The Effect of Generalisation on the Inadequacy of the Mode

Bryan Eikema
University of Amsterdam
b.eikema@uva.nl

Abstract

The highest probability sequences of most neural language generation models tend to be degenerate in some way, a problem known as the inadequacy of the mode. While many approaches to tackling particular aspects of the problem exist, such as dealing with too short sequences or excessive repetitions, explanations of why it occurs in the first place are rarer and do not agree with each other. We believe none of the existing explanations paint a complete picture. In this position paper, we want to bring light to the incredible complexity of the modelling task and the problems that generalising to previously unseen contexts bring. We argue that our desire for models to generalise to contexts it has never observed before is exactly what leads to spread of probability mass and inadequate modes. While we do not claim that adequate modes are impossible, we argue that they are not to be expected either.

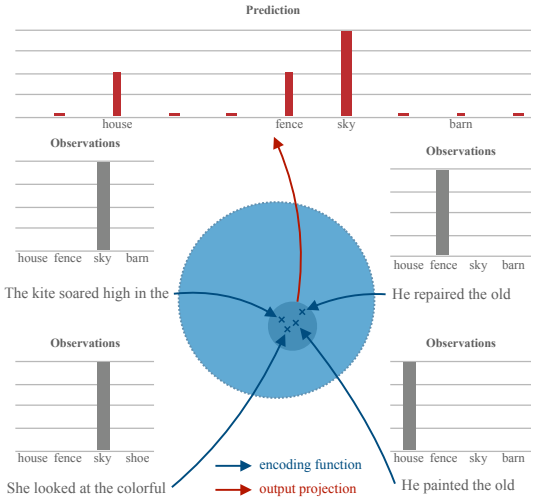

Figure 1: To generalise, a neural network needs to encode different contexts similarly to reduce data sparsity. While this could be because of actual linguistic similarity, it could also be the result of learned independence assumptions or other pragmatic reasons to bundle contexts, such that their observations contribute to the same next-word distribution. This can introduce spread, as all observations need to be well represented within the distributions, and as a result may lead to inadequate modes for some contexts.

1 Introduction

Neural language generators have made tremendous advances in recent years and are commonplace across natural language processing (NLP) tasks. An observation that has been made across use cases of such models, however, is that the highest probability sequences tend to be of low quality, such as being too short, containing excessive repetition or copying the input (Ott et al., 2018; Stahlberg and Byrne, 2019; Holtzman et al., 2020). This observation is known under several names, such as the inadequacy of the mode, the probability-quality paradox and the bad mode problem.

Many approaches for tackling particular aspects of the problem exist: altering beam search to do intentional sub-optimal search and/or adding length penalties (Wu et al., 2016; Koehn and Knowles, 2017), using different decoding algorithms more robust to inadequate modes such as (truncated) sampling (Holtzman et al., 2020) or minimum Bayes risk (Eikema and Aziz, 2022), or training and fine-tuning the model to discourage certain degenerate behaviours (Shi et al., 2020; Zhao et al., 2023). Explanations of why it occurs, however, are much rarer. Existing explanations are few, most only attempt to explain specific degeneracies, and more general explanations do not agree with each other on what the cause of the problem is. Meister et al. (2022) hypothesise that the inadequacy of the mode is an inherent property of natural language and human communication, while Yoshida et al. (2023) instead attribute it to a fundamental shortcoming of the most commonly employed training algorithm in the presence of noisy training data.

Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024), pages 87–92

We believe none of these explanations paint a complete picture of the causes behind the inadequacy of the mode. In this position paper, we want to bring light to the role of generalisation. We hypothesise that while our models in principle should be able to produce sequence distributions with adequate modes, in order to generalise well to unseen contexts, compromises are made that lead to inadequate modes. In particular, the mapping of different contexts into a similar representational space, such that they jointly contribute observations to the same next-word distribution, may not always be because of actual linguistic equivalence of those contexts (see Figure 1). The result of this is the introduction of spread in those probability distributions and potentially also inadequate modes, especially when such spread is compounded to the sequence level. Therefore, while adequate modes could occur, they are not to be expected, especially if we want models to work well for previously unseen contexts.

2 The Inadequacy of the Mode

Neural language generation models consist of some neural network architecture, typically a Transformer (Vaswani et al., 2017), parameterising a distribution over discrete token sequences. This means that, possibly given some context x (*e.g.* a source sentence or instruction), a neural network parameterised with parameters θ predicts a distribution over sequences $P_\theta(Y|x)$. In practice, the random sequence is factorised without conditional independence assumptions into predicting next-word distributions using the chain rule of probability, and thus splitting Y into a sequence of tokens (*e.g.* words or sub-words) within some pre-defined vocabulary. The outcome space of the random sequence Y consists of all sequences that can be formed from tokens within this vocabulary ending with a special end-of-sequence token. Parameters θ are estimated using maximum likelihood estimation (MLE), which aims to capture the statistics of the training data. Regularisation techniques such as dropout (Srivastava et al., 2014) are employed along with a parameter bottleneck (a finite number of parameters) to generalise towards unseen contexts.

The inadequacy of the mode is the phenomenon that for a well-trained neural generation model, distributions $P_\theta(Y|x)$ are such that the highest probability mass put on any individual sequence (*i.e.* the mode of the distribution) is typically a sequence that is degenerate in some obvious way: an empty sequence (Stahlberg and Byrne, 2019), a too short translation (Koehn and Knowles, 2017), a copy of the input (Ott et al., 2018), etc. This degeneracy is also not typically exclusive to the mode of the distribution, rather the set of the highest scoring sequences up to some point are degenerate in some way and are thus of lower quality. Therefore, this phenomenon is sometimes also called the probability-quality paradox (Meister et al., 2022). While the individual sequences on which the model puts the highest probability mass are typically inadequate in some way, collectively most sequences can still be reasonable and cumulatively they may be much more probable than degenerate sequences (Eikema and Aziz, 2020). That is, oftentimes the mode and other high probability sequences do not get high probability mass in absolute terms, but sequence distributions simply exhibit high spread.

This results in decoding algorithms focused on finding the highest scoring sequence to have to severely limit the strength of the search (the beam search curse (Koehn and Knowles, 2017)) or having to alter the decoding objective to explicitly avoid known degeneracies (*e.g.* using a length penalty to avoid empty or too short translations). Other works or entire fields (mainly open-ended generation tasks) change the decoding algorithm to be better suited to these kinds of probability distributions (Eikema and Aziz, 2022; Suzgun et al., 2023) or use stochastic generation procedures such as truncated sampling instead (Fan et al., 2018; Holtzman et al., 2020; Meister et al., 2023). Some works change the training algorithm or fine-tune models with the goal of alleviating the inadequacy of the mode (Shi et al., 2020; Zhao et al., 2023). These works have some success, but they only alleviate the problem somewhat, rather than leading to models with consistently adequate modes.

Some research has shown that the extent of the inadequacy of the mode varies for different tasks and contexts. Stahlberg et al. (2022) show that higher variation in references is predictive of the degree with which the inadequacy of the mode is observed. Riley and Chiang (2022) similarly show using an artificial setup to constrain the amount of context available that the constrainedness of the task has an influence on the inadequacy of the mode. Yang et al. (2018) show that the problem of producing too short translations in NMT is worse for longer inputs. Eikema and Aziz (2020) show that the inadequacy of the mode is likely much worse

for out-of-domain and low-resource settings.

3 Existing Explanations

To our knowledge, there exist two attempts at explaining the inadequacy of the mode generally across generation tasks and types of degeneracies.

3.1 The Expected Information Hypothesis

The first strand of works (Meister et al., 2022, 2023) revolves around the observation that ground-truth sequences often get assigned an amount of information (the negative log probability) under the model close to the entropy of the model, both locally in next-word distributions (Meister et al., 2023) as well as on the sequence level (Meister et al., 2022). The authors hypothesise that this is an inherent property of human language and communication, optimising for reliability and efficiency. The authors therefore claim that as result, if our models are a good approximation of the underlying linguistic production process they would mimic this, placing human-like text at an amount of information content around the entropy. For high entropy models, the mode often has an amount of information content far away from the entropy. Hence, this would explain the inadequacy of the mode as an inherent property of human language, that good models therefore tend to capture as well. The authors coin this the expected information hypothesis.

3.2 Noisy Training Data

A second hypothesis attributes the inadequacy of the mode to noisy training data (Yoshida et al., 2023). So-called low-entropy distractors, a small amount of consistent noise present in the training data, such as for example copies of the input, can even in a perfect MLE fit in the non-parametric limit lead to inadequate modes. Assuming that the optimal distribution over sequences is uniform over valid ground-truth sequences[1] (as response to an input or continuations of a prompt), any noise that is more frequent than the uniform probability assigned to those sequences becomes modal, *i.e.* take on the mode of the distribution. Hence, inputs or prompts that allow for more variability suffer more from inadequate modes, as the probability assigned to each valid sequence is lower and noise rates are thus relatively higher.

[1]In practice, if we collected enough responses per input, this distribution is likely not uniform as some responses are more likely than others. However, for the argument this assumption is not crucial.

4 Where Existing Hypotheses Fall Short

While these two hypotheses could provide a partial explanation, we do not believe they paint a full picture of the problem.

While we won't contradict the claim that the human language generation process may have an inadequate mode as hypothesised by Meister et al. (2022), we do not believe that there is sufficient evidence for the claim that neural language generators should also exhibit such properties as a result. Well-curated training datasets would not contain inadequate sequences and given sufficient modelling capacity and data collected *for a single input or prompt* to be representative of human variability, there is no reason to believe that our models are unable to capture these empirical distributions perfectly in theory. Modes of neural language generators in grammatical error correction (Stahlberg et al., 2022) are known to be adequate, so we know that our models are in principle capable of exhibiting adequate modes. The inadequacy of the mode tells us, however, that this does not happen in practice when training models on more complex tasks and settings where generalisation towards unseen contexts is desired. Therefore, we cannot fully refute the expected information hypothesis either. Instead, we will provide an alternative hypothesis of why such settings lead to inadequate modes.

Training data noise may well be a problem as is claimed by Yoshida et al. (2023). It is plausible that a consistent type of noise that occurs marginally across inputs at a high rate becomes modal, especially for inputs or prompts that allow for a larger amount of variability. This is also in line with observations made in the literature, where less-constrained tasks and inputs that require longer sequences to be generated suffer more from the inadequacy of the mode (Stahlberg et al., 2022; Riley and Chiang, 2022; Yoshida et al., 2023), and the presence of copy noise (where the output is a copy of the input) in the training data is known to lead to beam search errors of the kind (Ott et al., 2018).

However, this does not cover all cases. Even when empty sequences are excluded from the training data, as is likely to happen nevertheless during pre-processing, Shi et al. (2020) observe that empty sequences are still assigned higher probability than references. Furthermore, the theoretical scenario where many continuations are observed per context is not at all like the typical setting, where often only a single continuation is observed per context.

5 The Role of Generalisation

We think it's worthwhile considering how tremendously difficult the task is that we are facing in modelling natural language. Essentially, for any conditional distribution $P_\theta(Y = y|x)$[2] that is being learned, there typically is only 1 observation for any x in the training data. An exact maximum likelihood solution in the non-parametric limit would concentrate all probability mass on the observed y, which would hardly be a good representation of the variability in human language. For a test input x^*, we likely even have never seen a single observation, meaning that the model has to fill in the entire sequence distribution from scratch.

Now, of course, we rely on the factorisation of our model as well as the contextual encoding power of neural networks to reduce data sparsity to a level that it can learn patterns from text data. While factorisation alone does not help reduce data sparsity much, it does break up the task of predicting a sequence distribution into predicting many smaller conditional probability distributions over the next word Y_j in the sequence given a context $c_j = (x, y_1^{j-1})$. This context is encoded by the neural network, which is where it has the ability to reduce data sparsity.

Most contexts c_j are likely unique within the training data, even after factorisation. The neural network, however, encodes a context into a d-dimensional continuous vector from which the probability distribution over the next word is predicted. Given a finite d, many such continuous vectors may map into practically equivalent next-word distributions. The neural network can use this fact to its advantage by mapping different contexts that are similar, from a next-word prediction standpoint, into similar representations.

It can for example learn conditional independencies, essentially only truly encoding a few components of the context into the continuous representation. This already reduces data sparsity, as now many similar contexts may contribute observations to the same next-word distribution if they share the right components within the prefix. We know, however, that neural networks can also learn to relate semantically similar words, even if their surface forms differ (Mikolov et al., 2013). This would allow even more contexts to contribute to the same conditional probability distribution, as seemingly completely different contexts (in terms of tokens present), may be mapped to similar representations.

Nothing in the model inherently enforces this to happen, however, and given sufficient modelling power the MLE solution may be able to perfectly fit the training data (*i.e.* overfit), likely without being able to usefully extrapolate to unseen contexts. Therefore, it is likely that modelling bottlenecks such as the size of the neural network, the use of dropout and other regularisation techniques is what is enforcing the model to "bundle" observations together by encoding different contexts similarly (see Figure 1).

Now, why would this affect the inadequacy of the mode? In order to reduce data sparsity to a manageable level, the network needs to encode different contexts similarly. There is no guarantee that contexts that are encoded similarly and thus predict similar next-word distributions are actually linguistically similar. That is, independence assumptions and perceived equivalence by the network does not need to mean that human continuations of the sequence are distributed identically for those contexts. We hypothesise that some contexts are mapped similarly for pragmatic reasons (*e.g.* because some observations in the training data overlap) or randomly due to artefacts of optimisation. It is also likely that some contexts still have not seen sufficient observations to make well-informed predictions.[3]

Ultimately, the bundling of similar but not truly (linguistically) equivalent contexts in combination with data sparsity for some contexts would result in probability mass of the sequence distribution being spread. While spread does not need to result in an *inadequate* mode, it also doesn't guarantee an *adequate* mode. When similarly encoded contexts are not predictive of actual human variation in the continuations of the sequence, it is unlikely that all those contexts get an adequate mode under the model, especially when all of these errors compound to the sequence level. Also, when probability mass is spread over many sequences (and thus each sequence gets smaller probability), other random artefacts of training such as some marginal noise in the training data as suggested by Yoshida

[2] We will assume that there is some input context x given at the start of generation. However, this is not central to our argument and a very similar argument can be made without having such an x available.

[3] It is also known that approximate MLE is "zero-avoiding", meaning it prefers to spread probability mass to cover all observations rather than concentrate it. This comes from an equivalence between MLE and minimising a KL-divergence from the model distribution to the data generation process.

et al. (2023) are more likely to be a problem.

6 Conclusion

In this position paper, we hypothesise that generalisation could play a crucial role in leading to inadequate modes in natural language generation systems. We leave it to future work to empirically validate or disprove this hypothesis. It would be particularly interesting to see whether the aforementioned "bundling" of observations can be detected throughout and after training of language generation models.

All in all, we argue that one should not expect their models to exhibit adequate modes for previously unseen inputs. Instead, one should expect to observe a considerable amount of uncertainty. While we do not rule out that we will eventually obtain models that exhibit adequate modes, it seems that we are still quite far away from that, given that the recent increase in model and data size with large language models has not lead to adequate modes yet (Yoshida et al., 2023). Accepting this uncertainty allows us to develop techniques to robustly generate from our models and properly evaluate them.

Acknowledgements


This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 (UTTER).

References

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. 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.

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*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. 2022. On the probability–quality paradox in language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 36–45, Dublin, Ireland. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*.

Darcey Riley and David Chiang. 2022. A continuum of generation tasks for investigating length bias and degenerate repetition. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 426–440, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xing Shi, Yijun Xiao, and Kevin Knight. 2020. Why neural machine translation prefers empty outputs.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? 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 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Felix Stahlberg, Ilia Kulikov, and Shankar Kumar. 2022. Uncertainty determines the adequacy of the mode and the tractability of decoding in sequence-to-sequence models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 8634–8645, Dublin, Ireland. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.

Davis Yoshida, Kartik Goyal, and Kevin Gimpel. 2023. Map's not dead yet: Uncovering true language model modes by conditioning away degeneracy.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*.