A Well-Composed *Text* is Half Done!
Semantic Composition Sampling for Diverse Conditional Generation

Anonymous ACL submission

Abstract

We propose Composition Sampling, a simple but effective method to generate higher quality diverse outputs for conditional generation tasks, compared to previous stochastic decoding strategies. It builds on recently proposed planning-based neural generation models that are trained to first create a composition of the output using an entity chain and then continue to generate conditioned on the entity chain and the input (Narayan et al., 2021). Our approach avoids text degeneration by first sampling a composition in the form of an entity chain and then using beam search to generate the best possible text grounded to the entity chain. Experiments on summarization (CNN/DailyMail and XSum) and SQuAD question generation tasks, using a wide variety of automatic metrics and human-based evaluation, demonstrate that Composition Sampling is currently the best available decoding strategy for generating diverse meaningful outputs. We further introduce a novel automatic measure for jointly evaluating diversity and faithfulness in summaries.

Haman Written Summary: Chelsea star Eden Hazard is set to make his 100th top-flight appearance. Santi Cazorla should hit the same milestone when Arsenal meet Burnley. Both players have impressed since moving to the Premier League in 2012. Hazard has more goals this season but Cazorla has one more assist. Sportsmail's reporters choose the player who has excited them the most.

[✓] Beam Search: Eden Hazard and Santi Cazorla are both set to make their 100th Premier League appearances this weekend. Both players have been hugely influential since they moved to London. Here, Sportsmail's reporters choose the player they most enjoy seeing in action.

[✗] Nucleus Sampling: Eden Hazard and Santi Cazorlag will each make their 100th Premier League appearance this weekend. nightstandapplication.com. Sportsmail's hovercraft reporters choose their man of the match countermeasures.

[✗] Nucleus Sampling: By making their 100th Premier League appearances this weekend, Eden Hazard and Santi Cazor halibut will set new records. Here, Anna Coren and Dominic King select their favourites.

[✓] Composition Sampling: (Composition: Eden Hazard | Santi Cazorla | Chelsea | Arsenal | Premier League ||| London | 2012 |||) Eden Hazard and Santi Cazorla are set to make their 100th appearances for Chelsea and Arsenal respectively in the Premier League this weekend. Both players have been hugely influential since they moved to London in the summer of 2012. But who has been the most exciting import to watch?

[✓] Composition Sampling: (Composition: Chelsea | Eden Hazard | Arsenal | Santi Cazorla ||| Sportsmail ||| London) Chelsea's Eden Hazard and Arsenal's Santi Cazorla will both make 100th appearances this weekend. Sportsmail's reporters pick the player they most enjoy seeing in action. Both players have been hugely influential since moving to London.

Figure 1: Human written summary, single-best predicted summary using beam search (beam size 8), nucleus sampled ($p = 0.95$) diverse summaries and our composition sampled diverse summaries, for the CNN/DailyMail article shown in Figure 5 in Appendix. We highlight spans in orange that are not faithful to the input article.

1 Introduction

In many NLG tasks, it is important to be able to generate multiple diverse outputs from a model. Tasks like summarization (Mani, 2001; Nenkova and McKeown, 2011) and question generation (Zhou et al., 2017) exhibit one-to-many relationships; there can be multiple semantically diverse summaries or questions for the same source, and it can be useful for a model to generate multiple outputs. Yet, the primary focus of recent research in NLG has been on improving the quality of single-best outputs (Raffel et al., 2019; Lewis et al., 2019; Dong et al., 2019; Zhang et al., 2019; Narayan et al., 2021), generating high quality diverse outputs is still an unsolved problem (Hashimoto et al., 2019; Zhang et al., 2021). This is particularly challenging in conditional generation, where diversity in the target sequences should not come at the cost of correctness or faithfulness; for example, alternate summaries are not valuable if they are unfaithful to the input document(s) (Maynez et al., 2020; Kryscinski et al., 2020). In this work, we investigate decoding methods for generating semantically diverse and faithful summaries and questions for input documents.

Beam search (Li et al., 2016b; Wiseman et al., 2017) has proven to be successful for single-best generation (Rush et al., 2015; Barrault et al., 2020; Meister et al., 2020), but struggles with generation of multiple diverse outputs (Vijayakumar et al.,

2016). Stochastic sampling strategies, such as top-k sampling (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2020), are better at generating more diverse sequences than beam search, but are not suitable for conditional generation as they degenerate,[1] producing output that is not faithful to the source. Figure 1 exposes degeneration in summary output using nucleus sampling.

To address these shortcomings, we propose *Composition Sampling*, a simple but effective hybrid decoding method for diverse and faithful conditional generation. It builds on recently proposed planning-based neural generation models (Narayan et al., 2021) that are trained to first plan a semantic composition of the target using an entity chain and then continue generating the target conditioned on the entity chain and the input. Composition sampling first samples a diverse composition in the form of an entity chain and then uses beam search to generate the best possible sequence grounded to the sampled entity chain. Unlike top-k or nucleus sampling, composition sampling avoids degeneration by inducing diversity in semantic composition, and not in the surface form directly.

When extensively evaluated on two popular single document summarization datasets, CNN/DailyMail highlight generation (Hermann et al., 2015) and XSum extreme summarization (Narayan et al., 2018), and SQuAD question generation (Rajpurkar et al., 2016; Zhou et al., 2017), we find that composition sampling is most effective in generating relevant, faithful, and diverse summaries or questions. When assessed by humans for faithfulness, composition sampled summaries were as faithful as the best summaries produced using beam search, on both XSum and CNN/DailyMail. In comparison, nucleus sampled summaries were far less faithful. Additionally, according to our newly proposed metric for jointly evaluating diversity and faithfulness in summaries, composition sampled diverse sets of summaries achieved the highest scores on both XSum and CNN/DailyMail.

Our main contributions are as follows: (i) We introduce Composition Sampling, a simple but effective diverse decoding method for conditional generation, enabled by a planning-based approach to generation and Nucleus Sampling. (ii) We introduce pairwise *Self-Entailment* and *Self-BERTScore*

to compute semantic diversity in generated outputs. Our newly introduced metrics complement lexical diversity-based measures (Self-BLEU; Zhu et al., 2018; Alihosseini et al., 2019) and assess whether a diverse set of outputs are contextually dissimilar (Self-BERTScore) or do not entail each other (Self-Entailment). (iii) Finally, we introduce a novel measure jointly "**E**valuating **D**iversity a**N**d f**A**ithfulness" (EDNA) for Summarization; EDNA is designed to capture whether all summaries in a diverse set of summaries of a document are entailing the document and summaries themselves do not entail each other. Our results demonstrate that EDNA correlates positively with both human judgements for faithfulness and diversity.[2]

2 Background

Conditional generation tasks such as summarization (See et al., 2017), data-to-text Generation (Wiseman et al., 2017) and machine translation (Bahdanau et al., 2015), are typically modeled using attention-based encoder-decoder architectures (Bahdanau et al., 2015; Gu et al., 2016; Vaswani et al., 2017). The encoder first encodes the input text d and then the decoder predicts the output $s_{1:n}$ (as the translation or the summary of d) one token at a time as $p(s_i|s_1, \ldots, s_{i-1}; d)$, where, n is the output length and s_i is the ith token in the output. Often these models benefit from large scale task-agnostic pretraining (Song et al., 2019; Radford et al., 2018; Lewis et al., 2019; Rothe et al., 2020a; Raffel et al., 2019; Zhang et al., 2019).

Planning-Based Conditional Generation Recently, Narayan et al. (2021) proposed a *planning-based approach to neural summarization*; the decoder learns to generate a composition $c_{1:m}$ of the output summary s as $p(c_j|c_1, \ldots, c_{j-1}; d)$, and, then the same decoder continues generating the summary s as $p(s_i|s_1, \ldots, s_{i-1}; c; d)$ conditioned on the input document d and the composition $c_{1:m}$ with m being the composition length. Narayan et al. (2021) proposed the use of an entity chain for composition c where the entities ought to be observed in the summary s. During inference, the model takes input document d and generates the concatenated composition and summary sequences $c; s$, instead of directly generating the summary s; c and s are prefixed with special markers "[CONTENT]" and "[SUMMARY]", respectively, as shown in Figure 2. If

---

[1] Holtzman et al. (2020) used the term 'degeneration' to identify generated texts that were generic, repetitive and awkward for story continuation. These issues are less common in our case of conditional generation. Here, with the term 'degeneration' we mostly mean that the generated texts are unfaithful or inconsistent to the input.

[2] Our models and code for Composition Sampling will be released at `anonymized.com` under Apache License, Version 2.0.
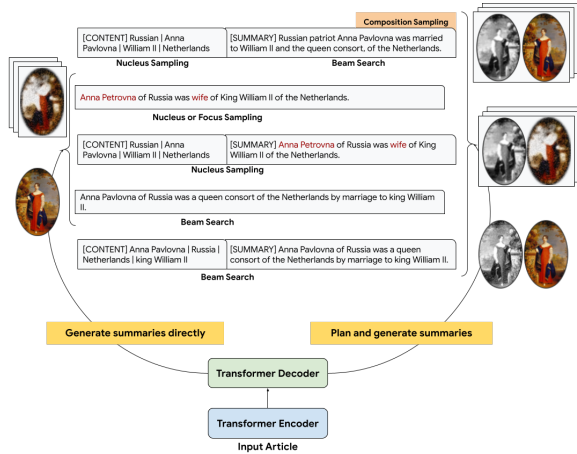
Figure 2: Illustration of composition sampling and other decoding strategies with the vanilla generation models and with the entity-driven planning-based generation models. The term 'composition' is inspired from the quote "*A Well-Composed Painting is Half Done*" from French painter Pierre Bonnard. Images in black-and-white are early sketches or composition of the painting in color. Nucleus or focus sampling often lead to hallucinations; corresponding images in color are blurred to illustrate this. (Credit: The image of "Anna Pavlovna of Russia" is taken from Wikipedia.)

s consists of multiple sentences, the sentence markers "|||" are used to mark sentence boundaries in c.

Entity-level planning provides a lot of flexibility in constructing the entity chain that could in turn affect the quality of the summary, for example, by dropping hallucinated entities from the predicted chain during inference one can control hallucinations in generated summaries (*Planning with Constraints*; Narayan et al., 2021). This planning based approach was applied to summarization, but it can be easily adapted to other conditional generation tasks.

Maximization-Based Decoding The best output $\hat{s}$ with the highest likelihood from these models can be rendered by solving a maximization-based decoding objective: $\hat{x} = \arg\max_x p(x|d)$, where x is either the predicted output text s (for models without planning) or the concatenation of the predicted composition and the output text $c; s$ (for models with planning). A standard practice is to use *beam search* (Tillmann and Ney, 2003; Li et al., 2016b; Wiseman et al., 2017) as solving the objective for the optimal sequence from neural sequence models is not tractable (Chen et al., 2018).

Stochastic Sampling Strategies for Diverse Decoding Sampling-based strategies have been widely used to induce diversity in language models.

Temperature sampling uses a temperature to skew the distribution towards high probability tokens at each decoding step (Ackley et al., 1985; Ficler and Goldberg, 2017; Fan et al., 2018), while top-k sampling truncates the distribution to only keep top k high probability tokens (Fan et al., 2018; Holtzman et al., 2018; Radford et al., 2019). A popular alternative, *nucleus sampling*, similar to top-k sampling, truncates the tail of the distribution but chooses the value of k dynamically (Holtzman et al., 2020). At each decoding step, it samples high-probable tokens from a nucleus N; N is defined as the smallest subset of tokens from the vocabulary V with the cumulative probability $p' \geq p$, where p is the pre-specified mass of the nucleus.

Aralikatte et al. (2021) introduced *Focus Sampling* to promote diversity in summarization models. It constructs a subset $V_k \subseteq V$ by sampling k source-relevant and topical tokens from the vocabulary distribution. The standard beam search decoding is then used to generate a high quality summary limiting itself to V_k. However, the authors show that focus sampling is very sensitive to k, while increasing k improves generation quality, it comes at the cost of decreasing diversity.

3 Composition Sampling

Composition Sampling is a novel hybrid stochastic-maximization-based decoding method enabled by planning-based generation models (Narayan et al., 2021) to generate fluent, faithful and diverse texts for conditional generation. The key idea behind it is to leverage planning-based generation models to sample high quality diverse output compositions c_{sample} in the form of entity chains (shown as prefixed with "[CONTENT]" in Figure 2); we employ nucleus sampling to sample diverse output compositions from $p(c|d)$, where d is the input text. Building on the foundation of these diverse compositions, we employ beam search to generate most-likely diverse output s (shown as prefixed with "[SUMMARY]" in Figure 2), given the input text d and the composition c_{sample} as $p(s|c_{\text{sample}}; d)$.

Why Sample Entity Chains? Unlike top-k or nucleus sampling, composition Sampling avoids degeneration by introducing diversity in composition, and not directly in the surface form. For this to effectively work, the choice of c needs to be well correlated with an underlying notion of "semantic composition", which we want to "diversify"; if c_1 and c_2 are two semantic compositions for input d such that $c \neq c'$, then two summaries $s_1 =$
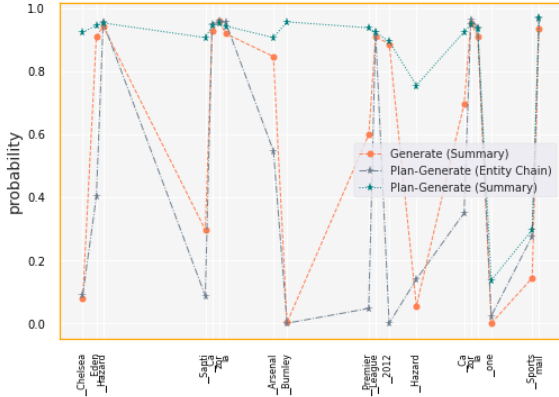
Figure 3: Probabilities of generating entities in the human written reference summary "*Chelsea star Eden Hazard is set to make his 100th top-flight appearance. Santi Cazorla should hit the same milestone when Arsenal meet Burnley. Both players have impressed since moving to the Premier League in 2012. Hazard has more goals this season but Cazorla has one more assist. Sportsmail's reporters choose the player who has excited them the most.*" shown in Figure 1 for the input article in Figure 5. We present these entity probabilities when predicting them directly in the summary (Generate, Summary), or, when first predicting them in the entity chain "*Chelsea | Eden Hazard ||| Santi Cazorla | Arsenal | Burnley ||| Premier League | 2012 ||| Hazard | Cazorla ||| Sportsmail*" during planning (Plan-Generate, Entity Chain) and then predicting them in the summary (Plan-Generate, Summary). We use PEGASUS (Zhang et al., 2019) finetuned models.

$\arg\max_s p(s|c_1; d)$ and $s_2 = \arg\max_s p(s|c_2; d)$ are bound to be diverse. In our work, we have chosen entity chains to model semantic compositions; entity chains have been widely studied to model entity-level lexical cohesion (Barzilay and Elhadad, 1997) and coherence (Halliday and Hasan, 1976; Azzam et al., 1999) in text. Also, entity chains are unique to d, and thus can be easily distinguished from compositions for other inputs.

Hypothesis 1: *If the semantic composition c of the output text s corresponds to entity chains, then learning $p(c|d)$ is much easier than learning $p(s|d)$; d is the input. Hence, we can sample from $p(c|d)$ with higher confidence than sampling directly from $p(s|d)$, and then compute $\arg\max_s p(s|c; d)$.*

We demonstrate the effectiveness of entity chains as a choice for c using the example in Figure 3 for Summarization. The negative log likelihood of generating the summary s from scratch without planning $(-\log p(s|d))$ and the negative log likelihood of generating the entity plans c with planning $(-\log p(c|d))$ are 121.18 and 46.95, respectively, hence the model will be much more confident when sampling from $p(c|d)$ than when sampling directly from $p(s|d)$.

Why Grounded Generation? The generation of s is inherently grounded to its entity composition c; following Narayan et al. (2021) the entity chains are extracted from their targets during training. Hence, once the hard part of planning the composition is done, the model is less perplexed during generation of the output.

In Figure 3, the planning-based model is more confident in predicting entities than its counterpart without planning; perplexities of predicting entities in the summary with and without planning are 0.24 and 1.36, respectively, and perplexities of generating the whole summary with and without planning are 1.15 and 1.48, respectively. In fact, despite the increased length of the target $(c_{1:m}; s_{1:n}$ instead of $s_{1:n})$ in the planning-based model, we find that the perplexity of predicting the whole sequence $(c_{1:m}; s_{1:n})$ using the planning-based model is lower than predicting the summary only without any planning, due to grounding (1.16 vs 1.48). Overall, $p(c; s|d)$, the planning-based approach learns a more confident distribution compared to the approach without planning, $p(s|d)$, at each decoding step. For the example in Figure 3, the average cumulative probabilities for the top 15 tokens in the vocabulary distribution at each decoding step are 0.283 for $p(s|d)$ and 0.433 for $p(c; s|d)$.

Why Sampling with Constraints? Finally, entity chains provide a very effective knob for entity-level content modification in abstractive generation. Composition sampling utilizes this property to draw entity-level faithful compositions which can be further constrained to avoid entity degeneration, e.g., entities that are not in the input.

We assess composition sampling for its ability to generate semantically diverse summaries (§4) and questions (§5) for input documents.

4 Single Document Summarization

4.1 Datasets and Models

We evaluate our decoding strategy on two popular single document summarization datasets: CNN/DailyMail highlight generation (Hermann et al., 2015) and XSum extreme summarization (Narayan et al., 2018), using their original train/validation/test splits. Inputs and outputs were truncated to 512 and 128 for XSum, and, 1024 and

4

256 for CNN/DailyMail.[3]

We experiment with state-of-the-art pretrained models for summarization: PEGASUS (Zhang et al., 2019) and planning-based FROST (Narayan et al., 2021). PEGASUS finetuned models generate summaries directly, whereas FROST finetuned models must generate both the entity chain followed by the summary. For both we experiment with the large transformer architectures (Vaswani et al., 2017) with $L = 16$, $H = 1024$, $F = 4096$, $A = 16$ (568M parameters), where L denotes the number of layers for encoder and decoder Transformer blocks, H for the hidden size, F for the feed-forward layer size, and A for the number of self-attention heads. Since this paper is proposing a decoding strategy, there is no need to train new summarization models. We use the publicly available PEGASUS and FROST checkpoints. Training details and model hyperparameters can be found in Zhang et al. (2019) and Narayan et al. (2021).

All models were decoded with a beam size of 8 and a length-penalty of 0.8. For nucleus sampling and composition sampling we use a nucleus probability p of 0.95.[4] For focus sampling (Aralikatte et al., 2021), we use $k = 10,000$.

4.2 Evaluation Metrics

We assess our decoding strategy for likelihood, fluency, relevance, faithfulness and diversity, using both automatic and human evaluation. FROST models predict a summary plan in the form of an entity chain, followed by a summary. All evaluations, except likelihood, are done on the summary, the predicted entity chains are stripped out. For each diverse decoding strategy, we sample 5 times for each test document; we report on the average for each document.

Sequence Likelihood We report the perplexity of generated sequence (i.e., entity chains concatenated with their summaries for planning based models and summaries only for others) using various decoding strategies.

Lexical Fluency and Relevance We report ROUGE-L F1 scores (Lin and Hovy, 2003) against reference summaries.[5]

Semantic Relevance We report *BERTScore* (Zhang et al., 2020) which computes the contextual similarity between a candidate and its reference.

Faithfulness We follow Maynez et al. (2020) and report on textual entailment (Pasunuru and Bansal, 2018; Falke et al., 2019; Kryscinski et al., 2020). In particular, we report the probability of a summary entailing (*Entailment*) its input document using a classifier trained by fine-tuning an uncased BERT-Large pretrained model (Devlin et al., 2019) on the Multi-NLI dataset (Williams et al., 2018).

We further assess faithfulness by humans. Our annotators, proficient in English, were tasked to read the document carefully and then grade its summary on a scale of 1-4 (*fully unfaithful*, *somewhat unfaithful*, *somewhat faithful* and *fully faithful*); a summary is "fully faithful" if all of its content is fully supported or can be inferred from the document. We collected 3 ratings for each (document, summary) pair and report the average of all assigned labels (1-4) to a system. When either of *somewhat unfaithful* or *somewhat faithful* were selected, annotators were asked to also specify what was faithful or unfaithful in the summary, to improve annotator agreement.

Diversity We report the number of times (out of 5), a decoding technique is able to generate a completely new summary (*Unique*). We also report on *Self-BLEU* (Zhu et al., 2018; Alihosseini et al., 2019) measuring lexical diversity in the generated summaries. We consider all pairs of summaries out of 5 sampled summaries, for each pair we compute the BLEU score (Papineni et al., 2002) considering one summary as a hypothesis and the others as a reference. We report the average BLEU score as the Self-BLEU of the document. The lower the Self-BLEU for a decoding strategy is, the better it is in generating more diverse set of summaries.

We propose two novel measures to capture semantic diversity in summaries: *Self-Entailment* and *Self-BERTScore*. Similar to Self-BLEU, we compute Entailment score and BERTScore for each possible pair of summaries (out of 5), respectively and report on their averages. A lower value of Self-Entailment shows that the generated summaries do not entail each other. Similarly, a lower value of Self-BERTScore shows that the decoding technique is able to generate more contextually dissimilar summaries.

---

[3] We also experimented with MultiNews (Fabbri et al., 2019), a multi-document summarization dataset. Results can be found in Appendix (Table 7).

[4] Additional results of the diverse summary generation with other values of p for random sampling, nucleus sampling and composition sampling can be found in Appendix (Figure 11).

[5] Additional ROUGE-1 and ROUGE-2 scores can be found in Appendix (Table 5 and 6). We lowercased candidate and reference summaries and used `pyrouge` with parameters "-a -c 95 -m -n 4 -w 1.2."

Model	XSum RL	CNN/DMail RL
Single-best with Beam Search		
GSum	36.67	42.48
CTRLsum	–	**42.50**
Focus (PEGASUS)	37.46	39.90
PEGASUS	39.40	40.98
FROST	**39.76**	42.01
FROST (c_{drop})	37.20	41.99
Diverse Decoding		
Focus (PEGASUS)	34.97	–
Nucleus (PEGASUS)	30.99	33.46
Nucleus (FROST)	32.49	35.49
Composition (FROST)	**36.98**	38.69
Composition+Constraints (FROST)	35.89	**39.28**

Table 1: ROUGE-L (RL) results on the full test sets comparing different decoding techniques. The top block shows results from maximization-based decoding of single-best summaries from SOTA models: Focus (PEGASUS) (Aralikatte et al., 2021), CTRLsum (He et al., 2020), GSum (Dou et al., 2020), PEGASUS (Zhang et al., 2019) and FROST (Narayan et al., 2021). FROST with c_{drop} is where the predicted entity chain is modified to keep only the supported entities to generate more faithful summaries. The bottom block shows results from diverse decoding, for each row we sample 5 times for each test document and report on the average for each document. We report results from Focus (Aralikatte et al., 2021), Nucleus (Holtzman et al., 2020), our Composition sampling techniques. For Composition+Constraints, sampled entity chain is modified to keep only the supported entities. The best results in each block are bold-faced. See Appendix Table 5 for more comparisons.

We further assess diversity by humans. Our annotators, proficient in English, were asked to read two summaries (out of 5 sampled summaries) and then grade the pair on a scale of 1-4 (*identical*, *somewhat identical*, *somewhat diverse* and *diverse*); the document was not shown in this assessment. Two summaries are *identical* if they are semantically same, the same information may be presented differently in the case of *somewhat identical*. *Somewhat diverse* pair may introduce 1 or 2 new concepts or topics in one of the summary that are not present in the other summary. *Diverse* pair should introduce few new concepts or topics in each of the summaries. We collected 3 ratings for each pair of sampled summaries and report the average of all assigned labels (1-4) to a system. This assessment was only done with the single-sentence XSum summaries, in the future work we will explore how to do this effectively for the multi-sentence CNN/DailyMail summaries.

Diversity and Faithfulness For summarization, diverse summaries are not meaningful if they are not faithful to the input. We propose EDNA, a novel measure jointly "**E**valuating **D**iversity a**N**d f**A**ithfulness" in summaries, reporting the harmonic mean between Entailment and (1 - Self-Entailment). The higher the number the better the technique is overall at generating faithful and diverse summaries. The reason EDNA relies on Self-Entailment to measure the diversity component is because the faithfulness metric is also based on Entailment. This means that both faithfulness and diversity components will be mapped into a score in similar output spaces (i.e., they are both values between 0 and 1 obtained through the same trained model), making it more likely that they will be properly balanced when mixed.

4.3 Summarization Results

Table 1 presents ROUGE results on the full XSum and CNN/DailyMail test sets comparing diverse decoding methods to their single-best beam Search counterparts. Table 2 presents more detailed faithfulness and diversity results, but they are done on challenge sets consisting of 50 documents for each of XSum and CNN/DailyMail summaries. We construct these challenge sets by randomly selecting documents whose reference summaries have non-extractive entity chains in them; an entity chain is fully extractive if all entities in it can be found in the input document. Narayan et al. (2021) have found that models struggle to generate faithful summaries for documents with data-divergence issues (Dhingra et al., 2019). The same challenge sets were used for human evaluations for faithfulness and diversity.

Composition Sampling is not as Performance Diminishing as Nucleus Sampling, in terms of ROUGE The single-best summary for the planning based model FROST achieves ROUGE (RL) performance of 39.76 on XSum, whereas, nucleus sampled and composition sampled diverse summaries achieves ROUGE scores of 32.49 (average drop of 7.27) and 36.98 (average drop of 2.78), respectively. Similarly, for CNN/DailyMail, nucleus sampling drops ROUGE performance by an average of 6.51 points, compared to an average drop of 3.28 points with composition sampling, for FROST models. Nucleus sampling is even more derogatory for non-planning based summarization models, such as PEGASUS; we see a drop of average 8.59 and 7.30 ROUGE points for XSum and CNN/DailyMail summaries, respectively. These gaps are slightly larger for the results in Table 2; this is expected due to their highly abstractive nature of reference summaries in the challenge sets.

		Models	ppl	With Ref.		Faithfulness		Uniq	S-BL	Diversity			Div+Faith
				RL	BSc	Ent	Human			S-Ent	S-BSc	Human	EDNA
XSum	Single	Focus (PEGASUS)	–	34.23	0.70	0.24	2.19	–	–	–	–	–	–
		PEGASUS	0.51	40.69	0.76	0.40	2.52	–	–	–	–	–	–
		FROST (c; s)	0.31	40.90	0.75	0.37	2.63	–	–	–	–	–	–
		FROST (c_{drop}; s)	0.71	33.75	0.70	0.44	2.78	–	–	–	–	–	–
	Diverse	Focus (PEGASUS)	–	29.19	0.66	0.23	1.88	2.6	89.51	0.62	0.91	1.84	0.09
		Nucleus (PEGASUS)	1.47	31.10	0.68	0.24	2.00	**5.0**	**26.22**	**0.10**	0.68	**3.11**	0.30
		Nucleus (FROST)	0.83	33.81	0.71	0.22	2.11	**5.0**	31.08	**0.10**	0.71	3.08	0.27
		Composition (FROST)	**0.51**	**36.95**	**0.73**	0.27	2.37	4.7	58.94	0.17	0.79	2.73	0.30
		Composition+Constraints (FROST)	0.74	33.87	0.70	**0.43**	**2.75**	3.5	76.87	0.40	0.84	2.25	**0.35**
CNN/DailyMail	Single	PEGASUS	0.35	36.09	0.65	0.70	3.78	–	–	–	–	–	–
		FROST (c; s)	0.30	39.03	0.66	0.72	3.74	–	–	–	–	–	–
		FROST (c_{drop}; s)	0.37	38.87	0.66	0.79	3.94	–	–	–	–	–	–
	Diverse	Nucleus (PEGASUS)	1.39	28.99	0.62	0.62	3.08	**5.0**	**26.99**	0.03	**0.63**	–	0.70
		Nucleus (FROST)	1.04	31.58	0.63	0.56	3.08	**5.0**	29.60	0.03	0.64	–	0.66
		Composition (FROST)	0.52	35.06	**0.64**	0.59	3.45	**5.0**	58.60	0.04	0.71	–	0.66
		Composition+Constraints (FROST)	**0.46**	35.07	0.64	**0.73**	**3.89**	4.9	62.81	0.07	0.72	–	**0.78**

Table 2: Likelihood, faithfulness and diversity results on 50 documents sampled from XSum and CNN/DailyMail each. We report on perplexity (ppl), Entailment (Ent), Uniqueness (Uniq), Self-BLEU (S-BL), Self-Entailment (S-Ent), Self-BERTScore (S-BSc) and EDNA scores, along with ROUGE (RL) and BERTScore (BSc) for comparison. We also report on human judgements for faithfulness and diversity. Additional R1 and R2 numbers can be found in Appendix (Table 6). Best results in the diverse block for each dataset are bold faced. Scores for single-best decoded summaries are also presented for comparison. Focus (PEGASUS) diverse predictions on the CNN/DailyMail are not available. The lower the Self-* metric is, the better the decoding method is in generating diverse outputs.

Metrics	Faithfulness	Diversity
ROUGE-L	0.197	-0.164
BERTScore	0.209	-0.195
Entailment	**0.588**	-0.067
1 - Self-BLEU	-0.208	**0.880**
1 - Self-Entailment	-0.187	0.771
1 - Self-BERTScore	-0.198	0.873
EDNA	0.482	0.174

Table 3: Spearman's correlation coefficient of different metrics with human assessments.

Composition Sampling with Constraints (Composition+Constraints) performs poorly than the vanilla composition sampling on XSum, in terms of ROUGE. This is due to the data divergence issue in XSum summaries (Maynez et al., 2020); due to their extreme abstractive nature XSum reference summaries require a model to hallucinate factual content, that is not necessarily faithful to the input (see examples of XSum summaries in Appendix Figure 4). Composition+Constraints only keeps the supported entities in the sampled plans, hence generated XSum summaries diverge from their reference summaries. This is not the case with the CNN/DailyMail dataset which is mostly extractive and we see ROUGE performance improves with Composition+Constraints in Table 1.

Composition Sampling Makes More Confident Diverse Predictions than Nucleus Sampling Perplexity for FROST predictions increases from 0.31 to 0.83 for nucleus sampling, but only to 0.51 for composition sampling, on XSum. PEGASUS shows an even larger increment in perplexity (from 0.51 to 1.47) for nucleus sampling. Similar patterns are observed for CNN/DailyMail summaries.

Composition+Constraints is more perplexed when generating XSum summaries due to data divergence issues in their reference summaries, as explained earlier; perplexity increased from 0.51 to 0.74 when compared to its vanilla counterpart. Interestingly, Composition+Constraints is almost as confident in generating diverse summaries as the constrained beam decoding (FROST, c_{drop}; s) in generating the single-best summary (perplexities of 0.71 vs 0.74) for XSum. Unsurprisingly, Composition+Constraints is more confident in generating CNN/DailyMail summaries than its counterpart (0.46 vs 0.52) due to their extractive nature.

Composition+Constraints is Most Effective in Generating Meaningful Diverse Summaries It is no surprise that nucleus sampling is able to generate the most diverse summaries on both XSum and CNN/DailyMail (achieving best scores for Self-BLEU, Self-Entailment, Self-BERTScore and diversity human assessment), however these summaries perform poorly on faithfulness measures. Composition+Constraints is most effective in generating faithful summaries, as demonstrated automatically (the best entailment scores of 0.43 on XSum and 0.73 on CNN/DailyMail) and by humans (the highest rating of 2.75 on XSum and 3.89 on CNN/DailyMail, out of 4), that are diverse (the highest EDNA scores of 0.35 on XSum and 0.78 on CNN/DailyMail).

We also carried out pairwise comparisons for human assessments for faithfulness and diversity (using one-way ANOVA with post-hoc Tukey HSD tests; $p < 0.01$). The differences between Nucleus

(PEGASUS) and Nucleus (FROST) were insignificant for both assessments. Nucleus (PEGASUS) was also not significantly more faithful than Focus (PEGASUS) for XSum summaries. All other pairwise differences were significant for both the faithfulness and diversity judgements.

Our results demonstrate that nucleus or focus sampling is not reliable for generating meaningful diverse summaries, composition sampling is the best available decoding strategy for the purpose. Figure 1 presents summaries from different decoding strategies for a CNN/DailyMail article. Other example predictions for both XSum and CNN/DailyMail articles can be found in Appendix (Figure 4, 6, 7 and 8).

Spearman's Correlation. We estimate Spearman's correlation coefficients of different metrics with the faithfulness and diversity assessments. Results are presented in Table 3. In line with finding in (Maynez et al., 2020; Kryscinski et al., 2019), the textual entailment scores are best correlated with the faithfulness scores (moderate, $0.40 \leq r \leq 0.59$). Like Self-BLEU, our newly proposed measures, Self-Entailment and Self-BERTScore, to capture semantic diversity in summaries are also strongly correlated with the diversity scores. Our newly proposed EDNA is currently the best way to jointly measure diversity and faithfulness, compared to other measures.

5 Question Generation

Dataset and Metrics Question generation is often studied as the task of generating a question from a passage-answer pair (Zhou et al., 2017). We experiment on SQuAD (Rajpurkar et al., 2016) and use the same split of Zhou et al. (2017) consisting of 86,635, 8,965, and 8,964 source-target pairs for training, validation, and test, respectively.[6]

We follow Cho et al. (2019) and report on BLEU-4 (Top-1, the single-best accuracy), Oracle (Top-5, the best accuracy among the 5-sampled hypotheses) and Self-BLEU (similar to as defined in §4).

Results Results are presented in Table 4. Like our summarization experiments, composition sampling is not as performance diminishing as nucleus sampling, in terms BLEU. The single-best question using beam search for FROST achieves a BLEU of 21.04, whereas, top-1 nucleus and composition sampled diverse questions achieve BLEU scores

	Model	BLEU-4 Top-1	Oracle Top-5	Pairwise S-BLEU
Single	NQG++	13.27	–	–
	MCP+GSA	16.85	–	–
	PEGASUS	**22.17**	–	–
	FROST	21.04	–	–
Diverse	top-k Sampling	11.53	17.65	45.99
	Diverse Beam Search	13.38	18.30	74.80
	Mixture Decoder	15.17	21.97	58.73
	Mixture Selector (Cho et al.)	15.67	22.45	59.82
	Mixture Selector (Wang et al.)	15.34	21.15	54.18
	Nucleus (PEGASUS)	12.05	24.72	30.64
	Nucleus (FROST)	10.64	22.49	**25.50**
	Composition (FROST)	**18.77**	**26.60**	74.89

Table 4: Comparison of diverse generation methods on question generation. We experiment with PEGASUS and FROST finetuned models with different decoding strategies. We compare them against the pointer-generator based model NQG++ (Zhou et al., 2017), MCP+GSA (Zhao et al., 2018), top-k Sampling (Fan et al., 2018), Diverse Beam Search (Vijayakumar et al., 2018), Mixture Decoder (Shen et al., 2019) and Mixture Content Selection (Cho et al., 2019; Wang et al., 2020) models; all these baseline numbers are taken from Wang et al. (2020). The best results in each block are bold-faced.

of 10.64 (average drop of 10.40 points) and 18.77 (average drop of 2.27 points only), respectively. Nucleus sampled questions achieve the best pairwise diversity scores (Self-BLEU of 25.50), but very low BLEU Top-1 score of 10.64. Composition sampled questions are less diverse then other methods, but outperform all baselines on Top-1 and Oracle metrics. Poor diversity (in terms of Self-BLEU) in composition sampled questions can be attributed to two limitations: (i) SQuAD questions are mostly extractive, and (ii) questions are generated conditioned on the passage *and* the answer spans; leaving limited scope for models to generate diverse questions. An example in Appendix (Figure 10) demonstrates the effectiveness of composition sampling in generating accurate and diverse questions compared to other sampling methods.[7]

6 Conclusion

We proposed Composition Sampling, a simple yet effective decoding method for faithful and diverse conditional generation. Our method is straightforward to implement and does not require any external systems to augment the input during inference. We introduced Self-Entailment and Self-BERTScore, to compute semantic diversity in summaries, and, EDNA, for jointly measuring faithfulness and diversity in summaries.

---

[6]We also experimented with the split of Du et al. (2017). Results can be found in Appendix (Table 8).

[7]Detailed comparative error analysis to Cho et al. (2019) and Wang et al. (2020) were not possible as these predictions are not publicly available.

Ethical Considerations

The nature of text generation leads to multiple ethical considerations when applied to applications. The main failure mode is that the model can learn to mimic target properties in the training data that are not desirable.

Faithfulness and Factuality Since models create new text, there is the danger that they may neither be faithful to the source material nor factual. This can be exacerbated when the data itself has highly abstractive targets, which require the model to generate words not seen in the source material during training. This often leads the model to generate content inconsistent with the source material (Maynez et al., 2020; Kryscinski et al., 2020; Gabriel et al., 2021).

Trustworthy Data If the data itself is not trustworthy (comes from suspect or malicious sources) the model itself will naturally become untrustworthy as it will ultimately learn the language and topics of the training data. For instance, if the training data is about Obama birther conspiracies, and the model is asked to generate information about the early life of Obama, there is a risk that such false claims will be predicted by the model.

Bias in Data Similarly, biases in the data around gender, race, etc., risk being propagated in the model predictions, which is common for most NLP tasks. This is especially true when the models are trained from non-contemporary data that do not represent current norms and practices (Blodgett et al., 2020).

The above considerations are non-malicious, in that the model is merely learning to behave as its underlying source material. If users of such models are not aware of these issues and do not account for them, e.g., with better data selection, evaluation, etc., then the generated text can be damaging.

Generation models can also be misused in malicious ways. These include generating fake news, spam, and other text meant to mislead large parts of the general population.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6078–6095, Online. Association for Computational Linguistics.

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Coreference and Its Applications*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421, Online. Association for Computational Linguistics.

Yining Chen, Sorcha Gilroy, Andreas Maletti, Jonathan May, and Kevin Knight. 2018. Recurrent neural networks as weighted language recognizers. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2261–2271, New Orleans, Louisiana. Association for Computational Linguistics.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. 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 3121–3131, Hong Kong, China. Association for Computational Linguistics.

Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards diverse and natural image descriptions via a conditional GAN. *CoRR*, abs/1703.06029.

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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13042–13054. Curran Associates, Inc.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2018. Generating multiple diverse responses for short-text conversation. *CoRR*, abs/1811.05696.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

10

Short Papers), pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016a. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

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.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4252–4273, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Siva Reddy, and Shay B. Cohen. 2016. Paraphrase generation from latent-variable PCFGs for semantic parsing. In *Proceedings of the 9th International Natural Language Generation conference*, pages 153–162, Edinburgh, UK. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with Learned Entity Prompts for Abstractive Summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

11

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

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. 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 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report, OpenAI*.

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. *CoRR*, abs/1901.07291.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020a. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020b. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

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

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

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5926–5936. PMLR.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Christoph Tillmann and Hermann Ney. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7371–7379. AAAI Press.

12

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2134–2143, Online. Association for Computational Linguistics.

Nathaniel Weir, João Sedoc, and Benjamin Van Durme. 2020. COD3S: Diverse generation with discrete semantic signatures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5199–5211, Online. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *CoRR*, abs/1908.04319.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, Brussels, Belgium. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, Virtual Conference, Formerly Addis Ababa Ethiopia.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. *CoRR*, abs/1704.01792.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A   Related Work

Due to standard likelihood training, approximate decoding objectives or lack of diversity in references, neural text generation models tend to generate outputs with high fluency but low diversity.

Majority of the focus has been on open ended generations, e.g., story completion (Fan et al., 2018), response generation in dialogue (Adiwardana et al., 2020; Gao et al., 2018), causal generation (Weir et al., 2020) and prompt completion (Tevet and Berant, 2021), for generating diverse responses. This is partly due the fact that there is a considerable degree of freedom in what needs to be responded to the input, and the focus has been on how to generate more diverse and human-like outputs. Vijayakumar et al. (2018) and Kulikov et al. (2019) diversify beam search, using a task-specific scoring function, or constrain beam hypotheses to be sufficiently different. Others avoid text degeneration by truncating the unreliable tail of the probability distribution at each decoding step, either by sampling from the top-k tokens (*Top-k Sampling*; Fan et al., 2018) or by sampling from a dynamic nucleus of tokens with the bulk of the probability mass (*Nucleus Sampling*; Holtzman et al., 2020). Others modify the training objective to make the distribution sparse (Martins et al., 2020) or assign lower probability to unlikely generations (Welleck et al., 2019).

For close-ended generation, most work focuses on generating diverse questions (Narayan et al., 2016; Dong et al., 2017; Sultan et al., 2020; Wang et al., 2020) or paraphrases (Li et al., 2016a; Dai et al., 2017; Xu et al., 2018; Cao and Wan, 2020). For summarization, Cho et al. (2019) uses a mixture of experts to sample different binary masks

13

Model	XSum R1/R2/RL	CNN/DailyMail R1/R2/RL
Single-best with Beam Search		
RoBERTaShare	38.52/16.12/31.13	39.25/18.09/36.45
MASS	39.75/17.24/31.95	42.12/19.50/39.01
BART	45.14/22.27/37.25	44.16/21.28/40.90
GSum	45.40/21.89/36.67	**45.94**/22.32/42.48
UniLM	–/–/–	43.33/20.21/40.51
T5	–/–/–	43.52/21.55/40.69
ProphetNet	–/–/–	44.20/21.17/41.30
CTRLsum	–/–/–	45.65/**22.35/42.50**
Focus (PEGASUS) $_{d \to t_d \to s}$	45.31/22.75/37.46	42.95 20.79 39.90
PEGASUS $_{d \to s}$	47.56/24.87/39.40	44.05/21.69/40.98
FROST $_{d \to c;s}$	**47.80/25.06/39.76**	45.11/22.11/42.01
FROST $_{d \to c_{drop};s}$	44.94/21.58/37.20	45.08/22.14/41.99
Diverse Decoding with Focus, Nucleus and Composition Sampling		
Focus (Focus (PEGASUS))$_{d \to t_{sample} \to s_{beam}}$	42.76/19.89/34.97	–/–/–
Nucleus (PEGASUS)$_{d \to s_{sample}}$	38.49/16.57/30.99	36.27/15.10/33.46
Nucleus (FROST)$_{d \to c_{sample};s_{sample}}$	40.26/17.83/32.49	38.49/15.71/35.49
Composition (FROST)$_{d \to c_{sample};s_{beam}}$	**45.12/22.24/36.98**	41.76/18.94/38.69
Composition+Constraints (FROST)$_{d \to c_{sample,drop};s_{beam}}$	43.82/20.35/35.89	**42.37/19.48/39.28**

Table 5: All ROUGE results on the full test sets comparing different decoding techniques and SOTA models. Only ROUGE-L results were presented in Table 1. Additional results from SOTA models such as RoBERTaShare (Rothe et al., 2020b), MASS (Song et al., 2019), BART (Lewis et al., 2019), UniLM (Dong et al., 2019), T5 (Raffel et al., 2019) and ProphetNet (Qi et al., 2020) are addded here. See the caption of Table 1 for more details. The best results in each block are bold-faced.

	Models	With Reference R1/R2/RL
XSum – Single	Focus (PEGASUS)	41.20/20.30/34.23
	PEGASUS	49.49/28.43/40.69
	FROST ($c; s$)	49.12/28.35/40.90
	FROST ($c_{drop}; s$)	41.15/19.66/33.75
XSum – Diverse	Focus (PEGASUS)	36.58/16.32/29.19
	Nucleus (PEGASUS)	38.91/18.43/31.10
	Nucleus (FROST)	41.96/20.77/33.81
	Composition (FROST)	**45.88/23.74/36.95**
	Composition+Constraints (FROST)	41.81/19.61/33.87
CNN/DailyMail – Single	PEGASUS	38.50/15.04/36.09
	FROST ($c; s$)	41.89/17.54/39.03
	FROST ($c_{drop}; s$)	41.82/17.96/38.87
CNN/DailyMail – Diverse	Nucleus (PEGASUS)	31.57/10.62/28.99
	Nucleus (FROST)	34.62/11.78/31.58
	Composition (FROST)	**37.89**/14.88/35.06
	Composition+Constraints (FROST)	37.79/**15.07/35.07**

Table 6: Full ROUGE results on 50 documents sampled from XSum and CNN/DailyMail each for the results presented in Table 2. Additional ROUGE-1 and ROUGE-2 results are provided in this table.

Model	MultiNews R1/R2/RL
Single-best with Beam Search	
PEGASUS $_{d \to s}$	47.52/18.72/24.91
FROST $_{d \to c;s}$	43.12/16.93/22.49
Diverse Decoding, Average of five runs	
Nucleus (FROST)$_{d \to c_{sample};s_{sample}}$	39.50/12.94/19.50
Composition (FROST)$_{d \to c_{sample};s_{beam}}$	42.47/15.43/21.43
Composition+Constraints (FROST)$_{d \to c_{sample,drop};s_{beam}}$	42.37/15.78/21.90
Diverse Decoding, Best of five runs	
Nucleus (FROST)$_{d \to c_{sample};s_{sample}}$	44.40/16.86/23.03
Composition (FROST)$_{d \to c_{sample};s_{beam}}$	46.98/19.34/24.96
Composition+Constraints (FROST)$_{d \to c_{sample,drop};s_{beam}}$	46.71/19.55/25.36

Table 7: ROUGE results on **Multi-News** (Fabbri et al., 2019) multi-document summarization test set comparing different decoding techniques. The datasets contains 56k articles as well as multi-line human-written summaries in total from the site newser.com.

on the source sequence for diverse content selection for summarization and (Aralikatte et al., 2021) biases the decoder proactively to generate output from a set of tokens which are topically-relevant to the input.

Model	BLEU-4 Top-1	Oracle Top-5	Pairwise S-BLEU
Single-best with Beam Search			
PEGASUS	**21.52**	–	–
FROST	19.98	–	–
Diverse Decoding			
Nucleus (PEGASUS)	12.60	24.45	31.23
Nucleus (FROST)	10.98	22.61	**26.36**
Composition (FROST)	**17.28**	**25.03**	75.81

Table 8: We also experimented with the split of Du et al. (2017) of SQuAD (Rajpurkar et al., 2016) for question generation, consisting of 70,484, 10,570, and 11,877 examples for training, validation, and test, respectively. The best results in each block are bold-faced.

GOLD: Walsall have signed defender Luke Leahy on a two-year contract from Scottish Championship side Falkirk.

Input: Leahy, 24, scored 12 goals in 158 appearances with Falkirk, having joined the club from Rugby Town in 2012. The left-back made 38 appearances last season, helping the club finish second in the Scottish second tier before they lost to Dundee United in the play-offs. He joins Walsall on a free transfer after his contract expired and is the League One club's first summer signing. Find all the latest football transfers on our dedicated page.

Single-best summaries

[✗] **Focus (PEGASUS)** $(d \rightarrow t_d \rightarrow s)$: Walsall have signed Falkirk defender Alex Leahy on a two-year deal.
[✗] **PEGASUS** $(d \rightarrow s)$: Walsall have signed defender Paddy Leahy from Scottish Championship side Falkirk on a three-year deal.
[✗] **FROST** $(d \rightarrow c; s)$:[CONTENT] Walsall | Falkirk | Liam Leahy | two [SUMMARY] Walsall have signed Falkirk defender Liam Leahy on a two-year deal.
[✓] **FROST** $(d \rightarrow c_{drop}; s)$: [CONTENT] Walsall | Falkirk | Leahy [SUMMARY] Walsall have signed Falkirk defender Leahy on a free transfer.

Focus Sampling: Focus (PEGASUS) $(d \rightarrow t_{sample} \rightarrow s)$

[✗] $s_1 \rightarrow$ Welsall have signed defender Adebayu " Adebayu "eahy on a two-year deal following his departure from Scottish Championship club Falkiri.
[✗] $s_2 \rightarrow$ Welsall have signed defender Adebayu " Adebayu "eahy on a two-year deal from Scottish Championship club Falkock.
[✗] $s_3 \rightarrow$ Welsall have signed defender Adebayu " Adebayu "eahy on a two-year deal from Scottish Championship club Falkock.
[✗] $s_4 \rightarrow$ Welsall have signed defender Adebayu Leahys from Scottish Championship club Falk Falkiri for an undisclosed fee on a three-year deal.
[✗] $s_5 \rightarrow$ Welsall have signed defender Adebayu " Adebayu "eahny on a two-year deal following his departure from Scottish Championship club Falkock.

Nucleus Sampling: PEGASUS $(d \rightarrow s_{sample})$

[✗] $s_1 \rightarrow$ Walsall have signed defender Adam Leahy from fellow Scottish Championship side Falkirk on a two-year contract.
[✗] $s_2 \rightarrow$ Walsall have signed defender Matt Leahy on a two-year deal from Falkirk.
[✗] $s_3 \rightarrow$ Walsall have signed Falkirk full-back Tyrone Leahy for an undisclosed fee.
[✗] $s_4 \rightarrow$ Walsall have signed defender Jason Leahy from Scottish Championship club Falkirk.
[✗] $s_5 \rightarrow$ Walsall have signed Driscoll defender Chris Leahy for an undisclosed fee from Scottish Championship side Falkirk.

Nucleus Sampling: FROST $(d \rightarrow c_{sample}; s_{sample})$

[✗] $c_1; s_1 \rightarrow$ [CONTENT] Walsall | Rory Leahy | Falkirk [SUMMARY] dawned on Walsall as they signed defender Rory Leahy on a season-long loan from Falkirk.
[✗] $c_2; s_2 \rightarrow$ [CONTENT] Walsall | Falkirk | Liam Leahy [SUMMARY] Walsall have signed Falkirk defender Liam Leahy.
[✗] $c_3; s_3 \rightarrow$ [CONTENT] Falkirk | Wade Leahy | Walsall [SUMMARY] Former Falkirk defender Wade Leahy has joined Walsall for an undisclosed fee.
[✗] $c_4; s_4 \rightarrow$ [CONTENT] Walsall | Todd Leahy | Scottish Championship | Falkirk [SUMMARY] Walsall have signed defender Todd Leahy from Scottish Championship side Falkirk.
[✗] $c_5; s_5 \rightarrow$ [CONTENT] Walsall | Greg Leahy | Scottish Championship | Falkirk | two [SUMMARY] Walsall have signed defender Greg Leahy from Scottish Championship side Falkirk on a two-year contract.

Composition Sampling: FROST $(d \rightarrow c_{sample}; s)$

[✗] $c_1; s_1 \rightarrow$ [CONTENT] Walsall | Rory Leahy | Falkirk [SUMMARY] Walsall have signed defender Rory Leahy from Falkirk.
[✗] $c_2; s_2 \rightarrow$ [CONTENT] Walsall | Falkirk | Liam Leahy [SUMMARY] Walsall have signed Falkirk defender Liam Leahy.
[✗] $c_3; s_3 \rightarrow$ [CONTENT] Falkirk | Wade Leahy | Walsall [SUMMARY] Falkirk defender Wade Leahy has joined Walsall.
[✗] $c_4; s_4 \rightarrow$ [CONTENT] Walsall | Todd Leahy | Scottish Championship | Falkirk [SUMMARY] Walsall have signed defender Todd Leahy from Scottish Championship side Falkirk.
[✗] $c_5; s_5 \rightarrow$ [CONTENT] Walsall | Greg Leahy | Scottish Championship | Falkirk | two [SUMMARY] Walsall have signed defender Greg Leahy from Scottish Championship side Falkirk on a two-year deal.

Composition Sampling with Constraints: FROST $(d \rightarrow c_{sample,drop}; s)$

[✓] $c_1; s_1 \rightarrow$ [CONTENT] Walsall | Leahy | Falkirk [SUMMARY] Walsall have signed defender Leahy from Falkirk.
[✓] $c_2; s_2 \rightarrow$ [CONTENT] Walsall | Falkirk | Leahy [SUMMARY] Walsall have signed Falkirk defender Leahy on a free transfer.
[✓] $c_3; s_3 \rightarrow$ [CONTENT] Falkirk | Leahy | Walsall [SUMMARY] Falkirk defender Leahy has joined Walsall on a free transfer.
[✓] $c_4; s_4 \rightarrow$ [CONTENT] Walsall | Leahy | Scottish | Falkirk [SUMMARY] Walsall have signed defender Leahy from Scottish side Falkirk.
[✓] $c_5; s_5 \rightarrow$ [CONTENT] Walsall | Leahy | Scottish | Falkirk [SUMMARY] Walsall have signed defender Leahy from Scottish side Falkirk.

Figure 4: An example input article, its human written summary, and the model predictions including diverse summaries for the XSum dataset. We highlight spans in orange that are not faithful to the input document.

Input Article: Chelsea's Eden Hazard and Arsenal's Santi Cazorla are set to reach a Premier League milestone this weekend when they each make their 100th appearance. Both players have been hugely influential since they moved to London in the summer of 2012, but who has been the most exciting import to watch? Here, Sportsmail's reporters choose the player they most enjoy seeing in action. Eden Hazard (L) and Santi Cazorla are both set to make their 100th Premier League appearance this weekend. Lee Clayton. Cazorla has wonderful balance. So does Hazard. Cazorla scores important goals. So does Hazard. Cazorla is two-footed. So is Hazard. Cazorla dances past opponents. So does Hazard. So, while there is not a lot to choose between them and Hazard is likely to get the most picks in this article, I am going for Cazorla. It's a personal choice. He is a wonderful footballer. I have paid to watch them both (and I will pay to watch them both again), but the little Spanish magician edges it for me. VERDICT: CAZORLA. Cazorla, pictured in action against Burnley, has been an influential part of Arsenal's midfield this season. Ian Ladyman. I remember when Manchester City baulked at paying Hazard's wages when the Belgian was up for grabs in 2012. Back then City thought the young forward had a rather high opinion of his own worth for a player who was yet to play in a major European league. In the early days of his time at Chelsea, it looked as though City may have been right. He showed flashes of brilliance but also looked rather too easy to push off the ball. Roll forward to 2015, however, and the 24-year-old has developed in to one of the most important players in the Barclays Premier League. Brave, strong and ambitious, Hazard plays on the front foot and with only one thought in this mind. Rather like Cristiano Ronaldo, he has also developed in to the type of player ever defender hates, simply because he gets back up every time he is knocked to the ground. He would get in every team in the Premier League and is one of the reasons Chelsea will win the title this season. VERDICT: HAZARD. Hazard controls the ball under pressure from Stoke midfielder Stephen Ireland at Stamford Bridge. Dominic King. It has to be Hazard. I saw him play for Lille twice in the season before he joined Chelsea – once against St Etienne, the other was what proved to be his final appearance against Nancy. He scored two in the first match, a hat-trick the latter and played a different game to those around him. He hasn't disappointed since arriving here and I love the nonchalance with which he takes a penalty, his low centre of gravity and the way he can bamboozle defenders. If there is such a thing as £32million bargain, it is Hazard. VERDICT: HAZARD. Hazard celebrates after scoring a fine individual goal in Chelsea's 3-2 win against Hull in March. Nick Harris. Now this is a tricky one because while Eden Hazard will frequently embark on a dribble or dink in a pass that will make you nod in appreciation, he'll also miss a penalty and make you groan. Whereas the older Cazorla, less flashy but no less of a technical master, is to my mind more of a fulcrum, more important relatively to the sum of Arsenal's parts than Hazard is to Chelsea. You'll gasp at Hazard but Cazorla's wow factor is richer. That's not to dismiss either: both are brilliant footballers, contributing goals, assists and flair. Any neutral would bite your hand off to have either playing in your team. Forced to pick though, it's Cazorla, for his consistency and crucially doing it in the biggest games. Exhibit A would be Manchester City 0 Arsenal 2 in January; goal, assist, all-round brilliance, against a big team, at an important time. VERDICT: CAZORLA. Cazorla scores from the penalty spot in Arsena's 2-0 away win at Manchester City in January. Riath Al-Samarrai. Eden Hazard for me. Cazorla is an utter delight, a little pinball of a man who is probably the most two-footed player I've seen. Put him in a tight space and then you see what makes him rare among the best. But Hazard is the top player in the Premier League, in my opinion. This is the sixth of his eight seasons as a professional where he has reached double figures and yet he offers so much more than goals (36 in 99 in the Premier League for Chelsea). He can beat a man and, better still, you sense he likes doing it. Technically, his passing and shooting are excellent and he also has a mind capable of sussing out the shapes and systems in front of him. That intelligence, more specifically.

Figure 5: An example CNN/DailyMail article for summaries presented in Figure 1.

Human Written Summary: Nick Clegg made the admission in a rare joint interview with his wife Miriam. Lib Dem said she decided against moving into 'government mansion'. 'Discussion's a rather grand word for Miriam basically saying no,' he joked. Miriam claims he has put 'country above party' at 'great personal cost'. Tonight: Spotlight Nick Clegg tonight (Thursday) on ITV at 7.30pm.

Single-best summaries using Beam Search

[✓] PEGASUS ($d \rightarrow s$): Lib Dem leader reveals balance of power lies with wife Miriam Gonzalez Durantez. Couple invited cameras into the kitchen of their £1.5million family home. They revealed why they decided to remain with their three sons in London. Mrs Clegg said: 'It has been the best decision for all of us and if I may, also for Nick as well because you can see how politicians sometimes can get in a bubble and be completely distant from normal society'. Mr Clegg insisted he felt more full of 'vim and vigour' than he did in the start of the coalition. Focus groups conducted by the programme found Mr Clegg was characterised as the 'chihuahua in a handbag' of the government. He said: 'If you know you disagree with someone, there's no tension. I suspect meetings between Tony Blair and Gordon Brown were so fractious because they kind of should have agreed but they didn't'.

[✓] FROST ($d \rightarrow c; s$): [CONTENT] Nick Clegg | Miriam ||| Lib Dem | Whitehall ||| ||| Couple | ITV | Tonight ||| Putney | London [SUMMARY] Nick Clegg reveals balance of power lies with his wife Miriam in joint interview. Lib Dem leader was given option of moving family into Whitehall mansion. But the move was vetoed by his high flying lawyer wife after last election. Couple invited the cameras into their kitchen for ITV's Tonight programme. They were seen drinking wine and cooking in the kitchen of their home in Putney, south west London.

[✓] FROST ($d \rightarrow c_{\text{drop}}; s$): [CONTENT] Nick Clegg | Miriam ||| Lib Dem | Whitehall ||| ||| ITV | Tonight ||| Putney | London [SUMMARY] Nick Clegg reveals balance of power lies with his wife Miriam in joint interview. Lib Dem leader was given option of moving family into Whitehall mansion. But the move was vetoed by his high flying lawyer wife. The couple invited the cameras into their kitchen for ITV's Tonight programme. They were seen drinking white wine and cooking in their home in Putney, London.

Figure 6: Human written summary and single-best summaries using Beam Search for the CNN/DailyMail article in Figure 9.

Nucleus Sampling: PEGASUS ($d \rightarrow s_{\text{sample}}$)

[✗] $s_1 \rightarrow$ Nick Clegg reveals balance902 of power lies with his wife Miriam Gonzalez Durantez irish. Couple performed joint interview to promote new series on ITV News Tonight. Cleggs spoke of pressure of moving from their Putney home into a taxpayer-funded mansion.

[✓] $s_2 \rightarrow$ Lib Dem leader and wife Miriam give TV interview to ITV's Tonight program. Co-hosts have been pictured drinking white wine and cooking paella. They explained why she vetoed family heading to Whitehall mansion. Husband quipped: 'It's a grand word for Miriam basically saying no'.

[✗] $s_3 \rightarrow$ Lib Dem leader admitted wife Miriam has the final say over family life. Couple chose not to move their three Laundry to Whitehall home earlier this May.

[✗] $s_4 \rightarrow$ Nick Clegg and his wife Miriam Gonzalez Durantez open up in TV interview. Lib Dem leader revealed she Bloomberg-style 'discussions' in their home. Couple revealed they opted not to stay with their sons in their £1.5m house.

[✗] $s_5 \rightarrow$ Liberal Democrats leader revealed balance of power lies 30-plus metres away. He brought cameras into family home due to Cameron and Miliband controversies. Lib Dem leader joked that wife Miriam vetoed their move to Whitehall.

Nucleus Sampling: FROST ($d \rightarrow c_{\text{sample}}; s_{\text{sample}}$)

[✗] $c_1; s_1 \rightarrow$ [CONTENT] Liberal Democrats | Nick Clegg | Miriam Gonzalez Durantez ||| Putney | London ||| Cleggs ||| ITV ||| Couple [SUMMARY] Liberal Democrats leader Nick Clegg reveals balance of power with wife Miriam Gonzalez Durantez in joint interview. They invited cameras into kitchen of £1.5million family home in Putney, south west London. Cleggs are seen trying white wine as they discuss family life and girlfriends. They wereFurness on ITV programme and said they chose home to protect family. Couple say choosing home stopped them veering off from wider society 'in a bubble'

[✓] $c_2; s_2 \rightarrow$ [CONTENT] Lib Dem | ITV | Tonight | Miriam Gonzalez Durantez ||| ||| Couple | Putney | London [SUMMARY] Lib Dem leader appeared on ITV's Tonight programme with wife Miriam Gonzalez Durantez. He was given the option of moving his family into a grace-and-favour government mansion but was vetoed. Couple invite cameras into family home in Putney, south west London to talk about family life.

[✗] $c_3; s_3 \rightarrow$ [CONTENT] Lib Dems | Miriam ||| Couple | ITV | Tonight ||| Putney | London ||| bestseller | Miliband [SUMMARY] Lib Dems leader revealed balance of power lies with wife Miriam. Couple invited cameras into kitchen of their home for ITV's Tonight programme.Asked why they kept the family home Galore in Putney, south west London. Documentary follows millions-selling bestseller's rave over Miliband'!!

[✗] $c_4; s_4 \rightarrow$ [CONTENT] Clegg | Putney ||| ||| ||| Lib Dem [SUMMARY] Mrs Clegg explains why the family stayed in their £1.5million home in Putney 1929. Comparing their relationship to that of a different marriage, she said: 'We just stand together and it's not any more of a difficulty than what many others have to go through'. Revealingly, suggests that although no longer planning a political showdown they are closer than they have ever been. Senior Lib Dem says he has learned how to ignore populist attacks.

[✗] $c_5; s_5 \rightarrow$ [CONTENT] Lib Dem | ITV | Tonight ||| David Cameron | Lib Demaceae ||| Nick | Miriam Gonzalez Durantez [SUMMARY] Lib Dem leader allowed ITV's Tonight news programme into kitchen of their home. David Cameron elegant kitchens sometimes look like Lib Demaceae. Nick's high flying wife Miriam Gonzalez Durantez reveals balance of power is with her.

Figure 7: Nucleus sampled diverse summaries for the CNN/DailyMail article in Figure 9. We highlight spans in orange that are not faithful to the input document.

Composition Sampling: FROST $(d \rightarrow c_{\text{sample}}; s)$

[✓] $c_1; s_1 \rightarrow$ [CONTENT] Liberal Democrats | Nick Clegg | Miriam Gonzalez Durantez ||| Putney | London ||| Cleggs ||| ITV ||| Couple [SUMMARY] Liberal Democrats leader Nick Clegg reveals balance of power lies with wife Miriam Gonzalez Durantez. The couple invited the cameras into the kitchen of their £1.5million family home in Putney, south west London. Cleggs were seen drinking white wine and cooking in the kitchen as they chatted about their family life. Focus groups conducted by the ITV programme found he was characterised as the 'chihuahua in a handbag' of the government. Couple believe that choosing to remain in their family home has stopped them from becoming cut off from the rest of the world.

[✓] $c_2; s_2 \rightarrow$ [CONTENT] Lib Dem | ITV | Tonight | Miriam Gonzalez Durantez ||| ||| Couple | Putney | London [SUMMARY] Lib Dem leader appeared on ITV's Tonight programme with wife Miriam Gonzalez Durantez. He was given the option of moving his family into a grace-and-favour government mansion - but the move was vetoed by his wife. Couple invited the cameras into the kitchen of their £1.5million family home in Putney, south west London.

[✓] $c_3; s_3 \rightarrow$ [CONTENT] Lib Dems | Miriam ||| Couple | ITV | Tonight ||| Putney | London ||| bestseller | Miliband [SUMMARY] Lib Dems leader reveals balance of power lies with wife Miriam in joint interview. Couple invited the cameras into their kitchen for ITV's Tonight programme. They were seen drinking wine and cooking in their £1.5million home in Putney, south west London. Interview comes after bestseller's row over Miliband's small kitchen.

[✓] $c_4; s_4 \rightarrow$ [CONTENT] Clegg | Putney ||| ||| ||| Lib Dem [SUMMARY] Mr Clegg and his wife invited the cameras into the kitchen of their Putney home. They were seen drinking wine and cooking as they chatted about their family life. The couple were asked why they decided to remain in their family home. Lib Dem leader was given the option of moving his family into a government mansion.

[✗] $c_5; s_5 \rightarrow$ [CONTENT] Lib Dem | ITV | Tonight ||| David Cameron | Lib Demaceae ||| Nick | Miriam Gonzalez Durantez [SUMMARY] Lib Dem leader reveals balance of power in joint interview with wife on ITV's Tonight. David Cameron's large country kitchen and Lib Demaceae' small kitchen criticised. Nick and Miriam Gonzalez Durantez reveal why they stayed at home.

Composition Sampling with Constraints: FROST $(d \rightarrow c_{\text{sample,drop}}; s)$

[✓] $c_1; s_1 \rightarrow$ [CONTENT] Liberal Democrats | Nick Clegg | Miriam Gonzalez Durantez ||| Putney | London ||| Cleggs ||| ITV ||| [SUMMARY] Liberal Democrats leader Nick Clegg revealed balance of power lies with wife Miriam Gonzalez Durantez. The couple invited the cameras into the kitchen of their £1.5million family home in Putney, south west London. The Cleggs were seen drinking white wine and cooking in the kitchen of their home. ITV programme found he was characterised as the 'chihuahua in a handbag' of the government. He also insisted he felt more full of 'vim and vigour' than he did in the start of the coalition.

[✓] $c_2; s_2 \rightarrow$ [CONTENT] Lib Dem | ITV | Tonight | Miriam Gonzalez Durantez ||| ||| Putney | London [SUMMARY] Lib Dem leader appeared on ITV's Tonight programme with wife Miriam Gonzalez Durantez. He was given the option of moving his family into a grace-and-favour government mansion - but the move was vetoed by his wife. The couple invited the cameras into the kitchen of their £1.5million family home in Putney, south west London.

[✓] $c_3; s_3 \rightarrow$ [CONTENT] Lib | Miriam ||| ITV | Tonight ||| Putney | London ||| Miliband [SUMMARY] Lib Dem leader reveals balance of power lies with wife Miriam in joint interview. The couple invited the cameras into their kitchen for ITV's Tonight programme. They were seen drinking wine and cooking in their £1.5million home in Putney, south west London. Comes after Miliband was widely mocked for posing with wife in his kitchen.

[✓] $c_4; s_4 \rightarrow$ [CONTENT] Clegg | Putney ||| ||| ||| Lib Dem [SUMMARY] Mr Clegg and his wife invited the cameras into the kitchen of their Putney home. They were seen drinking wine and cooking as they chatted about their family life. The couple were asked why they decided to remain in their family home. Lib Dem leader was given the option of moving his family into a government mansion.

[✓] $c_5; s_5 \rightarrow$ [CONTENT] Lib Dem | ITV | Tonight ||| David Cameron | Lib ||| Nick | Miriam Gonzalez Durantez [SUMMARY] Lib Dem leader reveals balance of power in joint interview with wife on ITV's Tonight. Comes after David Cameron invited cameras into Lib Dem leader's country kitchen. Nick and Miriam Gonzalez Durantez were seen drinking wine and cooking.

Figure 8: Composition sampled diverse summaries for the CNN/DailyMail article in Figure 9. We highlight spans in orange that are not faithful to the input document.

Input Article: It is a conversation that will be familiar to couples across the country. What one spouse thinks is a 'discussion', the other understands they are being overruled. In a joint interview with his high flying lawyer wife Miriam Gonzalez Durantez, Nick Clegg revealed the balance of power lies where many long suspected: with her. After the last election, Mr Clegg was given the option of moving his family into a grace-and-favour government mansion - but the move was vetoed by his wife. After controversies over David Cameron's large country kitchen and Ed Miliband's small second kitchen, the couple invited the cameras into the kitchen of their £1.5million family home in Putney, south west London for ITV's Tonight programme. Scroll down for video. Home: In a revealing joint interview, Liberal Democrats leader Nick Clegg (pictured) admitted his wife Miriam (right) makes the big decisions in their household. Mr Clegg is seen in the documentary drinking wine as his wife explains why she chose not to move her family into a government property. They revealed why they decided to remain with their three sons Antonio, Alberto, and Miguel, in the family home instead of making the move to Whitehall. Miriam, who uses her maiden name Gonzalez Durantez, told ITV News Political Editor Tom Bradby: 'We had a lot of pressure at the time to go to one of the houses of the government. 'We discussed and thought the best thing would be for the children to stay here. Revealingly, Mr Clegg quipped: 'Discussion's a rather grand word for Miriam basically saying no.' But he quickly added: 'You were so right, you were so right.' However, the couple believe that choosing to remain in their family home has stopped them from becoming cut off from the rest of the world. Mrs Clegg said: 'If you look at it with perspective it has been the best decision for all of us and if I may, also for Nick as well because you can see how politicians sometimes can get in a bubble and be completely distant from normal society and I think if you're in your house in your neighbourhood, it's much easier really.' The couple were asked why they decided to remain with their three sons Antonio, Alberto, and Miguel, in their £1.5million family home in Putney, south west London. The couple believe that choosing to remain in their family home has stopped them from becoming cut off from the rest of the world. Asked how they coped with the 'terrific kicking' given to her husband she said she didn't take it 'too seriously'. 'Just like any other marriage, we just stand together and it's not any more of a difficulty than what many others have to go through and you know. You should never take it too seriously.' And if he wanted five more years Mr Clegg said: 'Ten, 15, 20 why not! In for a penny, in for a pound.' He also insisted he felt more full of 'vim and vigour' than he did in the start of the coalition. Focus groups conducted by the programme found Mr Clegg was characterised as the 'chihuahua in a handbag' of the government. When asked what kind of drink he was the participants settled on Babycham. Asked how they coped with the 'terrific kicking' given to her husband, Mrs Clegg said she didn't take it 'too seriously' The Cleggs were seen drinking white wine and cooking paella in the kitchen of their home as they chatted about their family life. Honest: 'Discussion's a rather grand word for Miriam basically saying no,' Mr Clegg (left) joked during the interview. Ed Miliband was widely mocked after he posed with wife Justine in this picture, which turned out to be a second kitchen in his north London home used for 'tea and snacks' David Cameron invited the cameras into his Oxfordshire home, where he revealed he did not plan to stand for a third term. Mr Clegg sought to explain why his relations with the Prime Minister always seemed to be so cordial. He said: 'If you know you disagree with someone, there's no tension. I suspect meetings between Tony Blair and Gordon Brown were so fractious because they kind of should have agreed but they didn't. 'When David Cameron and I sit in a meeting, as we do week in week out, we kind of know that our starting point is that we come from different vantage points...' He claimed not to read all newspapers, and had learned how to ignore attacks form his opponents. 'It sounds glib but I actually think you can't take it too seriously otherwise you spend all your time reacting to stuff and you just have to laugh at it because some of it is faintly silly.' Mrs Clegg added that their close bond as a family has protected from the political brickbats. 'From my point of view if I spend my time thinking about whatever a specific person may has said, I don't have any time to do what I want to do.

Figure 9: An example CNN/DailyMail article for summaries presented in Figures 6, 7 and 8.

GOLD Question: What does the Premier of Victoria need to lead in the Legislative Assembly?

Context with Answer (in boldface): Answer: **most seats** <n> Context: The Premier of Victoria is the leader of the political party or coalition with the **most seats** in the Legislative Assembly. The Premier is the public face of government and, with cabinet, sets the legislative and political agenda. Cabinet consists of representatives elected to either house of parliament. It is responsible for managing areas of government that are not exclusively the Commonwealth's, by the Australian Constitution, such as education, health and law enforcement. The current Premier of Victoria is Daniel Andrews.

Single-best summaries

[✓] **PEGASUS** $(d \rightarrow s)$: How many seats does the Premier of Victoria have in the Legislative Assembly?

[✓] **FROST** $(d \rightarrow c; s)$: [CONTENT] Premier | Victoria | Legislative Assembly [SUMMARY] What does the Premier of Victoria have in the Legislative Assembly?

Nucleus Sampling: PEGASUS $(d \rightarrow s_{\text{sample}})$

[✓] $s_1 \rightarrow$ The Premier of Victoria would have how many seats in the Legislative Assembly?

[✗] $s_2 \rightarrow$ What is the politician MP expect to have in Legislative Assembly?

[✗] $s_3 \rightarrow$ Aside from being the leader of a political party or coalition, how is the Premier of Victoria Geometry of the Legislative Assembly?

[✗] $s_4 \rightarrow$ How many Legislative Assembly seats is the Premier of Victoria?

[✗] $s_5 \rightarrow$ What are the Legislative Assembly seats?

Nucleus Sampling: FROST $(d \rightarrow c_{\text{sample}}; s_{\text{sample}})$

[✗] $c_1; s_1 \rightarrow$ [CONTENT] criteria | Premier | Victoria | Coalition [SUMMARY] What is a Varied criteria for a Premier of Victoria to possess in a Coalition?

[✓] $c_2; s_2 \rightarrow$ [CONTENT] Premier | Victoria | leader | party | coalition | Legislative Assembly [SUMMARY] The Premier of Victoria isThe leader of the political party or coalition with to what in the Legislative Assembly?

[✓] $c_3; s_3 \rightarrow$ [CONTENT] number | Legislative Assembly | seats | Premier [SUMMARY] What is the number of Legislative Assembly seats that the Premier holds?

[✗] $c_4; s_4 \rightarrow$ [CONTENT] piece | legislature | leader | party | mixture | members [SUMMARY] What piece of the legislature does the leader of the party have a mixture of members?

[✓] $c_5; s_5 \rightarrow$ [CONTENT] Premier | Victoria | Legislative Assembly [SUMMARY] What does the Premier of Victoria have in the Legislative Assembly

---

Composition Sampling: FROST $(d \rightarrow c_{\text{sample}}; s)$

[✓] $c_1; s_1 \rightarrow$ [CONTENT] Premier | Victoria | Legislative Assembly [SUMMARY] What does the Premier of Victoria have in the Legislative Assembly?

[✓] $c_2; s_2 \rightarrow$ [CONTENT] Premier | party | coalition | Legislative Assembly [SUMMARY] The Premier of the political party or coalition has what in the Legislative Assembly?

[✓] $c_3; s_3 \rightarrow$ [CONTENT] Premier | Victoria | leader | party | Legislative Assembly [SUMMARY] The Premier of Victoria is the leader of the political party with what in the Legislative Assembly?

[✓] $c_4; s_4 \rightarrow$ [CONTENT] Premier | Victoria | party | coalition [SUMMARY] What does the Premier of Victoria have in his political party or coalition?

[✓] $c_5; s_5 \rightarrow$ [CONTENT] Premier | Victoria | leader | party | coalition | Legislative Assembly [SUMMARY] The Premier of Victoria is the leader of the political party or coalition with what in the Legislative Assembly?

Figure 10: An example input passage with answer in boldface, its human written question, and the model predictions including diverse questions for the SQuAD Question Generation dataset. We highlight spans in orange that are not accurate with respect to the input context.
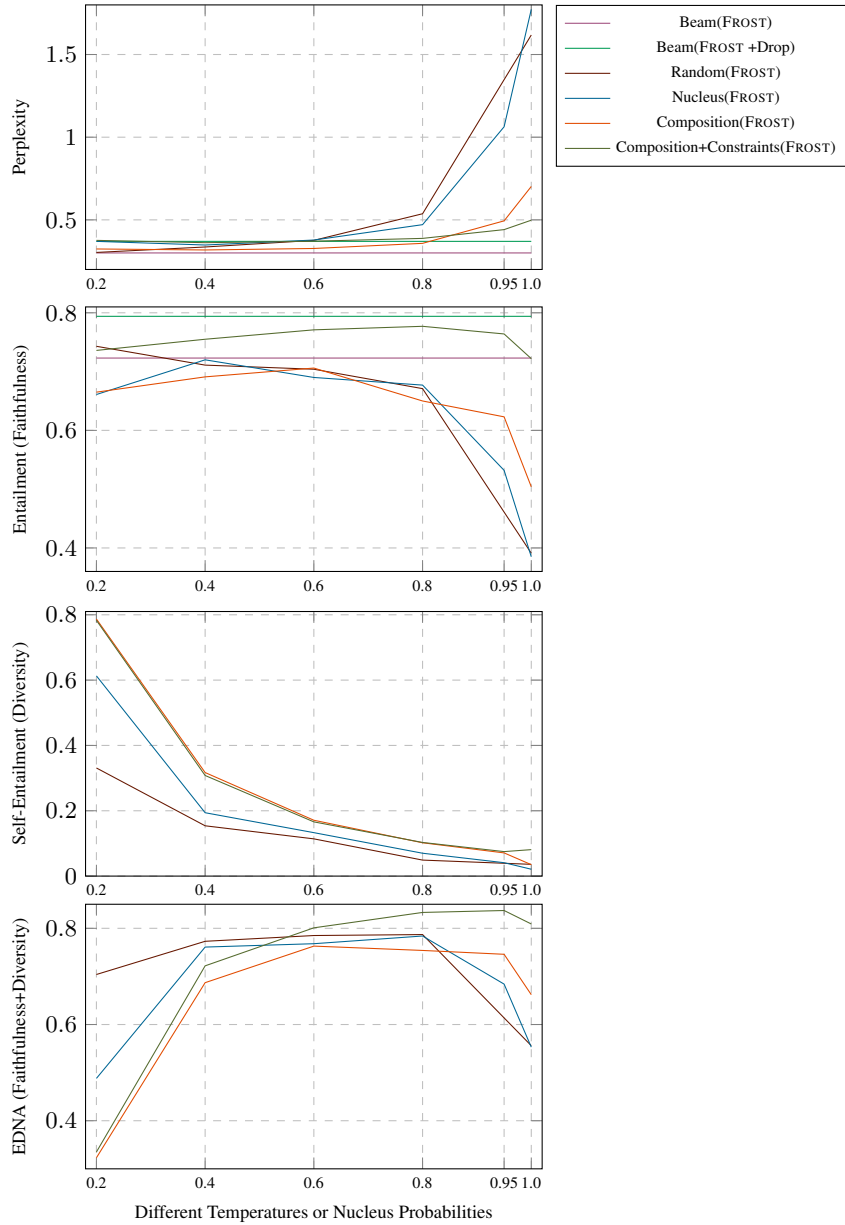
Figure 11: Perplexity, entailment, self-entailment and EDNA scores on the CNN/DailyMail challenge set (Table 2) with varying temperatures (for random sampling) and nucleus Probabilities (for nucleus and composition sampling). For each diverse decoding strategy, we sample 5 times for each document; we report on the average for each document.