Decoding, Fast and Slow: A Case Study on Balancing Trade-Offs in Incremental, Character-level Pragmatic Reasoning

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

Abstract

Recent work has adopted models of pragmatic reasoning for the generation of informative language in, e.g., image captioning. We propose a simple but highly effective relaxation of fully rational decoding, based on an existing incremental and character-level approach to pragmatically informative neural image captioning. We implement a mixed, 'fast' and 'slow', speaker that applies pragmatic reasoning occasionally (only word-initially), while unrolling the language model. In our evaluation, we find that increased informativeness through pragmatic decoding generally lowers quality and, somewhat counter-intuitively, increases repetitiveness in captions. Our mixed speaker, however, achieves a good balance between quality and informativeness.

1 Introduction

Kahneman (2011) famously said that humans have two ways of thinking (along with others theories on human information processing, e.g., Schneider and Shiffrin, 1977): one way is *fast*, automatic and intuitive, the other is a *slow*, controlled, and explicit way of reasoning. This distinction also arises in research on human language processing: slow processes of reasoning that allow speakers to adapt their utterances very flexibly and strategically to a given context are central to theories of pragmatics (Searle, 1969; Grice, 1975; Clark, 1996). Yet, speakers are known to produce utterances in context quickly and easily, which has been a central concern in, e.g., psycholinguistics and experimental pragmatics (Keysar et al., 1998; Galati and Brennan, 2010; Degen and Tanenhaus, 2016, 2019). Similarly, models of pragmatic reasoning and their applications in NLP face the challenge that fully rational language generation is computationally costly or even intractable (Reiter, 1991; White et al., 2020).

Recent work on pragmatics in NLP has taken interest in the Rational Speech Acts (RSA) model (Frank and Goodman, 2012) which resulted in implementations of frameworks that model the generation of informative language with so-called *rational speakers* (Andreas and Klein, 2016; Fried et al., 2018; Shen et al., 2019). Cohn-Gordon et al. (2018) use an image captioning set-up inspired by classical reference games (see Figure 1) to show that a 'slow' rational speaker which reasons internally about the informativeness of utterances generated by a plain neural language model is communicatively more effective than a 'fast' literal speaker that produces the most likely utterance for the target as predicted by the language model. More generally, recent work in NLG has shown a lot of interest in reasoning or decoding methods that extend neural generation models with additional objectives that cannot be easily achieved by decoding the underlying neural language model with greedy or beam search (e.g., Li et al., 2016; Vedantam et al., 2017; Vijayakumar et al., 2018; Ippolito et al., 2019; Holtzman et al., 2020; Tam, 2020).

Reasoning schemes like RSA provide an attractive, since explicit and theoretically motivated, way of incorporating linguistically plausible, communicative strategies into a neural generation framework. At the same time, however, RSA and various related decoding methods have been found to not achieve a good balance between different dimensions of output quality. For instance, Ippolito et al. (2019) investigates a range of decoding methods that aim at increasing the lexical diversity of image captions or responses in dialogue and report on a very clear quality-diversity trade-off: the more the decoding procedure (e.g., sampling) increases diversity and deviates from the predictions of the underlying language model, the more the generated expressions decrease in quality. Recently, Schüz et al. (2021) found similar trade-offs for decod-

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 371–376,
Aberdeen, Scotland, UK, 20-24 September 2021.

S_0	a group of people riding on the backs of horses
S_1	two brown hornes grazing in a fenced grassy field
S_x	two horses in a field in front of a field

Figure 1: Captions for the target image (large), generated by a literal (S_0), rational (S_1) and mixed speaker (S_x), with rationality parameter $\alpha = 5$ and beam search. The captions by S_1 and S_x are more discriminative ("field"), but contain repetitions and out-of-vocabulary words ("hornes").

ing word-level image captioning models with RSA and Vedantam et al. (2017)'s discriminative beam search.

Next to these trade-offs, rational speakers in RSA, which apply complex recursive reasoning using an internal listener and speaker, incur a high computational cost, particularly in generation setups with large candidate spaces where exhaustive search is not tractable. Therefore, recent works have implemented incremental decoding schemes that reason about discriminativeness at *every* time-step, during unrolling the language model (Vedantam et al., 2017; Cohn-Gordon et al., 2019). Cohn-Gordon et al. (2018)'s character-level approach fully confronts pragmatic reasoning: the neural language model captions images in a character-by-character fashion such that each character can be internally scored for its informativeness by the rational speaker. While this incremental generation and reasoning scheme makes it possible to search a large space of potential utterances, it is still extremely costly as the internal, recursive reasoning of the speaker is applied at every character.

In this paper, we propose a very simple but highly efficient relaxation of fully rational and incremental decoding with RSA: we propose a *mixed speaker* that switches between literal and rational inference, that is, between 'fast' and 'slow' decoding, while unrolling the language model. This speaker applies pragmatic reasoning at *particular* time steps during decoding rather than at every time-step. Extending Cohn-Gordon et al. (2018)'s character-level RSA, our mixed speaker is rational only when generating the first character of a new word and uses fast literal decoding for the remaining steps in the sequence. Adopting Schüz et al. (2021)'s evaluation setting that combines quality, informativeness and diversity, we find that Cohn-Gordon et al. (2018)'s original, fully incremental character-level generation approach produces captions that are not only more informative and globally more diverse, but also have lower quality, contain more repeated words as well as more out-of-vocabulary words. Generally, the mixed speaker that switches between fast and slow decoding is computationally more efficient and achieves a good balance in these various trade-offs maintaining, most notably, a better balance between quality and informativeness than a fully rational speaker.

2 Models

2.1 Image Captioning Model

We use Lu et al. (2017)'s adaptive attention model. The model's encoder uses a pre-trained CNN to represent images as feature vectors (we used ResNet152). A single LSTM layer with rectifier activation transforms the feature vectors into new vectors v^g. We concatenate the vector v^g with the word embedding vector w_t as the input for the decoder. Conditioned on image feature vector v_g and the hidden state vector h_{t-1} of the encoder from the previous time step, the attention module generates an attention map vector c_t. A single layer neural network transforms c_t and current hidden state vector h_t into a new vector, the final layer is a softmax over the vocabulary. While Lu et al. (2017) trained word level image captioning models, we trained a character-level model with the same architecture.

2.2 Pragmatic Reasoning

In the RSA model, a so-called rational speaker reasons about how an utterance would be understood by a listener, i.e., whether it allows the identification of the target. The speaker and listener are given a set of images W and one image $w^* \in W$ is known to the speaker as the target (see Figure 1). The rational speaker in RSA has an internal *literal speaker* who produces utterance candidates, i.e., a conditional distribution $S_0(u|w)$ which, in the simplest case, assigns equal probability to all true utterances $u \in U$ and zero probability to false utterances. A *pragmatic listener* L_0 then discriminates between

images given the utterance, 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 as:

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

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, see Cohn-Gordon et al. (2018) for further details. Essentially, the literal speaker S_0 corresponds to the standard formulation of the image captioning task, i.e., it generates descriptions for single target images. In contrast to this, the pragmatic speaker S_1 considers the respective context, i.e., the distractor images, during decoding.

We use Cohn-Gordon et al. (2018)'s character-incremental implementation of RSA that uses a character-level captioning model as the literal speaker S_0 and applies recursive pragmatic reasoning at each time-step, during unrolling a character-level neural speaker. While Cohn-Gordon et al. (2018) only reported results on decoding RSA with beam search, we compare greedy and beam search.

A Mixed Speaker While previous studies on RSA for neural generation have implemented fully rational speakers that apply pragmatic reasoning over the entire utterance or incrementally at each time-step of the decoding process, we propose a new scheme for decoding with RSA which we call a *mixed speaker*. This speaker is both literal and rational (or 'fast' and 'slow'), using different levels of reasoning during incremental decoding. We define our mixed speaker (S_x) to be: (i) rational (S_1) when generating the first character of a new word, i.e., at the beginning of the sequence or after generating a whitespace and (ii) literal (S_0) when generating other characters, i.e., not at the beginning of a word. This captures the intuition that the speaker can (and should) in many cases rely on its language model, e.g., when continuing words in a morphologically well-formed way. We test whether pragmatic reasoning at the beginning of words is enough to push the model towards being more informative, by giving more probability to initial letters of discriminative words. For instance, when the speaker describes a *big* and *yellow* object, pragmatic reasoning will be needed only at the beginning of the word, discriminating between *b* and *y*, depending on properties of the distractor objects.

3 Character-level Experiment

Our experiments compare three different speakers: the literal (or 'fast') speaker S_0, which simply decodes the language model, the rational (or 'slow') speaker S_1, which reasons at every time step, and the mixed (or 'fast and slow') speaker S_x.

3.1 Evaluation

Our evaluation setting is similar to Schüz et al. (2021), who investigated global diversity in pragmatic reasoning with word-level captioning models. Here, in addition, we analyze the repetitiveness of generated captions and evaluate informativeness in similar ways as in Cohn-Gordon et al. (2018), instead of using an external cross-modal retrieval model as in Schüz et al. (2021).

Data We performed experiments using the MSCOCO data set (Lin et al., 2014), with 82,783 and 40,504 images in the training and validation sets, respectively. Each image is annotated with around five captions. Following Cohn-Gordon et al. (2018), we train our speaker and listener models on distinct training splits. Because of this, we randomly split the training set into halves for model training. For evaluation, we randomly selected 1,000 images from the MSCOCO validation set.

Informativeness Following Cohn-Gordon et al. (2018), we train a listener model to predict target images for captions produced by a speaker. Given a set of potential target images and a generated caption, the listener ranks the images in terms of their likelihood. If the target image (i.e., the input of the speaker) is on top, the caption is accurate (reported as listener accuracy, L_{acc}, in Table 1). The clusters of potential target images were compiled based on caption similarity: For each target image, we select the two images as distractors whose annotated captions have the highest Jaccard similarity with the annotated captions of the target image.

Quality Evaluation We assess the quality of generated captions in terms of CIDEr scores (Vedantam et al., 2015), measuring the overlap with human captions. Since our model generates captions character by character, we report the absolute number of out-of-vocabulary types and tokens (OOV types and tokens) where we treat every

		α	Inform. L_{acc}	Quality CIDEr	Type Vocab IV	Type Vocab OOV	Token Vocab IV	Token Vocab OOV	Diversity TTR_{cap}	Diversity TTR_{cap_c}	Diversity TTR_{cap_2}	Diversity TTR_g
greedy	S_0	-	54.8	0.668	376	4	11273	5	0.760	0.870	0.916	0.166
	S_1	1	63.0	0.559	444	7	13136	7	0.714	0.822	0.884	0.177
	S_1	3	66.9	0.506	549	8	12990	8	0.720	0.819	0.881	0.201
	S_1	5	68.9	0.462	620	20	12846	25	0.723	0.817	0.884	0.212
	S_x	1	61.5	0.575	443	6	13127	8	0.718	0.825	0.888	0.175
	S_x	3	65.1	0.524	491	6	13014	6	0.723	0.820	0.888	0.189
	S_x	5	65.5	0.493	529	9	12998	10	0.728	0.824	0.890	0.198
beam	S_0	-	54.1	0.778	303	0	10369	0	0.841	0.930	0.964	0.160
	S_1	1	63.1	0.704	348	3	10359	3	0.826	0.915	0.952	0.171
	S_1	3	68.5	0.589	428	17	10360	32	0.796	0.872	0.927	0.193
	S_1	5	70.4	0.481	486	72	10730	199	0.769	0.839	0.902	0.209
	S_x	1	61.8	0.718	341	2	10497	2	0.828	0.918	0.952	0.170
	S_x	3	64.8	0.652	373	3	10580	3	0.812	0.899	0.939	0.180
	S_x	5	66.9	0.606	413	8	10694	8	0.797	0.884	0.927	0.190

Table 1: Evaluation of informativeness (listener accuracy), quality (CIDEr and OOV types and tokens), local diversity (TTR_{cap}, TTR_{cap_c}, TTR_{cap_2}) and global diversity (TTR_g) for literal (S_0), rational (S_1), mixed (S_x) speakers.

word token (occurring between whitespaces) that did not occur in the training set as an OOV token. In-vocabulary (IV) token and type counts are provided for comparison.

Diversity and Repetitiveness We measure diversity using different type-token ratios (TTR) (Youmans, 1990; van Miltenburg et al., 2018). TTR_g is calculated *globally* as the total number of types divided by the total number of tokens as in van Miltenburg et al. (2018) and Schüz et al. (2021). In contrast to this, TTR_{cap} is computed *locally* as the arithmetic mean of the TTR values for individual captions. While TTR_g reflects the general lexical diversity, TTR_{cap} is an indicator for word repetitions in captions. We supplement this with TTR_{cap_c} which is analog to TTR_{cap} but on captions filtered for stop words, i.e., indicating repetitions of content words. Finally, TTR_{cap_2} is based on bigrams and thus indicates the repetition of word combinations or phrases.

3.2 Informativeness–Quality Trade-Off

The results in Table 1 show that there is a systematic trade-off between informativeness and quality in character-level image captioning. All speakers that use some level of reasoning, S_1 or S_x with different α-values, achieve a higher listener accuracy but lower quality in terms of CIDEr than S_0. Beam search is generally beneficial for caption quality, according to the CIDEr scores shown in Table 1. However, it seems to interact with highly rational S_1 in unfortunate ways and leads to a drastic increase of the number of OOVs (see Figure 1 for an example). Here, RSA fails to achieve a good balance with its underlying language model. Our mixed speaker achieves a much better trade-off between quality and informativeness, especially in combination with beam search: $S_{x,\alpha=5}$ clearly outperforms $S_{1,\alpha=5}$ in CIDEr (more than 12 points improvement) and number of OOVs (only 8 types as compared to 72 for S_1), while its listener accuracy is only 3 points lower. From this, we conclude that occasional, word-initial pragmatic reasoning is highly effective and offers a decent balance between informativeness and quality.

Example 1 illustrates further, more general aspects of an informativeness–quality trade-off. S_0 and S_x generate more informative captions by integrating the background of the image as a discriminative feature (*"field"*). However, other features are also added, some of them inaccurate (e.g., *"grazing"*, *"fenced"*). This shows general limitations of this approach: Since discriminativity is the primary concern in RSA, other problems can arise, such as semantic inadequacies.

3.3 Local–Global Diversity Trade-Off

Results in Table 1 point to further trade-offs of pragmatic decoding, which generally lead to more repetitive captions being generated. In comparison to the literal speakers S_0, S_1 and S_x have lower TTR_{cap}, TTT_{cap_c} and TTR_{cap_2} scores. The mixed speaker attenuates this effect, but has still lower local TTR as the fully literal speaker. This should not happen in theory, since it is questionable whether repeating is a useful strategy for making utterances more discriminative. It could even be seen as a strategy that violates the Gricean maxim of relevance (Grice,

1975) – see the caption of S_x in Figure 1 for a representative example. Interestingly, however, increases in local repetitiveness are combined with increases in global diversity. Thus, speakers which are more repetitive also use a larger vocabulary (S_1 and S_x use more absolute types and have higher TTR_g). RSA counteracts the tendency of beam search to globally reduce the vocabulary, but, here, greedily decoded speakers achieve higher numbers of types and TTR_g. This indicates that RSA might be a useful approach to generating, e.g., less frequent words when the communicative context makes this necessary, as previously suggested in Schüz et al. (2021).

4 Conclusion

In this paper we have replicated Cohn-Gordon et al. (2018)'s approach to character-level pragmatic image captioning and have evaluated it not only with respect to informativeness, but also quality, repetitiveness, and global diversity of the model outputs. Our results show various trade-offs in character-level captioning: models that are more informative and rational produce captions that are of lower quality and contain more out-of-vocabulary words. In terms of diversity, we find that character-level RSA *increases* the amount of word repetitions within captions. Interestingly, at the same time, it also increases global diversity and leads to a larger vocabulary being exploited by the underlying language model. This analysis fully confirms and extends findings on word-level decoding methods facing different types of trade-offs as a result of additional objectives introduced into the generation process at the decoding stage (e.g., Li et al., 2016; Vijayakumar et al., 2018; Ippolito et al., 2019; Schüz et al., 2021).

Our analysis also shows that these trade-offs can be countered by our mixed, 'fast' and 'slow', speaker that applies reasoning in a simple, word-initial fashion. Future work could explore further ways of controlling *when* reasoning is needed during decoding and generalize this to word-level decoding. As character-level image captioning is arguably not state-of-the-art, some of the effects that we report in this case study might not generalize to more powerful – especially word-level – models. Nevertheless, we believe that the trade-offs observed in this pilot study could be explored in other task that require the generation of long sequences (e.g., image paragraphs, longer responses in an interaction) and that the effectiveness of mixed pragmatic decoding might be an interesting avenue for such tasks.

References

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.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

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.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2019. An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 81–90.

Judith Degen and Michael K Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive science*, 40(1):172–201.

Judith Degen and Michael K Tanenhaus. 2019. Constraint-based pragmatic processing. *Handbook of Experimental Semantics and Pragmatics*.

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

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. 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 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Alexia Galati and Susan E. Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62:35–51.

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.

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.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Macmillan.

Boaz Keysar, Dale J. Barr, and William S. Horton. 1998. The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, 7:46–49.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv: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.

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.

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.

Ehud Reiter. 1991. A new model of lexical choice for nouns. *Computational Intelligence*, 7(4):240–251.

Walter Schneider and Richard M Shiffrin. 1977. Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, 84(1):1.

Simeon Schüz, Ting Han, and Sina Zarrieß. 2021. Diversity as a by-product: Goal-oriented language generation leads to linguistic variation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422, Singapore and Online. Association for Computational Linguistics.

John Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text 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 4060–4067, Minneapolis, Minnesota. Association for Computational Linguistics.

Yik-Cheung Tam. 2020. Cluster-based beam search for pointer-generator chatbot grounded by knowledge. *Computer Speech & Language*, 64:101094.

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.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Julia White, Jesse Mu, and Noah Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

Gilbert Youmans. 1990. Measuring lexical style and competence: The type-token vocabulary curve. *Style*, 24:584–599.