Generalizing Back-Translation in Neural Machine Translation

Miguel Graça[1†] **Yunsu Kim**[1] **Julian Schamper**[1†] **Shahram Khadivi**[2] **Hermann Ney**[1]
[1]Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{surname}@i6.informatik.rwth-aachen.de
[2]eBay, Inc., Aachen, Germany
{skhadivi}@ebay.com

Abstract

Back-translation — data augmentation by translating target monolingual data — is a crucial component in modern neural machine translation (NMT). In this work, we reformulate back-translation in the scope of cross-entropy optimization of an NMT model, clarifying its underlying mathematical assumptions and approximations beyond its heuristic usage. Our formulation covers broader synthetic data generation schemes, including sampling from a target-to-source NMT model. With this formulation, we point out fundamental problems of the sampling-based approaches and propose to remedy them by (i) disabling label smoothing for the target-to-source model and (ii) sampling from a restricted search space. Our statements are investigated on the WMT 2018 German ↔ English news translation task.

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2014; Vaswani et al., 2017) systems make use of back-translation (Sennrich et al., 2016a) to leverage monolingual data during the training. Here an inverse, target-to-source, translation model generates synthetic source sentences, by translating a target monolingual corpus, which are then jointly used as bilingual data.

Sampling-based synthetic data generation schemes were recently shown to outperform beam search (Edunov et al., 2018; Imamura et al., 2018). However, the generated corpora are reported to stray away from the distribution of natural data (Edunov et al., 2018). In this work, we focus on investigating why sampling creates better training data by re-writing the loss criterion of an NMT model to include a model-based data generator. By doing so, we obtain a deeper understanding of synthetic data generation methods, identifying their desirable properties and clarifying the practical approximations.

In addition, current state-of-the-art NMT models suffer from probability smearing issues (Ott et al., 2018) and are trained using label smoothing (Pereyra et al., 2017). These result in low-quality sampled sentences, which influence the synthetic corpora. We investigate considering only high-quality hypotheses by restricting the search space of the model via (i) ignoring words under a probability threshold during sampling and (ii) N-best list sampling.

We validate our claims in experiments on a controlled scenario derived from the WMT 2018 German ↔ English translation task, which allows us to directly compare the properties of synthetic and natural corpora. Further, we present the proposed sampling techniques on the original WMT German ↔ English task. The experiments show that our restricted sampling techniques work comparable or superior to other generation methods by imitating human-generated data better. In terms of translation quality, these do not result in consistent improvements over the typical beam search strategy.

2 Related Work

Sennrich et al. (2016a) introduce the back-translation technique for NMT and show that the quality of the back-translation model, and therefore resulting pseudo-corpus, has a positive effect on the quality of the subsequent source-to-target model. These findings are further investigated in (Hoang et al., 2018; Burlot and Yvon, 2018) where the authors confirm work effect. In our work, we expand upon this concept by arguing that the quality of the resulting model not only depends on the

† Now at DeepL GmbH.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 45–52
Florence, Italy, August 1-2, 2019.

data fitness of the back-translation model but also on how sentences are generated from it.

Cotterell and Kreutzer (2018) frame back-translation as a variational process, with the space of source sentences as the latent space. Their approach argues that the distribution of the synthetic data generator and the true translation probability should match. Thus it is invaluable to clarify and investigate the sampling distributions that current state-of-the-art data generation techniques utilize. A simple property is that a target sentence must be allowed to be aligned to multiple source sentences during the training phase. Several efforts (Hoang et al., 2018; Edunov et al., 2018; Imamura et al., 2018) confirm that this is in fact beneficial. Here, we unify these findings by re-writing the optimization criterion of NMT models to depend on a data generator, which we define for beam search, sampling and N-best list sampling approaches.

3 How Back-Translation Fits in NMT

In NMT, one is interested in translating a source sentence $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ into a target sentence $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. For this purpose, the translation process is modelled via a neural model $p_\theta(e_i|f_1^J, e_1^{i-1})$ with parameters θ.

The optimal optimization criterion of an NMT model requires access to the true joint distribution of source and target sentence pairs $Pr(f_1^J, e_1^I)$. This is approximated by the empirical distribution $\hat{p}(f_1^J, e_1^I)$ derived from a bilingual data-set $(f_{1,s}^{J_s}, e_{1,s}^{I_s})_{s=1}^S$. The model parameters are trained to minimize the cross-entropy, normalized over the number of target tokens, over the same.

$$L(\theta) = -\sum_{(f_1^J, e_1^I)} Pr(f_1^J, e_1^I) \cdot \frac{1}{I} \log p_\theta(e_1^I|f_1^J) \tag{1}$$

$$= -\sum_{(f_1^J, e_1^I)} \hat{p}(f_1^J, e_1^I) \cdot \frac{1}{I} \log p_\theta(e_1^I|f_1^J) \tag{2}$$

$$= -\frac{1}{S}\sum_{s=1}^{S} \frac{1}{I_s} \log p_\theta(e_{1,s}^{I_s}|f_{1,s}^{J_s}) \tag{3}$$

Target monolingual data can be included by generating a pseudo-parallel source corpus via, e.g. back-translation or sampling-based methods. In this section, we describe such generators as a component of the optimization criterion of NMT models and discuss approximations made in practice.

3.1 Derivation of the Generation Criterion

Eq. 1 is the starting point of our derivation in Eqs. 4-6. $Pr(f_1^J, e_1^I)$ can be decomposed into the true language probability $Pr(e_1^I)$ and true translation probability $Pr(f_1^J|e_1^I)$. These two probabilities highlight the assumptions in the scenario of back-translation: we have access to an empirical target distribution $\hat{p}(e_1^I)$ with which $Pr(e_1^I)$ is approximated, derived from the monolingual corpus $(e_{1,s}^{I_s})_{s=1}^S$. However, one lacks access to $\hat{p}(f_1^J|e_1^I)$. Generating synthetic data is essentially the approximation of the true probability of $Pr(f_1^J|e_1^I)$. It can be described as a sampling distribution[1] $q(f_1^J|e_1^I; p)$ parameterized by the target-to-source model p.

$$L(\theta) = -\sum_{(f_1^J, e_1^I)} Pr(f_1^J, e_1^I) \cdot \frac{1}{I} \log p_\theta(e_1^I|f_1^J) \tag{4}$$

$$= -\sum_{e_1^I} Pr(e_1^I) \cdot \frac{1}{I} \sum_{f_1^J} Pr(f_1^J|e_1^I) \cdot \log p_\theta(e_1^I|f_1^J) \tag{5}$$

$$= -\sum_{e_1^I} \hat{p}(e_1^I) \cdot \frac{1}{I} \sum_{f_1^J} q(f_1^J|e_1^I; p) \cdot \log p_\theta(e_1^I|f_1^J) \tag{6}$$

This derivation highlights an apparent condition that the generation procedure $q(f_1^J|e_1^I; p)$ should result in a distribution of source sentences similar to the true data distribution $Pr(f_1^J|e_1^I)$. Cotterell and Kreutzer (2018) show a similar derivation hinting towards an iterative wake-sleep variational scheme (Hinton et al., 1995), which reaches similar conclusions.

Following this, we formulate two issues with the back-translation approach: (i) the choice of generation procedure q and (ii) the adequacy of the target-to-source model p. The search method q is responsible not only for controlling the output of source sentences but also to offset the deficiencies of the target-to-source model p.

An implementation for q is, for example, beam search where q is a *deterministic* sampling procedure, which returns the highest scoring sentence according to the search criterion:

$$q_{\text{beam}}(f_1^J|e_1^I; p) = \begin{cases} 1, & f_1^J = \underset{\hat{J}, f_1^{\hat{J}}}{\text{argmax}} \left\{ \frac{1}{\hat{J}} \log p(f_1^{\hat{J}}|e_1^I) \right\} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

[1] The properties of a probability distribution hold for $q(f_1^J|e_1^I; p)$.

Sampling as described by Edunov et al. (2018) would be simply the equality

$$q_{\text{sample}}(f_1^J|e_1^I;p) = p(f_1^J|e_1^I). \quad (8)$$

3.2 Approximations

Applications of back-translation and its variants largely follows the initial approach presented in (Sennrich et al., 2016a). Each target authentic sentence is aligned to a single synthetic source sentence. This new dataset is then used as if it were bilingual. This section is dedicated to the clarification of the effect of such a strategy in the optimization criterion, especially with non-deterministic sampling approaches (Edunov et al., 2018; Imamura et al., 2018).

Firstly, the sum over all possible source sentences in Eq. 6 is approximated by a restricted search space of N sentences, with $N = 1$ being a common choice. Yet, the cost of *generating* the data and *training* on the same scales linearly with N and it is unattractive to choose higher values.

Secondly, the pseudo-corpora are static across training, i.e. the synthetic sentences do not change across training epochs, which appears to cancel out the benefits of sampling-based methods. Correcting this behaviour requires an on-the-fly sentence generation, which increases the complexity of the implementation and slows down training considerably. Back-translation is not affected by this approximation since the target-to-source model always generates the same translation.

The approximations are shown in Eq. 9 with a fixed pseudo-parallel corpus where $e_{1,s}^{I_s}$ is aligned to N source sentences $(f_{1,s,n}^{J_{s,n}})_{n=1}^{N}$.

$$L(\theta) \approx -\sum_{s=1}^{S} \frac{1}{N \cdot I_s} \sum_{n=1}^{N} \log p_\theta\left(e_{1,s}^{I_s}|f_{1,s,n}^{J_{s,n}}\right) \quad (9)$$

We hypothesize that these conditions become less problematic when large amounts of monolingual data are present due to the law of large numbers, which states that repeated occurrences of the same sentence e_1^I will lead to a representative distribution of source sentences f_1^J according to $q(f_1^J|e_1^I;p)$. In other words, given a high number of representative target samples, Eq. 9 matches Eq. 6 with $N = 1$. This shifts the focus of the problem to find an appropriate search method q and generator p.

4 Improving Synthetic Data

In this section, we discuss how the known generation methods $q(f_1^J|e_1^I;p)$ fail in approximating $Pr(f_1^J|e_1^I)$ due to modelling issues of model p and consider how the generation approach q can be adapted to compensate p.

We base our remaining work on the approximations presented in Section 3.2 and consider $N = 1$ synthetic sentences. The reasoning for this is twofold: (i) it is the most attractive scenario in terms of computational costs and (ii) the approximations lose their influence with large target monolingual corpora.

4.1 Issues in Translation Modelling

With sampling-based approaches, one does not only care about whether high-quality sentences get assigned a high probability, but also that low-quality sentences are assigned a low probability.

Label smoothing (LS) (Pereyra et al., 2017) is a common component of state-of-the-art NMT systems (Ott et al., 2018). This teaches the model to (partially) fit a uniform word distribution, causing unrestricted sampling to periodically sample from the same. Even without LS, NMT models tend to smear their probability to low-quality hypotheses (Ott et al., 2018).

To showcase the extent of this effect, we provide the average cumulative probabilities of top-N words for NMT models, see Section 5.2, trained with and without label smoothing in Figure 1. The distributions are created on the development corpus. We observe that training a model with label smoothing causes a re-allocation of roughly 7% probability mass to all except the top-100 words. This re-allocation is not problematic during beam search, since this strategy only looks at the top-scoring candidates. However, when considering sampling for data generation, there is a high likelihood that one will sample from the space of low probability words, creating non-parallel outputs, see Table 4.

4.2 Restricting the Search Space

Changing the search approach q is less arduous than changing the model p since it does not involve re-training the model. Restricting the search space to high-probability sentences avoids the issues highlighted in Section 4.1 and provides a middle-ground between unrestricted sampling and beam search.

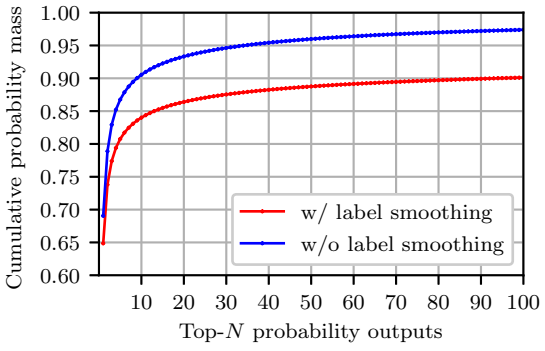

Figure 1: Cumulative probabilities of the top-N word candidates as estimated on newstest2015 English → German with and without label smoothing. See section 5.2 for descriptions of the models.

Edunov et al. (2018) consider top-k sampling to avoid the aforementioned problem, however, there is no guarantee that the candidates are confident predictions. We propose two alternative methods: (i) restrict the sampling outputs to words with a minimum probability and (ii) weighted sampling from the N-best candidates.

4.2.1 Restricted Sampling

The first approach follows sampling directly from the model $p(\cdot|e_1^I, f_1^{j-1})$ at each position j, but only taking words with at least $\tau \in [0, 0.5)$ probability into account. Afterwards, another softmax activation[2] is performed only over these words by masking all the remaining ones with large negative values. If no words have over τ probability, then the maximum probability word is chosen. Note that a large τ gets closer to greedy search ($\tau \geq 0.5$) and a lower value gets near to unrestricted sampling.

$$q(f|e_1^I, f_1^{j-1}; p) = \tag{10}$$
$$\begin{cases} \text{softmax}\big(p(f|e_1^I, f_1^{j-1}), C\big), & |C| > 0 \\ 1, & |C| = 0 \wedge \\ & f = \underset{f'}{\text{argmax}}\Big\{p(f'|e_1^I, f_1^{j-1})\Big\} \\ 0, & \text{otherwise} \end{cases}$$

with $C \subseteq V_f$ being the subset of words of the source vocabulary V_f with at least τ probability:

$$C = \big\{f \mid p(f|e_1^I, f_1^{j-1}) \geq \tau\big\} \tag{11}$$

and $\text{softmax}\big(p(f|e_1^I, f_1^{j-1}), C\big)$ being a soft-max normalization restricted to the elements in C.

[2] Alternatively an L1-normalization would be sufficient.

4.2.2 N-best List Sampling

The second approach involves generating a list of N-best candidates, normalizing the output scores with a soft-max operation, as in Section 4.2.1, and finally sampling a hypothesis.

The score of a translation is abbreviated by $s(f_1^J|e_1^I) = \frac{1}{J} \log p(f_1^J|e_1^I)$.

$$q_{\text{nbest}}(f_1^J|e_1^I; p) = \tag{12}$$
$$\begin{cases} \text{softmax}\big(s(f_1^J|e_1^I), C\big), & f_1^J \in C \\ 0, & \text{otherwise} \end{cases}$$

with $C \subseteq \mathbb{D}_{src}$ being the set of N-best translations found by the target-to-source model and $\mathbb{D}_{src}$ being the set of all source sentences:

$$C = \underset{\mathcal{D} \subset \mathbb{D}_{src}: |\mathcal{D}| = N}{\text{argmax}} \Big\{ \sum_{f_1^J \in \mathcal{D}} s(f_1^J|e_1^I) \Big\}. \tag{13}$$

5 Experiments

5.1 Setup

This section makes use of the WMT 2018 German ↔ English [3] news translation task, consisting of 5.9M bilingual sentences. The German and English monolingual data is subsampled from the deduplicated NewsCrawl2017 corpus. In total 4M sentences are used for German and English monolingual data. All data is tokenized, true-cased and then preprocessed with joint byte pair encoding (Sennrich et al., 2016b)[4].

We train Base Transformer (Vaswani et al., 2017) models using the Sockeye toolkit (Hieber et al., 2017). Optimization is done with Adam (Kingma and Ba, 2014) with a learning rate of 3e-4, multiplied with 0.7 after every third 20k-update checkpoint without improvements in development set perplexity. In Sections 5.2 and 5.3, word batch sizes of 16k and 4k are used respectively. Inference uses a beam size of 5 and applies hypothesis length normalization.

Case-sensitive BLEU (Papineni et al., 2002) is computed using the `mteval_13a.pl` script from Moses (Koehn et al., 2007). Model selection is performed based on the BLEU performance on newstest2015. All experiments were performed using the workflow manager Sisyphus (Peter et al., 2018). We report the statistical significance of

[3] http://www.statmt.org/wmt18/translation-task.html

[4] 50k merge operations and a vocabulary threshold of 50 are used.

	test2015	test2017	test2018
beam search	30.9*	31.9*	**40.1**
sampling	30.4*	31.0*	37.9*
w/o LS	30.4*	31.3*	37.9*
$\tau = 10\%$	**31.1***	**32.1***	39.8
50-best sampling	**31.1***	31.9*	39.8
reference	**32.6**	**33.5**	40.0

Table 1: BLEU[%] results for the controlled scenario. * denotes a p-value of < 0.01 w.r.t. the reference.

	Entropy	PPL	
	En → De	Train	test2015
beam search	2.60	2.74	5.77
sampling	3.13	9.07	5.55
w/o LS	**2.93**	**5.17**	**5.31**
$\tau = 10\%$	2.66	3.34	5.61
50-best sampling	2.62	2.84	5.70
reference	**2.91**	**5.18**	**4.50**

Table 2: IBM-1 model entropy and perplexity (PPL) on the training and development set for the controlled scenario using different synthetic generation methods.

our results with MultEval (Clark et al., 2011). A low p-value indicates that the performance gap between two systems is likely to hold given a different sample of a random process, e.g. an initialization seed.

5.2 Controlled Scenario

To compare the performance of each generation method to natural sentences, we shuffle and split the German → English bilingual data into 1M bilingual sentences and 4.9M monolingual sentences. This gives us a reference translation for each sentence and eliminates domain adaptation effects. The generator model is trained on the smaller corpus until convergence on BLEU, roughly 100k updates. The final source-to-target model is trained from scratch on the concatenated synthetic and natural corpora until convergence on BLEU, roughly 250k updates for all variants.

Table 1 showcases the translation quality of the models trained on different kinds of synthetic corpora. Contrary to the observations in Edunov et al. (2018), unrestricted sampling does not outperform beam search and once the search space is restricted all methods perform similarly well.

To further investigate this, we look at other relevant statistics of a generated corpus and the performance of the subsequent models in Table 2. These are the perplexities (PPL) of the model on the training and development data and the entropy of a target-to-source IBM-1 model (Brown et al., 1993) trained with GIZA++ (Och and Ney, 2003). The training set PPL varies strongly with each generation method since each produces hypotheses of differing quality. All methods with a restricted search space have a larger translation entropy and smaller training PPL than natural data. This is due to the sentences being less noisy and also the translation options being less varied. Unrestricted sampling seems to overshoot the statistics of natural data, attaining higher entropy values.

However, once LS is removed, the best PPL on the development set is reached and the remaining statistics match the natural data very closely. Nevertheless, the performance in BLEU lags behind the methods that consider high-quality hypotheses as reported in Table 1. Looking further into the models, we notice that when trained on corpora with more variability, i.e. larger translation entropy, the probability distributions are flatter. We explain the better dev perplexities with unrestricted sampling with the same reason for which label smoothing is helpful: it makes the model less biased towards more common events (Ott et al., 2018). This uncertainty is, however, not beneficial for translation performance.

5.3 Real-world Scenario

Previously, we applied different synthetic data generation methods to a controlled scenario for the purpose of investigation. We extend the experiments to the original WMT 2018 German ↔ English task and showcase the results in Table 3. In contrast to the experiments of Section 5.2, the distribution of the monolingual data now differs from the bilingual data. The models are trained on the bilingual data for 1M updates and then fine-tuned for further 1M updates on the concatenated bilingual and synthetic corpora.

The restricted sampling techniques perform comparable to or better than the other synthetic data generation methods in all cases. Especially on English → German, unrestricted sampling only produces statistical significant improvements over beam search when LS is replaced. Furthermore, restricting the search space via 50-best list sam-

	De → En		En → De	
	test2017	test2018	test2017	test2018
beam search	35.7	**43.6**	28.2	41.3
sampling	35.8	42.3*	28.6	41.5
w/o LS	35.9	42.5*	**29.1***	41.7
$\tau = 10\%$	35.9	43.0*	28.7*	41.6
50-best samp.	**36.0**	**43.6**	28.6*	**41.8***

Table 3: WMT 2018 German ↔ English BLEU[%] values comparing different synthetic data generation methods.
* denotes a p-value of < 0.01 w.r.t. beam search.

pling improves significantly in both test sets.

We observe that on German → English newstest2018 particularly, there is a large drop in performance when using unrestricted sampling. This is slightly alleviated by applying a minimum probability threshold of $\tau = 10\%$, but there is still a gap to be closed. This behaviour is investigated in the following section.

5.3.1 Scalability

A benefit of non-deterministic generation methods is the scalability in contrast to beam search. Under the assumption of a good fitting translation model, as argued in Section 3, sampling does appear to be the best option.

We compare different monolingual corpus sizes for the German → English task in Figure 2 on three different test sets. Particularly, newstest2018 shows the exact opposite behaviour from the remaining test sets: the amount of data generated via beam search improves the resulting model, whereas sampling improves the system by a small margin. Normal sampling has a general tendency to perform better with more data, but it saturates in two test sets (newstest2015 and newstest2018). Restricted sampling appears to be the most consistent approach, always outperforming unrestricted sampling and also always scaling with a larger set of monolingual data.

These observations are strongly linked to the properties of current state-of-the-art models, see Section 4.1 and experimental setup via e.g. the domain of the bilingual, monolingual and test data. Therefore, the high performance scaling with beam search in newstest2018 might be due to its *relatedness* to the training data as measured by the high BLEU values attained in inference.

5.4 Synthetic Source Examples

To highlight the issues present in unrestricted sampling, we compare the outputs of different generation methods in Table 4. The unrestricted sampling output hypothesizes a second sentence which is not related at all to the input sentence but generates a much longer sequence. The restricted sampling methods and the model trained without label smoothing provide an accurate translation of the input sentence. Compared to the beam search hypothesis, they have a reasonable variation which is indeed closer to the human-translated reference.

6 Conclusion

In this work, we link the optimization criterion of an NMT model with a synthetic data generator defined for both beam search and additional sampling-based methods. By doing so, we identify that the search method plays an important role, as it is responsible for offsetting the shortcomings of the generator model. Specifically, label smoothing and probability smearing issues cause sampling-based methods to generate unnatural sentences.

We analyze the performance of our techniques on a closed- and open-domain of the WMT 2018 German ↔ English news translation task. We provide qualitative and quantitative evidence of the detrimental behaviours and show that these can be influenced by re-training the generator model without label smoothing or by restricting the search space to not consider low-probability outputs. In terms of translation quality, sampling from 50-best lists outperforms beam search, albeit at a higher computational cost. Restricted sampling or the disabling of label smoothing for the generator model are shown to be cost-effective ways of improving upon the unrestricted sampling approach of Edunov et al. (2018).

Acknowledgments

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS"), the Deutsche Forschungsgemeinschaft (DFG; grant agreement NE 572/8-1,

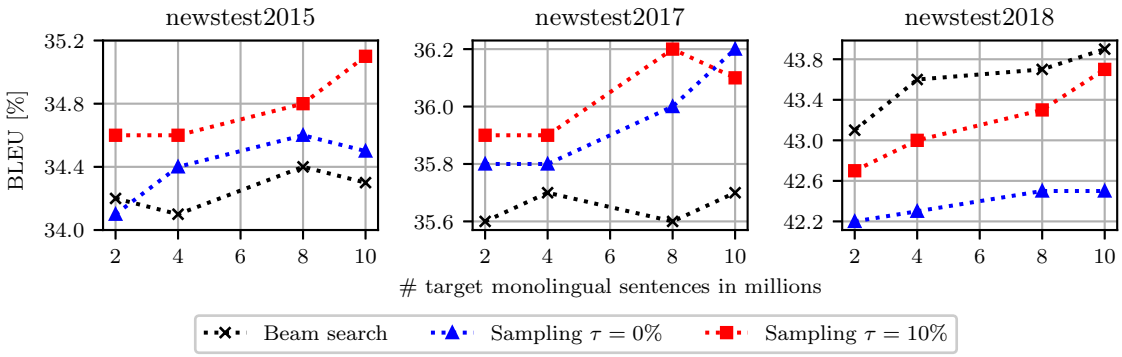

Figure 2: WMT 2018 German → English BLEU[%] values comparing different synthetic data generation methods with a differing size of synthetic corpus.

source	it is seen as a long sag@@ a full of surprises .
beam search	es wird als eine lange Geschichte voller Überraschungen angesehen .
sampling	es wird als eine lange S@@ aga voller Überraschungen angesehen . injury , Skepsis , Feuer) , Duschen verursach@@ ter Körper , Pal@@ ä@@ ste , Gol@@ fen , Flu@@ r und Mu@@ ffen , Diesel@@ - Total Bab@@ ylon , der durch@@ s Wasser und Wasser@@ kraft fliet .
w/o label smoothing	es wurde als eine lange Geschichte voller Überraschungen gesehen .
$\tau = 10\%$	es wird als lange S@@ age voller Überraschungen angesehen .
50-best sampling	es wird als eine lange S@@ age voller Überraschungen gesehen .
reference	er wird als eine lange S@@ aga voller Überraschungen angesehen .

Table 4: Random example generated by different methods for the controlled scenario of WMT 2018 German → English. @@ denotes the subword token delimiter.

project "CoreTec"), and eBay Inc. The GPU cluster used for the experiments was partially funded by DFG Grant INST 222/1168-1. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. Version 4.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 144–155.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 176–181.

Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv preprint arXiv:1806.04402*. Version 1.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*. Version 2.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*. Version 2.

Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. 1995. The" wake-sleep" algo-

rithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (WNMT 2018)*, pages 18–24.

Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. 2018. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (WNMT 2018)*, pages 55–63.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. Version 9.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, pages 19–51.

Myle Ott, Michael Auli, David Granger, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*. Version 4.

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 on Association for Computational Linguistics (ACL 2002)*, pages 311–318.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*. Version 1.

Jan-Thorsten Peter, Eugen Beck, and Hermann Ney. 2018. Sisyphus, a workflow manager designed for machine translation and automatic speech recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 84–89.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1715–1725.

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 (NIPS 2017)*, pages 6000–6010.