EntSUM: A Data Set for Entity-Centric Summarization

Mounica Maddela§*
Georgia Institute of Technology
mmadela3@cc.gatech.edu

Mayank Kulkarni*
Bloomberg
mkulkarni24@bloomberg.net

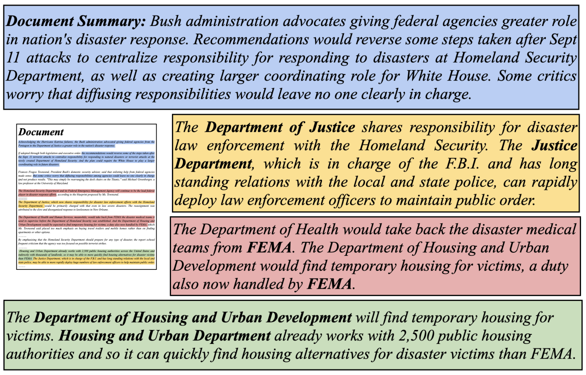
Daniel Preoţiuc-Pietro
Bloomberg
dpreotiucpie@bloomberg.net

Abstract

Controllable summarization aims to provide summaries that take into account user-specified aspects and preferences to better assist them with their information need, as opposed to the standard summarization setup which build a single generic summary of a document. We introduce a human-annotated data set (EntSUM) for controllable summarization with a focus on named entities as the aspects to control. We conduct an extensive quantitative analysis to motivate the task of entity-centric summarization and show that existing methods for controllable summarization fail to generate entity-centric summaries. We propose extensions to state-of-the-art summarization approaches that achieve substantially better results on our data set. Our analysis and results show the challenging nature of this task and of the proposed data set.[1][2]

1 Introduction

Automatic summarization is a core NLP problem that aims to extract key information from a large document and present it to the user with the role of assisting them to digest the core information in the document faster and more easily. However, each user may have a distinct information need and generating a single summary for a document is not suitable for all readers of the document. Recently, various setups for summarization were proposed such that user preferences can be taken into account in the summarization process. These include providing guidance signals such as summary length (Kikuchi et al., 2016), allowing users to provide terms of interest such as aspects (Amplayo et al., 2021) or entities (Fan et al., 2018) or providing users the flexibility to interact with the summary and explore new facets of interest (Avinesh et al., 2018). The development of such methods may be paramount in enabling the wide-spread usability of summarization technology. Figure 1 shows an example of a document, its generic summary and summaries controlled through salient named entities in the original document.



Document Summary: *Bush administration advocates giving federal agencies greater role in nation's disaster response. Recommendations would reverse some steps taken after Sept 11 attacks to centralize responsibility for responding to disasters at Homeland Security Department, as well as creating larger coordinating role for White House. Some critics worry that diffusing responsibilities would leave no one clearly in charge.*

The ***Department of Justice*** *shares responsibility for disaster law enforcement with the Homeland Security. The* ***Justice Department****, which is in charge of the F.B.I. and has long standing relations with the local and state police, can rapidly deploy law enforcement officers to maintain public order.*

The Department of Health would take back the disaster medical teams from ***FEMA****. The Department of Housing and Urban Development would find temporary housing for victims, a duty also now handled by* ***FEMA****.*

The ***Department of Housing and Urban Development*** *will find temporary housing for victims.* ***Housing and Urban Department*** *already works with 2,500 public housing authorities and so it can quickly find housing alternatives for disaster victims than FEMA.*

Figure 1: Example of a generic summary (blue), with three entity-centric summaries from EntSUM focusing on the entities in **bold**.

High quality reference data sets are needed to foster development and facilitate benchmarking. Most summarization data sets are obtained using opportunistic methods such as using abstracts written by editors or librarians when indexing documents. These are by default generic, thus not applicable to controllable summarization. Initial research in this area used small scale human annotations to compare between controllable and generic summarization methods (Fan et al., 2018; He et al., 2020), but these can be prone to biases or qualitative issues, offer only relative quality measurement and do not allow for replicable comparisons between multiple methods or model tuning.

Thus, this paper introduces a new data set for controllable summarization focusing on entities as control aspects given these are usually key aspects in documents and their summaries. The data set consists of 2,788 human-generated entity-centric summaries across 645 documents that are obtained

* Equal Contribution

§Work done during an internship at Bloomberg

[1] The data set is available at: https://zenodo.org/record/6359875

[2] The code is available at: https://github.com/bloomberg/entsum

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3355 - 3366
May 22-27, 2022 ©2022 Association for Computational Linguistics

using a strict quality control process mechanism involving several intermediate annotation steps which can be further used in modelling and analyses such as identifying sentences relevant to an entity. The summaries are elicited largely to merge the most important content in a coherent way, while maintaining factuality during the summary creation process.

Our data set demonstrates the distinct nature of the entity-centric summarization as opposed to generic summarization and that methods proposed to date for controllable summarization fail at this task. We propose adaptations of state-of-the-art extractive and abstractive summarization methods that significantly improve performance when compared to generic summaries. Our contributions are:

- the first annotated data set for controllable summarization with entities as targets for control (ENTSUM - Entity SUMmarization);
- quantitative data set analysis that highlights the challenges and distinctiveness of this task;
- evaluation of generic and also controllable summarization methods on the ENTSUM data set;
- adaptations of extractive and abstractive summarization methods for performing entity-centric summarization when trained with generic summaries only.

2 Related Work

Controllable summarization was proposed with the goal of allowing users to define high-level attributes of summaries such as length, source-style or entities (Fan et al., 2018). Methods relied on adapting existing summarization methods such as CNNs (Fan et al., 2018) or BART (He et al., 2020) by pre-pending the controls to the training data and presenting the target control only in inference. However, these methods were only evaluated by comparison to generic summarization methods using human judgments, which can suffer from biases and qualitative issues.

Closely related to controllable summarization, **guided summarization** also uses an input guidance variable in addition to the document when generating the summary (Dou et al., 2021). This is different to controllable summarization because the goal of the guidance signal is to generate an improved generic summary by using the guidance to increase faithfulness and quality. Guidance signals explored in past research include summary length (Kikuchi et al., 2016; Liu et al., 2018b; Sarkhel et al., 2020), keywords (Li et al., 2018; Saito et al., 2020), relations (Jin et al., 2020) or highlighted sentences (Liu et al., 2018a).

Opinion summarization is the task of automatically generating summaries for a set of reviews about a specific target and usually involves inferring the aspects of interest, predicting sentiment towards them and generating a summary from the extracted sentences (Kim et al., 2011; Angelidis and Lapata, 2018). Amplayo and Lapata (2021) studied zero-shot controllability to generate need-specific summaries for movie reviews and evaluated using human comparison judgments.

Contemporaneous to this work, controllable multi-document summarization for aspects in reviews was introduced (Angelidis et al., 2021; Amplayo et al., 2021). This work created two data sets used for testing, one focusing on six aspects in hotel reviews (SPACE) and another focusing on 18 aspects for product reviews (OPOSUM+), both obtained using a multi-step annotation process related to the one we use in this paper.

Interactive Summarization is a technique which aims to provide to an interactive faceted summarization of a set of documents and help the user inquire for more information via suggested or free-text queries (Avinesh et al., 2018; Shapira et al., 2021; Hirsch et al., 2021). This setup is focused on a multi-document scenario where relevant content to a target concept is retrieved, then fed to a generic abstractive summarization method.

Recently, Hsu and Tan (2021) proposed **decision-focused summarization**, where the goal is to summarize information across multiple documents with the goal of aiding a human to forecast an outcome.

3 The ENTSUM Data Set

This section details the collection and annotation process for data set creation. We focus on entities as the aspect to control because named entities are central actors in most news articles and entities are key aspects that make good summaries, together with events and facts. Initial work on controllable summarization considered entities as one of the target for controls (Fan et al., 2018; He et al., 2020).

Most large-scale summarization data sets were obtained opportunistically by mining existing sources of documents and their generic summaries expressed either as titles (Narayan et al., 2018), bullet points (Hermann et al., 2015) summaries created for indexing purposes (Sandhaus, 2008) or

TL;DR's created by scientific paper authors (Cachola et al., 2020). However, we could not identify any similar proxies for entity-centric summaries. Thus, we created the ENTSUM data set through a manual annotation process.

3.1 Task

Given a document and entity pair, where the entity is a named entity mentioned in the document, the goal of the annotation is to obtain a summary capturing important information about the entity in that document.

3.2 Data Collection and Preparation

Our entity-centric summarization data set consists of news articles from the *The New York Times Annotated Corpus* (NYT) (Sandhaus, 2008), which consists of 1.8 million articles written between 1987 and 2007. Around 650k articles in the corpus contain article summaries written by library scientists for indexing purposes. We choose to annotate documents from the NYT data set to enable comparison to generic summaries. We selected the NYT data set instead of other popular summarization data sets (e.g. CNN/DailyMail) because of the clarity of the data licensing terms on the NYT corpus for research purposes (Sandhaus, 2008).

We use the NYT test set as defined in (Kedzie et al., 2018) to sample the articles used in the ENTSUM data set, as we envision the data set will be used primarily for evaluation purposes. We removed documents with over 1500 words, as we found the majority of these are opinion articles not involving many entities. We split the rest of the documents into sentences and identified named entities using Flair, a high performing system for named entity recognition (Akbik et al., 2019) which identifies Organizations, Person and Location entities. We only select for annotation entities that are Organization and Persons because Locations are usually not salient to the document, thus do not play an active role in the article. From this set, we randomly sampled 10,000 entities spanning 693 documents.

3.3 Annotation Process

Summarization is a highly subjective task because the notion of salient information in a document is user-specific and task-dependent (Iskender et al., 2020). There has been relatively little work on the topic of designing annotation guidelines. The most common method to collect summaries is to ask annotators to summarize the document within a specific length limit (Harman and Over, 2004; Dang, 2006). However, such methods are prone to subjective bias with a low human agreement about the content in the summary (Li et al., 2021). Therefore, to ensure quality of the annotation process, we propose a multi-step approach to collect entity-centric summaries that has similarities to the collection method for opinion summarization (Angelidis and Lapata, 2018). Splitting the tasks in multiple steps allows us to ensure quality of the data set through adjudication across multiple annotations at each step which reduces error propagation across tasks. Figure 2 shows an overview of the four-step annotation process.

3.3.1 Entity Salience

The first tasks judges if an automatically extracted entity is really a named entity and how salient it is to the source document (Gamon et al., 2013a,b; Dojchinovski et al., 2016; Trani et al., 2016). We do this to keep only salient entities for generating summaries, as others are not important targets for entity-centric summaries and may not have enough related content to produce a summary.

Given an article and an entity in the article, we asked the annotators to rate the salience of the entity with respect to the article on a four point scale ranging from not salient (1), through low salience (2), medium salience (3) and high salience (4), similar to Trani et al. (2016).

We collected 2 independent annotations for each entity and increased redundancy up to 5 if there was disagreement. We take the salience rating as the average of all individual ratings. We observe that entities with an average rating < 1.5 are generally mentioned once in the document and, therefore can not have a meaningful summary. We remove these entities, resulting in 3,846 entities. We further grouped the entity mentions from each document using substring matching because multiple entity strings can refer to the same entity (e.g. *Barack Obama – Obama*). After grouping, we obtain 2,788 entities to use in subsequent tasks.

3.3.2 Salient Sentence Extraction

The second task aims to identify all sentences in the article that are salient to the target entity. To facilitate the process, we displayed all sentences in a document in a tabular format and premarked sentences that contain the given entity mention. The annotators can add additional sentences or remove existing ones. We also asked the annotators to keep

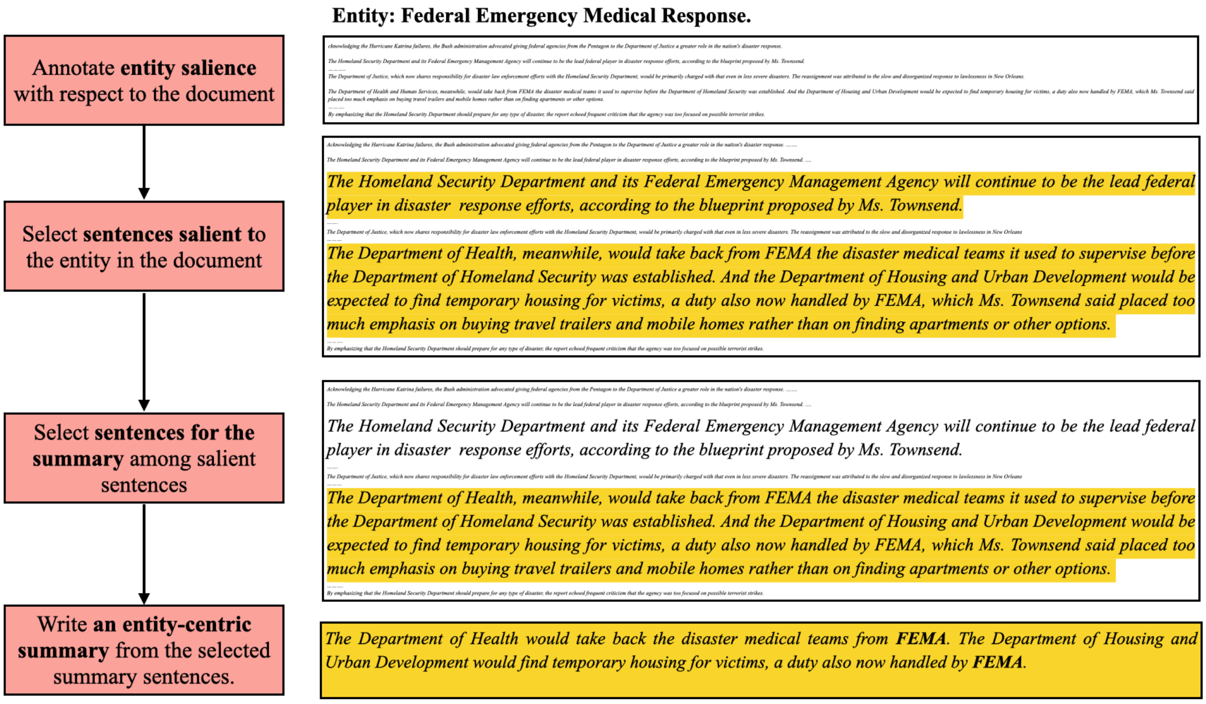

Entity: Federal Emergency Medical Response.
Annotate entity salience with respect to the document
Select sentences salient to the entity in the document
Select sentences for the summary among salient sentences
Write an entity-centric summary from the selected summary sentences.
The Homeland Security Department and its Federal Emergency Management Agency will continue to be the lead federal player in disaster response efforts, according to the blueprint proposed by Ms. Townsend.
The Department of Health, meanwhile, would take back from FEMA the disaster medical teams it used to supervise before the Department of Homeland Security was established. And the Department of Housing and Urban Development would be expected to find temporary housing for victims, a duty also now handled by FEMA, which Ms. Townsend said placed too much emphasis on buying travel trailers and mobile homes rather than on finding apartments or other options.
The Homeland Security Department and its Federal Emergency Management Agency will continue to be the lead federal player in disaster response efforts, according to the blueprint proposed by Ms. Townsend.
The Department of Health, meanwhile, would take back from FEMA the disaster medical teams it used to supervise before the Department of Homeland Security was established. And the Department of Housing and Urban Development would be expected to find temporary housing for victims, a duty also now handled by FEMA, which Ms. Townsend said placed too much emphasis on buying travel trailers and mobile homes rather than on finding apartments or other options.
The Department of Health would take back the disaster medical teams from FEMA. The Department of Housing and Urban Development would find temporary housing for victims, a duty also now handled by FEMA.

Figure 2: Annotation pipeline of ENTSUM

Metric	Overall	Entity Type		Entity Salience	
		PER	ORG	Medium	High
Number of Salient Entities (**Task 1**)	2788	1741	1047	2100	688
Sentences with entity mentions	3.95	4.21	3.46	3.36	5.65
Entity Salient Sentences (**Task 2**)	5.80	6.34	5.02	4.95	8.56
Entity-Centric Summary Sentences (**Task 3**)	2.49	2.59	2.28	2.33	2.66
Summary word length (**Task 4**)	81.7	84.9	76.1	78.6	88.2
Summary char length (**Task 4**)	444.3	458.1	421.7	432.1	482.9

Table 1: Statistics for the output of each task in our entity-centric summary annotation pipeline, overall and across entity types and salience scores as annotated in Task 1. **PER** and **ORG** refer to “Person” and “Organization” entity types respectively.

		Avg. summary len.			Avg. article len.			Compression Ratio		% novel ngram	
Data set	Size	sents.	word	char.	sents.	word	char.	article	salient	unigram	bigram
NYT	41,265	4.9	117	677	36.9	1021	5471	0.12	–	11.5	39.5
CNNDM	312,085	3.7	56	297	33.1	782	3998	0.089	–	13.3	49.95
ENTSUM	2788	2.5	81	444	34.4	1002	5319	0.09	0.62	0.82	5.93

Table 2: Comparison of the existing document summarization data sets with ENTSUM. We report the corpus size, average article and summary length (in terms of words, sentences, and characters), and percentage of novel n-grams in the summary when compared to the article. We also report the compression ratio of the summary with respect to the original article text and the entity-specific salient text selected by annotators.

the salient sentences as complete as possible by including the sentences that resolve any references in the initially selected sentences.

We collected three annotations for each document and entity pair resulting in three annotations for all sentence and entity pairs. We assigned each sentence a binary label (salient to the entity or not) using majority vote across the three annotations.

Table 1 shows the average number of salient sentences (5.80) is much higher than the average number of premarked sentences (3.95), indicating this task resulted in an expansion from only using the sentences that explicitly mention the target entity.

3.3.3 Entity-Centric Summary Sentences

The third task aims to identify the sentences in the article that are used to make up the entity-centric summary. We display the sentences of the document in a tabular format with the salient sentences extracted from the previous task highlighted and allowed the annotators to select only from these sentences. We instructed the annotators to first select up to 3 sentences and add up to 3 more sentences if these are needed to provide context.

3.3.4 Entity-Centric Summary

The final task is to write a coherent summary for the entity in the document of up to 150 words using the summary sentences selected previously. This task was performed together with the third task, as they are tightly coupled, to limit cognitive load and to be able to control for quality by comparing selected summary sentences.

As this is a labor intensive task, we collected two annotations for a subset of the target entities (867 out of 2,788) to measure agreement. We provide both summaries in the data set release in order to facilitate evaluation with multiple references. The annotated summary sentences represent only 41.3% of all salient sentences across all the tasks. Table 1 shows the annotation statistics.

We note the output of each task is released with the ENTSUM data set and can be used when training models, for separate tasks or as auxility tasks in a multi-task learning setup.

3.4 Data Quality

We devised multiple tasks to accomplish our goal of ensuring quality throughout the annotation process and to make the complex and subjective task of summarization easier for annotators. We adjudicate annotations across multiple annotators to reduce error propagation, wherein if one task has wrong annotations, the subsequent tasks will have the error propagated.

We use our internal annotation platform for obtaining annotations. The annotation was performed using a group of English-speaking vendors who were hired and trained for completing this task through training sessions and performed the task independently from each other. We do not collect any private information from the annotators and do not release the identity of the annotators together with the data. We conducted several training sessions and initial rounds with the annotators, the results of which were discarded, to ensure the annotators are proficient in the task. The training rounds included 100 items for the first two tasks and 50 items for the latter two for all annotators.

We perform multiple annotations for the upstream tasks. For the entity salience task which is a four-way classification task, we elicit 2 annotations for each item and, if these disagree, we increase redundancy to up to 5 annotations if there is no majority (2 annotations – 6261 items; 3 annotations – 3318 items; 5 annotations – 421 items). For the salient sentence extraction task, we elicit 3 annotations for each item and adjudicate annotations at the sentence level using majority vote.

We report inter-annotator agreement for each task. For the 4-way ordinal entity salience task we observe 0.709 interval Krippendorf's Alpha (Krippendorff, 2011), which corresponds to substantial agreement (Artstein and Poesio, 2008). The annotators agreed on a single annotation 62.6% of the time. For the salient sentence selection task, we compute inter-annotator agreement using Krippendorf's Alpha between binary sentence-level judgments and obtain a value of 0.744 Krippendorf's Alpha, which again indicates substantial agreement. All three annotators agreed on the same value for 88.4% of the sentences.

Selecting the summary sentences is a more subjective task, especially given that all sentences are salient to the target entity. Despite this, the inter-annotator agreement is of 0.539 Krippendorf's Alpha, which is considered good agreement. Finally, in the summary creation task, we compute ROUGE (Lin and Hovy, 2003) between the summaries and achieve the following values: ROUGE-1 = 71.7; ROUGE-2 = 62.6 and ROUGE-L = 69.0. We release both summaries in our data set where available, as these could be used as multiple references when computing evaluation metrics.

3.5 Data Analysis

Summary Statistics Table 2 presents summary statistics relevant to summarization data for the newly introduced ENTSUM data set, with the commonly used document generic summarization data sets CNN-DailyMail (CNNDM) and NYT. We note that summaries in ENTSUM are shorter than their generic counterparts in the NYT corpus, but longer than those in CNNDM, except for the number of sentences, which is expected as the summaries in CNNDM undergo the most compression as demonstrated by the article compression ratio. ENTSUM exhibits the lowest percentage of novel unigrams and bigrams, in line with how our annotation was set up to focus on integrating the original content in a coherent summary. The entity-specific salient text is significantly shorter than the entire document and, as a result, the summary contains the relevant content without requiring dramatic paraphrasing or compression.

Comparison to Generic Summaries Our hypothesis is that a new data set for entity-centric summarization is needed as entity-centric summaries do not align well with generic summaries. We compute ROUGE (Lin and Hovy, 2003) scores between the entity-centric summaries in ENTSUM and their corresponding generic summaries in the NYT corpus, with the following values: ROUGE-1 = 26.2, ROUGE-2 = 9.8 and ROUGE-L = 22.9. Low scores show there is low lexical and content overlap between the entity-centric summaries and their corresponding document summaries, demonstrating the distinctiveness of the entity-centric summarization task.

Entity Type and Salience Table 1 shows the task-specific statistics of ENTSUM by entity type and salience level separately. We note that the data set has more person entities than organizations and, on average, the related content and summaries associated to people is slightly longer. There are significantly more entities with medium salience values when compared to highly salient entities, which are an average slightly more than one for each document. We note that both sentences with entity mentions and salient sentences to the entities are substantially larger in number for highly salient entities, but there is just a small gap for the entity-centric summaries and sentences, which shows that more selection and compression was achieved for these highly salient entities.

Sentence Position Distribution Figure 3 shows the position distribution of entity salient and entity-centric summary sentences in the original document. The figure highlights that both types of sentences are more likely to be distributed at the start of the document, which is expected given we are only considering salient entities to the document. We see that sentences used for summaries are even more likely to be towards the start of the document. However, the sentence distribution is not very skewed, with hundreds of summary sentences being present even in position 20 or higher in the original document. This highlights the challenging nature of the data set.

4 Methods

For an initial modelling attempt for the ENTSUM data set, we evaluate all controllable summarization approaches proposed to date, generic summarization methods, strong heuristics for summarization and a couple of adaptations of state-of-the-art methods for abstractive (Dou et al., 2021) and extractive summarization (Liu and Lapata, 2019) to the entity-centric summarization task.

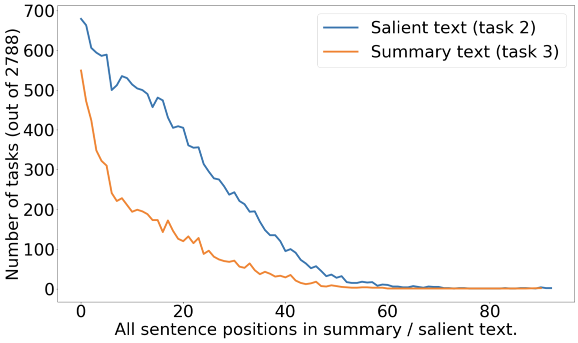


Figure 3: Distribution of sentence positions for salient and summary sentences.

Some of the methods described in this section involve detecting the entity mentions in documents unlabeled with entities in training and/or at inference time. For this, we use a combination of standard methods for NER based on Flair (Akbik et al., 2018) and their coreferent mentions as identified through the SpanBERT coreference system (Joshi et al., 2020).

4.1 Abstractive Methods

Abstractive summarization uses generation methods to express the content of the original document.

4.1.1 ConvNet for Controllable Summarization

We denote through **ConvNet** the first method for controllable summarization proposed in Fan et al. (2018). It adopts a CNN encoder-decoder model for summarization and is trained by replacing entities in the document with placeholders and prepending them to the document. At inference time, only the target entity is prepended to the summary to generate the entity-centric summary (Fan et al., 2018).

4.1.2 CTRLSum

CTRLSum (He et al., 2020) is a method based on BART (Lewis et al., 2020), a popular Transformer-based sequence-to-sequence model for summarization. CTRLSum is fine-tuned by prepending keywords, in this case all detected entity mentions, to the input document to control the summary (He et al., 2020). At inference time, only the target entity is prepended to the target document to generate the entity-centric summary.

4.1.3 GSum

GSum (Dou et al., 2021) is a document summarization framework that allows for using as input a guidance signal (e.g. keywords, sentences) along with the source document with the goal of improving the generic document summarization task through improving faithfulness. The model architecture consists of a Transformer (Vaswani et al., 2017) model initialized with BART (Lewis et al., 2020). The model has two encoders: one to encode the source document and the other to encode the guidance signal. The encoders share the embedding and the encoding layers except for the topmost layer. The decoder first attends to the guidance signal to select the part of the document to focus on and then attends to the document with these guidance-aware representations. The framework allows to include varied guidance signals and demonstrates improvements on generating generic summaries.

4.1.4 Adapting GSum for Entity-Centric Summarization

We adapt GSum to generate entity summaries by using the entity information as guidance signal. However, the original GSum implementation used a single generic summary as output for each input document, which is not suitable for our setup in which the output is conditioned on both the input document and the guidance signal (i.e. entity). In addition, we do not have access to gold entity mentions in training and inference and, because we only use ENTSUM in evaluation only, we do not have gold reference entity-centric summaries. We create proxies as above for the input and output in training as follows:

- for each training and testing (document, entity) pair, we feed the full document and as guidance input either the mention string (**GSum**$_{ent-name}$) or the sentences that mention the given entity (**GSum**$_{ent-sent}$) as detected by our NER and coreference approach previously described;
- the output summary for each (document, entity) training pair is obtained from the reference entity-agnostic summary as follows: (a) Select at most 3 sentences in the reference that mention the entity; (b) If we obtain less than 3 sentences in the previous step, then select the remaining sentences from the lead 3 sentences that mention the given entity.

Note this GSum setup can be used with gold entity mentions, sentences and output if ENTSUM data is used in training or development.

4.2 Extractive Methods

Extractive summarization methods aim to extract the segments (in this case, sentences) from the original document to form a summary.

4.2.1 Heuristics

Selecting the top sentences in a document is a strong heuristic for the document summarization tasks (Nallapati et al., 2017). We evaluate the following variants:

Lead3$_{ovr}$ is a generic summarization method that selects the first three sentences in the document irrespective of the target entity.

Lead3$_{ent}$ is the entity-aware summarization variant which selects the first three sentences in the document that mention the given entity, as inferred by our NER and coreference resolution approach.

4.2.2 BERTSum

BERTSum obtains near state-of-the-art results for extractive summarization (Liu and Lapata, 2019). The method uses the BERT (Devlin et al., 2019) encoder to generate representations for each sentence, then models the interactions between these sentences through a BERTSum summarization layer and then predicts the most important sentences from these as the sentences to be part of the generic summary. We evaluate on both all and top 3 predicted sentences to make fair comparisons with Lead3 baselines.

4.2.3 Adapting BERTSum for Entity-Centric Summarization

We adapt BERTSum in the training phase by restricting the input only to all the sentences containing the entity string mention and its coreferent mentions, instead of the entire source document. In training, the output entity-centric summary is constructed in a similar way to the GSum training procedure, where we use the generic summary to select top 3 sentences that mention the entity or otherwise up to 3 sentences that mention the entity.

4.2.4 Heuristics using Oracle Sentence Information

Most previous approaches make the realistic assumption that gold entity mentions or other entity-related annotations are not available at inference time. To explore the impact of these, we explore the following additional heuristics:

Oracle - Lead3$_{ent}$ (salient) uses as summary the first three salient sentences selected by annotators during the second step of the annotation pipeline.
Oracle - Lead3$_{ent}$ (summary) uses as summary the first three sentences selected by annotators for writing the summary.

We expect these to have high performance given the extractive nature of ENTSUM and that these tasks were a prerequisite to writing the summary.

5 Experimental Setup

5.1 Training Data

We train all non-entity-centric methods on the NYT corpus consisting of 44,382 training and 5,523 validation (document, summary) pairs as specified in Kedzie et al. (2018). However, this data set size increases to 464,339 training and 58,991 validation pairs when training the adapted GSum and BERTSum as each document contains multiple entities resulting in multiple <document, summary> pairs for a single document.

5.2 Implementation Details

We use the author's implementations for the following methods: CTRLSum,[3] BERTSum,[4] and GSum.[5] We reimplement the ConvNet method using the FairSeq library (Ott et al., 2019) as described in Fan et al. (2018). For all our implementations, we first train on the CNN DailyMail data set and compared to published numbers to ensure we are able to reproduce the original results and then retrain on the NYT data set for reporting our results on ENTSUM.

We experiment with various hyperparameter settings for each of the architectures but we find that the original hyperparamters used for training each of the CNN DailyMail models seem to be the most stable and produce the best results.

5.3 Evaluation

We automatically evaluate the quality of the generated summaries using unigram and bigram overlap (ROUGE-1 and ROUGE-2), which are a proxy for assessing informativeness and use the longest common subsequence (ROUGE-L) to measure fluency (Lin and Hovy, 2003). We also use BERTScore (Zhang et al., 2020) to compute a similarity score for each token in the generated summaries with each token in the reference summaries using contextualized word embeddings provided by BERT (Devlin et al., 2019). BERTScore incorporates semantic information behind sentences, thus can provide better evaluations for cases where ROUGE score fails to account for meaning-preserving lexical and semantic diversity. BERTScore showed to have better correlations with human judgments for natural language generation (Zhang et al., 2020). For the samples in ENTSUM where we have multiple reference summaries, we take the maximum ROUGE or BERTScore scores. We also report the average sentence and word lengths of the generated summaries to observe summary statistics for the behavior of the output, as automated metrics are sensitive to summary length.

6 Results

We benchmark all methods described above on the newly proposed ENTSUM data set in order to establish baseline performance of both abstractive and extractive methods for this new task and data set. Table 3 shows the automatic evaluation results.

The results show the following trends across all four evaluation metrics:

Entity-centric summarization is very different to generic summarization given that methods that do not take entity information into account (Lead3$_{ovr}$, GSum$_{ovr}$) perform significantly lower than the best methods in the same class which use entity information.

Previously introduced methods (ConvNet, CTRLSum) for controllable summarization can not perform well on entity-centric summarization with their results being over 17 BERTScore and 29 ROUGE-L lower than the proposed adaptation for abstractive summarization on entity-centric summaries. Further, these methods actually obtain lower results by 4.93 BERTScore and 7.43 ROUGE-L than the entity-agnostic GSum$_{ovr}$ method, which shows these methods are not effective at modelling entity-centric information through their training and inference process.

Our proposed adaptations to both abstractive and extractive methods perform well on entity-centric evaluation, despite they were trained on a data set that used proxies for entity-centric summaries. For extractive summarization BERTSum$_{ent-top3}$ performs better than

[3] https://github.com/salesforce/ctrl-sum

[4] https://github.com/nlpyang/BertSum

[5] https://github.com/neulab/guided_summarization

	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**Avg. Len** **Sent. / Word**
Extractive Summarization Methods					
Lead3$_{ovr}$	34.44	19.14	30.97	58.32	3.0 / 99.38
Lead3$_{ent}$	**68.41**	**60.51**	**65.03**	**80.08**	2.76 / 92.31
BERTSum$_{ovr}$	33.8	17.79	30.17	58.24	3.03 / 110.0
BERTSum$_{ovr-top3}$	33.6	17.6	29.9	57.99	2.94 / 105.78
BERTSum$_{ent}$ (Ours)	65.9	58.7	62.8	77.67	4.26 / 128.39
BERTSum$_{ent-top3}$ (Ours)	**67.8**	**59.7**	**64.4**	**77.89**	2.49 / 81.53
Abstractive Summarization Methods					
ConvNet	28.92	13.52	25.85	54.72	3.93 / 102.07
CTRLSum	32.50	17.58	29.87	58.07	4.33 / 110.69
GSum$_{ovr}$	40.29	24.87	37.3	63.00	3.60 / 74.31
GSum$_{ent-name}$ (Ours)	51.71	40.49	48.75	70.11	3.63 / 111.0
GSum$_{ent-sent}$ (Ours)	**61.45**	**52.04**	**58.37**	**75.87**	3.33 / 99.62
Methods using Oracle Entity Sentence Information					
Lead3$_{ent}$ (Salient)	75.67	69.28	72.39	85.14	2.73 / 91.31
Lead3$_{ent}$ (Summary)	85.22	80.49	82.21	91.48	2.53 / 86.0

Table 3: Automatic evaluation results of different summarization models on the ENTSUM data set. **Bold** typeface denotes the best performance within a class of methods.

BERTSum$_{ovr}$ by 34.23 ROUGE-L and by 19.65 on BERTScore, while for abstractive summarization GSum$_{ent-sent}$ is better than GSum$_{ovr}$ by 21.07 ROUGE-L and 12.87 BERTScore. We also see that the choice of guidance signal in the GSum framework is impactful, with using sentences with entities leading to 9.62 ROUGE-L and 5.76 BERTScore improvements over using the entity name.

Extractive approaches perform better than abstractive methods, which is expected due to the extractive nature of the ENTSUM data set, the gap between the best performing methods (BERTSum$_{ent-top3}$ and GSum$_{ent-sent}$) is clear, when using BERTScore (+2.02) which better estimates semantic similarity opposed to the n-gram matches used in ROUGE (+7.66 on ROUGE-2, +6.03 on ROUGE-L).

Lead3$_{ent}$ is a very strong baseline as expected, because this is a strong baseline for document summarization in general and especially because ENTSUM is by design a more extractive summarization data set.

Lead3 using oracle selected sentences perform much better than Lead3 and shows the benefits of selecting salient sentences (+7.36 ROUGE-L, +5.16 BERTScore) and the benefits of selecting the most important sentences used in writing the summary (further +9.82 ROUGE-L, +6.26 BERTScore compared to top salient sentences).

The absolute results also show there is **further room for improvement in entity-centric summarization** approaches, given that performance of automated methods still lags behind Lead3$_{ent}$, whereas this is currently surpassed by automated methods in generic summarization.

7 Conclusion

We introduced the first annotated data set (ENTSUM) for controllable summarization where entities are targets for control. We conducted a quantitative analysis of the newly created resource and highlighted how this is different to generic summarization methods. We used the ENTSUM data set for benchmarking state-of-the-art generic abstractive and extractive summarization methods, as well as initial methods for controllable summarization. Further, we proposed a new setup for learning entity-centric summaries from generic summarization data sets and, extending previous methods, demonstrated good performance on the newly proposed ENTSUM data set.

In the future, we aim to propose new methods for both extractive and abstractive summarization performance through modelling information about the document and the entity in a more complex way. We also plan to create a data set for entity-centric summarization that is more abstractive in nature.

Acknowledgements

We would like to thank Chen-Tse Tsai, Umut Topkara and the other members of the NLP team and the broader Bloomberg AI group, who provided invaluable feedback on the task framing and experiments. We wish to thank Wei Xu for supporting the collaboration. We are grateful to our annotators for their diligence in performing this annotation task.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2662–2672, Online. 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.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

PVS Avinesh, Benjamin Hättasch, Orkan Ozyurt, Carsten Binnig, and Christian M Meyer. 2018. Sherlock: A system for interactive summarization of large text collections. *Proceedings of the VLDB Endowment*, 11(12).

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Hoa Trang Dang. 2006. Duc 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, SumQA '06, page 48–55, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Milan Dojchinovski, Dinesh Reddy, Tomáš Kliegr, Tomáš Vitvar, and Harald Sack. 2016. Crowdsourced corpus with entity salience annotations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3307–3311, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

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

Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible, and Patrick Pantel. 2013a. Identifying salient entities in web pages. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2375–2380.

Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible, and Patrick Pantel. 2013b. Understanding document aboutness step one: Identifying salient entities. *Microsoft Research*.

Donna Harman and Paul Over. 2004. The effects of human variation in DUC summarization evaluation. In *Text Summarization Branches Out*, pages 10–17, Barcelona, Spain. Association for Computational Linguistics.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Eran Hirsch, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan. 2021. iFacetSum: Coreference-based interactive faceted summarization for multi-document exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 283–297, Online and Punta Cana, Dominican Republic. 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.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. Towards a reliable and robust methodology for crowd-based subjective quality assessment of query-based extractive text summarization. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 245–253, Marseille, France. European Language Resources Association.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Semsum: Semantic dependency guided neural abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8026–8033.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

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

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.

Lei Li, Wei Liu, Marina Litvak, Natalia Vanetik, Jiacheng Pei, Yinan Liu, and Siya Qi. 2021. Subjective bias in abstractive summarization.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018a. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

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, Zhiyi Luo, and Kenny Zhu. 2018b. 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, Brussels, Belgium. 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 *AAAI Conference on Artificial Intelligence*.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Itsumi Saito, Kyosuke Nishida, Kosuke Nishida, and Junji Tomita. 2020. Abstractive summarization with combination of pre-trained sequence-to-sequence and saliency models. *CoRR*, abs/2003.13028.

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.

Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2021. Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 657–677, Online. Association for Computational Linguistics.

Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. 2016. Sel: A unified algorithm for entity linking and saliency detection. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, pages 85–94.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.