL-based Constrained Markov Decision process with a mix of attributes. Hyun et al. (2022) devise an RL-based framework that incorporates both length and quality constraints in the reward function to generate multiple summaries of different lengths and according to the experimental results present in Hyun et al. (2022), the model is computationally ***expensive***.

Style. The generation of user-specific summaries has gained significant interest, but achieving distinct styles has posed an enduring challenge. These stylistic variations may encompass tone, readability control, or the modulation of user emotions. Style control aims to generate source-specific summaries (Fan et al., 2018a) by utilizing the convolutional encoder-decoder network.

Chawla et al. (2019) obtain formality-tailored summaries by utilizing the input-dependent reward function. The pointer-generator (See et al., 2017) network is used as the under-laying architecture and the loss function is modified with the addition of a formality-based-reward function. In another study, Jin et al. (2020a) attempt to control humor, romance, and clickbait in headlines using a multitask learning framework. By employing an inference style classifier, Cao and Wang (2021) adjust the decoder final states to obtain stylistic summaries. Moreover, they obtain lexical control by utilizing the word unit prediction that can directly constrain the output vocabulary. Similarly, Goyal et al. (2022) extend the decoder architecture to a mixture-of-experts version by using multiple decoders. The gating mechanism helps to obtain multiple summaries for a single source. However, the major limitation in this model is its ***manual*** gating mechanism. To control various fine-grained reading grade levels, Ribeiro et al. (2023) present three methods: instruction-prompting, reinforcement learning-based reward model, and look-ahead readability decoding approach.

Coverage. Managing the information granularity is essential to measure the semantic coverage between the source text and the summary. To regulate the granularity, Wu et al. (2021) introduce a two-stage approach, where the model incorporates a summary sketch, that encompasses user intentions and key phrases, serving as a form of weak supervision. They leverage a text-span-based conditional generation to govern the level of detail in the generated dialogue summaries. Zhong et al. (2022) propose a multi-granular event-aware summarization method composed of four stages: event identification, unsupervised event-based summarizer pretraining, event ranking, and summary generation by adding events as hints. Extraction of events from source text may ***lower*** the abstractiveness. Zhang et al. (2023b) use the hard and soft-prompting strategies to control the amount of extracted text from the source in the summary. Additionally, Huang et al. (2023) utilize the natural language inference models to improve the coverage.

Entity. Entity-centric summarization concentrates on producing a summary of a document that is specific to a given target entity (Hofmann-Coyle et al., 2022). Zheng et al. (2020) extract the named entities using a pre-trained BERT (Devlin et al., 2019) based model and feed both the article and the selected entities to a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) encoder-decoder model. In another study, Liu and Chen (2021) extract the entities (speakers and non-speaker entities) from a dialogue to form a planning sequence. The entities extracted are concatenated to the source dialogue for training the conditional BART-based model. This model introduces factual ***inconsistency*** due to paraphrasing from a personal perspective.

Maddela et al. (2022) extend the GSum (Dou et al., 2021) by feeding it either sentences or strings, which mention extracted entities as guidance. The model is an adapted version of BERTSum (Liu and Lapata, 2019), where only the sentences containing the entity string mention and its coreferent mentions are fed. Hofmann-Coyle et al. (2022) model entity-centric extractive summarization as a sentence selection task. Building upon BERTSum (Liu and Lapata, 2019), they use a BERT (Devlin et al., 2019) based encoders to represent the sentence and target entity pair and train with a contrastive loss objective to extract sentences most relevant to the target entities.

Structure. Generic datasets lack key elements for emphasizing specific aspects in the corresponding ground truth summaries. To address this limitation and emphasize summary structure, Shen et al. (2022b) achieve structure-controllable text generation by adding a control sequence at the beginning of the input text and treating summary generation as a standalone process. However, this approach has two main ***limitations***, 1) generated tokens are solely based on logits predictions without ensuring that the sequence satisfies the control signal, 2) Auto-regressive models face error propagation in generation due to self-attention, causing subsequent generations to deviate from the desired output. To overcome these challenges, the sentence beam-search (SentBS) (Shen et al., 2022a) approach produces multiple sentence options for each sentence and selects the best sentence based on both the control structure and the model's likelihood probabilities. In a related study, Zhong and Litman (2023) utilize predicted argument role information to control the structure in legal opinion documents. Additionally, in the work of Zhang et al. (2023b), the prompt of entity chains, representing an ordered sequence of entities, is used for pre-training and fine-tuning with a planning objective to control the summary structure.

Abstractivity. It measures the degree of textual novelty between the source text and summary. See et al. (2017) introduce a pointer-generator network to control the source copying via *pointing* and generate novel sentence formations by using *generator* mechanism. However, this scheme ***fails*** to generate higher abstraction levels. Kryściński et al. (2018) tackle this problem in two ways: 1) decompose the decoder into a contextual network to retrieve the relevant parts of the text and generate the summary by utilizing a pretrained model, 2) a mixed RL-based objective jointly optimizes the n-gram overlap with the ground truth summary. Similarly, Song et al. (2020) control the copying behavior by using a *mix-and-match* strategy to generate summaries with varying n-gram copy rates. Based on the *seen*, *unseen* words from the source text, the system controls the copying percentage by acting as a language modeling task. Moreover, methods such as ControlSum (Fan et al., 2018a) allow the users to explicitly specify the control attribute to facilitate better control. However, it does not provide any supervision on ***violating*** the controllability. To alleviate this issue, Chan et al. (2021) propose an RL-based framework on the constrained Markov decision process and introduced a reward to penalize the violation of attribute requirement.

Salience. This attribute captures the most important information in a document. In SummaRuNNer (Nallapati et al., 2017), salience is modeled as a feature in a classification objective. It uses GRU-based encoders and decoders to frame summarization as a text-to-binary sequence learning task at the sentence level (Bahdanau et al., 2014; Cho et al., 2014). A binary score is assigned to each sentence, indicating its membership in the summary. The system performs ***poorly*** on out-of-domain datasets. To retain key content from the source, Li et al. (2018) introduce a Key Information Guide Network, where keywords are identified by the TextRank algorithm with a modified attention mechanism that accommodates this key information as an additional input. However, it focuses mostly on informativeness ***ignoring*** coherence and readability features.

Deutsch and Roth (2023) model salience in terms of noun phrases using QA signals where the generation of the summary is conditioned on these identified phrases. This approach is ***not applicable*** to languages for which question generation and question answering models are not available. In long document CLS tasks, summarization systems often ***fail*** to respond to user queries. To resolve this issue, Pagnoni et al. (2023) propose a pre-training approach that involves two tasks of salient information identification from sentences having the highest self ROUGE score and a question generation system to generate questions whose answers are the salient sentences.

Role. Role-oriented dialogue summarization generates summaries for different roles/agents present in a dialogue (e.g. doctor and patient) (Liang et al., 2022). Lin et al. (2021) propose the CSDS dataset (see Section 4.3) and benchmark a variety of existing state-of-the-art summarization models for the task of generating agent and user surveys. They find that agent summaries generated by the existing methods ***lack*** key information, that needs to be extracted from dialogues of the other role. To bridge this gap, Lin et al. (2022) build a role-aware summarization model for two users (agent and user) present in the dataset. They use two separate decoders for generating the user and agent summaries by utilizing user and agent masks. A role attention mechanism is introduced to each decoder so that it can leverage the overall context by attending to the hidden states of the other role. Liang et al. (2022) use a role-aware centrality scoring model that computes role-aware centrality scores for each utterance, which measures the relevance between the utterance and the role prompts (signaling whether the summary is for the user or agent). This is then used to reweight the attention scores for each utterance, which is subsequently used by the decoder to generate the summary.

Diversity. Traditional decoding strategies, like beam search, excel at generating single summaries but often ***struggle*** to produce diverse ones. Techniques such as top-k and nucleus sampling are effective in generating diverse outputs but may sacrifice faithfulness. In response to these challenges, Narayan et al. (2022) introduce compositional sampling, a decoding method to obtain diverse summaries. This method initiates by planning a semantic composition (Narayan et al., 2021) of the target in the form of entity chains, and then leverages beam search to generate diverse summaries.

Topic. Long documents often cover multiple topics, and a generic summary might not fully encompass the diverse scope. Krishna and Srinivasan (2018) train a topic-conditioned pointer-generator network (See et al., 2017) by concatenating one hot encoding representation of the topic with the embedding of each token in the input document. However, news categories are used as the predefined topics, that ***limits*** the generalization to other tasks. To handle diverse topics, Tan et al. (2020) utilize external knowledge sources like Wikipedia and ConceptNet to create a weakly supervised summarization framework compatible with any encoder-decoder architecture. Suhara et al. (2020) propose an unsupervised method, where aspect-specific opinions are extracted from a set of reviews by a pre-trained opinion extractor, and the summary of the opinion is generated by a generator model trained to reconstruct the reviews from the opinions. Similarly, given a set of reviews for a product (e.g. Hotels), Amplayo et al. (2021) train a Multiple Instance Learning (MIL) model, to extract the predictions for aspect (like cleanliness) codes at the document, sentence, token level (Mukherjee et al., 2020). These predicted aspects transform the input such that relevant sentences and keywords along with aspect tokens are fed into the pre-trained T5 (Raffel et al., 2020) model.

Hsu and Tan (2021) introduce the task of generating decision-supportive summaries. The focus is on predicting future Yelp ratings from the set of reviews using a Longformer-based (Beltagy et al., 2020) regression model. They propose an iterative algorithm that selects the sentences of the summary from a set of representative sentences. Mukherjee et al. (2022) extend topic-focused summarization for multimodal documents by creating a joint image-text context vector.

6 Evaluation Strategies

This section catalogs and briefly describes the variety of automatic and human evaluation metrics that are being used to evaluate the summaries generated by the different methods studied in this paper.

6.1 Automatic Evaluation

The automatic evaluation metrics can be categorized based on how they are defined. We categorize the metrics into n-gram-based, language-model-based, and aspect-specific.

N-gram based evaluation metrics like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) are based on matching n-grams from candidate summaries to a set of reference summaries. ROUGE is the most widely used metric in CTS literature. **Language-model based** metrics are computed using Pre-trained Language Models (PLM) like BERT (Devlin et al., 2019) or BART (Lewis et al., 2019). One class of approach computes the distance between the PLM embeddings of the reference and the generated summary. Another way is based on computing the log probability of the generated text conditioned on input text as demonstrated in BARTScore (Yuan et al., 2021). **Summarization specific** metrics including ROUGE-WE (Ng and Abrecht, 2015), S^3 (Peyrard et al., 2017), Sentence Mover's Similarity (SMS) (Clark et al., 2019), SummQA (Scialom et al., 2019), BLANC (Vasilyev et al., 2020), and SUPERT (Gao et al., 2020), *(Lite)³Pyramid* (Zhang and Bansal, 2021) are prominent for controllable summary evaluation. **Aspect specific** metrics do not fall cleanly into either of the above-mentioned categories. These metrics focus on evaluating specific controllable aspects such as Flesh Reading Ease (Flesch, 1948), Gunning Fog Index, and Coleman Liau Index for readability, control correlation, and error rate (Zhang et al., 2023b) for topic, abstractivity and role attributes. Appendix B Table 7 describes more details about the automatic evaluation metrics.

6.2 Human Evaluation

Human evaluation is an indicator of the robustness and effectiveness of different summarization systems on specific aspects that cannot be directly captured by automatic evaluation metrics. These aspects include generic properties of a summary such as truthfulness (Song et al., 2020; Hyun et al., 2022), relevance (Goyal et al., 2022; He et al., 2022; Shen et al., 2022b), fluency (Narayan et al., 2022; Suhara et al., 2020), and readability (Cao and Wang, 2021; Kryściński et al., 2018) or specific properties such as completeness (Yu et al., 2021; Liu et al., 2022a) for length-controlled summaries, coverage (Mukherjee et al., 2020, 2022) for the entity, and topic-controlled summary generation. Broadly two kinds of scoring mechanisms are used for human evaluation: binary and rank-based. The rank-based scores usually range from 1 to 5. Despite these widely adapted mechanisms, human evaluation of summarization is challenging due to ambiguity and subjectivity. Aspects like coherence and fluency help mitigate ambiguity, but remain subjective to individual annotators. Accurately defining annotation descriptions is crucial, yet achieving a standardized approach across annotators remains difficult (Iskender et al., 2021; Ito et al., 2023). The details about different human evaluation metrics are detailed in Appendix B Table 8.

7 Challenges and Future Prospects

Generic vs specialized benchmarks. We observe that more than 75% of CTS works either utilize or alter the generic news summarization datasets to evaluate the controllable summarization. As shown in Table 2, out of the 10 categories, we could find CA-specific datasets for only seven categories. We envisage that conducting evaluations with specialized datasets that align closely with real-world application scenarios or user requirements will help better in assessing the practical utility, robustness, and performance of CTS. It is evident from our survey of CTS systems that evaluations are often confined to specific domains, like news, possibly due to the abundance of available datasets in that domain. However, this narrow focus limits the evaluation of the CTS model's robustness.

Standardization of metrics. The goal of the CTS task is to produce CA-specific summaries, warranting a metric tailored to capture the nuances of this particular attribute. We observe that comparing models for a specific CA-based CTS task is challenging due to the use of varying metrics, leading each study to redo evaluations for a fair comparison with prior work. Standardizing CA-specific evaluation metrics could offer a valuable solution.

Explainability. For effectively controlling user or application-specific attributes, it is imperative to leverage the understanding of the decision-making process within CTS systems. Also, this comprehension is essential for users or stakeholders, enabling them to discern how the system generates summaries from source text. This holds particular significance in applications where human decision-making or interpretation plays a pivotal role, such as in legal, medical, or financial domains. The existing CTS efforts lack proper emphasis on the explainability aspects, which can be readily addressed through the incorporation of suitable explanation methodologies (Abnar and Zuidema, 2020; Sundararajan et al., 2017; Lundberg and Lee, 2017).

Multi-lingual, multi-modal, and code-mixed CTS. The existing literature on CTS predominantly focuses on works in English, with only one study addressing the topic in a Japanese context. We could not find any studies and datasets related

	Controllable Attributes						
	Length	*Entity*	*Style*	*Abstractivity*	*Coverage*	*Saliency*	*Topic*
Fan et al. (2018a)	✓	✓	✓	✗	✗	✗	✗
Zhang et al. (2023b)	✓	✗	✗	✗	✓	✗	✓
Chan et al. (2021)	✓	✓	✗	✓	✗	✗	✗
See et al. (2017)	✗	✗	✗	✓	✓	✗	✗
Pagnoni et al. (2023)	✗	✓	✗	✗	✗	✓	✗
He et al. (2022)	✓	✓	✗	✗	✗	✗	✗
Nallapati et al. (2017)	✗	✗	✗	✓	✗	✓	✗

Table 5: Support for multiple controllable attributes across various models.

to multilingual and code-mixed CTS approaches. Moreover, the task of controllable summarization in multi-modal and multi-document settings remains largely unexplored, presenting unique challenges for models to address and offering avenues for intriguing research problems.

Multi-CA control. Even though, few of the works perform multi-attribute controllable summarization (Goyal et al., 2022; He et al., 2022; Zhang et al., 2023b), we observe that existing works predominantly investigate combinations of length and entity attributes (see Table 5). As a future research direction, it's essential to design models that consider other important combinations of control attributes, such as length, style, and saliency. Furthermore, creating standardized multi-CA benchmarks is crucial to facilitate the evaluations.

Reproducibility. In the detailed analysis outlined in Table 10, we note that 35% of research studies do not share their code publicly. Furthermore, 25% of the papers did not carry out any human evaluation, and among the remaining studies, 79% did not conduct Inter Annotator Agreement (IAA) assessments. The lack of reproducibility (Ito et al., 2023; Gao et al., 2023; Iskender et al., 2021) measures hinders the scientific community's ability to validate and build upon existing work. On the other hand, the human study component should be a must for a text summarization evaluation scheme, otherwise, we are potentially overlooking essential aspects of real-world applicability.

Standing on the shoulders of LLMs. The rise and success of large language models (LLMs) have opened up unparalleled possibilities for leveraging their capabilities across diverse stages of the Natural Language Processing (NLP) pipeline. In the context of CTS, LLMs can be fine-tuned to grasp context-specific nuances about CAs without the need for a dedicated training set. Additionally, when it comes to evaluating CTS models, LLMs can serve as effective substitutes for human experts or judges (similar to Liu et al. (2023)), offering an efficient method for assessing performance.

8 Conclusions

We present a comprehensive survey on controllable text summarization (CTS) by offering a detailed analysis, from formalizing various controllable attributes, classifying them based on shared characteristics, and delving into existing datasets, proposed models, associated limitations, and evaluation strategies. Moreover, we discuss the challenges and prospects, making it a helpful guide for researchers interested in CTS. We plan to keep the GitHub repository regularly updated with the latest CTS works.

9 Limitations

Although we attempt to conduct a rigorous analysis of existing literature on controllable summarization, some works might have been possibly left out due to variations in search keywords. Furthermore, due to limited space, our survey primarily concentrates solely on the high-level aspects of the approaches, omitting a very fine-grained experimental comparison. Finally, our exploration of multilingual works was limited as we encountered challenges in finding them, likely influenced by the relatively low attention from the research community. We aim to further investigate the potential reasons behind the challenges associated with multilingual CTS tasks.

10 Ethics statement

To uphold transparency and accountability, the papers utilized in this survey are detailed in Appendix D Table 10. We have provided a comprehensive set of papers, accompanied by our qualitative classification and annotations, enabling public scrutiny and examination. Moreover, to alleviate qualitative bias, each paper underwent review by at least three different individuals independently, aiming to minimize misclassification. We adhere to the same methodology to validate the presence of diverse observations in each paper. By incorporating these ethical considerations, we affirm our dedication to conducting research in an ethical and accountable manner.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: 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, Dublin, Ireland. Association for Computational Linguistics.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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 Transactions on Information Systems (TOIS)*, 40(1):1–24.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shuyang Cao and Lu Wang. 2021. Inference time style control for summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5942–5953, Online. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232.

Kushal Chawla, Balaji Vasan Srinivasan, and Niyati Chhaya. 2019. Generating formality-tuned summaries using input-dependent rewards. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 833–842, Hong Kong, China. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Daniel Deutsch and Dan Roth. 2023. Incorporating question answering-based signals into abstractive summarization via salient span selection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 575–588.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 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*, pages 4830–4842, Online. Association for Computational Linguistics.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

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

Angela Fan, David Grangier, and Michael Auli. 2018b. Controllable abstractive summarization. *ACL 2018*, page 45.

Thibault Févry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66.

Mingqi Gao, Jie Ruan, and Xiaojun Wan. 2023. A reproduction study of the human evaluation of role-oriented dialogue summarization models. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tanya Goyal, Nazneen Rajani, Wenhao Liu, and Wojciech Kryscinski. 2022. HydraSum: Disentangling style features in text summarization with multi-decoder models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 464–479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRL-sum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics.

Yuta Hitomi, Yuya Taguchi, Hideaki Tamori, Ko Kikuta, Jiro Nishitoba, Naoaki Okazaki, Kentaro Inui, and Manabu Okumura. 2019. A large-scale multi-length headline corpus for analyzing length-constrained headline generation model evaluation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 333–343, Tokyo, Japan. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ella Hofmann-Coyle, Mayank Kulkarni, Lingjue Xie, Mounica Maddela, and Daniel Preotiuc-Pietro. 2022. Extractive entity-centric summarization as sentence selection using bi-encoders. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 326–333, Online only. Association for Computational Linguistics.

Chao-Chun Hsu and Chenhao Tan. 2021. Decision-focused summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 117–132, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955.

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. 2023. SWING: Balancing coverage and faithfulness for dialogue summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia. Association for Computational Linguistics.

Dongmin Hyun, Xiting Wang, Chayoung Park, Xing Xie, and Hwanjo Yu. 2022. Generating multiple-length summaries via reinforcement learning for unsupervised sentence summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2939–2951, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, Online. Association for Computational Linguistics.

Takumi Ito, Qixiang Fang, Pablo Mosteiro, Albert Gatt, and Kees van Deemter. 2023. Challenges in reproducing human evaluation results for role-oriented dialogue summarization. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020a. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, Online. Association for Computational Linguistics.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020b. Semsum: Semantic dependency guided neural abstractive summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8026–8033.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338.

Kundan Krishna, Aniket Murhekar, Saumitra Sharma, and Balaji Vasan Srinivasan. 2018. Vocabulary tailored summary generation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 795–805.

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

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2023. Abstractive document summarization with summary-length prediction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 606–612.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 55–60, New Orleans, Louisiana. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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), pages 4582–4597, Online. Association for Computational Linguistics.

Xinnian Liang, Chao Bian, Shuangzhi Wu, and Zhoujun Li. 2022. Towards modeling role-aware centrality for dialogue summarization. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 43–50, Online only. Association for Computational Linguistics.

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

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Puyuan Liu, Xiang Zhang, and Lili Mou. 2022a. A character-level length-control algorithm for non-autoregressive sentence summarization. *Advances in Neural Information Processing Systems*, 35:29101–29112.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment, may 2023. *arXiv preprint arXiv:2303.16634*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2022b. Length control in abstractive summarization by pretraining information selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, Dublin, Ireland. Association for Computational Linguistics.

Yizhu Liu, Zhiyi Luo, and Kenny 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*, pages 4110–4119.

Zhengyuan Liu and Nancy F. Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Conference on Empirical Methods in Natural Language Processing*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

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 (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Takuya Makino, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. 2019. Global optimization under length constraint for neural text summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048, Florence, Italy. Association for Computational Linguistics.

Dhruv Mehra, Lingjue Xie, Ella Hofmann-Coyle, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2023. EntSUMv2: Dataset, models and evaluation for more abstractive entity-centric summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5538–5547, Singapore. Association for Computational Linguistics.

Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1825–1828.

Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. Topic-aware multimodal summarization. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 387–398, Online only. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany. Association for Computational Linguistics.

Courtney Napoles, Matthew R Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction (AKBC-WEKEX)*, pages 95–100.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. 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*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Narges Nazari and MA Mahdavi. 2019. A survey on automatic text summarization. *Journal of AI and Data Mining*, 7(1):121–135.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Paul Over and James Yen. 2004. An introduction to duc-2004. *National Institute of Standards and Technology*.

Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2023. Socratic pretraining: Question-driven pretraining for controllable summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12737–12755, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Leonardo FR Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. *arXiv preprint arXiv:2310.10623*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Ritesh Sarkhel, Moniba Keymanesh, Arnab Nandi, and Srinivasan Parthasarathy. 2020. Interpretable multi-headed attention for abstractive summarization at controllable lengths. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6871–6882, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

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

Satoshi Sekine and Chikashi Nobata. 2003. A survey for multi-document summarization. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 65–72.

Chenhui Shen, Liying Cheng, Lidong Bing, Yang You, and Luo Si. 2022a. SentBS: Sentence-level beam search for controllable summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10256–10265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022b. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022c. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.

Kaiqiang Song, Bingqing Wang, Zhe Feng, and Fei Liu. 2021. A new approach to overgenerating and scoring abstractive summaries. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1392–1404, Online. Association for Computational Linguistics.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8902–8909.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Sho Takase and Naoaki Okazaki. 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, Volume 1 (Long and Short Papers)*, pages 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics.

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Junli Wang, Chenyang Zhang, Dongyu Zhang, Haibo Tong, Chungang Yan, and Changjun Jiang. 2024. A recent survey on controllable text generation: a causal perspective. *Fundamental Research*.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. Association for Computational Linguistics.

Zhongyi Yu, Zhenghao Wu, Hao Zheng, Zhe XuanYuan, Jefferson Fong, and Weifeng Su. 2021. LenAtten: An effective length controlling unit for text summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 363–370, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023b. Macsum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

Changmeng Zheng, Yi Cai, Guanjie Zhang, and Qing Li. 2020. Controllable abstractive sentence summarization with guiding entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5668–5678, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ming Zhong, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. 2022. Unsupervised multi-granularity summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4980–4995, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Yang Zhong and Diane Litman. 2023. Strong–structure controllable legal opinion summary generation. *arXiv preprint arXiv:2309.17280*.

A Survey papers selection criteria

We used keywords such as "controllable summarization", "text summarization" and "text generation" for selecting the initial pool of 105 papers. We selected the majority of papers from the reputed databases including the ACL Anthology[1], ACM Digital library[2], Google Scholar[3], which are known for hosting peer-reviewed articles that meet high academic standards. Among these 105 papers, six papers are pertinent to CTS, albeit they have not undergone peer review. Additionally, 23 papers touch upon the summarization aspect to some extent, although they may not be directly aligned with controllable summarization. Furthermore, we have excluded 15 papers as they primarily discuss controllable text generation or focus on enhancing the summarization task without specifically controlling any CTS attributes. Post to applying the above three filters we are left with 61 peer-reviewed and relevant papers to CTS. We have listed the filtration details in Table 6).

Criteria	Number of papers
arXiv version	6
Not relevant	23
Enhancement	15
Relevant	**61**
Total	105

Table 6: Survey papers filtration criteria.

B Evaluation Approaches

We have listed the automatic and human evaluation methodologies along with their respective metric details in Table 7 and Table 8. The automatic evaluation metrics are categorized into three groups: embedding-based, n-gram-based, and miscellaneous. Additionally, we present a compilation of papers organized by aspects, each associated with the relevant metrics, along with concise descriptions. As for human evaluation, we specify the corresponding metrics and provide definitions based on the attributes under consideration.

C Model Descriptions

As outlined in Table 9, we augment novel contributions, utilized dataset, and the corresponding limitation for each paper, all aligned with the respective controllable attribute.

D Survey papers checklist explanation

To underscore the comprehensiveness of our survey, as mentioned in Table 10, we include 23 features for each paper. For easier understanding, we briefly describe each feature in the master table below.

- *Paper:* Citation of the paper.
- *Year:* Year of the publication.
- *Venue:* Paper publishing conference or journal.

[1] https://aclanthology.org/
[2] https://dl.acm.org/
[3] https://scholar.google.com/

Automatic Evaluation				
Type of metric	**Attribute**	**Papers**	**Metrics**	**Description**
Embedding-based (Language Model)	General	Lin et al. (2022), Liang et al. (2022)	MoverScore	Computed using pretrained language models, either by computing similarity scores between reference and generated text embeddings or through likelihood computation of the generated text.
		Song et al. (2020), Shen et al. (2022a), Cao and Wang (2021) , Deutsch and Roth (2023), Chan et al. (2021), Pagnoni et al. (2023), Lin et al. (2022), Liang et al. (2022), Narayan et al. (2022), Zhong and Litman (2023), Ribeiro et al. (2023), Shen et al. (2022c), Maddela et al. (2022), Lin et al. (2021)	BERTScore	
		Huang et al. (2023)	BartScore	
		Zheng et al. (2020)	Bert-Reo	
	Readability	Luo et al. (2022)	Masked Noun Phrase-based Text Complexity, Ranked NP Based Text Complexity, Masked Random Token-Based Text Complexity	
Ngram Based	General	Lin et al. (2022), (Liang et al., 2022), Jin et al. (2020a), Narayan et al. (2022)	BLEU	These metrics are based on matching ngram tokens between reference and generated summaries
		All except* Zhang et al. (2023b), Goldsack et al. (2023), Cao and Wang (2021), Hsu and Tan (2021), Hofmann-Coyle et al. (2022)	ROUGE	
		Jin et al. (2020a), Sarkhel et al. (2020)	METEOR	
		Jin et al. (2020b)	Word Mover's Distance	
Miscellaneous	Length	Goyal et al. (2022), Kwon et al. (2023), Liu et al. (2018), Chan et al. (2021)	Absolute Length, Compression Ratio, Length Variance, Var, Bin Percentage	Non-normative metrics proposed by authors to evaluate specific controlled aspect
	Entity	Chan et al. (2021)	QA-F1	
		Narayan et al. (2021)	Entity Planning, Entity Specificity	
	Topic, Speaker, Length, Extractiveness, Specificity	Zhang et al. (2023b)	Control Correlation, Control Error Rate	
	Abstractiveness, Degree of Specificity	Goyal et al. (2022)	Abstractiveness, Degree of specificity	
	Readability	Goyal et al. (2022), Cao and Wang (2021)	Dale-Chall	
		Ribeiro et al. (2023)	Flesch Reading Ease, Gunning Fog Index, Coleman Liau Index	

Table 7: Automatic evaluation metrics for controllable summarization, "General" refers to all controllable attributes.

- *Controllable attribute:* Controllable attribute(s) concentrated in the paper.
- *Controlling more than one aspect:* Whether the paper handles more than one controllable aspect or not?
- *Model type:* Type of the model used in the paper such as encoder-decoder, encoder, or decoder architecture.
- *Training strategy:* Training approaches employed to perform CTS task.
- *Approach:* Type of the training approach employed to perform CTS task.
- *Code access:* Whether the code is publicly accessible or not?
- *Code link:* Address of the public repository.
- *Dataset:* Dataset utilized in the paper.
- *Source:* Source of the dataset used in the paper.
- *Nature of the data:* Dataset creation/acquisition strategy.
- *Data release:* Public availability of the dataset.
- *Domain:* The corresponding domain of the dataset.
- *Data link:* Public repository link to the dataset.
- *Metric name:* Name of the metric used in the paper.
- *Proposed new metric:* Names of the proposed new automatic evaluation metrics.
- *Human evaluation:* Human evaluation performed or not?
- *Metric names:* Name of the metrics used to perform human evaluation.
- *IAA:* Whether Inter Annotator Agreement assessment performed or not?
- *Limitation:* Any limitations of the paper mentioned or not?
- *Reproducibility:* Rate the reproducibility of the paper.

Human Evaluation			
Aspect(s)	**Papers**	**Metrics**	**Short description**
Abstractivity, Length, Style, Topic, Coverage Role, Entity	Sarkhel et al. (2020), Song et al. (2020) Liu et al. (2022b), Cao and Wang (2021) Amplayo et al. (2021), Wu et al. (2021), Jin et al. (2020b), Kwon et al. (2023), Lin et al. (2022), Liu and Chen (2021), Zheng et al. (2020), Bahrainian et al. (2021), Suhara et al. (2020), Lin et al. (2021)	Informativeness	Has the summary covered key content of the input text?
Structure, Length, Entity, Salience, Topic	Kwon et al. (2023)	Conciseness/Granularity	Is the key information presented in a crisp way?
Structure, Style, Topic Length, Entity, Abstractivity, Coverage, Role, Diversity	Goyal et al. (2022), Tan et al. (2020) Yu et al. (2021), Shen et al. (2022b), Song et al. (2020), Liu et al. (2022b), Févry and Phang (2018), Zheng et al. (2020), Shen et al. (2022a), Cao and Wang (2021), Amplayo et al. (2021), Hyun et al. (2022), Chan et al. (2021), Jin et al. (2020b), Liu et al. (2022a), Lin et al. (2022), Zhong et al. (2022), Jin et al. (2020a), Lin et al. (2021)	Fluency/Grammaticality	Are the sentences in a summary grammatically correct?
Role, Topic, Diversity	Narayan et al. (2022), Suhara et al. (2020), Lin et al. (2022), Lin et al. (2021), Mukherjee et al. (2020)	Non-redundancy/Diversity	Is the summary conveying diverse information?
Topic	Krishna et al. (2018)	Contextual Appropriateness	Is the substituted word more readable in the summary?
Style, Diversity	Goyal et al. (2022), Narayan et al. (2022), Chan et al. (2021), Jin et al. (2020b), Zhong et al. (2022), Huang et al. (2023)	Faithfulness/Factuality	Does the summary present factually correct content with respect to the source?
Style, Topic. Entity, Structure	Goyal et al. (2022), Zhong and Litman (2023)	Coherence	Is the summary composed of correlated sentences?
Style, Structure, Length, Entity, Abstractivity, Coverage, Topic, Diversity	Goyal et al. (2022), He et al. (2022), Shen et al. (2022b), Chan et al. (2021), Krishna and Srinivasan (2018), Shen et al. (2022a), Cao and Wang (2021), Zhong et al. (2022), Jin et al. (2020a), Kryściński et al. (2018), Luo et al. (2022), Huang et al. (2023)	Relevance	Does the summary contain relevant information regarding the user provided attribute (topic/entity)?
Abstractivity, Length	Song et al. (2020), Hyun et al. (2022), Févry and Phang (2018), Huang et al. (2023)	Truthfulness/Fidelity	Has the summary successfully preserved the meaning of the original text?
Length, Entity	He et al. (2022), Yu et al. (2021)	Accuracy/Correctness	Is the information in the summary accurate?
Style	Kryściński et al. (2018), Ribeiro et al. (2023), Cao and Wang (2021)	Readability	Is the text inside the summary readable?
Length	Yu et al. (2021), Liu et al. (2022a)	Completeness	Does the summary contain incomplete text?
Topic, Structure	Zhong and Litman (2023), Mukherjee et al. (2020), Mukherjee et al. (2022)	Coverage	Does the summary include all the topics or aspects defined in the source?

Table 8: Human evaluation metrics for controllable text summarization.

From the master table, we have represented our observations in Figures 2, 3, 4, 5, 6.

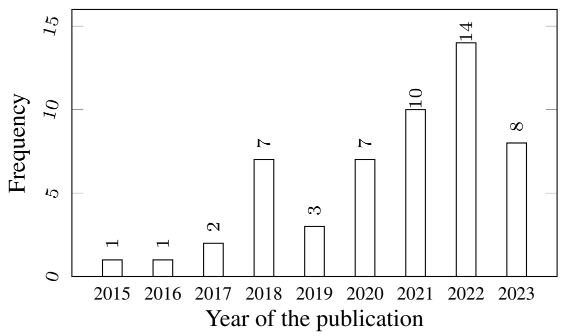

Figure 2: Year-wise papers published in CTS to handle various controllable attributes.

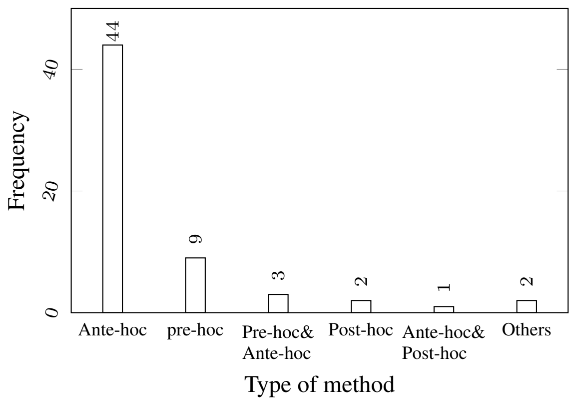

Figure 3: Various training approaches utilized to perform CTS tasks.

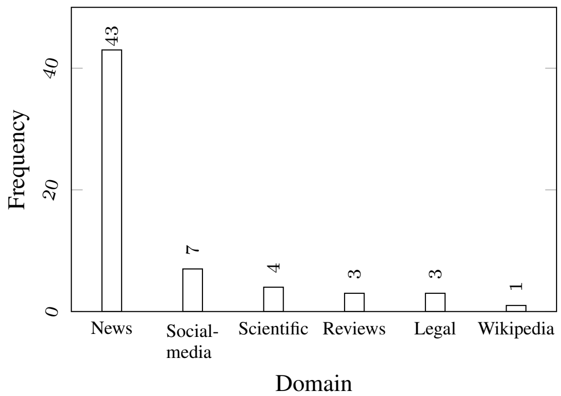

Figure 4: Domains utilized in CTS; most of the existing CTS tasks build on news domain data due to ease in accessibility.

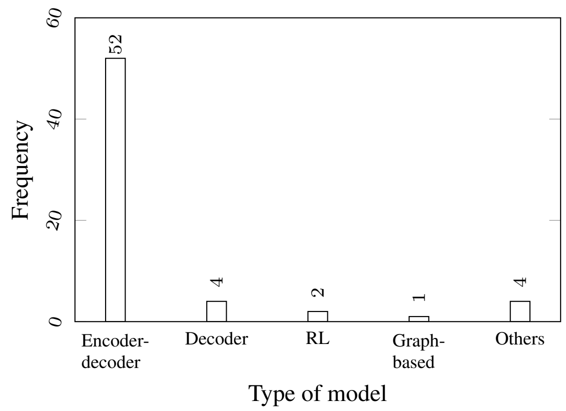

Figure 5: Type of models used in CTS; the majority of the models fall under standard sequence-to-sequence architecture.

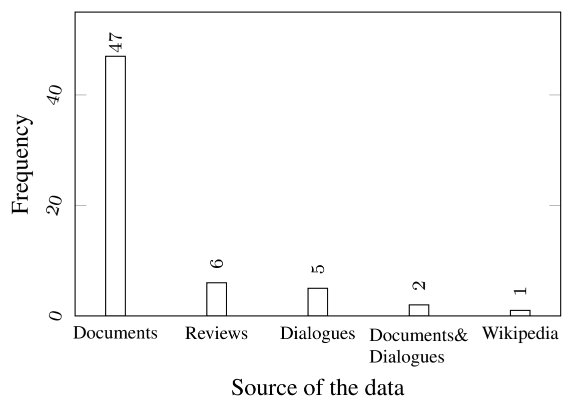

Figure 6: Source of the datasets used for the CTS task. The majority of the data samples are of ‘document’ type.

Aspect	Paper	Novel contribution	Dataset(s)	Limitations
Structure	Shen et al. (2022b)	Prepend structure prompt to the input	MRed	Subsequent generations deviate from the desired output
	Shen et al. (2022a)	Sentence-beam approach	MRed	Decoding methods significantly impact performance
	Zhong and Litman (2023)	Utilize predicted-role argument to control the structure	CanLII	Computationally expensive
Abstractivity	See et al. (2017)	Pointer-generator network	CNNDM	Failed to achieve higher abstraction and ineffective in core text selection
	Kryściński et al. (2018)	Decouples the decoder into a contextual network and mixed RL objective to encourage abstraction	CNNDM	Less readable summaries
	Song et al. (2020)	Mix-and-match strategy to generate summaries with various degree of copying levels	Gigaword, NEWSROOM	Poor performance in cross-domain settings
	Chan et al. (2021)	RL-based framework on constrained markov decision process to penalize the violation of control requirement	CNNDM, NEWSROOM	Poor performance for highly abstractive targets
Diversity	Narayan et al. (2022)	Compositional sampling decoding method	CNNDM, XSum	Generates unfaithful summaries for highly abstractive targets
Style	Fan et al. (2018b)	Convolutional encoder-decoder to generate stylistic summaries by adding the source prompt to the input	CNNDM, DUC2004	Repetitive and longer summaries
	Chawla et al. (2019)	RL-based method to generate formality-tailored summaries	CNNDM, Webis-TLDR-17	Poor performance in informal summary settings
	Jin et al. (2020a)	Multi-task learning framework with style-dependent layer normalization and style-guided encoder attention	NYT, CNN, Humor, Romance, Clickbait corpus	Poor performance on English Gigaword dataset
	Cao and Wang (2021)	Novel decoding methods: decode state adjustment, word unit prediction based	Hyperpartisan News detection dataset	-
	Goyal et al. (2022)	Mixture of experts strategy	CNNDM, XSum, NEWSROOM	Manual gating mechanism
	Luo et al. (2022)	Readability control of bio-medical documents	LS, PLS	Fail to handle fine-grained readability control
	Ribeiro et al. (2023)	Fine-grained readability control	CNNDM	Style insights may not generalize beyond English newswire datasets
Coverage	Wu et al. (2021)	A two-stage control generation strategy	SAMSUM	-
	Zhong et al. (2022)	Unsupervised framework to multi-granularity summary generation	Multi-NEWS, arXiv, DUC2004	Events extraction from source may effect the abstractiveness
	Huang et al. (2023)	Utilize the NLI models to improve the coverage	DIALOGSUM, SAMSUM	Partially addressing the factuality problem
Role	Lin et al. (2022)	Decoders for user and agent summaries and attention divergence loss for the same topic	CSDS, MC	-
	Liang et al. (2022)	Role aware centrality scores to reweight context representations for decoding	CSDS, MC	-
Entity	Zheng et al. (2020)	Controllable neural network with guiding entities	Gigaword, DUC 2004	Performance poorer than SOTA models
	Liu and Chen (2021)	Graph convolutional network based coreference fusion layer and entity conditioned Summary Generation	SAMSUM	Paraphrasing introduces factual inconsistencies in person-specific summaries
	Hofmann-Coyle et al. (2022)	Model as a sentence selection task using transformer based biencoder with a cosine similarity based loss and adapting contrastive loss	EntSUM	-
Salience	Nallapati et al. (2017)	Summarization as a sentence selection task with salience as a feature using sequence-to-sequence model	CNNDM	Poor performance on out-of-domain datasets
	Li et al. (2018)	Key information guided network with modified attention	CNNDM	Coverage mechanism not implemented
	Deutsch and Roth (2023)	Model salience in terms of noun phrases by incorporating QA signals	CNNDM, DUC-2004	Performance relies on question generation and answering models
	Pagnoni et al. (2023)	Unsupervised pretraining involving salient sentence selection	QMSum, SQuALITY	Computationally expensive
Length	Kikuchi et al. (2016)	Remaining words provided as additional input to decoder	Gigaword	Poor performance on DUC-2004
	Fan et al. (2018b)	Convolutional encoder-decoder, summary length grouping into bins and the source document prepend with length bin's value	CNNDM	Fails to generate summaries of arbitrary lengths
	Liu et al. (2018)	Remaining number of tokens replaced by characters at the decoder	CNNDM, DMQA	Fails to generalize to new control aspects at test time
	Févry and Phang (2018)	Unsupervised denoising auto-encoder for the task of sentence compression and the decoder provided with an additional input of the remaining summary length at each time step	Gigaword	Unfaithful summary generation in some cases
	Makino et al. (2019)	Global minimum risk training optimization method under length constraint	CNNDM, Mainichi	Fails to control length
	Sarkhel et al. (2020)	Multi-level summarizer with a multi-headed attention mechanism using a series of timestep independent semantic kernels	MSR Narratives and Thinking-Machines	Fail to encode desired length
	Takase and Okazaki (2019)	Extension to the sinusoidal positional embeddings to preserve the length constraint with length-difference positional encoding and length-ratio positional encoding	JAMUS corpus (Japanese)	Poor performance when desired target length is unseen
	Yu et al. (2021)	Concatenate the length context vector with the decoder hidden state and other attention vectors	CNNDM	Incomplete shorter summary generation
	Song et al. (2021)	Confidence driven generator trained on a denoising objective with a decoder only architecture with masked source and summary tokens	Gigaword, NEWSROOM	Poor performance on large datasets
	Chan et al. (2021)	Used a reinforcement learning based Constrained Markov Decision Process to control length along with constraints on a mix of attributes such as abstractiveness and covered entity	CNNDM, NEWSROOM DUC-2002	Length control only at word level
	Liu et al. (2022a)	Dynamic programming algorithm based on the Connectionist Temporal Classification model	Gigaword, DUC2004	Poor performance compared to autoregressive models
	Goyal et al. (2022)	Mixture-of-expert model with multiple decoders	CNNDM, XSum, NEWSROOM	No insights about style diversity in non-English and non-newswire datasets
	He et al. (2022)	A generic framework using keywords	CNNDM, arXiv, BIGPATENT	High reliance on the quality of extracted keywords
	Liu et al. (2022b)	Length aware attention model adapting the source encodings	CNNDM, XSum	Performance directly proportional to the summary length
	Zhong et al. (2022)	Events identification with unsupervised summary generation	GranuDUC, MultiNews, DUC2004, arXiv	Fails to capture abstractness due to event extraction
	Hyun et al. (2022)	RL based framework incorporating both the length and quality constraints in the reward function	DUC2004	Computationally expensive
	Kwon et al. (2023)	Summary length prediction task on the encoder side and encoded this information inserting a length-fusion positional encoding layer	CNNDM, NYT, WikiHow	Performance decreases with increase in summary length variance
	(Zhang et al., 2023b)	Hard prompt tuning and soft prefix tuning	CNNDM, QMSum	Low specificity in long generated summaries
Topic	Krishna and Srinivasan (2018)	RNN based attention model to generate multiple topic conditioned summaries	CNNDM	News categories provide predefined topics, limiting generalization to other tasks.
	Tan et al. (2020)	Extends topic based summarization to arbitrary topics, integrating external knowledge from ConceptNet and Wikipedia	CNNDM, MA News, All the News	-
	Suhara et al. (2020)	Framework for opinion summarization	HOTEL, Yelp	-
	Amplayo et al. (2021)	Multi-Instance Learning and a document preprocessing mechanism	SPACE, OPOSUM+	Incapable of handling unseen aspects
	Mukherjee et al. (2020)	Iterative sentence extraction algorithm	YELP	Poor performance in absence of attributes
	Mukherjee et al. (2022)	Topic-aware multimodal summarization system	MSMO	Output quality relies on data size

Table 9: CTS models descriptions and corresponding limitations.

	Paper	Publication info: Year	Venue	Controllable aspect: Controllable attribute	Controlling more than one aspect	Model details: Model type*	Training strategy	Approach	Code access	Code link	Dataset details: Dataset	Source	Nature of the data	Data release	Domain	Data link	Automatic evaluation: Metric name	Proposed New Metric	Human Evaluation: Human Evaluation	Metrics names	IAA Yes or No	Limitations: Mentioned?	Reproducibility: High/Medium/Not sure*
1	Chawla et al. (2019)	2019	CoNLL	Style	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM, Webis-TLDR-17 corpus	document	generic	no	news, social media	no	ROUGE	no	no	no	no	no	not sure
2	Goyal et al. (2022)	2022	EMNLP	Style	yes	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM, XSum, NEWSROOM	document	generic	no	news	no	ROUGE	abstractiveness, degree of specificity, length, readability	yes	relevance, coherence, grammaticality, factuality	no	yes	high
3	Song et al. (2021)	2021	NAACL	Length	no	decoder	ante-hoc	supervised	yes	Link	Gigaword, NEWSROOM	document	generic	yes	news	Link	ROUGE	no	yes	yes-no type questions	no	no	high
4	Song et al. (2020)	2020	AAAI	Abstractivity	no	decoder	ante-hoc	supervised	yes	Link	Gigaword, NEWSROOM	document	generic	no	news	no	ROUGE, BERTScore	no	yes	informativeness, grammaticality, truthfulness	no	no	high
5	Fan et al. (2018b)	2018	ACL-NMT(W)	Length, Style Entity	yes	encoder-decoder	pre-hoc	supervised	no	-	CNNDM	document	generic	no	news	no	ROUGE	no	yes	rater asked to preferred summary, binary evaluation	no	no	medium
6	Liu et al. (2022b)	2022	ACL	Length	no	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM, XSum	document	generic	no	news	no	ROUGE	no	yes	grammatically correct, informativeness, overall	no	no	high
7	Zhang et al. (2023b)	2023	TACL	Length, Topic Coverage	yes	encoder-decoder	pre-hoc	supervised	yes	Link	CNNDM, QMSum	document, dialogues	human annotated	yes	news, social media	Link	Control Error Rate, Control Correlation	control error rate, control correlation	yes	yes or no questions	yes	no	high
8	Takase and Okazaki (2019)	2019	NAACL	Length	no	encoder-decoder	ante-hoc	supervised	yes	Link	JAMUS	document	generic	no	news	no	ROUGE	no	no	no	no	no	medium
9	Shen et al. (2022a)	2022	EMNLP	Structure	no	encoder-decoder	post-hoc	supervised	yes	Link	Meta Review Dataset (MReD)	reviews	generic	no	scientific	no	ROUGE,BERTSCORE	no	yes	fluency, content relevance, structure similarity, decision correctness	no	yes	high
10	Goldsack et al. (2023)	2023	ACL-BIoNLP(W)	Style	no	-	-	-	-	-	PLOS and eLife	-	-	-	-	-	Relevance, Readability, Factuality	-	-	-	no	-	-
11	Cao and Wang (2021)	2021	NAACL	Style	no	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM	document	generic	no	news	no	Readability, Perplexity, BERTScore	readability, perplexity	yes	informativeness, fluency, relevance	no	no	high
12	Narayan et al. (2021)	2021	TACL	Structure	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM, XSum, SAMSum, and BillSum	document, dialogues	generic	no	news, legal	no	ROUGE	entity planning, specificity	yes	relevant, accurate, concise, fluent	no	no	medium
13	Amplayo et al. (2021)	2021	EMNLP	Topic	no	encoder-decoder	ante-hoc	supervised	yes	Link	SPACE, OPOSUM+	reviews	generic	no	social media	no	ROUGE	no	yes	informativeness,coherence, conciseness, fluency	no	no	high
14	Deutsch and Roth (2023)	2023	EACL	Saliency	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM, XSum, NYTimes	document	generic	no	news	no	ROUGE, BERTScore, QAEval	no	yes	selection of best summary (binary evaluation)	no	no	not sure
15	Wu et al. (2021)	2021	ALC-IJCNLP	Coverage	no	encoder-decoder	ante-hoc	supervised	yes	Link	SAMSum	dialogues	generic	no	social media	no	ROUGE	no	yes	factual consistency, informativeness	no	no	high
16	Hsu and Tan (2021)	2021	EMNLP	Topic	no	encoder-decoder	ante-hoc	supervised	yes	Link	Yelp's businesses, reviews, and user data	reviews	generic	no	restaurant reviews	Link	MSE	no	yes	-	no	no	high
17	Hyun et al. (2022)	2022	EMNLP	Length	yes	RL	ante-hoc	unsupervised	yes	Link	DUC2004	document	generic	no	news	no	ROUGE	no	yes	fluency, fidelity	yes	yes	high
18	Chan et al. (2021)	2021	TACL	Length, Entity, Abstractivity	yes	RL	ante-hoc	supervised	yes	Link	CNNDM, NEWSROOM, DUC-2002	document	generic	no	news	no	ROUGE, BERTSCORE, MoveScore, QA-F	yes	yes	fluency, entity relevance, faithfulness	yes	no	high
19	Krishna and Srinivasan (2018)	2018	NAACL-HLT	Topic	no	encoder-decoder	pre-hoc	supervised	no	-	CNNDM	document	generic	no	news	no	ROUGE	no	yes	topic relevance	no	no	not sure
20	Krishna et al. (2018)	2018	ACL	Topic	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM	document	generic	no	news	no	ROUGE	no	yes	quality	yes	no	not sure
21	Rush et al. (2015)	2015	EMNLP	Length	no	encoder-decoder	ante-hoc	supervised	no	-	NYT, DUC2004	document	generic	no	news	no	ROUGE	no	no	no	no	no	high
22	Févry and Phang (2018)	2018	CoNLL	Length	no	encoder-decoder	ante-hoc	unsupervised	yes	Link	Gigaword	document	generic	no	news	no	ROUGE	no	yes	grammaticality, preserving original information	no	no	high
23	Li et al. (2018)	2018	NAACL-HLT	Saliency	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM	document	generic	no	news	no	ROUGE	no	no	no	no	no	not sure
24	Jin et al. (2020b)	2020	AAAI	Coverage	no	encoder-decoder	ante-hoc	supervised	no	-	Gigaword, DUC2004, MSR dataset	document	generic	no	news	Link	ROUGE	Word Mover's Distance, Bert Score	yes	faithfulness, informativeness, fluency	no	no	not sure
25	Liu et al. (2022a)	2022	NeurIPS	Length	no	Non Auto Regressive Model	ante-hoc	supervised	yes	Link	Gigaword, DUC2004	document	generic	no	news	no	ROUGE	no	yes	overall quality, completeness/fluency	no	yes	medium
26	Kwon et al. (2023)	2023	EACL	Length	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM, NYT, WikiHow	document	generic	no	news	no	ROUGE, Length Variance	no	yes	informativeness, conciseness	no	yes	not sure
27	Pagnoni et al. (2023)	2023	ACL	Entity, Saliency	yes	encoder-decoder	ante-hoc	supervised	yes	Link	QMSum and SQuALITY	document	generic	no	news	no	ROUGE, BERTSCORE, levenshtein distance	no	yes	unambiguity, conciseness	no	yes	high
28	Hofmann-Coyle et al. (2022)	2022	AACL	Entity	no	encoder-decoder	ante-hoc	supervised	no	-	EntSum	document	generic	no	news	no	F1	no	no	no	no	no	not sure
29	Lin et al. (2022)	2022	ACL	Role	no	encoder-decoder	ante-hoc	supervised	yes	Link	CSDS, MC	dialogues	generic	no	customer service data, medical inquiries	Link	ROUGE, BertScore, BLEU, MoverScore	no	yes	informativeness, non-redundancy, fluency	yes	no	high
30	Zhong et al. (2022)	2022	EMNLP	Coverage	yes	encoder-decoder	ante-hoc	unsupervised	no	-	GranuDUC, MultiNews, DUC2004, Arxiv	document	generic	yes	news, scientific papers	Link	ROUGE	BERTScore	yes	fluency, relevance, faithfulness	no	yes	medium
31	Liang et al. (2022)	2022	AACL	Role	yes	encoder-decoder	ante-hoc	supervised	no	-	CSDS, MC	dialogues	generic	no	customer service data, medical inquiries	Link	ROUGE, BLEU, BERTScore, MoverScore	no	no	no	no	no	not sure
32	Liu and Chen (2021)	2021	EMNLP	Entity	no	encoder-decoder	ante-hoc	supervised	no	Link	SAMSum	dialogues	generic	no	social media	no	ROUGE	no	yes	factual consistency, informativeness	no	no	medium
33	Zheng et al. (2020)	2020	COLING	Entity	no	encoder-decoder	ante-hoc	supervised	yes	Link	Gigaword, DUC2004	document	generic	no	news	no	BERT-REO, ROUGE	BERT-REO	yes	informativeness, grammaticality, coherence	no	no	high
34	Huang et al. (2023)	2023	EACL	Coverage	no	encoder-decoder	ante-hoc	supervised	yes	Link	DIALOG-SUM, SAMSUM	dialogues	generic	no	social media	no	BARTScore, ROUGE, FactCC	no	yes	recall, precision, faithfulness	yes	yes	high
35	Mukherjee et al. (2022)	2022	AACL	Topic	no	encoder-decoder	ante-hoc	supervised	yes	Link	MSMO	document	semi-automatic	yes	news	Link	ROUGE	no	yes	Coverage, Grammar, Topic-Aware-Text	no	yes	high
36	Jin et al. (2020a)	2020	ACL	Style	yes	encoder-decoder	ante-hoc	supervised	yes	Link	NYT, CNN	document	generic	no	news	no	BLEU, ROUGE, METEOR, CIDEr	no	yes	relevance, attractiveness, language fluency, style strength	no	no	high
37	See et al. (2017)	2017	ACL	Abstractivity, Coverage	yes	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM	document	generic	no	news	no	ROUGE	no	no	no	no	no	high
38	Kryściński et al. (2018)	2018	EMNLP	Abstractivity	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM	document	generic	no	news	no	ROUGE	no	yes	readibility, relevance	no	no	not sure
39	Bahrainian et al. (2021)	2021	ACM	Topic	no	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM, AMI , ICSI, ADSE	document	generic	no	news., meetings	no	ROUGE	no	yes	informativeness, readability	yes	no	high
40	Kikuchi et al. (2016)	2016	EMNLP	Length	no	encoder-decoder	ante-hoc, post-hoc	supervised	yes	Link	DUC2004, Gigaword	document	generic	no	news	no	ROUGE	no	no	no	no	no	high
41	Liu et al. (2018)	2018	EMNLP	Length	no	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM, DMQA	document	generic	no	news	no	ROUGE, VAR, Similarity	no	no	no	no	no	high
42	Makino et al. (2019)	2019	ACL	Length	no	encoder-decoder	ante-hoc	supervised	no	-	CNNDM, Mainichi	document	generic	no	news	no	ROUGE	no	yes	post-editing-time	no	no	not sure
43	Yu et al. (2021)	2021	ACL	Length	no	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM	document	generic	no	news	no	ROUGE	no	yes	Correctness, completeness fluency	no	no	high
44	Sarkhel et al. (2020)	2020	COLING	Length	no	encoder-decoder	ante-hoc	supervised	no	-	MSR Narratives, Thinking-Machines	document	generic	no	social media, science/education	no	ROUGE, METEOR	no	yes	presence of key facts	no	no	not sure
45	He et al. (2022)	2022	EMNLP	Length, Entity	yes	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM, arXiv, BIGPATENT	document	generic	no	news, scientific data	no	ROUGE	no	yes	Control Accuracy, Control Relevance	no	yes	high
46	Tan et al. (2020)	2020	EMNLP	Topic	no	encoder-decoder	ante-hoc	supervised	yes	Link	CNNDM, MA News	document	generic	no	news	no	ROUGE	no	yes	Accuracy , Informativeness, Fluency	yes	no	high
47	Narayan et al. (2022)	2022	ACL	Diversity	no	encoder-decoder	post-hoc	supervised	yes	Link	CNNDM, XSum, SQuAD)	document	generic	no	news	no	ROUGE,BertScore, SELF-BLEU	no	yes	faithfulness and diversity	no	no	medium
48	Zhong and Litman (2023)	2023	IJCNLP-AACL	Structure	no	encoder-decoder	ante-hoc	supervised	no	-	CanLII	document	generic	no	legal	no	ROUGE, BERTScore, structure similarity.	no	yes	coherence, coverage	no	yes	not sure
49	Suhara et al. (2020)	2020	ACL	Topic	yes	extractor + generator(decoder)	ante-hoc	unsupervised	yes	Link	Hotel, Yelp	reviews	generic	no	reviews	no	ROUGE	no	yes	informativeness, coherence, non-redundancy, content-support	no	no	high
50	Ribeiro et al. (2023)	2023	EMNLP	Style	no	encoder-decoder	pre-hoc, ante-hoc	supervised	yes	Link	CNNDM	document	generic	no	news	no	ROUGE, BERTScore, FRE, GFI, CLI	yes	yes	most readable and least readable	no	yes	high
51	Nallapati et al. (2017)	2017	AAAI	Abstractivity, Saliency	yes	GRU-RNN	ante-hoc	supervised	yes	Link	CNN/DM, DUC2002	document	generic	no	news	no	ROUGE	no	no	no	no	no	high
52	Hitomi et al. (2019)	2019	INLG	Length	no	encoder-decoder	pre-hoc, ante-hoc	supervised	no	-	JAMUS	document	generic	yes	news	Link	ROUGE	no	no	no	no	no	medium
53	Shen et al. (2022c)	2022	NIPS	Coverage	no	encoder-decoder	pre-hoc	supervised	yes	Link	Multi-LexSum	document	Human-annotated	yes	Legal	Link	ROUGE,BERTSCORE	no	yes	Rate the generation (0-3 scale)	no	yes	high
54	Maddela et al. (2022)	2022	ACL	Entity	no	encoder, encoder-decoder	pre-hoc, ante-hoc	supervised	yes	Link	CNNDM, NYT	document	generic	yes	news	Link	ROUGE, BERTScore	no	no	no	no	no	high
55	Bahrainian et al. (2022)	2022	ACL	Topic	no	encoder-decoder, decoder	-	supervised	no	-	NEWTS	document	Human-annotated	yes	news	Link	ROUGE	no	yes	Binay evaluation	yes	no	high
56	Hayashi et al. (2021)	2021	TACL	Topic	no	encoder, textRank	pre-hoc	supervised	yes	Link	WikiAsp	Wikipedia	Automatic	yes	Wikipedia	Link	ROUGE	no	yes	no	no	no	high
57	Luo et al. (2022)	2022	EMNLP	Style	no	encoder-decoder	pre-hoc	supervised	no	-	TS and PLS	document	Automatic	yes	Scientific	Link	ROUGE, FKG, CLI,ARI, RNTPC	yes	yes	Relevance, Grammar, Coherence	no	yes	high
58	Ahuja et al. (2022)	2022	ACL	Topic	no	encoder, encoder-decoder	pre-hoc	supervised	yes	Link	ASPECTNEWS	document	Automatic	yes	news	Link	ROUGE	no	no	no	no	no	Medium
59	Mukherjee et al. (2020)	2020	SIGIR	Topic	no	Graph-based	ante-hoc	Unsupervised	yes	Link	Tourism Reviews	document	Automatic	yes	Tourism	Link	ROUGE	no	yes	aspect-coverage, readability, diversity	no	no	high
60	Lin et al. (2021)	2021	EMNLP	Role	no	all	pre-hoc	supervised	yes	Link	CSDS	reviews	Human-annotated	yes	reviews	Link	ROUGE, BERTScore	no	yes	Informativeness, non-redundancy, fluency, matching rate	yes	no	high
61	Shen et al. (2022b)	2022	ACL	Structure	no	encoder-decoder	pre-hoc	supervised	yes	Link	MReD	reviews	Human-annotated	yes	Scientific	Link	ROUGE	no	yes	Fluency, content relevance, structure, decision correctness	no	no	high

Table 10: Master survey table (*marked fields are filled to best of our understanding based on the available information).