Generalizing Back-Translation in Neural Machine Translation

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi* and Hermann Ney

{surname}@i6.informatik.rwth-aachen.de

skhadivi@ebay.com*

WMT 2019, August 1st 2019

Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Germany

***eBay Inc., Aachen, Germany**

Back-Translation

Back-translation (BT) [Sennrich & Haddow+ 16]

- State-of-the-art way to use monolingual target corpora
- Generate target-to-source translations to obtain synthetic data
- Recent variants:
  - Sampling [Edunov & Ott+ 18, Imamura & Fujita+ 18]
  - Sum over N-best [Zhang & Liu+ 18]

This work

- A general formulation for all BT variants: the role of synthetic data in NMT
- Clarifies the advantage of sampling approaches over beam search
- Highlights deficiencies in SOTA models and proposes solutions for them

Training Criterion of NMT

Notations

- Source sentence $f_1^J = f_1 \ldots f_j \ldots f_J$, target sentence $e_1^I = e_1 \ldots e_i \ldots e_I$
- Distributions: Pr (true), $\hat{p}$ (empirical from data), p_θ (model)

Training criterion $L(\theta)$ for parameters θ: cross-entropy

Standard scenario

$$
\begin{aligned}
L(\theta) &= -\sum_{(f_1^J, e_1^I)} Pr(f_1^J, e_1^I) \cdot \frac{1}{I} \log p_\theta(e_1^I | f_1^J) \\
&\approx -\sum_{(f_1^J, e_1^I)} \hat{p}(f_1^J, e_1^I) \cdot \frac{1}{I} \log p_\theta(e_1^I | f_1^J)
\end{aligned}
$$

- Empirical distribution $\hat{p}(f_1^J, e_1^I)$

Back-translation scenario

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

- Known target distribution $\hat{p}(e_1^I)$
- How to approximate $Pr(f_1^J | e_1^I)$?

General Formulation of Training with Back-Translation

$$
\begin{aligned}
L(\theta) &= -\sum_{e_1^I} Pr(e_1^I) \cdot \frac{1}{I} \sum_{f_1^J} Pr(f_1^J|e_1^I) \cdot \log p_\theta(e_1^I|f_1^J) \\
&\approx -\sum_{e_1^I} \hat{p}(e_1^I) \cdot \frac{1}{I} \sum_{f_1^J} q(f_1^J|e_1^I; p_\Omega) \cdot \log p_\theta(e_1^I|f_1^J)
\end{aligned}
$$

Synthetic data generation procedure $q(f_1^J|e_1^I; p_\Omega)$

- **Uses a target-to-source translation model $p_\Omega(f_1^J|e_1^I)$**
- **Can be designed to correct deficiencies of p_Ω**
- **Intractable to enumerate all possible sentences ($\sum_{f_1^J}$)**

Desired properties

- **Approximates $Pr(f_1^J|e_1^I)$ well**
- **High weights to representative hypotheses ("sample efficiency")**
  - **Due to restricted sample size, often just one**

Why is beam search inappropriate?

$$q_{\text{beam}}(f_1^J|e_1^I;p_\Omega) = \begin{cases} 1, & f_1^J = \underset{\hat{J},\hat{f}_1^{\hat{J}}}{\text{argmax}}\left\{\frac{1}{\hat{J}}\log p_\Omega(\hat{f}_1^{\hat{J}}|e_1^I)\right\} \\ 0, & \text{otherwise} \end{cases}$$

Consider the scenario of word translation when synonyms are available:

Natural data		$\rightarrow$	Synthetic data	
Pr(hound | Hund)	= 49%		Pr(hound | Hund)	= 0%
Pr(dog | Hund)	= 51%		Pr(dog | Hund)	= 100%

- Every occurrence of "Hund" will be translated to "dog"
- Beam search collapses to the most likely translation option

Consequences:

- Biases the distribution of words in the synthetic corpus
- Results in oversimplified corpora [Burlot & Yvon 18]

Sampling from Target-to-source Model

Unrestricted sampling [Edunov & Ott^{+} 18, Imamura & Fujita^{+} 18]

$$q_{\text{sample}}(f_1^J|e_1^I; p_\Omega) = p_\Omega(f_1^J|e_1^I)$$

- **Does not enforce a bias, based on the choice of q**
- **Relies completely on a good fitting $p_\Omega(f_1^J|e_1^I)$**

In practice...

- **NMT models smear probability mass to low quality hypotheses [Ott & Auli^{+} 18]**
  - **Hurts sample efficiency**
- **Label smoothing increases the probability of low quality hypotheses**

$$L(\theta) = -\frac{1}{J}\sum_{j=1}^{J}\sum_{f\in V}\left[\alpha\cdot\frac{1}{|V|} + (1-\alpha)\delta_{f,f_j}\right]\cdot\log p_\theta(f|e_1^I, f_1^{j-1})$$

  - **Larger variability: good for regularization, bad when sampling from it**

The Middle-ground: Restricting the Search Space

Only consider high probability hypotheses:

- **Thresholded sampling**
  - **Sample from $p_\Omega(f|e_1^I, f_1^{j-1})$ only if probability is over $\tau \in (0, 1)$**
  - **Marginal overhead on top of standard sampling**
- **N-best list sampling**
  - **Sample a sentence from N-best list according to the model scores**
  - **Computational resources grow linearly w.r.t. N**
- **Top-k sampling [Edunov & Ott$^+$ 18]:**
  - **Still allows low probability sentences to be sampled**

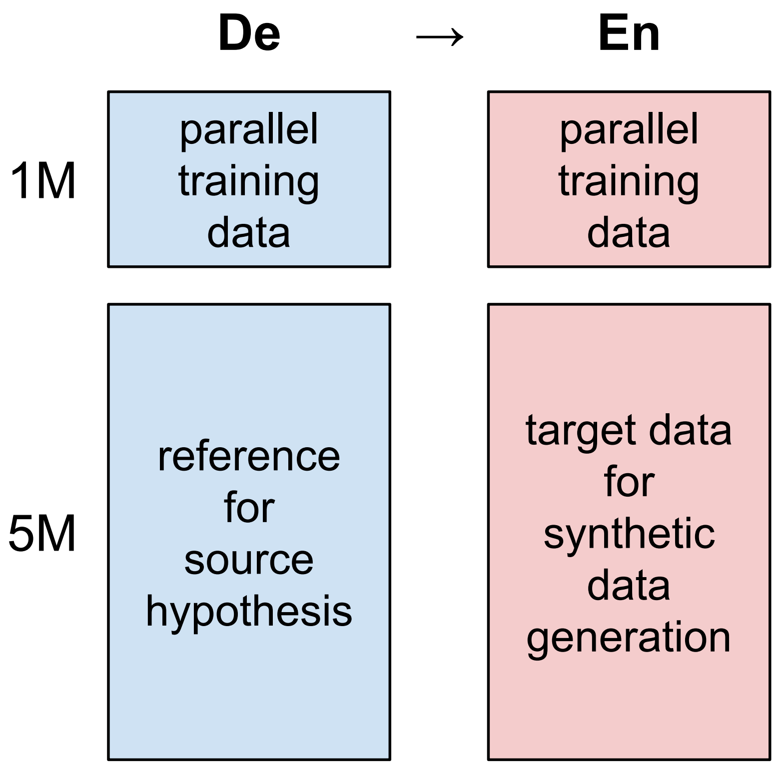
Benefits:

- **Dodge low probability hypotheses $\rightarrow$ Higher sample efficiency**
- **Still profit from the variability of sampling**

Experimental Setup: Controlled Scenario

WMT 2018 German $\leftrightarrow$ English news translation task

- **Original parallel training data: around 6M sentence pairs**



Controlled scenario

- **Parallel training data: subsample 1M sentence pairs from the original parallel data**
- **Remaining 5M sentence pairs**
  - **Target: use as monolingual data for synthetic data generation**
  - **Source: reference for the generated hypothesis**
    - **Upper bound for synthetic data quality**

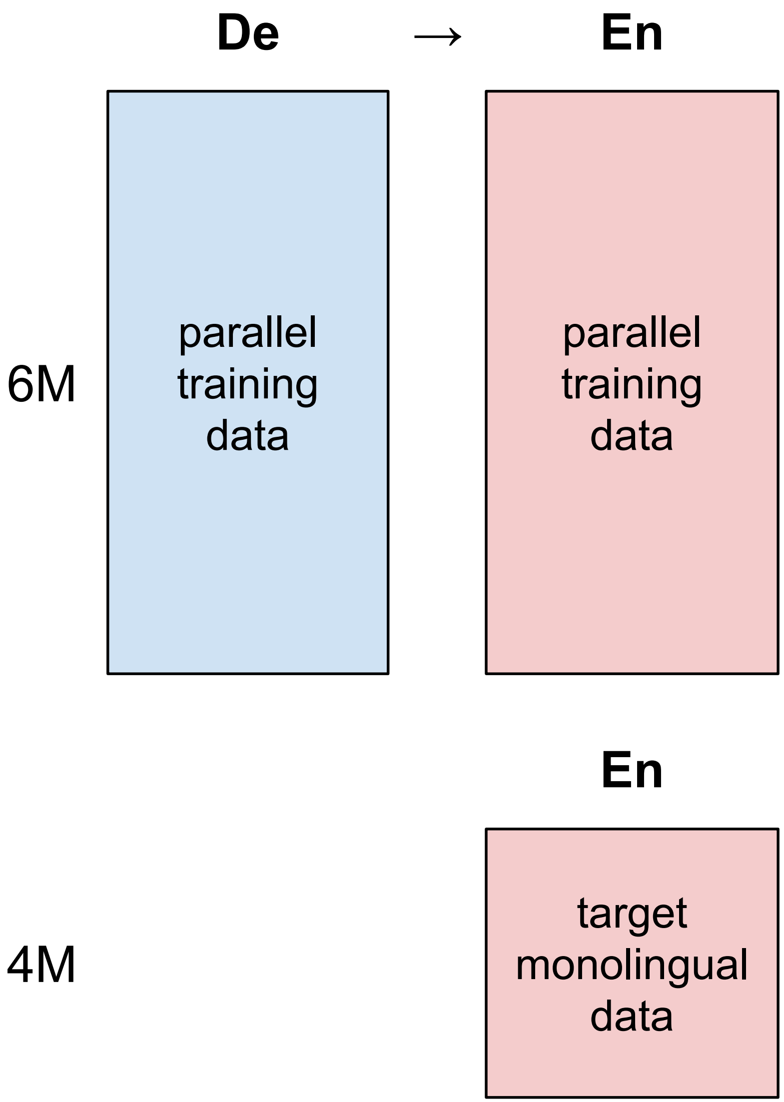
Results: Controlled Scenario

- Entropy of IBM-1 lexicon model: variability of word-by-word translations

	Entropy	PPL		BLEU [%]	
Source hypothesis	En $\rightarrow$ De	Train	test2015	test2015	test2017
Beam search ($b = 5$)	2.60	2.74	5.77	30.9	31.9
Unrestricted sampling	3.13	9.07	5.55	30.4	31.0
+ without label smoothing	2.93	5.17	5.31	30.4	31.3
Thresholded sampling ($\tau = 0.1$)	2.66	3.34	5.61	31.1	32.1
N-best list sampling ($N = 50$)	2.62	2.84	5.70	31.1	31.9
Reference	2.91	5.18	4.50	32.6	33.5

- Unrestricted sampling lags behind beam search considerably
- Statistics for the data are well matched for sampling without label smoothing
- Restricted search space makes the sampling more effective
- Clear inconsistency between PPL and BLEU

Experimental Setup: Real-world Scenario



Real-world scenario

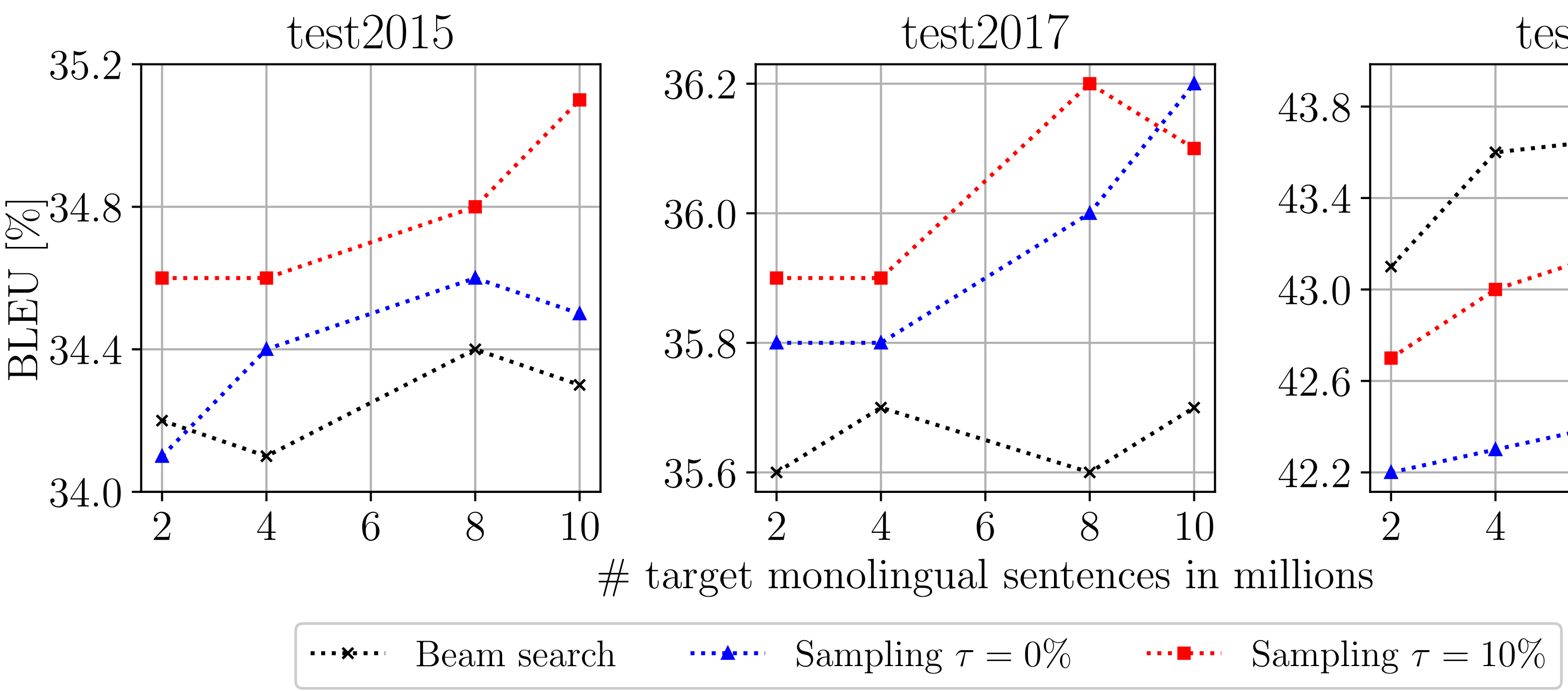
- **Parallel training data: full original parallel data (6M sentence pairs)**
- **Additional monolingual data: 4M sentences from NewsCrawl 2017**

Results: Real-world Scenario

Source hypothesis	De $\rightarrow$ En (Bleu[%]) test2017	De $\rightarrow$ En (Bleu[%]) test2018	En $\rightarrow$ De (Bleu[%]) test2017	En $\rightarrow$ De (Bleu[%]) test2018
Baseline	33.4	39.5	26.9	39.4
Beam search ($b = 5$)	35.7	43.6	28.2	41.3
Unrestricted sampling	35.8	42.3	28.6	41.5
+ without label smoothing	35.9	42.5	29.1	41.7
Thresholded sampling ($\tau = 0.1$)	35.9	43.0	28.7	41.6
N-best list sampling ($N = 50$)	36.0	43.6	28.6	41.8

- Unrestricted sampling: large drop in performance on De$\rightarrow$En test2018
  - Consistent improvements by removing label smoothing
- N-best list sampling: best in 3 of 4 test sets

Scalability of Sampling Methods



- **Beam search: not scalable except test2018**
- **Unrestricted sampling: only scales in test 2017**
- **Thresholded sampling: always scales**

Conclusion

Generalizing back-translation

- **Synthetic data generation is not the same as decoding/inference!**
- **Main goal: match the true translation probability $Pr(f_1^J|e_1^I)$**
  - **Approximated by sampling from a target-to-source model: $q(f_1^J|e_1^I;p_\Omega)$**

What can we do (consistently) better?

- **No label smoothing in training the target-to-source model**
- **Sample instead of beam search: better & faster!**
  - **Restrict the search space of the sampling**

Thank you for your attention

Yunsu Kim

kim@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

References

[Burlot & Yvon 18] F. Burlot, F. Yvon: Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pp. 144–155, 2018. 5

[Edunov & Ott+ 18] S. Edunov, M. Ott, M. Auli, D. Grangier: Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018. Version 2. 2, 6, 7

[Imamura & Fujita+ 18] K. Imamura, A. Fujita, E. Sumita: 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)*, pp. 55–63, 2018. 2, 6

[Ott & Auli+ 18] M. Ott, M. Auli, D. Granger, M. Ranzato: Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*, 2018. Version 4. 6

[Sennrich & Haddow+ 16] R. Sennrich, B. Haddow, A. Birch: Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 86–96, 2016. 2

[Zhang & Liu+ 18] Z. Zhang, S. Liu, M. Li, M. Zhou, E. Chen: Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

Back-translation generation: English → German samples

Source: it is seen as a long saga full of surprises.

Reference: er wird als eine lange Saga voller Überraschungen angesehen.

Beam search: es wird als eine lange Geschichte voller Überraschungen angesehen.

Sampling: es wird als eine lange Saga voller Überraschungen angesehen. injury, Skepsis, Feuer), Duschen verursachter Körper ...

Sampling w/o LS: es wurde als eine lange Geschichte voller Überraschungen gesehen.

Restricted sampling: es wird als lange Sage voller Überraschungen angesehen.

50-best sampling: es wird als eine lange Sage voller Überraschungen gesehen.

Back-translation generation: English → German samples

Source: in our opinion, this should also be the motto of a hotel.

Reference: wir meinen, dass dieser Spruch auch in einem Hotel gelten sollte.

Beam search: das sollte unserer Meinung nach auch das Motto eines Hotels sein.

Sampling: das sollte auch meiner Ansicht nach ein vorzüglicher Wunsch boote Tragfähigkeit.

Restricted sampling: das sollte auch unserer Ansicht nach das Motto eines Hotels sein.

Back-translation generation: English → German samples

Source: something else that needs to be improved in future is the House's internal democracy.

Reference: noch etwas, das sich in Zukunft verbessern ließe, ist die Demokratie im Innern dieses Hauses.

Beam search: ein weiterer Punkt, der in Zukunft verbessert werden muss, ist die innere Demokratie des Parlaments.

Sampling: ein weiteres, künftig verbessertes Element ist die integrierte Demokratie des Europäischen ganzer Aufbauwerks.

Restricted sampling: eine weitere Verbesserung muss künftig in der internen Demokratie des Parlaments bestehen.

Translation model hyperparameters

Training parameters:

- Glorot initialization
- Maximum sequence length: 100
- Learning rate: $3 \cdot 10^{-4}$
- Decay learning rate by 30% after every 3 checkpoints without improvement
- Gradient clipping whenever value is over 1

Model parameters:

- 6 layer Transformer model and word embedding size: 512
- Attention heads: 8
- Feed-forward projection dimension: 2048
- Dropout throughout architecture: 10%
- Label smoothing 0.1
- Tied source and target embeddings and output layer

Translation model update strategies

German $\rightarrow$ English controlled scenario (word batch size 16k):

- **All translation models: to convergence**

German $\leftrightarrow$ English (word batch size 4k):

- **Back-translation model: 1M updates**
- **Translation model:**
  - **without synthetic data: 1M updates**
  - **with synthetic data: fine-tune model without synthetic data for 1M updates**

Sampling measures: Word-by-word sampling

Sample a word f_j from $p_\Omega(\cdot|f_1^{j-1}, e_1^I)$ until sentence end is reached or $J = 2 \cdot I$:

- $q(f_1^J|e_1^I; p_\Omega) = p_\Omega(f_1^J|e_1^I)$

Restricted sampling:

$$q(f|e_1^I, f_1^{j-1}; p_\Omega) = \begin{cases} \text{softmax}\big(p_\Omega(f|e_1^I, f_1^{j-1}), C\big), & |C| > 0 \\ 1, & |C| = 0 \wedge \\ & f = \operatorname*{argmax}_{f'}\Big\{p_\Omega(f'|e_1^I, f_1^{j-1})\Big\} \\ 0, & \text{otherwise} \end{cases}$$

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

$$C = \{f \mid p_\Omega(f|e_1^I, f_1^{j-1}) \geq \tau\}$$

Sampling measures: N-best list sampling

Sample from N-best list weighted by hypothesis score:

- **score:** $s(f_1^J|e_1^I) = \frac{1}{J^\alpha} \log p_\Omega(f_1^J|e_1^I)$
- **assign 0 probability to the non-N best candidates**

Sentence probability:

$$q(f_1^J|e_1^I;p_\Omega) = \begin{cases} \textbf{softmax}(s(f_1^J|e_1^I), C), & f_1^J \in C \\ 0, & \textbf{otherwise} \end{cases}$$

with $C \subseteq \mathbb{D}_{src}$ being the set of N-best translations:

$$C = \underset{\mathcal{D}\subset\mathbb{D}_{src}:|\mathcal{D}|=N}{\operatorname{argmax}} \left\{ \sum_{f_1^J \in \mathcal{D}} s(f_1^J|e_1^I) \right\}$$