DESCGEN: A Distantly Supervised Dataset
for Generating Abstractive Entity Descriptions

Weijia Shi[1], Mandar Joshi[1], Luke Zettlemoyer[1]
[1]Paul G. Allen School of Computer Science
Engineering, University of Washington, Seattle, WA
{swj0419, mandar90, lsz}@cs.washington.edu

Abstract

Short textual descriptions of entities provide summaries of their key attributes and have been shown to be useful sources of background knowledge for tasks such as entity linking and question answering. However, generating entity descriptions, especially for new and long-tail entities, can be challenging since relevant information is often scattered across multiple sources with varied content and style. We introduce DESCGEN: given mentions spread over multiple documents, the goal is to generate an entity summary description. DESCGEN consists of 37K entity descriptions from Wikipedia and Fandom, each paired with nine evidence documents on average. The documents were collected using a combination of entity linking and hyperlinks to the Wikipedia and Fandom entity pages, which together provide high quality distant supervision. The resulting summaries are more abstractive than those found in existing datasets, and provide a better proxy for the challenge of describing new and emerging entities. We also propose a two-stage extract-then-generate baseline and show that there exists a large gap (19.9% in ROUGE-L) between state-of-the-art models and human performance, suggesting that the data will support significant future work.[1]

1 Introduction

Entity knowledge has been shown to play an important role in various applications including language modeling (Peters et al., 2019), open-domain question answering (Xu et al., 2016), and dialogue generation (Qin et al., 2019). Recent studies suggest that such entity knowledge can be provided by simple textual descriptions (Chen et al., 2019), which can be incorporated to improve downstream task performance (Nie et al., 2018; Logeswaran

Doc 1
...Are bitcoins, then, really worth anything? According to **Carl Menger's** subjective theory of value, they are worth whatever individuals choose to believe they are worth. It is clear that many individuals value this new medium of exchange highly...
Doc 2
...The Austrian School of Economics has its roots outside of Austria — particularly in the French economists Jean Baptiste Say and Claude-Frederic Bastiat. The Austrian School proper began with **Carl Menger**, who challenged the British labor theory of value. To learn more about Austrian Economics go to the website of The Ludwig von Mises Institute...
Doc 3
...Karl Menger was born on January 13, 1902, in Vienna. His father was the famous Austrian economist **Carl Menger** (1840–1921) who was one of the founders of marginal utility theory....
Entity Description
Carl Menger (February 23, 1840 – February 26, 1921) was an Austrian economist and the founder of the Austrian School of economics. He contributed to the development of the marginal utility theory and to the formulation of a subjective theory of value.

Table 1: An example from DESCGEN exhibiting the diversity of source documents and the abstractive nature of the entity description summaries.

et al., 2019). However, manually curating entity descriptions is labor-intensive and it is challenging to keep pace with the ever growing emergence of new entities. In this paper, we present a new dataset DESCGEN for automatically generating entity descriptions from relevant documents and mentions, which provides high quality supervision for a highly abstractive version of this task that targets early description of new entities as they emerge. For example, in Table 1, machines are required to generate a description of *Carl Menger*, given multiple documents mentioning him.

DESCGEN contains 37K entity descriptions extracted from Wikipedia and Fandom[2]. Fandom

---

allows us to capture the key challenge of generating descriptions for emerging entities that are not in Wikipedia because they are less popular or have just been introduced to the public. To obtain source documents of the entities, we collect web documents and news articles where entity mentions are linked using web hyperlinks or an entity linker. Our dataset is *distantly supervised* in that these heuristically collected documents are not guaranteed to contain all the facts required to generate the description—as would be seen for natural text collections describing emerging entities. We also carefully annotate a subset of 1,000 examples to support more reliable evaluation (see Table 2 for dataset statistics).

Unlike multi-document summarization that makes the assumption that a set of documents to be summarized are written on the same topic (Zopf et al., 2016), DESCGEN only assumes that source documents mention the entity. In contrast to an existing entity summarization benchmark (Liu et al., 2018, WikiSum), DESCGEN is more abstractive and better approximates challenges faced when describing new entities. Section 4.4 provides more details on these comparisons. Overall, our documents for generating a description can cover a much wider range of topics as well as text genres, including news, blog posts, and scientific articles. For instance, the documents 1 and 2 mentioning *Carl Menger* in Figure 13 discuss topics on bitcoins and the Austrian School of Economics.

Finally, we also propose a two-stage method that first extracts salient sentences relevant to the entity and then abstracts them into a description. We test a range of models to establish baseline results with both automatic and human evaluation. The best model based on BART (Lewis et al., 2020b) achieves 28.2% in the ROUGE-L F measure with a significant gap compared to the human performance 48.1%, suggesting there was great room for future improvement. In summary, our contributions include:

- We propose a new dataset DESCGEN that includes challenging, abstractive entity summaries. Our dataset contains over 37K pairs of entity descriptions and their associated documents, along with a human-annotated subset of 1,000 pairs.

- We conduct an extensive analysis of properties of the dataset and identify its challenges—extractive content selection from large

	Wikipedia	Fandom
Entities	26,585	11,366
Documents	177,454	170,204
Input size	11,568	1,872
Output size	53	32
Human-authored descriptions	598	403

Table 2: Basic statistics for DESCGEN. Input size and output size refer to the average number of words in the description and source documents respectively.

amounts of text and abstractive generation from it, particularly for emerging entities.

- We present a two-stage method and benchmark various models on our dataset, aiming to facilitate future work on this dataset.

2 Related work

Existing Entity Description Generation Task and Dataset Previous works (Novikova et al., 2017; Cheng et al., 2020; Trisedya et al., 2020) mainly take as input some structured data such as knowledge graphs to generate entity descriptions. However, knowledge graphs, often mined from text corpora, are overwhelmingly incomplete on real-world entities and may not be updated in real-time (Dong et al., 2014). Therefore, we focus on generating descriptions from natural language sources such as web texts and news because they are often primary sources for entities and have better coverage of entities across multiple domains. DESCGEN is most related to WikiSum, a recent dataset for generating Wikipedia summaries from textual sources (Liu et al., 2018). WikiSum source documents primarily come from high-quality articles cited in the Wikipedia pages which makes their data more extractive (Section 4.4). In contrast, we collect our source documents heuristically using web texts and news, providing a better proxy for emerging entities where high-quality citation sources may not be available. In addition, their evaluation is conducted only on distantly supervised test data. However, our experiments demonstrate that manually annotated data allows for much better evaluation of model performance (Table 7).

Multi-document summarization aims to condense a cluster of thematically-related documents into a short and informative summary. A wide range of multi-document summarization datasets have been built for the Document Understanding and Text Analysis Conferences (Over and Yen,

2004; Owczarzak and Dang, 2011), news (Fabbri et al., 2019), events (Gholipour Ghalandari et al., 2020) and Wikipedia summaries (Liu et al., 2018). Recent work has studied both extractive (Yasunaga et al., 2017; Nallapati et al., 2017; Tohalino and Amancio, 2018) and abstractive summarization (Banerjee et al., 2015; Chali et al., 2017; Nayeem et al., 2018). However, existing datasets typically are not entity focused and assume the input documents are at least loosely centered around a coherent topic or event.

Wikipedia generation Our work is also related to research on generating Wikipedia articles. For instance, Sauper and Barzilay (2009) learn to build content templates using an integer linear program to generate full articles. Similarly, Banerjee and Mitra (2016) generate Wikipedia pages by building a topic classifier to assign web retrieved contents into relevant sections. We focus on a different task – generating a short text description that can identify and best summarize an entity.

3 Dataset Collection

Task definition Given a collection of documents $D = \{D_i | i = 1...n\}$ with mentions linked to the same entity e, the goal is to generate a description of e. For example, Table 13 shows a description of an entity (*Carl Menger*) and three source documents with mentions.

Distant supervision We make use of existing knowledge bases, such as Wikipedia and Fandom, to collect entity descriptions. To obtain source documents and mentions for each entity, we use a combination of hyperlinks to Wikipedia pages and an entity linker that links entity mentions in text. Our dataset is *distantly supervised* in that these heuristically collected documents are not guaranteed to contain all the facts required to generate the description. To analyze the quality of distant supervision, we collect a smaller *verified* set of entity descriptions using human annotators. In contrast with our work, WikiSum (Liu et al., 2018) used documents cited in the Wikipedia pages or web pages returned by Google as source documents to generate Wikipedia lead sections. Because high-quality citation sources constitute a substantial part of overall documents (75%), their dataset is less abstractive than DESCGEN and unsuited for emerging entities where citations are not available.

Sources We paired entity descriptions with source documents from three sources: Wikilinks, RealNews, and Fandom using distant supervision. To capture the challenge of emerging entities, we retrieve source documents that are not in Wikipedia using Wikilinks and RealNews. We also include specialized entities in Fandom that do not have Wikipedia pages. For quality control, we filter out entities for which the unigram recall of the entity description against its concatenated source documents is lower than 0.6.

3.1 Distantly supervised data collection

Wikilinks Wikilinks (Singh et al., 2012) is a large dataset designed for cross-document coreference. It consists of non-Wikipedia web pages (discovered using the Google search index) containing entities that are hyperlinked to Wikipedia. For each entity, we retrieve a collection of web pages in Wikilink with the anchor text linked to it and use the lead section of target Wikipedia page as its description. We further parse the HTML texts of the web pages and extract contents as source documents.

Real News To expand the collection of source documents, we extract entity mentions in RealNews (Zellers et al., 2019), a large corpus of news articles from Common Crawl. We first conduct a longest prefix match between the entity surface form and text tokens via trie, a prefix tree structure that supports efficient string searching. More specifically, we build a trie of entity names where each node is a word and its children indicate all possible continuations from the prefix. After retriving candidates for entity mentions, we use an off-the-shelf entity linking model (Gupta et al., 2017) to rank the candidates and add the corresponding news articles as source documents of the rank-1 candidate.

Fandom Fandom[3] is a collection of encyclopedias, centered around particular subjects and themes such as movies, TV shows, and games. It contains specialized entities that require domain experts with background knowledge to make edits. Entities and their source documents can be automatically extracted by internal links. We filter out entities and only keep those without Wikipedia pages, which can be viewed as new or emerging entities. The description of the entity is extracted

---

[3]https://www.fandom.com/

417

Entity source		Train	Dev	Test
Wikipedia	Distant	21,267	2,659	2,659
(Wikilinks + Real news)	Verified	-	299	299
Fandom	Distant	9,092	1,137	1,137
	Verified	-	202	201

Table 3: Number of entities for train, dev and test set.

Metrics	R-1	R-2	R-L	METEOR
IAA	45.8	36.1	47.7	23.3

Table 4: Inter-annotator agreement (IAA) for human-authored descriptions.

from the lead section of its Fandom page. We collect data from the 32 largest Fandom Wikis.

3.2 Human-authored entity descriptions

Entity descriptions extracted from Wikipedia and Fandom have been authored and edited by multiple community contributors largely independently of our source documents. We collected additional entity descriptions via Upwork,[4] a freelancing platform, to better analyze how descriptions sourced from documents in our dataset contrast with those from Wikipedia and Fandom. We provided the entity and its source documents to annotators on Upwork, and asked them to write the entity descriptions. The annotators are also asked to mark sentences they used to write the description. Each entity was assigned to 2 annotators. We collected 500 entity descriptions for dev examples and 500 descriptions for test examples.

We control the quality of the crowdsourced descriptions by filtering annotators who produced low-quality descriptions. We ask every candidate to annotate the same 20 examples and use two criteria for narrowing down candidates: (1) missing key information in descriptions (2) unjustified information in descriptions that cannot be inferred from source documents alone. Eventually, we filtered out 4 annotators and accepted 7 qualified annotators. The total annotation cost was around $3500.

3.3 Experimental setup

All 37K entity description and document pairs in the dataset are randomly split into train, development and test sets. In addition to automatically collected descriptions from Wikipedia and Fandom, we use the human-authored descriptions (Section 3.2) as verified subsets into dev and test splits. Table 3 shows basic statistics of the final dataset. We report model performance on automatically collected descriptions (distant) and human-authored descriptions (verified).

The next section provides a detailed analysis of the data quality, including annotator agreement and
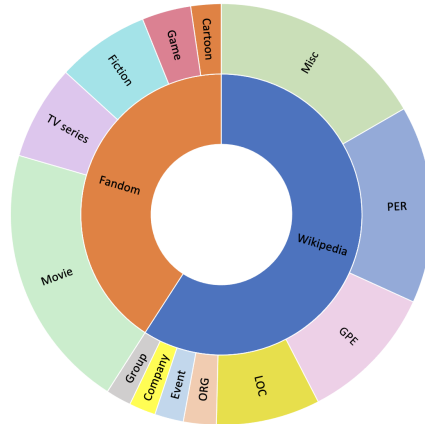

Figure 1: Distribution of entity domains (outer level) and knowledge sources (inner level).

other aggregate statistics.

4 Dataset Analysis

An analysis of the data shows that DESCGEN contains a high proportion of emerging entities from diverse domains, and is more extractive compared to other multi-document summarization datasets.

4.1 Statistics

Table 2 shows data statistics. DESCGEN contains about 37K entity descriptions from Wikipedia and Fandom. On average, each entity has nine source documents. We can see that 36% percent of entities come from Fandom, and therefore have never had a Wikipedia page written about them.

Domain diversity Figure 1 shows that DESC-GEN covers a diverse set of entity domains. For analysis, we associate entities in Wikipedia with domains (GPE, LOC, PER, ORG, EVENT, COMPANY, GROUP and MISC) by querying the DBPedia knowledge-base (Lehmann et al., 2015). Each entity in Fandom is manually categorized into 5 domains: movie, game, fiction, TV series and cartoon based on its source Wiki. An analysis of baseline performance by entity type and domain (Section 7.3) reveals a notable drop for less popular domains such as Games and Fiction, highlighting generalization challenges.

---

[4]https://www.upwork.com/

418

	R-1	R-2	R-L
Wikipedia	34.7	17.8	35.8
Fandom	45.6	27.8	44.5

Table 5: Rouge results on human reference against Wikipedia/Fandom descriptions.

Category	Paraphrasing	Missing info.	Extra details
Wikipedia	29	16	22
Fandom	32	15	26

Table 6: Number of times a human-authored description is classified into error categories with Wikipedia/Fandom descriptions as reference. The sample size is 40.

4.2 Inter-annotator agreement

Each entity in the verified subset has two descriptions written by two annotators. Following previous work (Chen et al., 2015), we quantify inter-annotator agreement on descriptions by treating one of the descriptions as the prediction and the other as the reference to compute ROUGE (Lin, 2004) and METEOR (Denkowski and Lavie, 2014). Table 4 shows high inter-annotator agreement of 47.7 in terms of ROUGE-L.

We additionally measure the agreement on content selection using sentences marked by annotators. In particular, agreement is achieved when both annotators selected the exact same sentences in all source documents for an entity. Cohen's Kappa is 0.38, which indicates high agreement (Brennan and Prediger, 1981) considering the strict criterion of reaching agreement.

4.3 Comparison between human-authored and Wikipedia/Fandom descriptions

To understand how human-authored descriptions differ with Wikipedia and Fandom descriptions in terms of content and style, we compare them using automatic metrics (ROUGE) and manual evaluation.

ROUGE Table 5 shows the averaged ROUGE scores of human-authored descriptions against Wikipedia and Fandom descriptions. Human-authored descriptions have higher word overlap with Wikipedia descriptions than with Fandom descriptions.

Pairwise comparison Can humans distinguish between Wikipedia/Fandom and human-authored descriptions? We have two human assessors evaluate 50 randomly sampled pairs of human-authored and Wikipedia/Fandom descriptions in a blind pairwise comparison, and ask them to classify descriptions into two categories: human-authored or Wikipedia/Fandom. The classification accuracy in Wikipedia and Fandom is 64.4% and 61.1% respectively and the inter-annotator agreement is 0.67 in Cohen's Kappa. The relatively low classification accuracy suggests that there is no substantial

quality and style difference in human-authored and Wikipedia/Fandom descriptions.

Quality analysis of distant supervision We are interested in understanding if automatically gathered documents can provide enough signals for writing the entity descriptions. To study the quality of distant supervision, we manually analyze 40 human-authored descriptions that have low n-grams overlap with Wikipedia/Fandom descriptions, in terms of *paraphrasing* (does the human-authored description express the same meaning but use different words?), *missing information* (does the human-authored description miss any information in Wikipedia/Fandom description?) and *extra details* (does the human-authored description contain extra details not included in the Wikipedia/Fandom description?). We use Wikipedia and Fandom descriptions as the ground truth and classify each human-authored description into one or more categories. The results are shown in Table 6. We find that the difference between the two sources of descriptions are mainly caused by paraphrasing and missing information. This suggests that even for entities that have very different human-authored and extracted descriptions, most of the information in the Wikipedia/Fandom descriptions is present in the documents.

4.4 Extraction vs abstraction

Generating entity descriptions involves extracting essential information about the entity and condensing them into a short description. To measure how much DESCGEN requires paraphasing and compressing, we quantify the extractive nature of our dataset by the measuring extractive fragment coverage and density defined in Grusky et al. (2018). Extractive fragment coverage computes the percentage of words in summary that appear in source documents:

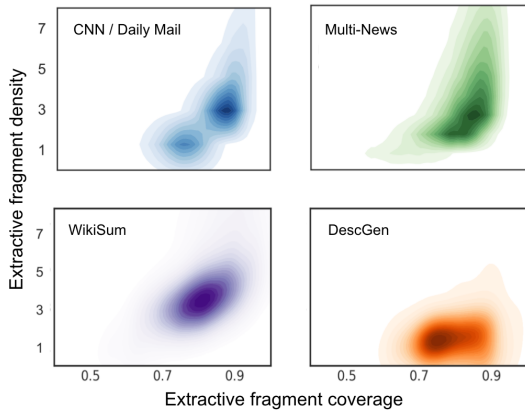$$Coverage(A, S) = \frac{1}{|S|} \sum_{f \in F} |f|$$

Figure 2: Density and coverage on different datasets. Large variability in y-axis reflects the variation in average length of shared token sequences.

where A is a concatenation of the source documents, S is the description and F is the set of shared token sequences in A and S. Likewise, extractive fragment density is related to the average length of shared token sequences. For example, an entity description with high coverage and low density shares many individual words with source documents but almost no long phrases.

$$Density(A, S) = \frac{1}{|S|} \sum_{f \in F} |f|^2$$

We compare our dataset with several multi-document summarization datasets, including CNN / Daily Mail, Multi-News (Fabbri et al., 2019) and WikiSum (Liu et al., 2018). Figure 2 presents the density and coverage distribution. The density of Multi-News, CNN / Daily Mail and WikiSum are high, showing that there is much copying of long sequences with respect to source documents. DE-SCGEN shows high coverage but low density, suggesting it is not common to copy long sequences and the data overall is much more abstractive.

5 Baselines

In this section, we introduce several new baseline methods, building on state-of-the-art pre-trained models. The input documents can be long (Section 8), making it computationally infeasible to train end-to-end models. We instead introduce a pipelined approach to generate an entity description in two stages. In the first extractive stage, a selector is used to identify representative sentences relevant to the entity from multiple source documents. In the second abstractive stage, a neural generation

model is used to fuse the selected sentences to a description of the entity. We compare a number of different approaches for each stage, as summarized in the subsections below.

5.1 Extractive stage

Trivial concatenates all sentences that mention the entity, along with one sentence before and after each. The content is truncated to the first 1,000 tokens to fit the token limit of models in the abstractive stage.

Cheating ranks sentences according to their unigram recall against the description and selects the top 15 sentences. This heuristic demonstrates the effect of extraction on final performance.

BERT (Devlin et al., 2019) with a classifier uses a linear layer stacked on top of the BERT outputs and predict whether a sentence should be selected. The model is trained on our training dataset in which sentences are labeled by the cheating method.

5.2 Abstractive stage

We compare three pre-trained language generation models, including BART (Lewis et al., 2020b), T5 (Raffel et al., 2019) and MARGE (Lewis et al., 2020a) to generate abstractive entity descriptions. We fine-tuned these models on our training dataset in a sequence-to-sequence fashion.

T5 is a text-to-text transformer pre-trained on a multi-task mixture of unsupervised and supervised tasks. We consider models of two sizes: base and large containing 220M and 770M parameters respectively. We use the Hugging Face version.[5]

BART introduces a denoising autoencoder combining a bidirectional encoder and auto-regressive decoder. It is trained by reconstructing text corrupted with a noising function. We consider the base model with 139M parameters.

MARGE is a multi-lingual sequence-to-sequence model trained by reconstructing target documents retrieving paraphrased documents in other languages. It has around 960M parameters.

6 Experiments

6.1 Evaluation metrics

Following other summarization tasks, we evaluate the quality of generated descriptions by ROUGE

---

[5]https://github.com/huggingface/transformers

Extract.	Abstract.	Distant supervision						Verified					
		Dev			Test			Dev			Test		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Trivial	BART	24.5	11.3	22.1	23.4	10.7	23.8	27.2	14.2	26.4	27.1	15.9	26.6
	T5-Base	21.7	9.3	20.6	21.1	10.5	21.1	25.1	12.8	24.7	24.9	13.2	23.7
	T5-Large	23.9	12.7	23.4	24.2	11.1	23.5	27.7	15.9	27.2	26.9	15.6	27.3
	MARGE	23.2	10.6	21.8	23.0	10.2	22.1	26.4	13.9	25.8	26.2	14.0	25.8
BERT	BART	**26.9**	13.9	**27.6**	**26.3**	13.2	**26.6**	**28.9**	**16.9**	27.3	26.7	**16.4**	**28.2**
	T5-Base	23.4	10.1	23.9	23.0	11.6	24.4	24.9	11.7	24.1	25.0	12.2	24.8
	T5-Large	26.8	**15.1**	27.4	25.4	**14.8**	25.9	27.1	16.6	**27.5**	**27.3**	16.1	27.3
	MARGE	25.1	13.8	26.2	24.9	11.9	25.0	26.7	15.7	25.9	26.3	14.8	25.8
Human Performance		40.7	21.9	39.9	39.1	21.8	39.3	45.2	36.7	48.7	45.3	35.4	48.1

Table 7: Experimental results of different baselines evaluated on distantly supervised and verified dev/test sets.

Method	Distant supervision				Verified			
	Dev		Test		Dev		Test	
	Uni.	Bi.	Uni.	Bi.	Uni.	Bi.	Uni.	Bi.
Trivial	60.5	23.8	59.9	23.4	78.8	50.4	76.9	43.2
BERT	65.1	26.1	66.9	27.7	80.4	50.6	77.5	43.8
Cheating	72.4	31.7	72.3	31.4	81.6	51.9	79.2	44.6

Table 8: Unigram (Uni.) recall (%) and bigram (Bi.) recall (%) for extractive methods.

Models	BART	T5-Large	T5-base
Non-redundancy	3.8	3.5	3.6
Fluency	4.6	4.7	4.6
Informativeness	3.5	3.2	3.1
Faithfulness	2.7	2.5	2.6

Table 9: Manual evaluation scores on a scale from 1 (very poor) to 5 (very high). All these models use BERT in the extractive stage.

F1-score (Lin, 2004), which measures the overlap of unigram (R-1), bigram (R-2), and the longest matching sequence of words (R-L). In addition, we evaluate content selection by unigram and bigram recall to assess the importance of the extractive stage. Lastly, in addition to automatic evaluation, we also conduct human evaluation for non-redudancy, fluency, informativeness, and accuracy.

6.2 Experimental results

Automatic evaluation In Table 8, we report the experimental results in the extractive stage. We observe that BERT consistently outperforms the unsupervised method Trivial, suggesting that training a model to predict sentence relevance can bring in immediate improvement in content selection. Meanwhile, the performance of BERT still lags behind the upper bound defined by Cheating by 1.7-7.3% in unigram.

Table 7 presents ROUGE scores of various baselines in the abstractive stage. T5-large and BART show similar performance and outperform other models for both distant supervision and verified subsets, by a small margin. Increasing model size from T5-base (220M) to T5-large (770M) parameters leads to a relatively large performance gain. The human baseline is superior to all the models and maintains a R-L score over 33 in distant supervision and 48 in the verified subset. The large gap between the human baseline and the best-performing model shows there is much room for future work.

Manual evaluation We present two human assessors with source documents and descriptions generated from different abstractive models and asked them to rate descriptions in terms of *non-redundancy* (does the description avoid repeating information?), *fluency* (Is the description well-formed and gramatically correct?), *informativeness* (does the description capture the salient information about the entity?) and *faithfulness* (Is the description faithful to the source text?). We compared BART, T5-Large, and T5-Base. For each model, we selected 100 descriptions and showed outputs of models to assessors side by side without revealing which model generates them. The score for each description was averaged between two assessors. As can be seen from Table 9, BART shows strong performance on all dimensions, except for fluency. Overall, all three models can generate fluent descriptions (high *fluency*) but struggle with producing accurate statements (low *faithfulness*). In most cases of low faithfulness, we observe that the model directly copies words from the input that are not relevant to the entity as part of the description or synthesize information that are not directly inferable from the input.

	Wikipedia description
	Wikipedia description
	Carl Menger (February 23, 1840 – February 26, 1921) was an Austrian economist and the founder of the Austrian School of economics. He contributed to the development of the marginal utility theory and to the formulation of a subjective theory of value.
	Human-authored description
	Carl Menger is an Austrian economist and one of founders of Marginal Utility Theory. He challenged the British labor theory of value and proposed subjective theory of value. He founded the Austrian School of Economics.
	BART
	Carl Menger was an Austrian economist and one of the founders of marginal utility theory.
	T5-Base
	Carl Menger was born on January 13, 1902, in Vienna. He was one of the founders of marginal utility theory.
	T5-Large
	Carl Menger was an Austrian economist.
	MARGE
	Carl Menger (born January 13, 1902) was an Austrian economist.

Table 10: Entity descriptions for *Carl Menger* generated by different models. Red text indicates incorrect information in predictions while green text indicates information in the Wikipedia and human-authored descriptions that was not covered by any of the model predictions.

7 Analysis

In this section, we perform qualitative and quantitative analysis of baseline results to better understand strengths and weaknesses of models, and hypothesize avenues for future work.

7.1 Case study

A qualitative analysis of model predictions suggests that these models tend not to generate novel words in the description, and mostly copy words from the original text. The entity-centric nature of DESCGEN makes extractive content selection difficult as evidenced by the gap between BERT extraction and the Cheating model (Section 6.2). For example, Table 10 shows the model-generated entity descriptions for *Carl Menger* using source documents from Table 13. BART, one of the best performing baselines, generates a description that has highest overlap with the Wikipedia description, but it still misses some important facts. T5-Base and MARGE confuse Carl Menger and his son, and incorrectly include information that does not describe the target entity.

Models	Name-only		Regular	
	Fandom	Wiki.	Fandom	Wiki.
BART	12.7	16.6	27.5	28.4
T5-Base	12.5	16.2	25.8	24.5
T5-Large	11.7	16.8	26.1	27.6

Table 11: Rouge-L scores for models evaluated on the verified test set. Name-only and regular refer to models using only the entity name as the input and models using source documents respectively.

Wikipedia	ROUGE-L	**Fandom**	ROUGE-L
GPE	28.6	Movie	28.1
LOC	28.5	Game	22.5
PER	23.7	Fiction	25.3
ORG	26.4	Cartoons	26.4
Event	25.6	TV series	27.6
Group	20.2		
Company	21.4		

Table 12: ROUGE-L scores for BERT+BART evaluated on different entity domains in the verified test set.

7.2 Entity knowledge in pre-trained models

BART, T5, and MARGE are language models pretrained on text corpora including Wikipedia and Common Crawl. The parameters of the models appear to contain substantial linguistic and factual information (Petroni et al., 2019; Peters et al., 2018). In particular, we wonder if entity-related knowledge is captured in the pretraining stage and investigate the following questions: (a) Can the model memorize entity descriptions in pretraining stage? (b) Does the memorized knowledge improve model performance on generating entity descriptions?

To investigate the questions, we test the model's ability to write a description given only the entity name instead of source documents. We train the model on our training dataset to adapt to the style of Wikipedia in a similar way. The results are shown in Table 11. Considering the name-only baselines, we can see that all of them perform worse on Fandom entities than Wikipedia entities. However, the regular baselines perform similarly on Fandom and Wikipedia. This result suggests that facts about entities learnt in pretraining stage have much less influence on model performance when source documents are provided.

7.3 Entity type

To understand how the performance of the models varies with different types of entities, we report the performance breakdown for different entity types in Table 12. Among domains in Wikipedia, our model obtains low scores on group and company, suggest-

ing that they are more challenging than other domains. In Fandom, entities from the game domain prove to be most difficult.

In summary, our analysis suggests there is room for improvement in extractive content selection and abstractive generation, particularly for new and emerging entities from less popular domains.

8 Conclusion

In this work, we introduce DESCGEN, a new dataset for generating entity descriptions from mentions. DESCGEN contains 37K pairs of entity descriptions from Wikipedia and Fandom, and 481K automatically gathered source documents based on distant supervision. We also present a clean human-authored subset of 1,000 pairs for test. We show that, as compared to existing benchmarks, DESCGEN requires more abstractive summaries, which we argue better approximate the challenge of describing emerging entities. We also show that the performance of state-of-art models is far from human levels, suggesting that our task remains a significant challenge with room for improvement. Our study points to an interesting research direction on modeling entity knowledge from contexts. We hope it will facilitate future work on incorporating entity knowledge into downstream tasks and generating descriptions for emerging entities.

References

Siddhartha Banerjee and Prasenjit Mitra. 2016. Wikiwrite: Generating wikipedia articles automatically. In *IJCAI*, pages 2740–2746.

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Multi-document abstractive summarization using ilp based multi-sentence compression. In *IJCAI*.

Robert L Brennan and Dale J Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699.

Yllias Chali, Moin Tanvee, and Mir Tafseer Nayeem. 2017. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 418–424, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mingda Chen, Zewei Chu, Yang Chen, Karl Stratos, and Kevin Gimpel. 2019. Enteval: A holistic evaluation benchmark for entity representations. 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 421–433.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv e-prints*, pages arXiv–1504.

Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. 2020. ENT-DESC: Entity description generation by exploring knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1187–1197, Online. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

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.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *NAACL-HLT*, pages 708–719.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity Linking via Joint Encoding of Types, Descriptions, and Context. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. 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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. *arXiv preprint arXiv:1906.07348*.

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.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Feng Nie, Yunbo Cao, Jinpeng Wang, Chin-Yew Lin, and Rong Pan. 2018. Mention and entity description co-attention for entity disambiguation. In *AAAI*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Paul Over and James Yen. 2004. An introduction to duc-2004.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with KB retriever. 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 133–142, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Christina Joan Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012*, 15.

Jorge V Tohalino and Diego R Amancio. 2018. Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications*, 503:526–539.

Bayu Trisedya, Jianzhong Qi, and Rui Zhang. 2020. Sentence generation for entity description with content-plan attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9057–9064.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on Freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336, Berlin, Germany. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9054–9065.

Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler. 2016. The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1535–1545.

A Appendix

A.1 Experimental Details

All the abstractive models are initialized from the pretrained models. The BART, T5-base and T5-large are adopted by the huggingface framework (Wolf et al., 2020). The MARGE model is adopted by the official authors (Lewis et al., 2020a). We apply the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e - 08$. The learning rate is selected from {1e-3, 0.5e-3, 1e-4, 0.5e-4, 1e-5, 0.5e-5}. The best learning rate for BART, T5-base, T5-large and MARGE is 1e-5, 1e-5, 0.5e-5,0.5e-4. We use beam searching with beam-size 5 as decoding algorithm, which is selected from {5, 10, 15, 20}. We use the batch size of 5 for all models due to memory limit.

A.2 More examples

See next page.

Doc 1
...It sometimes gets confusing in the global village , where technology, finance, cross-cultural interactions, and expanding ethnic diasporas are tearing apart the relationship between borders and making multiple identities possible. Hence, Ang Lee is a Taiwanese artist who directs American films, but he is also an American film director of Chinese movies. As a member of the Sinosphere, enlarged by fifty million overseas Chinese, Ang is not only a creative individual who makes our world more interesting and prosperous. He also helps to bridge between nations and cultures and to produce a Sino-American synergy that is more conducive to peace than a contingency of Chinese and U.S. diplomats...

Doc 2
...The Life of Pi. One of the most interesting film adaptations set for release in 2012 is Brokeback Mountain fame. Suraj Sharma, who has no previous acting experience, will play the central character, Piscine Patel. Based on the novel by Yann Martel, it is being brought to the big screen by Ang Lee...

Doc 3
Comic character Hulk is Dr. Bruce Banner, who becomes a green monster with powerful strength after an experiment went bad, or well, depending on who you ask. In 2003, director Ang Lee's film Hulk brought this character to the big screen, but was poorly received by Hulk's fans...

Wikipedia Description
Ang Lee, (born October 23, 1954, P'ing-tung county, Taiwan), is an Taiwan-born film director who transitioned from directing Chinese films to major English-language productions.

Human-authored Description
Ang Lee is a Taiwanese director who directs American and Chinese films. He is a director of the Life of Pi and Hulk and regarded as Second New Wave of Taiwanese directors.

BART
Ang Lee is a Taiwanese film director and screenwriter.

T5-base
Ang Lee is a Taiwanese film director.

T5-large
Ang Lee is a Taiwanese film directors and screenwriter

Doc 1
...In the summer of 1994, Arthur managed to get himself and his family (as well as Harry and Hermione) tickets for the 1994 Quidditch World Cup from Ludovic Bagman because Arthur had helped Otto Bagman, Ludo's brother, out of a minor scrape. Arthur was among the Weasleys who fetched Harry from the Dursley family via the Floo Network. While there, he expressed his fascination at various Muggle artefacts in the Dursley's house.The group sat in the Top Box, where they were confronted by the Malfoy family, who were there by a personal invitation from the Minister himself, though both Arthur and Lucius were able to restrain themselves out of respect for Cornelius Fudge...

Doc 2
...Before working at the Ministry, he was a Beater for both the Wimbourne Wasps and the English National Quidditch team. He had a brother named Otto Bagman. He also tended to play dirty when gambling and betting as he tried to find loopholes or even pay in fake money/gold...

Doc 3
...A lawn mower is found in the Muggle Studies classroom at Hogwarts School of Witchcraft and Wizardry. Arthur once helped Ludovic Bagman's brother, Otto Bagman, by smoothing over a problem involving a lawn mower enchanted with magical powers. As thanks, Ludo got Arthur prime tickets to the 1994 Quidditch World Cup...

Fandom Description
Otto Bagman was the brother of Ludovic Bagman. He once had a problem with a magical lawn mower, a Muggle artifact. Arthur Weasley helped him out with the problem, and was rewarded by Ludo with tickets to the 1994 Quidditch World Cup final.

Human-authored Description
Otto Bagman is the brother of Ludovic Bagman. He had a problem involving a lawn mower enchanted with magical powers. He was helped by Arthur and gave Arthur prime tickets to the 1994 Quidditch World Cup.

BART
Otto Bagman was a fictional character in the 1994 film Harry Potter.

T5-base
Otto Bagman was an English footballer who played for the Wimbourne Wasps and the English National Quidditch team. He also played dirty when gambling and betting as he tried to find loopholes or even pay in fake money.

T5-large
Otto Bagman was a brother of Ludovic Bagman.

Table 13: Examples of entity descriptions generated by our model. Red text indicates incorrect information in predictions while green text indicates information in the Wikipedia and human-authored descriptions that was not covered by any of the model predictions.