Diversity as a By-Product: Goal-oriented Language Generation Leads to Linguistic Variation

Simeon Schüz[1]*, **Ting Han**[2], **Sina Zarrieß**[1]
[1]Bielefeld University, [2]Artificial Intelligence Research Center, Tokyo
[1]{simeon.schuez,sina.zarriess}@uni-bielefeld.de,
[2]ting.han@aist.go.jp

Abstract

The ability for variation in language use is necessary for speakers to achieve their conversational goals, for instance when referring to objects in visual environments. We argue that diversity should not be modelled as an independent objective in dialogue, but should rather be a result or by-product of goal-oriented language generation. Different lines of work in neural language generation investigated decoding methods for generating more diverse utterances, or increasing the informativity through pragmatic reasoning. We connect those lines of work and analyze how pragmatic reasoning during decoding affects the diversity of generated image captions. We find that boosting diversity itself does not result in more pragmatically informative captions, but pragmatic reasoning does increase lexical diversity. Finally, we discuss whether the gain in informativity is achieved in linguistically plausible ways.

1 Introduction

When speakers converse, for instance, in and about a visual environment, their utterances are remarkably diverse: Analyzing a corpus of human descriptions of MSCOCO images, Devlin et al. (2015) find that 99% of the image captions are unique. More generally, it is well known that word usage in language data follows a Zipfian distribution (Zipf, 1937). In this paper, we take a closer look at linguistic diversity in image captioning, following van Miltenburg et al. (2018)'s notion of corpus-level *global diversity* as "the ability to use (many different combinations of) many different words".

Reproducing the diversity of natural language remains a key challenge in neural generation, despite all progress in recent years. Neural generation systems in various tasks, but most notably in image captioning (Vinyals et al., 2015) and conversation modeling (Vinyals and Le, 2015) have been found to produce bland, generic and repetitive utterances (Li et al., 2016b; Dai et al., 2017; van Miltenburg et al., 2018; Ippolito et al., 2019). This lack of diversity in neural sequence-to-sequence models is often attributed to their standard training and decoding objective, i.e. likelihood, and the corresponding decoding method, i.e. beam search, which seems too biased towards highly probable and generic output (Li et al., 2016b; Vijayakumar et al., 2016; Shao et al., 2017; Kulikov et al., 2019; Holtzman et al., 2020). A commonly adopted solution is to relax the likelihood objective and sample candidate words during decoding, thereby introducing randomness into the generation process at testing time (Wen et al., 2015; Shao et al., 2017; Fan et al., 2018; Ippolito et al., 2019; Holtzman et al., 2020; Wolf et al., 2019; Panagiaris et al., 2021).

In this paper, we take a different perspective on diversity and argue that it should not result from *randomness* but from *principles* of intentional and goal-oriented language use, as formulated by e.g. Grice (1975) or Clark (1996). In particular, we hypothesize that linguistic variation in image descriptions should arise as a by-product from reasoning about different ways of referring to objects and scenes in coordination with an interlocutor. This builds upon a long tradition of linguistic research showing that speakers consider the pragmatic informativity of their lexical choices (Brown, 1958; Brennan and Clark, 1996; Grondelaers and Geeraerts, 2003; Coppock et al., 2020). For example, the more specific word "collie" might be preferred over the more common word "dog" when speakers need to unambiguously identify an entity in a context with other, similar entities (Cruse, 1977; Graf et al., 2016). Hence, in different contexts, the same types of entities could be described differently, resulting in higher diversity when considering all generated utterances.

*Work done while at Friedrich Schiller University Jena

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 411–422

With this in mind, we investigate whether linguistic diversity is triggered by simulating pragmatic objectives during the decoding of neural language models. We use recent approaches from discriminative and pragmatically informative captioning (Vedantam et al., 2017; Cohn-Gordon et al., 2018) that generate unambiguous descriptions of a target image in the context of distractor images and compare them to sampling- and search-based generation. To the best of our knowledge, no detailed comparison has yet been made between decoding strategies maximising diversity on the one and informativity on the other hand. We assess the effect of decoding along three dimensions: (i) likelihood, i.e. overlap with ground-truth captions, (ii) lexical diversity as in van Miltenburg et al. (2018) and (iii) pragmatic informativity measured in terms of the performance of a pre-trained image retrieval model (Faghri et al., 2018). We show that neither sampling methods nor beam search lead to higher pragmatic informativity compared to a greedy baseline, despite the higher diversity or likelihood to annotated ground-truth captions. Conversely, however, incorporating pragmatic objectives leads to increased diversity. Finally, we show that even simple pragmatic constraints lead to variation which is linguistically plausible.

2 Background

Criteria for high-quality and human-like descriptions of images have been discussed much in work on image captioning, pragmatics and dialogue. Besides conformity with ground truth annotations, suggestions include, for example, that descriptions should exhibit human-like diversity, sufficiently distinguish their target image from others and exhibit human-like strategies for referring (e.g. Dai and Lin, 2017; Luo et al., 2018; Liu et al., 2019; McMahan and Stone, 2020; Takmaz et al., 2020).

Diverse outputs are desirable in both open-ended dialogue and more constrained tasks like image captioning (Ippolito et al., 2019), and needed for, e.g., generating entertaining responses in chit-chat dialogues (Li et al., 2016a), responses with certain personality traits (Mairesse and Walker, 2011), or accounting for variation in referring expressions (Viethen and Dale, 2010; Castro Ferreira et al., 2016). In neural image captioning (Bernardi et al., 2016), various approaches have been presented to generate more diverse captions (e.g. Wang et al., 2016; Shetty et al., 2017; Dai et al., 2017; Wang et al., 2017; Li et al., 2018; Lindh et al., 2018; Dai et al., 2018; Chen et al., 2019; Deshpande et al., 2019; Liu et al., 2019; Wang et al., 2020). Ippolito et al. (2019) describe different decoding methods for increasing diversity in image captioning, e.g. Diverse Beam Search (Vijayakumar et al., 2016) or sampling from sets of candidate tokens. Not all methods are applicable in our setting, since the authors focus on local diversity, i.e., generating diverse sets of descriptions for individual stimuli (van Miltenburg et al., 2018). Hence, for this group of methods, we focus on the widely used sampling approaches Top-K (Fan et al., 2018) and Nucleus sampling (Holtzman et al., 2020), cf. Section 3.2.

Apart from diversity, recent work focused on generating more specific, accurate or detailed, yet (more or less) neutral descriptions (Liu et al., 2018; Dai and Lin, 2017; Luo et al., 2018; Vered et al., 2019). Other works have extended the task to pragmatically informative captioning, given a specific context (Andreas and Klein, 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018). Here, neural captioning models are trained on standard image description datasets and decoded, at testing time, to produce captions that discriminate target images from a given set of distractor images. This setting, which we adopt for our evaluation of pragmatic informativity, is very similar to the Referring Expression Generation (REG) task (Krahmer and van Deemter, 2011; Dale and Reiter, 1995; Yu et al., 2017). In our experiments we use the methods proposed by Vedantam et al. (2017) and Cohn-Gordon et al. (2018) (adapted to word level decoding), cf. Section 3.3.

To the best of our knowledge, recent work on pragmatics in neural generation has not looked explicitly at lexical diversity, although the ability to use a rich, human-like vocabulary and control lexical choice seems an important prerequisite to being able to discriminate a referent in a given context (Cruse, 1977). Inversely, most of the literature on diversity in image captioning does not explicitly analyze the underlying linguistic phenomena that cause diversity in image descriptions. However, some work discusses whether increased diversity facilitates the selection of the corresponding referent image from a large number of potential targets (Li et al., 2018; Liu et al., 2019; Chen et al., 2019). In particular, Lindh et al. (2018) bears certain similarities to our work, as the authors suggest that more specific captions lead to higher diversity. We

differ from this line of work in the following aspects: a) we focus on the decoding stage, b) our approach is linked more closely to pragmatic theory, as we generate captions that are not more specific in general, but more informative in a particular context, and c) we examine the relationship between informativity and diversity in more detail by systematically varying the contextual pressure through rationality parameters and inspecting further properties of the resulting captions.

3 Decoding Methods

A large number of decoding strategies for neural NLG has been developed recently (cf. Section 2). We focus on several representative decoding methods that target conceptually very different aspects of language use: likelihood, diversity and pragmatic informativity. These dimensions will be the basis of our analysis, as reflected in our evaluation criteria (see Section 4). Technically, the decoding methods are very generic and should be compatible with most neural NLG models.

3.1 Likelihood: Greedy and Beam Search

Greedy Search At each time step, the word with the highest probability is appended to the output sequence. Search terminates when the end token or the maximal sequence length is reached.

Beam Search keeps a fixed number of hypotheses and expands them simultaneously at each step (Graves, 2012). While this method allows for different modifications (Zarrieß and Schlangen, 2018), we use a standard approach: static beam widths, no pruning or length normalization, and terminate if the top candidate has the end token as its final segment or reaches the maximal sequence length.

3.2 Diversity: Nucleus and Top-K sampling

We take Nucleus (Holtzman et al., 2020) and Top-K sampling (Fan et al., 2018) as widely used examples of sampling-based methods aimed at increasing diversity. Both strategies are very similar in that they sample from truncated language model distributions, from which the tail of low-probability tokens have been removed that would potentially lead to flawed outputs. In each decoding step, a set of most probable next tokens is determined, from which one item is then randomly selected.

They differ, however, in how the distribution is truncated. Given a probability distribution over all candidate tokens at each time step, Top-K sampling always samples from a fixed number of k items; Nucleus sampling from the set of candidates that constitute the top-p part of the cumulative probability mass. As the probability distribution changes, the candidate pool expands or shrinks dynamically. This way, Nucleus sampling can effectively leverage the high probability mass and suppress the unreliable tail.

The initial probability distribution over candidate tokens can be shaped using a temperature parameter (Ackley et al., 1985). Subsequently, it is possible to either sample directly from this reshaped distribution or from a truncated section. Following Holtzman et al. (2020), at each time step we first shape a probability distribution with temperature t (where $t = 1.0$ results in the original distribution being unchanged), then apply Nucleus or Top-K sampling.

3.3 Pragmatics: RSA and ES Beam search

RSA Beam Search The RSA framework (Frank and Goodman, 2012) models informativity at the semantics-pragmatics interface, i.e. it provides a formalization of how pragmatically informative utterances can be derived from literal semantics using Bayesian inference. Cohn-Gordon et al. (2018) implemented RSA as a decoding strategy which integrates pragmatic factors into the iterative unrolling of recurrent generation models.

At the heart of the RSA approach, a *rational speaker* reasons about how an utterance would be understood by a listener, in order to assess whether the utterance allows the identification of the target. The speaker and listener are given a set of images W, out of which one image $w^* \in W$ is known to the speaker as the target image. This setup is illustrated in Figure 1. The rational speaker in RSA is based on a *literal speaker* who produces initial utterance candidates. In the simplest case, the literal speaker is a conditional distribution $S_0(u|w)$ which assigns equal probability to all true utterances $u \in U$ and zero probability to false utterances. The *pragmatic listener* L_0 then assesses the discriminative information of these candidates and is defined as follows:

$$L_0(w|u) \propto \frac{S_0(u|w) * P(w)}{\sum_{w' \in W} S_0(u|w') * P(w')}$$

where $P(w)$ is a prior over possible target images. The pragmatic speaker S_1 is defined in terms

Greedy	a clock tower with a clock on top of it	*Greedy*	a desk with a laptop and a desktop computer
$Nucleus_{p_{0.7}-t_{1.0}}$	a building that has a clock tower in the center	$Nucleus_{p_{0.7}-t_{1.0}}$	a desktop computer sitting on top of a desk
$ES - Beam_{\lambda 0.5}$	a tall clock tower with trees in the background	$ES - Beam_{\lambda 0.5}$	a cluttered cubicle with multiple computers and monitors
$RSA - Beam_{\alpha 1.0}$	a view of a tall clock tower with trees in the background	$RSA - Beam_{\alpha 1.0}$	an office cubicle with multiple computers and monitors

Figure 1: Example images with two distractors each. In both cases, ES and RSA captions lead to the correct identification of the target, the other captions are misleading (distractor images are selected by the retrieval model). The words "cluttered", "office" "cubicle" and "multiple" are not found in any of the *greedy* captions.

of the pragmatic listener:

$$S_1(u|w) \propto \frac{L_0(w|u)^\alpha * P(u)}{\sum_{u' \in U} L_0(w|u')^\alpha * P(u')}$$

where $P(u)$ is a uniform distribution over possible utterances U and $\alpha > 0$ is a rationality parameter determining the relative influence of the pragmatic listener in the rational speaker.

We adapted Cohn-Gordon et al. (2018)'s RSA implementation to our neural image captioning model. Importantly, we use RSA decoding with a word-level model, unlike the character-level approach in the original paper. RSA decoding can be embedded in either greedy or beam search decoding schemes. We use RSA with beam search. Crucially, in this case, beam search does not aim to maximize the literal predictions of the model (and thus the likelihood), but rather the joint speaker and listener predictions.

ES Beam Search Less grounded in pragmatic theory, the Emitter-Suppressor method (henceforth *ES*), as proposed by Vedantam et al. (2017), follows a similar idea as RSA decoding. Differences lie in a less strict distinction between speakers and listeners, and in reshaping the literal predictions of the model without Bayesian inference. In ES, a speaker (*emitter*) models a caption for a target image I_t in conjuction with a listener function (*suppressor*) that rates the discriminativeness of the utterance with regard to a distractor image. We adapted the approach of Vedantam et al. (2017) to apply ES with multiple distractor images. For this, we apply the speaker and listener functions to pairs of the target image and individual distractors, and then aggregate the resulting distributions:

$$\Delta(I_t, D) = \arg\max_s \sum_{\tau=1}^{T} \sum_{i=1}^{|D|} \log \frac{p(s_\tau|s_{1:\tau-1}, I_t)}{p(s_\tau|s_{1:\tau-1}, D_i)^{1-\lambda}}$$

where I_t is the target image and D the set of distractor images. D_i is the i-th image from this set. s is the caption for I_t in context of the distractor image D_i and T is the length of the resulting caption. λ is a trade-off parameter that determines the weight by which I_t and D_i are considered in the generation of s. For $\lambda = 1$ the model generates s with respect to I_t only, thus ignoring the context. The smaller the value of λ, the more D_i is weighted.

3.4 Differences between discriminative and sampling-based methods

In principle, both sampling-based and discriminative methods achieve their respective goals through deviation from the original predictions of the underlying captioning model. Hence, both can lead to more varied descriptions, i.e. different expressions for the same object types. In contrast, references

generated through greedy and beam search can be expected to be less variable. However, the underlying token probabilities assigned by the base model remain unchanged for Nucleus and Top-K sampling: Rather, a certain number of the highest ranked candidates is determined, from which a random draw is subsequently made. In RSA and ES, on the other hand, the literal model predictions are re-ranked deterministically through a pragmatic layer, resulting in higher ranks for tokens which are more discriminative in the respective context.

4 Experimental Set-Up

4.1 Research Hypotheses

Our hypothesis that diversity and conversational goals are connected leads us to different assumptions with regard to the evaluation results. First, it is widely described that captioning models trained with likelihood objectives struggle to generate diverse outputs. We hypothesize that discriminative decoding leads to controlled deviations from the underlying model predictions, and thus to a higher corpus-level diversity. Second, we expect the diversity induced by conversational and contextual constraints to be "meaningful" (Lindh et al., 2018): Since the linguistic variation results from contextual adjustments instead of random sampling, we suspect that diversity in ES and RSA is associated with higher informativity and thus improved retrieval results. In addition, since we consider linguistic variation through pragmatic reasoning to be linguistically plausible, we suspect parallels between the generated captions and human descriptions that aim to be informative. In particular, we expect to find evidence of linguistic strategies to increase informativity as described by Coppock et al. (2020).

4.2 Image Captioning Model

As a representative neural image captioning framework, we use Lu et al. (2017)'s adaptive attention model[1]. The model's encoder uses a pretrained CNN to represent images as feature vectors (we used ResNet152[2]). In addition to the spatial attention mechanism, the adaptive attention model includes a sentinel gate which allows it to decide whether to incorporate visual information or rely on the language model. We trained our model with a learning rate of 0.0004 for 42 epochs. The encoder CNN was fine-tuned after 20 epochs with the learning rate set to 0.0001.

4.3 Data

We performed experiments using the MSCOCO data set (Lin et al., 2014)[3]. It contains 82,783 images and 40,504 images in the training and validation sets respectively. Each image is annotated with around 5 different captions from humans. We rely on the widely used *Karpathy Split* (Karpathy and Li, 2015) for training and evaluation.

4.4 Evaluation Metrics

Likelihood We used the common COCO evaluation API[4] to calculate metrics for overlap between ground-truth and generated captions. We report BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016).

Diversity We use the metrics and implementation from van Miltenburg et al. (2018) to test the global diversity (i.e. vocabulary and word combinations with respect to the entire evaluation set) of our generated captions. We measure the type-token ratio for unigrams (TTR1) and bigrams (TTR2), the percentage of descriptions that do not appear in the training data (% novel), the number of types (Types) and the percentage of words used from the training data (% coverage). In addition, we calculate the average frequency rank of the generated types and tokens as compared to the training captions. We restrict the coverage and frequency ranks to the types accessible in the model vocabulary.

Informativity We test our captions for informativity using a pre-trained cross-modal retrieval model (Faghri et al., 2018). The model maps text and images into a common vector space; image retrieval is performed by assessing the cosine similarity between caption and image embeddings. Given a set of potential target images as well as generated captions as queries, we assess the informativity of our captions by measuring the recall R@1. Following Cohn-Gordon et al. (2018), the clusters of potential target images are compiled based on caption similarity. For each target image, we select the n images as distractors whose annotated captions have the highest Jaccard similarity with the annotated captions of the target image. We perform

[1] https://github.com/yufengm/Adaptive
[2] https://pytorch.org/docs/stable/torchvision
[3] https://cocodataset.org/
[4] https://github.com/cocodataset/cocoapi

Method	$BLEU_4$	CIDEr	SPICE	TTR1	TTR2	% nov.	Types	% cov.	avg. rank Types	avg. rank Tokens
Greedy	0.303	0.988	0.188	0.232	0.532	72.36	929	11.050	737.93	86.36
Beam	0.321	1.020	0.192	0.219	0.482	51.52	829	9.861	652.25	79.35
Top-$K_{k10-t0.7}$	0.231	0.813	0.168	0.268	0.627	87.18	1338	15.915	886.29	106.02
Top-$K_{k10-t1.0}$	0.173	0.673	0.153	0.296	0.694	94.54	1586	18.865	1022.73	126.34
Top-$K_{k25-t0.7}$	0.222	0.785	0.164	0.278	0.641	89.02	1482	17.616	971.38	113.08
Top-$K_{k25-t1.0}$	0.154	0.612	0.144	0.314	0.721	96.02	1857	22.077	1153.18	145.17
$Nucleus_{p0.7-t0.7}$	0.276	0.923	0.180	0.244	0.566	77.92	1088	12.942	792.13	92.71
$Nucleus_{p0.7-t1.0}$	0.223	0.779	0.164	0.280	0.638	87.66	1546	18.389	1023.76	117.31
$Nucleus_{p0.9-t0.7}$	0.250	0.855	0.174	0.261	0.601	84.24	1319	15.677	904.59	101.89
$Nucleus_{p0.9-t1.0}$	0.165	0.623	0.144	0.325	0.723	93.96	2133	25.324	1362.44	168.11
ES-$Beam_{\lambda 0.7}$	0.290	0.919	0.179	0.257	0.569	67.40	1201	14.286	918.30	111.97
ES-$Beam_{\lambda 0.5}$	0.225	0.727	0.154	0.303	0.670	83.22	1619	19.258	1171.08	177.90
ES-$Beam_{\lambda 0.3}$	0.088	0.371	0.104	0.360	0.757	96.90	2225	26.454	1452.41	404.15
RSA-$Beam_{\alpha 0.5}$	0.291	0.951	0.183	0.234	0.521	62.86	966	11.490	753.70	88.52
RSA-$Beam_{\alpha 1.0}$	0.282	0.928	0.180	0.245	0.547	66.24	1033	12.287	767.66	92.83
RSA-$Beam_{\alpha 5.0}$	0.235	0.797	0.165	0.285	0.651	83.20	1356	16.118	950.74	123.10
Human	-	-	-	0.391	0.803	95.94	3704	43.642	2288.41	302.58

Table 1: Likelihood (BLEU, CIDEr, SPICE) and diversity metrics (type-token ratio, % novel captions, number of distinct types, % coverage of the training vocabular, average frequency rank for types and tokens with respect to the training captions) for decoding strategies

the evaluation with three setups ($n \in \{2, 4, 9\}$, see Figure 1 for an example with two distractors).

4.5 Decoding Parameters

For all decoding strategies, maximum length is set to 20 words per caption, excluding the $\langle start \rangle$ token. After decoding, the generated captions were cleaned of leftover $\langle end \rangle$ and $\langle unk \rangle$ tokens using regular expressions.

We use a static beam width of 5. For sampling-based decoding, we report results for different settings regarding the p and k thresholds as well as temperature t. In RSA and ES decoding, the rationality parameters α and λ determine the degree of pragmatic reasoning (cf. Section 3.3). We report results for different levels of rationality.

We generate the captions using the same clusters of target and distractor images that are used for listener evaluation (cf. Section 4.4). Since RSA and ES captions are generated given both target and distractor images, the number of distractors has a considerable influence. For better clarity, we only report results for settings with two distractors per target image when discussing quality and diversity.

5 Results

5.1 Likelihood and Diversity

In the following, we test our hypothesis that ES and RSA lead to more diverse captions. We further compare how discriminative and sampling-based decoding affects likelihood and diversity scores.

The results in Table 1 show that pragmatic reasoning does increase the diversity of generated captions as compared to a greedy baseline. Importantly, this is related to the degree of pragmatic influence: Higher rationality values systematically increase TTR, number of word types, coverage and the rate of novel captions, as well as the average frequency of types and tokens with respect to the training captions. Therefore, for higher α values (RSA) or lower λ (ES) the size of the used vocabulary increases, including a higher proportion of lower frequency words. This strengthens the hypothesis that pragmatic constraints are indeed amplifying the diversity of linguistic utterances. At the same time, ES and RSA substantially decrease BLEU, CIDEr and SPICE as compared to greedy and beam search.

Nucleus and Top-K sampling exhibit similar patterns in terms of likelihood and diversity. Higher values for p, k and t systematically lead to increased diversity scores across metrics, accompanied by lower likelihood scores. In contrast to the methods described above, beam search leads to increases in likelihood but generally lower diversity values. Rather unsurprisingly, the human baseline outperforms all methods and parameter settings in most diversity metrics. The only exception is ES ($\lambda = 0.3$) with higher average token ranks and more novel captions, but also the lowest overall likelihood scores.

Method	Recall		
	2 Dist.	4 Dist.	9 Dist.
Greedy	68.42	56.98	44.34
Beam	66.98	55.22	42.56
Top-K$_{k10-t0.7}$	67.92	56.30	44.00
Top-K$_{k10-t1.0}$	66.66	54.90	42.78
Top-K$_{k25-t0.7}$	66.14	55.48	43.50
Top-K$_{k25-t1.0}$	67.00	55.50	42.62
Nucleus$_{p0.7-t0.7}$	67.38	55.76	43.88
Nucleus$_{p0.7-t1.0}$	66.58	55.64	43.14
Nucleus$_{p0.9-t0.7}$	67.32	56.00	43.62
Nucleus$_{p0.9-t1.0}$	66.46	55.02	43.00
ES-Beam$_{\lambda 0.7}$	78.00	66.58	54.02
ES-Beam$_{\lambda 0.5}$	85.66	74.98	61.86
ES-Beam$_{\lambda 0.3}$	89.94	80.46	68.02
RSA-Beam$_{\alpha 0.5}$	70.84	59.24	46.56
RSA-Beam$_{\alpha 1.0}$	74.18	63.32	50.16
RSA-Beam$_{\alpha 5.0}$	82.02	71.74	58.16
Human	67.00	56.96	46.58

Table 2: R@1 retrieval scores, using generated captions as queries. ES and RSA show the best results, further improving with higher rationalities.

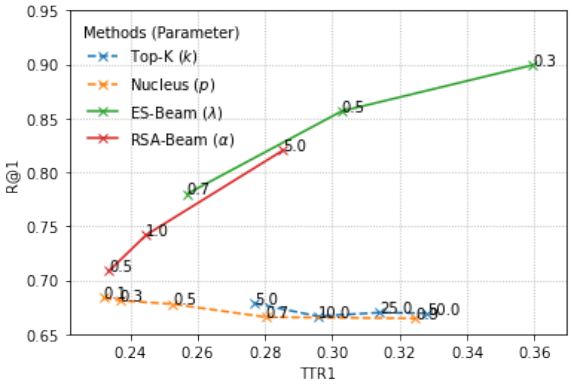

Figure 2: R@1 (2 distractors) and TTR1 scores for Top-K sampling, Nucleus sampling, ES and RSA, with different settings for k, p, λ and α. For ES and RSA, increases in TTR1 are accompanied by higher R@1. For sampling-based methods, R@1 is largely unaffected.

Generally, we observe that increase in diversity goes along with lower likelihood results and vice versa. This resembles the quality-diversity trade-off as described e.g. by Ippolito et al. (2019); Wang and Chan (2019).

5.2 Informativity

In the following, we replicate the results of Vedantam et al. (2017); Cohn-Gordon et al. (2018) using the state-of-the-art retrieval model from Faghri et al. (2018) and investigate whether variation through pragmatic reasoning or sampling leads to more informative captions.

Here, RSA and ES have a clear advantage as they are conditioned on the target and distractor images whereas the other strategies decode the caption by looking only at the target image (see Section 3). Thus, unsurprisingly, we find that these strategies clearly outperform all other decoding methods in terms of R@1 scores. This holds for all parameters and distractor settings. Remarkably, both ES and RSA surpass the human baseline in this regard. The results in Table 2 thus replicate the results from Vedantam et al. (2017); Cohn-Gordon et al. (2018). It is noteworthy that even low rationality levels ($\alpha = 0.5$ or $\lambda = 0.7$) improve the recall[5].

For Nucleus and Top-K sampling, none of the configurations lead to improved pragmatic informativity over the greedy baseline, even though they clearly improve diversity (cf. Table 1, as discussed above). Beam search also decreases informativity as compared to greedy search. Perhaps unsurprisingly, the higher the number of distractors, the lower are the scores for all decoding strategies. Still, the recall is well above the random level in all cases, which demonstrates the general capability of our used captioning and retrieval models.

In summary, this shows substantial differences between the kind of linguistic variation caused by sampling-based and discriminative decoding methods: Whereas both types of methods result in higher lexical diversity and lower overlap to human annotations, sampling-based diversity does not seem to naturally lead to higher pragmatic informativity (illustrated in Figure 2).

6 Linguistic Strategies in Pragmatic Decoding

The results discussed above show that pragmatic reasoning during decoding results in both increased diversity and informativity of captions. This suggests that the phenomenon of linguistic diversity can be integrated, at least to some extent, into well-established theories of intentional and goal-oriented language use (Grice, 1975; Clark, 1996).

Figure 1 shows two different ways, in which variation of literal image descriptions leads to higher informativity: Re-conceptualizing and re-describing entities mentioned in the literal caption in a way that distinguishes them from similar entities in distractor images, or describing further objects and elements, which are present in the target image but not in the distractor images. Changing "clock

[5] Cohn-Gordon et al. (2018) used $\alpha = 5.0$

Method	% ADJ	% N	% V	WN dist.
Greedy	3.90	35.65	8.09	8.096
Beam	4.75	36.40	9.19	7.886
Top-K$_{k10-t0.7}$	5.18	35.19	7.89	8.159
Top-K$_{k10-t1.0}$	6.30	34.28	7.93	8.147
Top-K$_{k25-t0.7}$	5.43	34.83	8.02	8.165
Top-K$_{k25-t1.0}$	6.57	34.16	8.30	8.177
Nucleus$_{p0.7-t0.7}$	4.52	35.49	7.97	8.143
Nucleus$_{p0.7-t1.0}$	5.50	35.06	7.93	8.153
Nucleus$_{p0.9-t0.7}$	4.76	35.34	8.08	8.143
Nucleus$_{p0.9-t1.0}$	6.30	34.49	8.62	8.147
ES-Beam$_{\lambda 0.7}$	5.93	36.58	9.12	8.048
ES-Beam$_{\lambda 0.5}$	7.97	37.17	8.96	8.258
ES-Beam$_{\lambda 0.3}$	14.14	39.79	9.85	8.478
RSA-Beam$_{\alpha 0.5}$	5.26	34.98	8.32	7.889
RSA-Beam$_{\alpha 1.0}$	5.74	34.93	8.48	7.937
RSA-Beam$_{\alpha 5.0}$	7.93	35.01	8.61	8.141
Human	7.32	34.82	9.16	8.227

Table 3: Distribution of POS tags in the generated captions and mean distance for generated nouns from WordNet root (2 distractors for ES and RSA)

tower" to "tall clock tower" can be seen as refining the description; switching "desk" to "office cubicle" as re-conceptualizing parts of the scene in favour of more informative categories. The inclusion of "trees in the background" states an example of additional distinctive elements.

In human annotations, the informativity of unambiguous referring expression is achieved e.g. by increasing lexical specificity or adding descriptive modifiers (Coppock et al., 2020). To explore those strategies in our captions, we measure the average distance of generated nouns from the WordNet root, as a rough approximation of specificity, and accumulate the POS tags for the generated captions, both using off-the-shelf models from the SpaCy library. The results are shown in Table 3.

Regarding lexical specificity, beam search appears to generate more general nouns in comparison to the greedy baseline. In contrast, sampling-based methods lead to a more specific vocabulary. However, neither does this specificity translate to improved retrieval results (cf. Section 5.2), nor does changing the parameters seem to have much impact. For ES and RSA, higher α or lower λ settings systematically lead to a higher specificity for nouns, as well as improved retrieval results. The average specificity for RSA with low rationality is surprisingly low, which could be due to the beam search scheme in which reasoning is integrated. Whereas there doesn't seem to be a systematic relation between rationality and the ratio of nouns and verbs, we observe a higher ratio for adjectives if rationality is increased. However, we should note that e.g. ES ($\lambda = 0.3$) generates more ungrammatical sentences, which may affect the POS tagger. Also, this extends to sampling-based methods, where more adjectives are produced if the parameters are tuned towards higher diversity.

Taken together, the higher average specificity of nouns and greater proportion of adjectives are consistent with the linguistic devices described by Coppock et al. (2020). Although future work should explore this in more detail, this suggests that linguistic variation in ES and RSA corresponds, at least to some degree, to plausible strategies for achieving communicative goals.

7 Discussion and Conclusion

Our findings show that pragmatic reasoning in neural generation adds an interesting dimension to the analysis and modeling of lexical diversity in neural image captioning. Although not aiming at diversity itself, ES and RSA lead to linguistic variation through simulated coordination with interlocutors, which in turn leads to increased lexical diversity (Section 5.1). Whereas this variation translates to improved informativity, this is not the case for sampling-based methods like Nucleus and Top-K sampling (Section 5.2). Further exploration revealed that discriminative decoding results in a higher rate of generated adjectives and a higher average specifity for nouns (Section 6), resembling linguistic strategies found in human annotations (Coppock et al., 2020). Therefore, pragmatic reasoning leads to linguistically meaningful variation, resulting in higher informativity due to linguistically plausible devices, and, from a global perspective, increased diversity. In this regard, linguistic diversity arises naturally from conversational goals and adaptations to contextual constraints.

We see great potential for future work in exploring linguistic variation in tasks related to and going beyond image captioning. First, the human annotations used here were produced in a relatively neutral communicative context. Hence, they differ from generated captions in terms of their communicative purpose and possibly do not reflect the full range of variation that speakers might use in more challenging tasks. Thus, similar studies could be made on e.g. referring expressions (Yu et al., 2017) or other datasets that record longer interactions centered on images (Takmaz et al., 2020). Second,

as discriminative image captioning captures only partial aspects of natural conversation, it could be investigated whether our findings apply to other dialogue tasks. Finally, other sources of variation should be considered, e.g. formality or individual characteristics of speakers (Geeraerts, 1994).

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – INST 275/363-1 FUGG.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, 22 6:1482–93.

Roger Brown. 1958. How shall a thing be called? *Psychological Review*, 65(1):14–21.

Thiago Castro Ferreira, Sander Wubben, and Emiel Krahmer. 2016. Towards proper name generation: a corpus analysis. In *Proceedings of the 9th International Natural Language Generation conference*, pages 222–226, Edinburgh, UK. Association for Computational Linguistics.

Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. 2019. Variational structured semantic inference for diverse image captioning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. 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 439–443.

Elizabeth Coppock, Danielle Dionne, Nathanial Graham, Elias Ganem, Shijie Zhao, Shawn Lin, Wenxing Liu, and Derry Wijaya. 2020. Informativity in image captions vs. referring expressions. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 104–108, Gothenburg. Association for Computational Linguistics.

D. A. Cruse. 1977. The pragmatics of lexical specificity. *Journal of Linguistics*, 13(2):153–164.

Bo Dai, Sanja Fidler, and Dahua Lin. 2018. A neural compositional paradigm for image captioning. In *Advances in Neural Information Processing Systems*, pages 658–668.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10695–10704.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press.

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.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Dirk Geeraerts. 1994. Varieties of lexical variation. In *Proceedings of the 6th EURALEX International Congress*, pages 78–83, Amsterdam, the Netherlands. Euralex.

Caroline Graf, Judith Degen, Robert XD Hawkins, and Noah D Goodman. 2016. Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions. In *Proceedings of the 38th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *the International Conference of Machine Learning (ICML) 2012 Workshop on Representation Learning*.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Stefan Grondelaers and Dirk Geeraerts. 2003. Towards a pragmatic model of cognitive onomasiology. In *Cognitive Approaches to Lexical Semantics*, pages 67–92. Mouton De Gruyter.

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*.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38.

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

Dianqi Li, Qiuyuan Huang, Xiaodong He, Lei Zhang, and Ming-Ting Sun. 2018. Generating diverse and accurate visual captions by comparative adversarial learning. In *Visually Grounded Interaction and Language (ViGIL) at NeurIPS 2018*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Annika Lindh, Robert J. Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D. Kelleher. 2018. Generating diverse and meaningful captions. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 176–187, Cham. Springer International Publishing.

Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. 2019. Generating diverse and descriptive image captions using visual paraphrases. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4239–4248.

Nelson F. Liu, Omer Levy, Roy Schwartz, Chenhao Tan, and Noah A. Smith. 2018. LSTMs exploit linguistic attributes of data. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 180–186, Melbourne, Australia. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

R. Luo, Brian L. Price, S. Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

François Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

Brian McMahan and Matthew Stone. 2020. Analyzing speaker strategy in referential communication. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 175–185, 1st virtual meeting. Association for Computational Linguistics.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. Generating unambiguous and diverse referring expressions. *Computer Speech & Language*, 68:101184.

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.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219, Copenhagen, Denmark. Association for Computational Linguistics.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *The IEEE International Conference on Computer Vision (ICCV)*.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. 2019. Joint optimization for cooperative image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jette Viethen and Robert Dale. 2010. Speaker-dependent variation in content selection for referring expression generation. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 81–89.

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

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5756–5766.

Qingzhong Wang and Antoni B Chan. 2019. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4195–4203.

Qingzhong Wang, Jia Wan, and Antoni B Chan. 2020. On diversity in image captioning: Metrics and methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. 2016. Diverse image captioning via grouptalk. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2957–2964. IJCAI/AAAI Press.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.

George Kingsley Zipf. 1937. Observations of the possible effect of mental age upon the frequency-distribution of words, from the viewpoint of dynamic philology. *The Journal of Psychology*, 4(1):239–244.