Decoding and Diversity in Machine Translation

Nicholas Roberts♠   **Davis Liang**◇   **Graham Neubig**♠   **Zachary C. Lipton**♠◇
♠Carnegie Mellon University; ◇Amazon AWS AI
{ncrobert, gneubig, zlipton}@cs.cmu.edu, liadavis@amazon.com

Abstract

Neural Machine Translation (NMT) systems are typically evaluated using automated metrics that assess the agreement between generated translations and ground truth candidates. To improve systems with respect to these metrics, NLP researchers employ a variety of heuristic techniques, including searching for the conditional mode (vs. sampling) and incorporating various training heuristics (e.g., label smoothing). While search strategies significantly improve BLEU score, they yield deterministic outputs that lack the diversity of human translations. Moreover, search can amplify socially problematic biases in the data, as has been observed in machine translation of gender pronouns. This makes *human-level BLEU* a misleading benchmark; modern MT systems cannot approach human-level BLEU while simultaneously maintaining human level translation diversity. In this paper, we characterize *distributional differences* between generated and real translations, examining the cost in diversity paid for the BLEU scores enjoyed by NMT. Moreover, our study implicates search as a salient source of known bias when translating gender pronouns.

1   Introduction

Neural Machine Translation (NMT) typically proceeds in the following two-stage pipeline: **(i)** train a conditional language model (using neural networks) by optimizing a probabilistic objective (the *modeling* stage); then **(ii)** produce predictions by *searching* for the mode (the hoped-for "best translation") of the conditional distribution (decoding either greedily or via beam search). We confirm that search is remarkably effective at maximizing BLEU. In fact, an NMT model trained for only $1/3$ of an epoch and decoded via search can match the BLEU score of a fully trained model decoded via sampling. Moreover, the fully trained model gains an additional 14 BLEU points when we decode deterministically via search instead.

However, due to search (either beam search or greedy decoding), NMT models are dialed to an extreme operating point of exhibiting zero variability (conditional on input) whereas human translations exhibit considerable variability. Yet NMT systems are often rated against "human-level" performance (calculated via BLEU on sentences with multiple available translations (Ott et al., 2018)), which makes this comparison misleading. Diversity in NMT is valuable for numerous reasons: **(a)** when translating between a non-gendered and gendered language, search will *disproportionately* choose the more frequent gender, conditioned on the input; **(b)** homogeneity can make NMT outputs monotonous and less engaging to users; **(c)** individuals who interact with language primarily through NMT might develop a warped exposure to that language. We stress that these issues are inherent to using deterministic search methods—*including both beam search and greedy decoding*—to recover high probability translations for the sake of optimizing BLEU score. In turn, the singular focus on improving BLEU leaves no incentive to address issues of diversity.

In this paper, we expose search as a cause of the lack of diversity in NMT outputs, as it relates to the translation of gender pronouns and a battery of other diversity metrics that we introduce. Specifically, we propose a panel of diversity diagnostics for NMT systems, measuring the distributional similarity (vs ground truth translations) of n-grams, sentence length, punctuation, and copy rates. We also examine domain confusion scores, using both linear discriminators with term frequency–inverse document
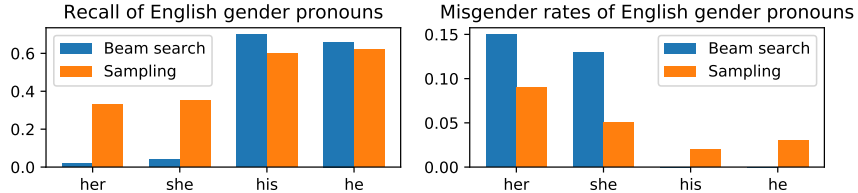
Figure 1: Recall and misgender rates of English gender pronouns made worse by beam search.

frequency (TF-IDF) features and BERT-based discriminators (Devlin et al., 2019) to distinguish between generated and real translations. Our study centers on the WMT 2017 German-English (De-En), English-German (En-De), and WMT 2014 French-English (Fr-En) datasets. We note that the focus on metrics like BLEU can cause researchers to disregard the consequences of ad hoc decisions that may improve BLEU while undermining any straightforward probabilistic interpretations of the learning objective. Minimizing cross entropy on the original targets corresponds to maximizing the likelihood of the data, but what can we say about the parameters that minimize cross entropy against label-smoothed targets? Examining the effect label smoothing on the diversity of NMT outputs, we find that when sampling, this results in poor performance in both BLEU score and our diversity diagnostics. When decoding via beam search the effect of label smoothing is minimal, which can lead to the negative consequences going unnoticed by the practitioner.

2   Related Work

BLEU score was designed to correlate with human judgements of translation quality (Papineni et al., 2002), although several studies have questioned this correspondence (Callison-Burch et al., 2006; Ma et al., 2019). Ott et al. (2018) explores the lack of diversity in NMT outputs and relate this to inherent uncertainty in the task. We note in passing that their study does not document the use of label smoothing, yet their observation that the "model distribution is too spread in hypothesis space" is an obvious consequence of label smoothing. We consider a broader set of diversity metrics and decoding strategies while using the same models and languages to facilitate conversation between these works. Furthermore, we explicitly consider the impact of label smoothing. Other studies propose decoding strategies to increase diversity, but lack comparisons to the ground truth distribution (Gimpel et al., 2013; Vijayakumar et al., 2016; Li and Jurafsky, 2016; Cho, 2016). It has been observed that NMT can often produce misgendered outputs, and solutions focusing on the modeling stage have been proposed, however, our study implicates search as a source of gender pronoun bias (Vanmassenhove et al., 2018; Saunders and Byrne, 2020).

Tevet and Berant (2020) propose a framework intended for tasks outside of NMT for evaluating diversity metrics relative to a "diversity parameter" used in decoding, expecting correlation between the metrics and the parameter. They consider as metrics the number of distinct n-grams from the model output, as done by Li et al. (2016), and BERT-based sentence similarity scores. In contrast, we additionally compare against the ground truth distribution of n-grams and use BERT as a discriminative model.

Müller et al. (2019) examine the role of label smoothing in NMT, claiming that improvements arise due to improved calibration of the model. Notably, Vaswani et al. (2017) employs label smoothing to improve beam search outputs at the expense of perplexity. Label smoothing may reduce overconfidence of predictions, which can be beneficial in light of known miscalibration issues in NMT models (Kumar and Sarawagi, 2019). However, we note here that label smoothing *is not a valid calibration procedure* and unsurprisingly, label-smoothed models remain miscalibrated (Wang et al., 2020). Most importantly, we find label-smoothing negatively impacts various human-level desiderata when sampling.

3   Experiments

Implementation details   We use the same convolutional architecture as Ott et al. (2018) in our experiments. For experiments using label smoothing, we set it to 0.1. We perform our analysis on the WMT'17 En-De dataset and present results in both directions. We repeat our analysis on the much larger WMT'14 En-Fr dataset, where we fix the task to be translation from French to English (the results of which can be found in Appendix C). Additional implementation details can be found in Appendix A. The sampling method used in our experiments involves randomly sampling from the softmax with some temperature at each time step. For beam search, we do not include any additional
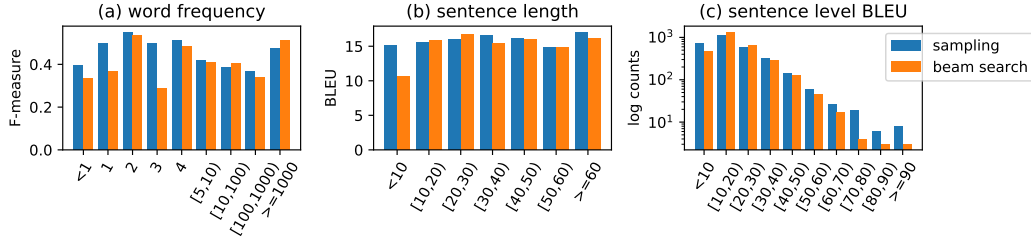
Figure 2: Left to right: Sampling has higher F1 for rare words, beam search achieves a lower BLEU score for shorter sentences, and sampling has a higher variance in BLEU scores than beam search.

penalties. For all sampling and search procedures, the output up to the current time step is passed as input to the decoder. We also note that as the sampling temperature approaches 0, the sampling procedure deterministically selects the argmax at each time step, which is equivalent to greedy search, and is also equivalent to beam search with a beam width of 1.

Beam search is biased towards selecting more frequent gender pronouns. We evaluate the bias of search toward more frequent tokens using a model trained to translate from German to English on the WMT'17 En-De dataset, and draw a direct comparison to sampling (i.e. not performing search at all) by fixing the BLEU score to a particular value for both. Our analysis for this is performed on the test set. We reduced the amount of training for the beam search decoded model to $1/3$ of an epoch to produce a model that achieves **the same** BLEU score (16.2) as the fully converged model decoded by sampling. We examine distributional differences including correct prediction rates of female pronouns, word frequency, sentence length, and the distribution of sentence-level BLEU scores between the two systems using compare-mt (Neubig et al., 2019). In this setting, we find that beam search underpredicts female gender pronouns. We find that the recall for the tokens 'she' and 'her' were significantly higher when decoding via sampling: 0.35 and 0.33, respectively, compared to 0.04 and 0.02 from beam search. In contrast, the recall of male pronouns from beam search was higher than that of sampling. Beam search also replaces 'she' and 'her' with male pronouns at higher rates than sampling, while never replacing male pronouns with female pronouns (see Figure 1). We also find that beam search attains a consistently lower F1 for rare words compared to sampling (Figure 2a). We show in Figure 2b that beam search achieves lower BLEU scores for shorter sentences, and in Figure 2c, we find that sampling results in higher a variance of BLEU scores than beam search, and in particular, more sentences with BLEU scores of 30 or higher.

Our diversity diagnostics reveal trade-offs between diversity and BLEU. We now focus on fully trained models on translation from English to German and compare the distributions of translations produced via beam search and sampling to reference translations. We present a variety of metrics to capture various aspects of the distributions and, in the spirit of Tevet and Berant (2020), compare various operating points along a spectrum mediated by dialing the softmax temperature to elucidate the trade-off between diversity and BLEU. In particular, we sample using softmax temperatures, T, ranging from 0 to 1 and compare this to beam search with a beam width, B, up to 10. We repeat these analyses using models trained with and without label smoothing, and note that the side effects of label smoothing are prominent when sampling, and minimal when decoding deterministically.

To assess similarity of n-gram frequency and sentence length between generated and human translations, we adopt the L1 distance, due to its simplicity and robustness to small changes in single elements (unlike KL divergence). We calculate the L1 distance between the normalized histograms for the model output on the validation set for a given decoding strategy and those of reference translations. For n-grams, we compare frequencies of n-grams and report results for $n = 1, 5$. When evaluating sentence length, all unique sentence lengths are considered. As a baseline, we compute the L1 distance between validation set partitions. We find that a temperature of 1 achieves L1 distance similar to the baseline (Figures 3a and 3b). For n-grams, we observe a trade-off between closeness in L1 distance to the ground truth distribution and BLEU score. Evaluating the L1 distance between the sentence length distributions, we find that both sampling and beam search produce outputs with a similar distribution of sequence lengths to ground truth candidates. For n-gram L1 distances, sampling results in a closer distributional similarity to the ground truth than beam search. In all cases, label smoothing results in a lower distributional similarity to the ground truth (Figure 3a,b,c).

We compare the total frequencies of a subset of the vocabulary in generated text to that of validation set references. We evaluate the frequencies of punctuation and of male and female gender pronouns
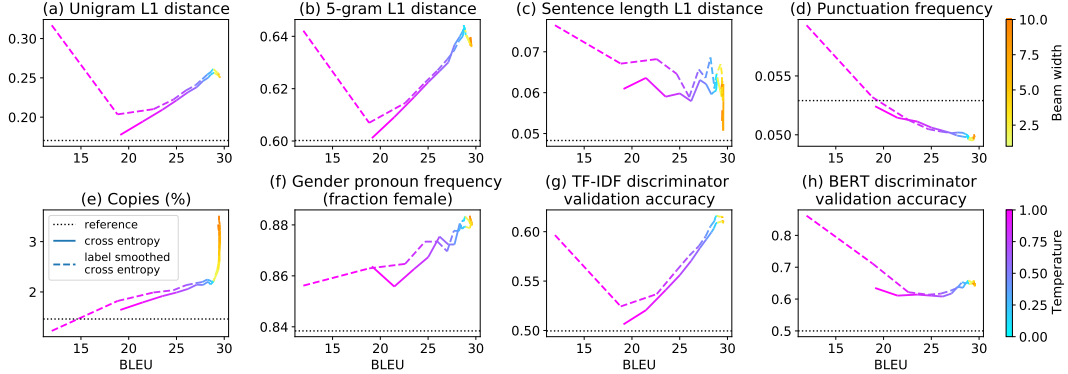
Figure 3: Distributional similarity to gold translations across temperatures for sampling and beam widths for beam search. Improvements in BLEU often come at the cost of distribution dissimilarity.

(see Appendix B). In Figure 3d, we show that punctuation frequency decreases relative to the reference as temperature decreases, and increasing beam width does little to improve this. Label smoothing at $T = 1.0$ actually increases punctuation frequency well above that of the reference. In Figure 3f, we evaluate the fraction of pronouns which are female. Decreasing temperature from 1 and applying beam search increases the fraction of German female gender pronouns. The German word 'sie' translates as 'she,' 'they,' or 'you' in English, yielding an effectively higher representation of female gender pronouns than male in the training set. Beam search is thus biased toward the more common gender pronoun, 'sie.' We see this same effect when translating from German to English and French to English, except in these cases, male pronouns are more represented in the training set. Hence in these cases, the bias is toward male pronouns, as shown in Appendix C. In both cases, beam search outputs are biased toward the more represented gender in the training set.

We similarly evaluate the rate of copies from the sentences in the source language to the output. Like Ott et al. (2018), we define the copy rate as the fraction of sentences with more than 50% unigram overlap, excluding punctuation and numbers. We compare these scores with the copy rate between the sources and references in the training set. Our findings in Figure 3e corroborate the findings of Ott et al. (2018) regarding the exacerbation of copy rates by beam search. We expand on this by showing that sampling at $T = 1.0$ results in copy rates which nearly match the copy rates measured on the training set. Hence the exacerbation of copy rates is specifically due to beam search.

Finally, we examine the ability for discriminators to distinguish between model outputs on the validation set and validation set reference translations. We first construct a dataset comprising generated translations from the model, labeled 'generated,' and reference translations labeled 'real,' which are taken from a different partition of the validation set. We use half of the resulting dataset to train a discriminator and we use the remaining half to evaluate the generalization performance of the discriminator. We train a logistic regression model using TF-IDF features and, using the same setup, fine-tune a BERT-based discriminator (108M parameters). We find that even logistic regression trained on TF-IDF features (Figure 3g) can classify samples generated by beam search above 60% accuracy but cannot distinguish between translations generated via sampling with temperature 1 and reference translations. Out of all the methods we evaluated, using label smoothing and sampling with temperature 1 produces the most discriminable output. A BERT-based discriminator, on the other hand (Figure 3h), can distinguish between generated and reference samples better than the linear discriminator. In both cases, outputs generated via sampling are harder to discriminate from ground truth translations compared to outputs generated via search, and outputs from label smoothed models are always easier to classify as 'generated.'

4 Conclusions

Examining distributional dissimilarity between the outputs of NMT systems under sampling, beam search, and natural translations, we find that beam search performs well by BLEU score, but there is a significant cost to be paid in naturalness and diversity, including a higher rate of misgendering of gender pronouns. Moreover, modifications to the objective undertaken to increase BLEU (here, label smoothing), can have unintended side effects that practitioners focused on BLEU might overlook. In future work, we plan to explore techniques to achieve the highest possible BLEU subject to constraints on the distributional similarity between generated and natural translations.

4

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Kyunghyun Cho. 2016. Noisy parallel approximate decoding for conditional recurrent language model.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Seattle, Washington, USA. Association for Computational Linguistics.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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 and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4694–4703. Curran Associates, Inc.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. *arXiv preprint arXiv:2004.02990*.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*.

A Implementation details

The convolutional sequence to sequence model used in our experiments consists of encoder and decoder networks, each of which contains several 'blocks' of convolutional layers (Gehring et al., 2017). All models are trained on 8 V100 GPUs using Fairseq-py (Ott et al., 2019), with learning rate 0.5 (for En-De and De-En) and 0.25 (for Fr-En), with a fixed learning rate schedule, a clip norm of 0.1, dropout of 0.2, and 4000 maximum tokens. All models were trained to convergence (up to 100 epochs), with the best checkpoint chosen based on validation set performance. We note that the sampling method used in our experiments involves randomly sampling from the softmax with a temperature parameter at each time step. For beam search decoding, we do not include any additional penalties on the search. For all sampling and search procedures, the output up to the current time step is passed as input to the decoder. We also note that as the sampling temperature approaches 0, the sampling procedure deterministically selects the argmax at each time step, which is equivalent to greedy search, and is also equivalent to beam search with a beam width of 1.

B Token subsets used for punctuation and gender pronoun frequency scores

Category	Subset
punctuation	. , ? ! " ' ... !!! ?! !? ; :
English female	she, her, hers, herself
English male	he, him, his, himself
German female	sie
German male	er

Table 1: Subsets used for punctuation and gender pronoun analysis

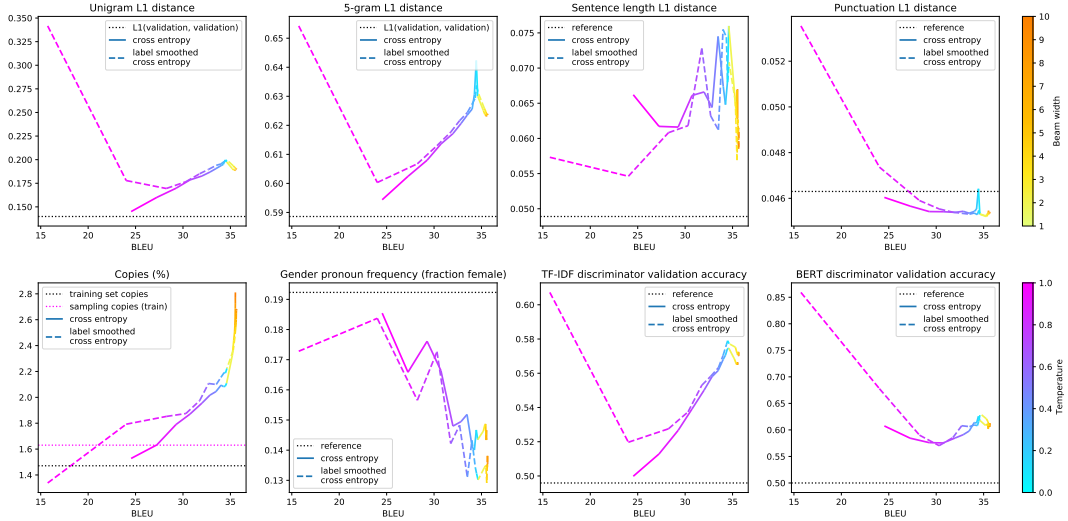# C Diversity diagnostics applied to De-En and Fr-En

Figure 4: Results from Figure 3 reproduced on the WMT 2017 English-German dataset, where the task is translation from German to English.
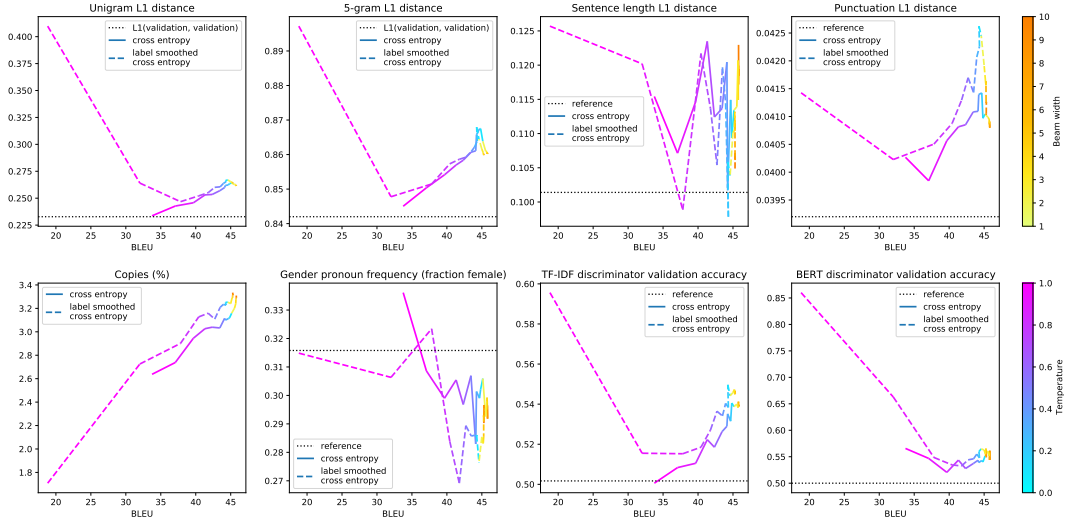

Figure 5: Results from Figure 3 reproduced on the WMT 2014 English-French dataset, where the task is translation from French to English.