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

Shashi Narayan
Google Research
shashinarayan@google.com

Gonçalo Simoes
Google Research
gsimoes@google.com

Yao Zhao
Google Brain
yaozhaoyz@google.com

Joshua Maynez
Google Research
joshuahm@google.com

Dipanjan Das
Google Research
dipanjand@google.com

Michael Collins
Google Research
mjcollins@google.com

Mirella Lapata
Google Research
lapata@google.com

Abstract

We propose Composition Sampling, a simple but effective method to generate diverse outputs for conditional generation of higher quality compared to previous stochastic decoding strategies. It builds on recently proposed plan-based neural generation models (Narayan et al., 2021) that are trained to first create a composition of the output and then generate by conditioning on it and the input. 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 this entity chain. Experiments on summarization (CNN/DailyMail and XSum) and question generation (SQuAD), using existing and newly proposed automatic metrics together with human-based evaluation, demonstrate that Composition Sampling is currently the best available decoding strategy for generating diverse meaningful outputs.

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 may be useful for a model to be able 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., 2020a; Narayan et al., 2021), while diversity remains an unsolved problem (Hashimoto et al., 2019; Zhang et al., 2021). This is particularly challenging in conditional generation, where diversity in the target sequence 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 text which is also faithful to its input focusing on two tasks, namely summarization and question generation.

Beam search (Li et al., 2016; Wiseman et al., 2017) has proven successful for single-best generation (Rush et al., 2015; Barrault et al., 2020; Meister et al., 2020), but struggles to generate diverse output (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 diverse sequences 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 generation models (Narayan et al., 2021) that are trained to first *plan* a semantic *composition* of the target and then generate the text conditioned on the composition *and* the input. Composition sampling first samples a 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, it avoids degeneration by instilling diversity in composition, rather than directly on the surface form.

Our contributions can be summarized as follows: (a) we introduce Composition Sampling, a simple yet effective decoding method for diverse conditional generation, which combines planning with stochastic sampling; (b) we propose several metrics to compute semantic diversity in generated text; our metrics are complementary to lexical diversity

[1] Holtzman et al. (2020) use the term 'degeneration' to describe automatically generated text that is generic, repetitive, and awkward for story continuation. These issues are less common in conditional generation. In our case, 'degenerate' refers to text unfaithful or inconsistent to the input.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1319 - 1339

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: (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: (Chelsea | Eden Hazard | Arsenal | Santi Cazorla ||| Sportsmail ||| London) Ch elsea'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), and diverse summaries with nucleus sampling ($p = 0.95$) and our composition sampling for a CNN/DailyMail article (shown in the Appendix, Figure 6). We highlight spans in orange that are not faithful to the input.

(e.g., Self-BLEU; Zhu et al. 2018; Alihosseini et al. 2019) and assess whether a set of diverse outputs are contextually dissimilar (*Self-BERTscore*; Zhang et al. 2020b) or non-entailing (*Self-Entailment*); and (c) finally, we introduce, EDNA, a novel metric aiming to "**E**valuate **D**iversity a**N**d f**A**ithfulness" for summarization by quantifying whether summaries in a diverse set are faithful to their input without entailing each other.

Evaluation on two popular summarization tasks, namely highlight generation (CNN/DailyMail; Hermann et al. 2015) and extreme summarization (XSum; Narayan et al. 2018), and question generation (SQuAD; Rajpurkar et al. 2016; Zhou et al. 2017), shows that composition sampling is most effective in generating diverse summaries or questions. When assessed by humans, composition sampled summaries are as faithful as the best summaries produced with beam search. In comparison, nucleus sampled summaries can be as diverse but far less faithful. Taken together our results demonstrate that Composition Sampling is currently the best available decoding strategy for generating diverse and meaningful output.[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 $s_{1:n}$ (e.g., the translation or 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., 2020; Raffel et al., 2019; Zhang et al., 2020a).

Plan-based Conditional Generation Narayan et al. (2021) develop a *plan-based* approach for neural summarization; their decoder generates a composition $c_{1:m}$ of target summary s as $p(c_j|c_1, \ldots, c_{j-1}; d)$, and then the same decoder produces s as $p(s_i|s_1, \ldots, s_{i-1}; c; d)$ conditioned on input d and composition $c_{1:m}$, with m being the composition length. Specifically, they adopt entity chains as the composition c of summary s, under the assumption that entities in the chain ought to be observed in the output summary. During inference, the model takes document d as input and generates $c; s$, the concatenation of composition and summary sequences, instead of generating s directly; c and s are prefixed with special markers "[CONTENT]" and "[SUMMARY]", respectively, as shown in Figure 2. If s consists of multiple sentences, markers"|||" denote sentence boundaries in composition c.

[2]Our checkpoints and spaCy annotation code are available at https://github.com/google-research/language/tree/master/language/frost.

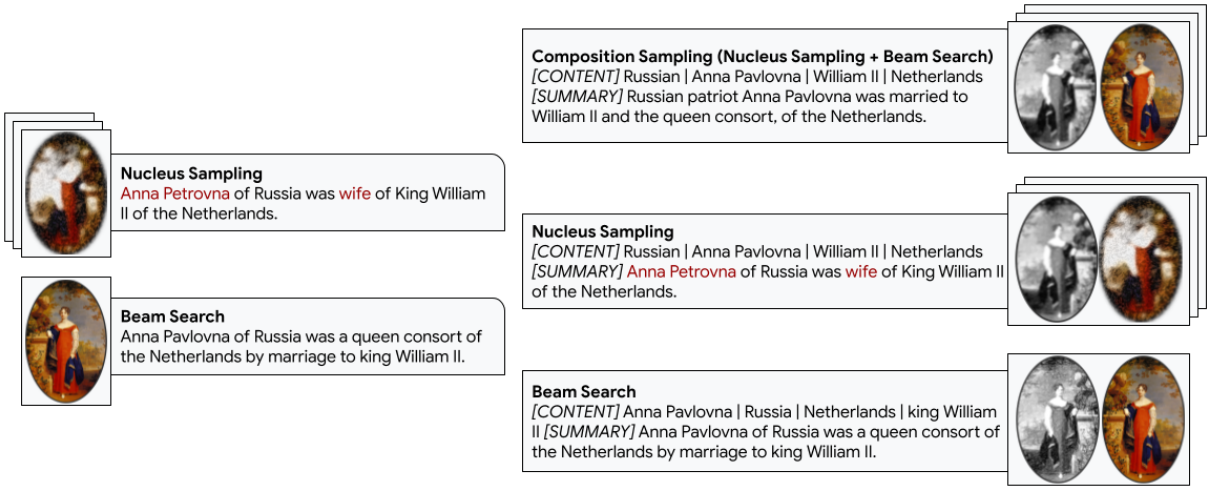


Figure 2: Illustration of composition sampling and other decoding strategies with vanilla and plan-based generation models. The term 'composition' is inspired from the quote "*A Well-Composed Painting is Half Done*" from French painter Pierre Bonnard. Images in black-and-white are early sketches or compositions of the painting in color. Nucleus or focus sampling often lead to hallucinations (highlight spans in red); corresponding color images are blurred to illustrate this. (Credit: The image of "Anna Pavlovna of Russia" is taken from Wikipedia.)

The approach allows to directly manipulate the content of summaries and their quality. For example, we might inspect the predicted chain during inference and drop entities which are not present in the input document, thereby controlling for hallucinations (Narayan et al., 2021). Outwith summarization, similar constraints can be easily adapted to other conditional generation tasks.

Maximization-Based Decoding In order to obtain the most likely output $\hat{s}$ from encoder-decoder models, we typically solve a maximization-based 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). It is standard practice to use *beam search* (Tillmann and Ney, 2003; Li et al., 2016; Wiseman et al., 2017) as solving the objective for the optimal sequence with neural sequence models is not tractable (Chen et al., 2018).

Stochastic Sampling 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 k high probability tokens (Fan et al., 2018; Holtzman et al., 2018; Radford et al., 2019). Similarly to top-k sampling, *nucleus sampling* (Holtzman et al., 2020) also truncates the tail of the distribution but chooses k dynamically. At each decoding step, it samples high-probable tokens from a nucleus N defined as the smallest subset of tokens from the vocabulary V with cumulative probability $p' \geq p$, where p is the pre-specified mass of the nucleus.

Aralikatte et al. (2021) introduce *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. Standard beam search decoding is then used to generate a summary limited to V_k. However, the authors show that focus sampling is very sensitive to k; increasing it improves generation quality but at the cost of diversity.

3 Composition Sampling

Composition Sampling is a novel hybrid method which combines stochastic sampling with maximization-based decoding, whilst leveraging plan-based generation (Narayan et al., 2021). Specifically, we employ nucleus sampling to obtain diverse compositions c_{sample} from $p(c|d)$ where d is the input text and c are entity chains (prefixed with "[CONTENT]" in Figure 2). We first employ nucleus sampling to obtain diverse compositions from $p(c|d)$, where d is the input text. And then employ beam search to generate the most-likely diverse output s (prefixed with "[SUMMARY]" in Figure 2), given input d and composition c_{sample} as

$p(s|c_{\text{sample}}; d)$. We will experimentally show that composition sampling enables the generation of fluent, faithful and diverse texts for conditional generation.

Why Entity Chains? Unlike top-k or nucleus sampling, composition sampling avoids degeneration by introducing diversity in composition, rather than directly on 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 $c_1 \neq c_2$, then two summaries $s_1 = \arg\max_s p(s|c_1; d)$ 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. Moreover, entity chains provide a very effective knob for content control in abstractive generation, e.g., compositions can be constrained to entities only present in the input document, thereby avoiding hallucinations and entity degeneration.

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 summarization example in Figure 3. The negative log likelihood of generating the summary s from scratch without planning ($-\log p(s|d)$) is 121.18, while the negative log likelihood of generating composition c with planning ($-\log p(c|d)$) is 46.95; hence, the model is much more confident when sampling from $p(c|d)$ than 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.

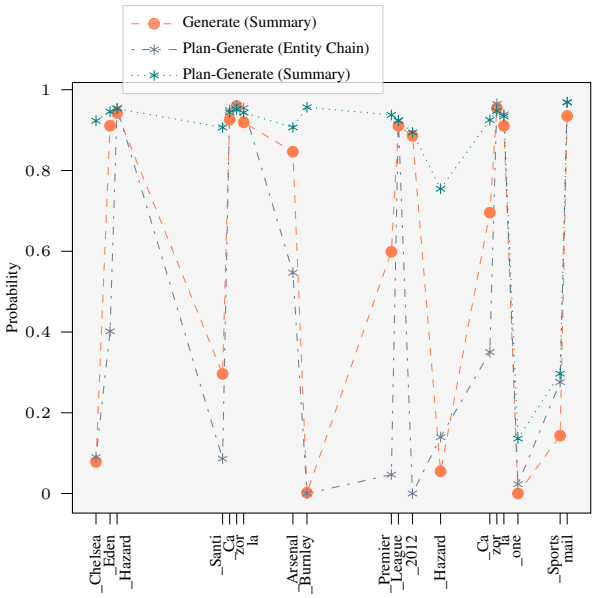


Figure 3: Probabilities of generating underlined entities in human written reference summary from Figure 1 (input article shown in Figure 6): when the summary is generated directly (Generate, Summary), when the entity chain "*Chelsea | Eden Hazard ||| Santi Cazorla | Arsenal | Burnley ||| Premier League | 2012 ||| Hazard | Cazorla | one ||| Sportsmail*" is predicted first during planning (Plan-Generate, Entity Chain), and when the entities are predicted in the summary after planning (Plan-Generate, Summary). All probabilities were computed with PEGASUS (Zhang et al., 2020a) fine-tuned models. The symbol '_' denotes start of token.

In Figure 3, the plan-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 in the plan-based model (i.e., $c_{1:m}; s_{1:n}$ instead of $s_{1:n}$), we find that the perplexity of predicting the longer sequence ($c_{1:m}; s_{1:n}$) is lower than predicting just the summary without any planning, due to grounding (1.16 vs 1.48). Overall, $p(c; s|d)$, the plan-based approach, learns a more confident distribution at each decoding step compared to no planing, i.e., $p(s|d)$. 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 $p(c; s|d)$.

In the following we assess composition sampling for its ability to generate semantically diverse output for two tasks, namely summarization (Sec-

tion 4) and question generation (Section 5).

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 the original train/validation/test splits. Inputs and outputs were truncated to 512 and 128 for XSum, and, 1,024 and 256 for CNN/DailyMail.[3]

We conduct experiments with state-of-the-art pretrained models for summarization, namely PEGASUS (Zhang et al., 2020a) and FROST (Narayan et al., 2021). Our PEGASUS finetuned model generates summaries directly, whereas FROST generates the entity chain followed by the summary. In both cases we use large transformer architectures (Vaswani et al., 2017) with $L = 16$, $H = 1,024$, $F = 4,096$, $A = 16$ (568M parameters), where L denotes the number of layers for encoder and decoder Transformer blocks, H is the hidden size, F the feed-forward layer size, and A 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. (2020a) and Narayan et al. (2021).

All models are decoded with a beam size of 8 and a length-penalty of 0.8. For nucleus sampling and composition sampling, we use nucleus probability p set to 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 plan in the form of an entity chain, followed by a summary. All evaluations, except likelihood, are done on the summary, while the predicted entity chains are stripped out. For each *diverse* decoding strategy, we sample 5 times for each test document and report the average.

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

[4]Results with different temperatures and nucleus probabilities for random sampling, nucleus sampling, and composition sampling are in Figure 4.

Sequence Likelihood We report the perplexity of the generated sequence (i.e., entity chains concatenated with their summaries for planning models and summaries only for the 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., 2020b) 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 a document and then grade its summary on a scale of 1–4 (*entirely unfaithful*, *somewhat unfaithful*, *somewhat faithful*, and *entirely faithful*); a summary is "entirely faithful" if its content is fully supported or can be inferred from the document. We collected 3 ratings for each (document, summary) pair; we report average system ratings (across documents). With summaries deemed "somewhat unfaithful" and "somewhat faithful", annotators were asked to also specify what was faithful or unfaithful, to improve agreement.

Diversity We report the number of times (out of five samples), a decoding technique is able to generate a completely new summary (*Unique*). We also use *Self-BLEU* (Zhu et al., 2018; Alihosseini et al., 2019) to measure lexical diversity in the generated summaries. We consider all pairs of summaries out of 5 samples, and for each pair we compute the BLEU score (Papineni et al., 2002) considering one summary as a hypothesis and the other 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 a more diverse set of summaries.

[5]We lowercased candidate and reference summaries and used `pyrouge` with parameters "-a -c 95 -m -n 4 -w 1.2."

	Model	XSum			CNN/DailyMail		
		R1	R2	RL	R1	R2	RL
Single	GSum (Dou et al., 2020)	45.40	21.89	36.67	**45.94**	22.32	42.48
	CTRLsum (He et al., 2020)	—	—	—	45.65	**22.35**	**42.50**
	FAME (Aralikatte et al., 2021)	45.31	22.75	37.46	42.95	20.79	39.90
	PEGASUS (Zhang et al., 2020a)	47.56	24.87	39.40	44.05	21.69	40.98
	FROST (Narayan et al., 2021)	**47.80**	**25.06**	**39.76**	45.11	22.11	42.01
	FROST$_{++}$ (Narayan et al., 2021)	44.94	21.58	37.20	45.08	22.14	41.99
Diverse	Focus (FAME)	42.76	19.89	34.97	—	—	—
	Nucleus (PEGASUS)	38.49	16.57	30.99	36.27	15.10	33.46
	Nucleus (FROST)	40.26	17.83	32.49	38.49	15.71	35.49
	Composition (FROST)	**45.12**	**22.24**	**36.98**	41.76	18.94	38.69
	Composition (FROST$_{++}$)	43.82	20.35	35.89	**42.37**	**19.48**	**39.28**

Table 1: Comparison of decoding strategies with ROUGE: single-best vs diverse decoding (we report averages over 5 samples). Best results in each block are bold-faced. See Table 5 in the Appendix for more comparisons.

We propose two additional measures to capture semantic diversity in summaries: *Self-Entailment* and *Self-BERTScore*. Similar to Self-BLEU, we compute the Entailment score and BERTScore for each possible pair of summaries, respectively and report the average. A lower Self-Entailment value suggests that the generated summaries do not entail each other. Analogously, a low Self-BERTScore value indicates that the decoding technique is able to generate more contextually dissimilar summaries.

We further assess diversity by humans. Our annotators, proficient in English, again read two summaries (out of five samples) and then graded 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 equivalent, while the same information may be presented differently in the case of "somewhat identical". A "somewhat diverse" pair may introduce one or two new concepts or topics in one summary. A "diverse" pair should introduce new concepts or topics in each summary. We collected 3 ratings for each pair and report their average. This assessment was only done with single-sentence XSum summaries, in future work we will explore how to do this effectively for multi-sentence 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 for "**E**valuating **D**iversity a**N**d f**A**ithfulness" in summaries. EDNA is the harmonic mean of Entailment and (1−Self-Entailment); higher values of EDNA imply more faithful and diverse summaries. The reason EDNA relies on Self-Entailment to measure diversity is because the faithfulness metric is also based on Entailment. This means that both components will be mapped to a score in a similar output space (i.e., they both yield values between 0 and 1 obtained through the same trained model), making it more likely to be properly balanced when mixed.

4.3 Results

Table 1 presents ROUGE results on the XSum and CNN/DailyMail test sets. The top block includes results for models which employ maximization-based decoding. GSum (Dou et al., 2020) is a state-of-the art system which decodes summaries guided by an an extractive model at test time. CTRLsum (He et al., 2020) controls the summarization output trough keywords and automatically extracted sentences. FAME (Aralikatte et al., 2021) uses a focus attention mechanism to encourage the decoder to proactively generate tokens that are similar or topical to the input document. As mentioned earlier FROST (Narayan et al., 2021) first generates an entity chain and then a summary while FROST$_{++}$ is a constrained variant which restricts the predicted entities to those present in the document. We also show results for a vanilla PEGASUS model (Zhang et al., 2020a) finetuned on our datasets.

The bottom block focuses on diverse decoding (we report averages across five samples). We show results with Focus sampling (Aralikatte et al., 2021) built on top of FAME, Nucleus sampling (Holtzman et al., 2020) with PEGASUS and FROST, and our Composition sampling.

Table 2 presents more detailed faithfulness and diversity results, on challenge sets consisting of 50 documents for each XSum and CNN/DailyMail

		Models	ppl	With Ref. RL	With Ref. BSc	Faithfulness Ent	Faithfulness Human	Diversity Uniq	Diversity S-BL	Diversity S-Ent	Diversity S-BSc	Diversity Human	Div+Faith EDNA
XSum	Single	FAME	—	34.23	0.70	0.24	2.19	—	—	—	—	—	—
		PEGASUS	0.51	40.69	0.76	0.40	2.52	—	—	—	—	—	—
		FROST	0.31	40.90	0.75	0.37	2.63	—	—	—	—	—	—
		FROST$_{++}$	0.71	33.75	0.70	0.44	2.78	—	—	—	—	—	—
	Diverse	Focus (FAME)	—	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 (FROST$_{++}$)	0.74	33.87	0.70	**0.43**	**2.75**	3.5	76.87	0.40	0.84	2.25	**0.35**
CNN/DM	Single	PEGASUS	0.35	36.09	0.65	0.70	3.78	—	—	—	—	—	—
		FROST	0.30	39.03	0.66	0.72	3.74	—	—	—	—	—	—
		FROST$_{++}$	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 (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 the 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 (FAME) diverse predictions on CNN/DailyMail are not available. The lower the Self-* metric is, the better the decoding method in generating diverse outputs.

summaries. We construct these challenge sets by randomly selecting documents whose reference summaries have non-extractive entity chains in them; an entity chain is 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 our human evaluations of faithfulness and diversity.

Composition Sampling is not as Performance Diminishing as Nucleus Sampling Single-best decoding for FROST achieves 39.76 ROUGE (RL) on XSum,; nucleus and composition sampling fare worse showing an average drop of 7.27 and 2.78, respectively. Similarly, for CNN/DailyMail, ROUGE drops for nucleus sampling by an average of 6.51 points, compared to an average drop of 3.28 points for composition sampling (with FROST). Nucleus sampling is even more damaging for non-plan based models, such as PEGASUS; we see an average drop of 8.59 and 7.30 ROUGE points on XSum and CNN/DailyMail. These gaps are slightly larger for the challenging subsets in Table 2 which is expected due to the highly abstractive nature of the reference summaries therein.

On XSum, composition Sampling with FROST$_{++}$ performs slightly worse than with vanilla FROST in terms of ROUGE. This is due to the extreme abstractive nature of the XSum reference summaries (Maynez et al., 2020); as a result, a model is required to hallucinate factual content, that is not necessarily faithful to the input (see examples of XSum summaries in the Appendix, Figure 5). But Composition(FROST$_{++}$) only keeps supported entities in the sampled plans giving rise to summaries which diverge from their reference. This is not the case with CNN/DailyMail which is mostly extractive and we see that ROUGE performance improves with Composition(FROST$_{++}$) in Table 1.

Composition Sampling is more Confident in its 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(FROST $_{++}$) is more perplexed when generating XSum summaries due to the reference summary divergence issue discussed earlier; perplexity increases from 0.51 to 0.74 compared to Composition(FROST). Interestingly, Composi-

tion(FROST$_{++}$) is almost as confident in generating diverse summaries as single-best beam decoding (FROST; perplexities of 0.71 vs 0.74 for XSum). Unsurprisingly, Composition(FROST$_{++}$) is more confident in generating CNN/DailyMail summaries than FROST (0.46 vs 0.52) due to their extractive nature.

Constrained Composition 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 assessed by humans); however these summaries perform poorly on faithfulness measures. Composition(FROST$_{++}$) is most effective in generating faithful summaries, as demonstrated automatically (with best entailment scores on XSum and CNN/DailyMail) and by humans (with highest ratings on XSum and CNN/DailyMail); these summaries are also diverse achieving highest EDNA scores on both summarization datasets.

We also examined whether models differ in terms of faithfulness and diversity as rated by our participants. We carried out pairwise comparisons using one-way ANOVA with post-hoc Tukey HSD tests ($p < 0.01$). The difference between Nucleus(PEGASUS) and Nucleus(FROST) is not significant. Nucleus(PEGASUS) was also not significantly more faithful than Focus(FAME) for XSum summaries. All other pairwise differences were significant for both faithfulness and diversity.

In sum, our results demonstrate that composition sampling is a better alternative to nucleus or focus sampling for generating meaningful diverse summaries. Figure 1 presents summaries from different decoding strategies for a CNN/DailyMail article. Other example predictions for XSum and CNN/DailyMail articles can be found in the Appendix (Figures 5–9).

Faithfulness and Diversity Metrics Correlate with Human Judgements We estimate the extent to which automatic metrics of faithfulness and diversity correlate with human ratings (using Spearman's rank correlation coefficient) in Table 3. In line with previous work (Maynez et al., 2020; Kryscinski et al., 2019), we find that entailment scores are best correlated with faithfulness (moderate, $0.40 \leq r \leq 0.59$). Like Self-BLUE, Self-Entailment and Self-BERTScore are also strongly correlated with diversity ratings. Compared to other metrics which capture a single dimension, EDNA is positively correlated with both dimensions of diversity *and* faithfulness.

Metric	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: Different automatic metrics and their correlation against human assessments using Spearman's rank coefficient.

Finally, in Figure 4, we plot faithfulness and diversity scores for different decoding strategies with varying temperatures and nucleus probabilities. We find that summaries sampled with Composition(FROST$_{++}$) are consistently more faithful than single-best Beam(FROST) summaries, but worse than summaries decoded with Beam(FROST$_{++}$). Summaries sampled with Composition(FROST$_{++}$) achieve the best EDNA score (with $p = 0.95$) amongst all diverse decoding strategies.

5 Question Generation

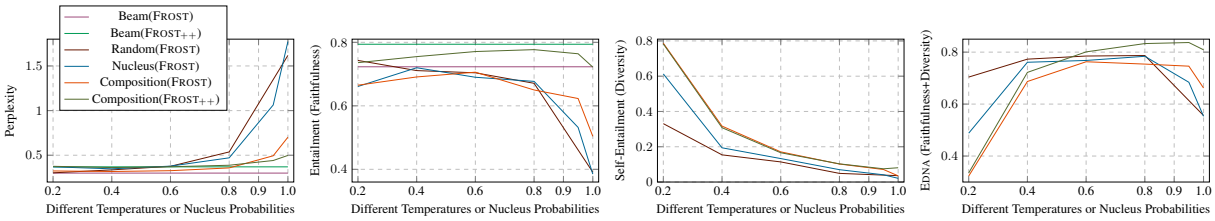
5.1 Dataset and Metrics

Question generation is often conceptualized 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 split of Zhou et al. (2017) consisting of 86,635, 8,965, and 8,964 source-target pairs for training, validation, and testing, respectively.[6] We follow Cho et al. (2019) and report BLEU-4 (Top-1, the single-best accuracy), Oracle (Top-5, the best accuracy among the 5-sampled hypotheses), and Self-BLEU (as defined in Section 4).

5.2 Results

For our question generation experiments we also compare models which employ single-best decoding against models which adopt diverse decoding techniques. The top block in Table 4 presents results for NQG++ (Zhou et al., 2017), a pointer generator-based model, CP+GSA (Zhao et al.,

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



a. b. c. d.

Figure 4: 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 per document and report the average.

	Models	T1	T5	S-BL
Single	NQG++ (Zhou et al., 2017)	13.27	—	—
	MCP+GSA (Zhao et al., 2018)	16.85	—	—
	PEGASUS (Zhang et al., 2020a)	**22.17**	—	—
	FROST (Narayan et al., 2021)	21.04	—	—
Diverse	top-k Sampling	11.53	17.65	45.99
	Diverse Beam Search	13.38	18.30	74.80
	Mixture Decoder (Shen et al.)	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)	17.16	**27.04**	61.68
	Composition (FROST$_{++}$)	**18.77**	26.60	74.89

Table 4: Comparison of different decoding techniques on question generation. We report on BLEU-4 Top-1 accuracy (T1) and Top-5 (T5), and Self-BLEU (S-BL). Results for diverse decoding comparison models are taken from Wang et al. (2020). Best results in each block are bold-faced.

2018), a model which combines a pointer mechanism with a gated self-attention encoder, and fine-tuned PEGASUS and FROST models. The second block in the table contains several diverse decoding approaches including top-k sampling (Fan et al., 2018), diverse beam search (Vijayakumar et al., 2018), mixture decoding (Shen et al., 2019) and mixture content selection (Cho et al., 2019; Wang et al., 2020). We compare these models against nucleus sampling with PEGASUS and FROST, and composition sampling with FROST.

As in our summarization experiments, we observe that composition sampling is not as performance diminishing as nucleus sampling, in terms BLEU. FROST achieves a BLEU of 21.04 (top-1) in the single-best decoding setting; in comparison, BLEU drops for nucleus sampling by 10.40 points (on average), and 2.27 points only for composition sampling (FROST$_{++}$). 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: (a) SQuAD questions are mostly extractive, and (b) questions are generated conditioned on the passage *and* the answer spans; leaving limited scope for models to generate diverse questions. An example in the Appendix (Figure 11) 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 system to augment the input during inference. Our experiments demonstrate that it is currently the best available decoding strategy for generating diverse and meaningful output. We also introduced Self-Entailment and Self-BERTScore, to automatically compute semantic diversity in summaries, and, EDNA, for jointly measuring faithfulness and diversity.

Acknowledgements

We thank the reviewers, the ARR action editor, and the senior area chair for their valuable feedback. We would like to thank Ryan McDonald, Ankur Parikh, and Slav Petrov for their insightful comments. Many thanks also to Ashwin Kakarla and his team for their help with the human evaluation.

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

Ethical Considerations

The nature of text generation leads to multiple ethical considerations when considering 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 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 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 and evaluation, 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 sections 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.

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.

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.

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, 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.

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 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.

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

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. 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.

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, 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.

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. 2020. 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.

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.

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.

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.

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 Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. 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.

	Models	XSum R1	XSum R2	XSum RL	CNN/DailyMail R1	CNN/DailyMail R2	CNN/DailyMail RL
Single	RoBERTaShare (Rothe et al., 2020)	38.52	16.12	31.13	39.25	18.09	36.45
	MASS (Song et al., 2019)	39.75	17.24	31.95	42.12	19.50	39.01
	BART (Lewis et al., 2019)	45.14	22.27	37.25	44.16	21.28	40.90
	GSum (Dou et al., 2020)	45.40	21.89	36.67	**45.94**	22.32	42.48
	UniLM (Dong et al., 2019)	—	—	—	43.33	20.21	40.51
	T5 (Raffel et al., 2019)	—	—	—	43.52	21.55	40.69
	ProphetNet (Qi et al., 2020)	—	—	—	44.20	21.17	41.30
	CTRLsum (He et al., 2020)	—	—	—	45.65	**22.35**	**42.50**
	FAME (Aralikatte et al., 2021)	45.31	22.75	37.46	42.95	20.79	39.90
	PEGASUS (Zhang et al., 2020a)	47.56	24.87	39.40	44.05	21.69	40.98
	FROST (Narayan et al., 2021)	**47.80**	**25.06**	**39.76**	45.11	22.11	42.01
	FROST$_{++}$ (Narayan et al., 2021)	44.94	21.58	37.20	45.08	22.14	41.99
Diverse	Focus (FAME)	42.76	19.89	34.97	—	—	—
	Nucleus (PEGASUS)	38.49	16.57	30.99	36.27	15.10	33.46
	Nucleus (FROST)	40.26	17.83	32.49	38.49	15.71	35.49
	Composition (FROST)	**45.12**	**22.24**	**36.98**	41.76	18.94	38.69
	Composition (FROST$_{++}$)	43.82	20.35	35.89	**42.37**	**19.48**	**39.28**

Table 5: Full set of ROUGE results on XSum and CNN/DailyMail test sets comparing different decoding techniques and SOTA models. Best results in each block are bold-faced.

		Models	With Reference R1	With Reference R2	With Reference RL
XSum	Single	Focus (FAME)	41.20	20.30	34.23
		PEGASUS	49.49	28.43	40.69
		FROST	49.12	28.35	40.90
		FROST$_{++}$)	41.15	19.66	33.75
	Diverse	Focus (FAME)	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 (FROST$_{++}$)	41.81	19.61	33.87
CNN/DailyMail	Single	PEGASUS	38.50	15.04	36.09
		FROST	41.89	17.54	39.03
		FROST$_{++}$	41.82	17.96	38.87
	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 (FROST$_{++}$)	37.79	**15.07**	**35.07**

Table 6: Full set of ROUGE results on 50 documents sampled from XSum and CNN/DailyMail (see also Table 2 in the main paper).

Models	R1	R2	RL
Single-best with Beam Search			
PEGASUS	47.52	18.72	24.91
FROST	43.12	16.93	22.49
Diverse Decoding, Average of five runs			
Nucleus (FROST)	39.50	12.94	19.50
Composition (FROST)	42.47	15.43	21.43
Composition (FROST$_{++}$)	42.37	15.78	21.90
Diverse Decoding, Best of five runs			
Nucleus (FROST)	44.40	16.86	23.03
Composition (FROST)	46.98	19.34	24.96
Composition (FROST$_{++}$)	46.71	19.55	25.36

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

Models	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)	16.62	**26.07**	62.47
Composition (FROST$_{++}$)	**17.28**	25.03	75.81

Table 8: We also experimented with the split of Du et al. (2017) for SQuAD (Rajpurkar et al., 2016) question generation, consisting of 70,484, 10,570, and 11,877 examples for training, validation, and testing, respectively. 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 ($d \rightarrow c_{\text{drop}}; s$): [CONTENT] Walsall | Falkirk | Leahy [SUMMARY] Walsall have signed Falkirk defender Leahy on a free transfer.

Focus Sampling: FAME

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

Nucleus Sampling: PEGASUS

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

Nucleus Sampling: FROST

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

$\mathbf{c_1}; \mathbf{s_1} \rightarrow$ [CONTENT] Walsall | Rory Leahy | Falkirk [SUMMARY] Walsall have signed defender Rory Leahy from Falkirk.
$\mathbf{c_2}; \mathbf{s_2} \rightarrow$ [CONTENT] Walsall | Falkirk | Liam Leahy [SUMMARY] Walsall have signed Falkirk defender Liam Leahy.
$\mathbf{c_3}; \mathbf{s_3} \rightarrow$ [CONTENT] Falkirk | Wade Leahy | Walsall [SUMMARY] Falkirk defender Wade Leahy has joined Walsall.
$\mathbf{c_4}; \mathbf{s_4} \rightarrow$ [CONTENT] Walsall | Todd Leahy | Scottish Championship | Falkirk [SUMMARY] Walsall have signed defender Todd Leahy from Scottish Championship side Falkirk.
$\mathbf{c_5}; \mathbf{s_5} \rightarrow$ [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 FROST$_{++}$

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

Figure 5: Example input article, its human written summary, and model predictions for the XSum dataset. We highlight spans in orange that are not faithful to the input. We use $c*$ and $s*$ to denote different compositions and their corresponding summaries.

Chelsea star Eden Hazard vs Arsenal playmaker Santi Cazorla: As duo prepare to reach 100 Premier League games, who has excited our experts the most since 2012?

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 6: Input 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: 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: [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$_{++}$: [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 7: Human written summary and single-best summaries using beam search for the CNN/DailyMail article in Figure 10.

Nucleus Sampling: PEGASUS
$\mathbf{s_1} \rightarrow$ 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. $\mathbf{s_2} \rightarrow$ 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'. $\mathbf{s_3} \rightarrow$ 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. $\mathbf{s_4} \rightarrow$ 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. $\mathbf{s_5} \rightarrow$ 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
$\mathbf{c_1}; \mathbf{s_1} \rightarrow$ [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' $\mathbf{c_2}; \mathbf{s_2} \rightarrow$ [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. $\mathbf{c_3}; \mathbf{s_3} \rightarrow$ [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'!! $\mathbf{c_4}; \mathbf{s_4} \rightarrow$ [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. $\mathbf{c_5}; \mathbf{s_5} \rightarrow$ [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 8: Diverse summaries with nucleus sampling for the CNN/DailyMail article in Figure 10. We highlight spans in orange that are not faithful to the input document. We use $c*$ and $s*$ to denote different compositions and their corresponding summaries.

Composition Sampling: FROST
$\mathbf{c_1}; \mathbf{s_1} \rightarrow$ [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.
$\mathbf{c_2}; \mathbf{s_2} \rightarrow$ [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.
$\mathbf{c_3}; \mathbf{s_3} \rightarrow$ [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.
$\mathbf{c_4}; \mathbf{s_4} \rightarrow$ [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.
$\mathbf{c_5}; \mathbf{s_5} \rightarrow$ [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: FROST$_{++}$
$\mathbf{c_1}; \mathbf{s_1} \rightarrow$ [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.
$\mathbf{c_2}; \mathbf{s_2} \rightarrow$ [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.
$\mathbf{c_3}; \mathbf{s_3} \rightarrow$ [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.
$\mathbf{c_4}; \mathbf{s_4} \rightarrow$ [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.
$\mathbf{c_5}; \mathbf{s_5} \rightarrow$ [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 9: Diverse summaries with composition sampling for the CNN/DailyMail article in Figure 10. We highlight spans in orange that are not faithful to the input document. We use $c*$ and $s*$ to denote different compositions and their corresponding summaries.

Inside the Clegg kitchen: Over white wine and paella Nick reveals how Miriam put her foot down and refused to swap their family home for a grace-and-favour property

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 10: CNN/DailyMail input article for the summaries presented in Figures 7–9.

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: How many seats does the Premier of Victoria have in the Legislative Assembly? **FROST**: [CONTENT] Premier | Victoria | Legislative Assembly [SUMMARY] What does the Premier of Victoria have in the Legislative Assembly?
Nucleus Sampling: PEGASUS
$\mathbf{s_1} \rightarrow$ The Premier of Victoria would have how many seats in the Legislative Assembly? $\mathbf{s_2} \rightarrow$ What is the politician MP expect to have in Legislative Assembly? $\mathbf{s_3} \rightarrow$ Aside from being the leader of a political party or coalition, how is the Premier of Victoria Geometry of the Legislative Assembly? $\mathbf{s_4} \rightarrow$ How many Legislative Assembly seats is the Premier of Victoria? $\mathbf{s_5} \rightarrow$ What are the Legislative Assembly seats?
Nucleus Sampling: FROST
$\mathbf{c_1}; \mathbf{s_1} \rightarrow$ [CONTENT] criteria | Premier | Victoria | Coalition [SUMMARY] What is a Varied criteria for a Premier of Victoria to possess in a Coalition? $\mathbf{c_2}; \mathbf{s_2} \rightarrow$ [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? $\mathbf{c_3}; \mathbf{s_3} \rightarrow$ [CONTENT] number | Legislative Assembly | seats | Premier [SUMMARY] What is the number of Legislative Assembly seats that the Premier holds? $\mathbf{c_4}; \mathbf{s_4} \rightarrow$ [CONTENT] piece | legislature | leader | party | mixture | members [SUMMARY] What piece of the legislature does the leader of the party have a mixture of members? $\mathbf{c_5}; \mathbf{s_5} \rightarrow$ [CONTENT] Premier | Victoria | Legislative Assembly [SUMMARY] What does the Premier of Victoria have in the Legislative Assembly
Composition Sampling: FROST
$\mathbf{c_1}; \mathbf{s_1} \rightarrow$ [CONTENT] Premier | Victoria | Legislative Assembly [SUMMARY] What does the Premier of Victoria have in the Legislative Assembly? $\mathbf{c_2}; \mathbf{s_2} \rightarrow$ [CONTENT] Premier | party | coalition | Legislative Assembly [SUMMARY] The Premier of the political party or coalition has what in the Legislative Assembly? $\mathbf{c_3}; \mathbf{s_3} \rightarrow$ [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? $\mathbf{c_4}; \mathbf{s_4} \rightarrow$ [CONTENT] Premier | Victoria | party | coalition [SUMMARY] What does the Premier of Victoria have in his political party or coalition? $\mathbf{c_5}; \mathbf{s_5} \rightarrow$ [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 11: Example input passage with answer in boldface, human written question, and 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. We use $c*$ and $s*$ to denote different compositions and their corresponding questions.