# When and Why is the Unsupervised Neural Machine Translation Useless?

**Yunsu Kim**
RWTH Aachen University, Aachen, Germany
kim@cs.rwth-aachen.de

EAMT 2020
November 3rd, 2020

# Unsupervised Machine Translation

Many recent works in unsupervised machine translation:

[Artetxe & Labaka+ 18b] [Lample & Denoyer+ 18] [Yang & Chen+ 18] [Lample & Ott+ 18]
[Kim & Geng+ 18] [Artetxe & Labaka+ 18a] [Ren & Zhang+ 19] [Artetxe & Labaka+ 19]
[Sun & Wang+ 19] [Conneau & Lample 19] [Pourdamghani & Aldarrab+ 19] [Song & Tan+ 19]
[Sen & Gupta+ 19] [Liu & Gu+ 20]

⋮

- Tested mostly on a **high-resource** language pair
  - German↔English, French↔English, ...
  - Linguistically similar source-target: already lots of bilingual corpora
- They **do not need unsupervised learning** in practice

**Question:** Is it useful also in low-resource, linguistically different language pairs?

# Our Experiments

| | | de-en | | ru-en | | zh-en | | kk-en | | gu-en | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | German | English | Russian | English | Chinese | English | Kazakh | English | Gujarati | English |
| Language family | | Germanic | Germanic | Slavic | Germanic | Sinitic | Germanic | Turkic | Germanic | Indic | Germanic |
| Alphabet Size | | 60 | 52 | 66 | 52 | 8,105 | 52 | 42 | 52 | 91 | 52 |
| Monolingual | Sentences | 100M | | 71.6M | | 30.8M | | 18.5M | | 4.1M | |
| | Words | 1.8B | 2.3B | 1.1B | 2.0B | 1.4B | 699M | 278.5M | 421.5M | 121.5M | 93.8M |
| Bilingual | Sentences | 5.9M | | 25.4M | | 18.9M | | 222k | | 156k | |
| | Words | 137.4M | 144.9M | 618.6M | 790M | 440.3M | 482.9M | 1.6M | 1.9M | 2.3M | 1.5M |

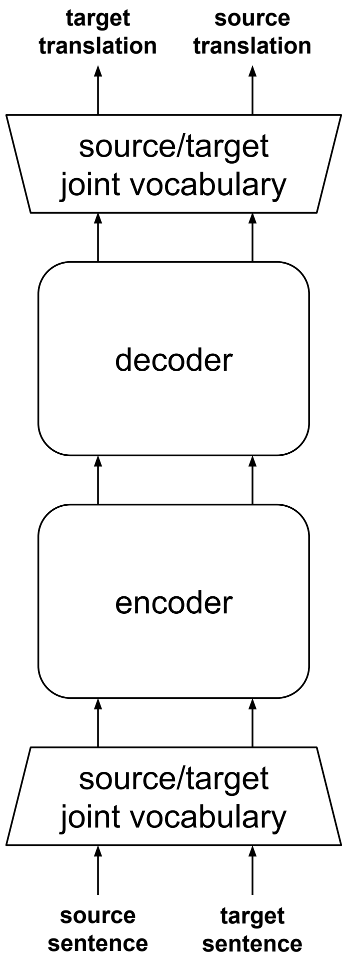
- Linguistically distant pairs: **ru-en**, **zh-en**, **kk-en**, **gu-en**
- Low-resource (bilingual data): **kk-en**, **gu-en**

Method: **XLM** [Conneau & Lample 19]

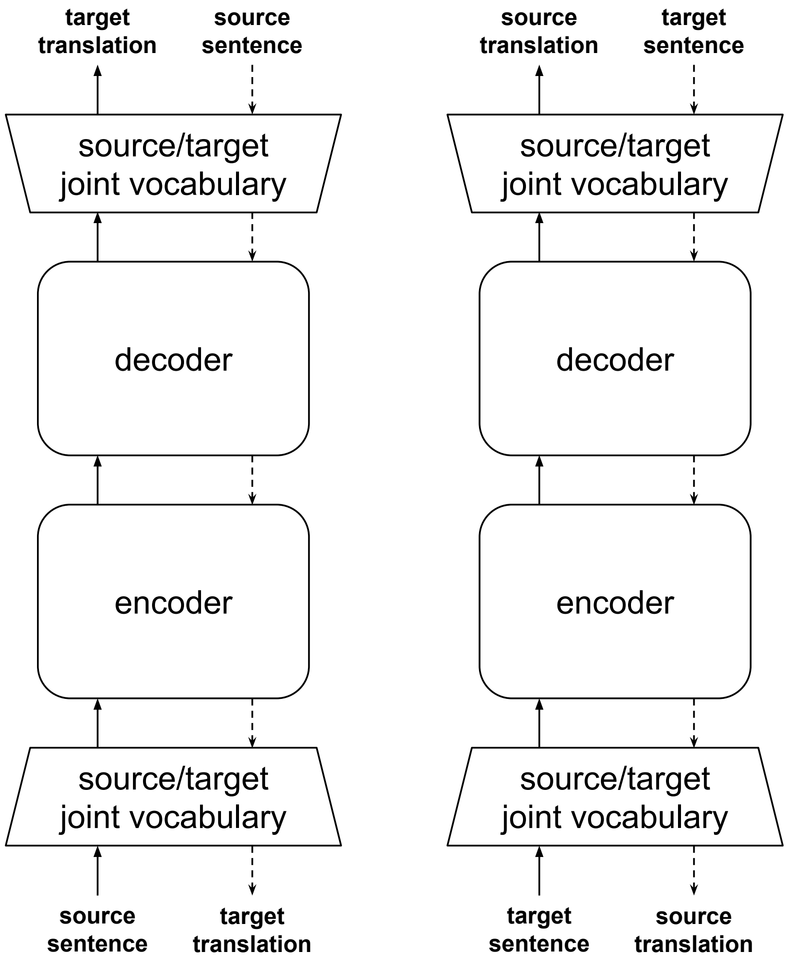
- Model: Transformer base
- Training: iterative back-translation + denoising autoencoder
- Initialization: cross-lingual masked LM

# Unsupervised NMT

## Bidirectional model



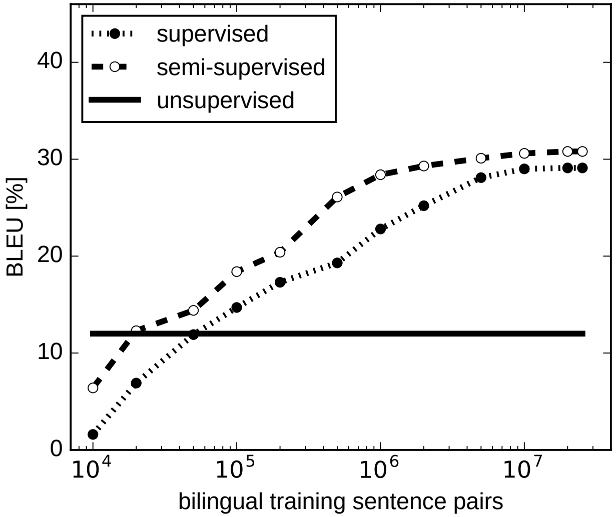

## Iterative back-translation



# Unsupervised vs. Supervised vs. Semi-supervised

| Approach | BLEU [%] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **de-en** | **en-de** | **ru-en** | **en-ru** | **zh-en** | **en-zh** | **kk-en** | **en-kk** | **gu-en** | **en-gu** |
| Supervised | 39.5 | 39.1 | 29.1 | 24.7 | 26.2 | 39.6 | 10.3 | 2.4 | 9.9 | 3.5 |
| Semi-supervised | 43.6 | 41.0 | 31.4 | 31.3 | 25.9 | 42.7 | 12.5 | 3.1 | 14.2 | 4.0 |
| Unsupervised | 23.8 | 20.2 | 12.0 | 9.4 | 1.5 | 2.5 | 2.0 | 0.8 | 0.6 | 0.6 |

- Unsupervised: much worse than (semi-)supervised

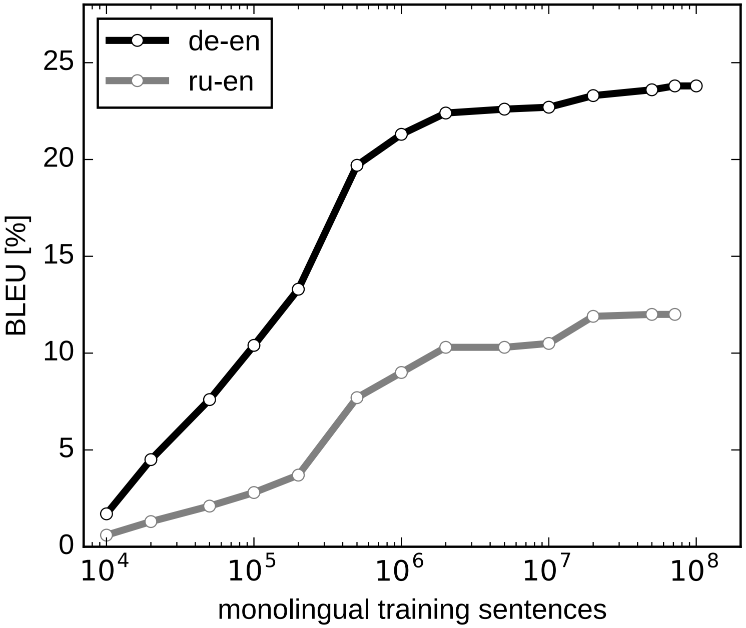



**ru-en**: When is the unsupervised learning useful?

- Only if bilingual data has less than **20k** lines

# Performance Factor: Training Data Size

How much monolingual data is needed for unsupervised NMT to work?




\# training sentences - performance

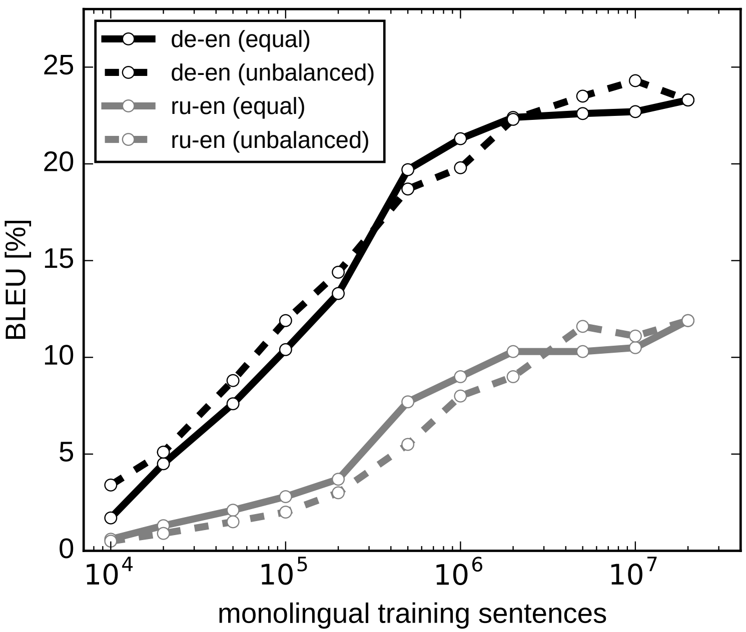
- 1M: already close to the best result
- 5M: starts to saturate
- 20M: no further improvment

**Important:** similarity of source/target languages (de-en > ru-en)

- Massive training data is not important

# Performance Factor: Unbalanced Training Data

What if the data size is largely different for source and target languages?




Source: varying (x-axis)
Target: fixed (20M sents)

- Oversizing one side has no effect
- Performance decided by the smaller side

**Important:** similar data distribution on source/target

# Performance Factor: Domain Similarity

| Domain (en) | Domain (de/ru) | BLEU [%] de-en | en-de | ru-en | en-ru |
|---|---|---|---|---|---|
| Newswire | Newswire | 23.3 | 19.9 | 11.9 | 9.3 |
| | Politics | 11.5 | 12.2 | 2.3 | 2.5 |
| | Random | 18.4 | 16.4 | 6.9 | 6.1 |

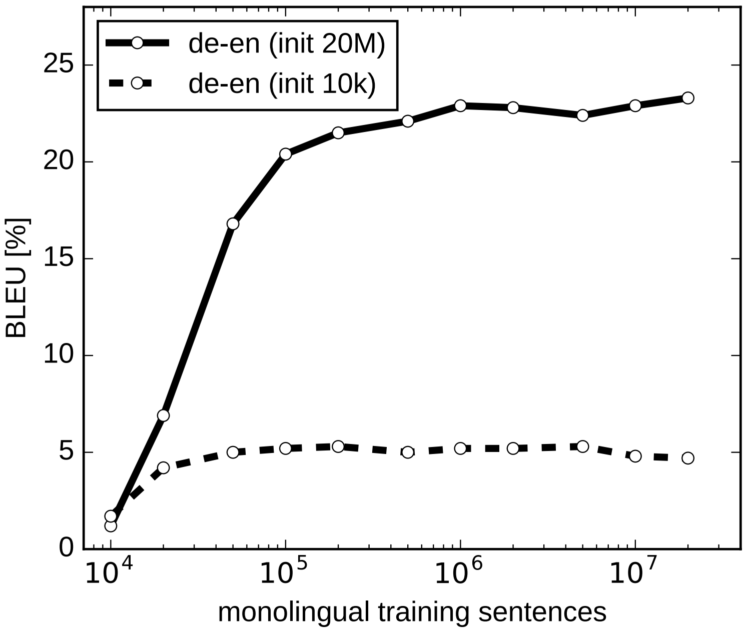
- Degenerates if domains do not match!

| Years (en) | Years (zh) | #sents (en/zh) | BLEU [%] zh-en | en-zh |
|---|---|---|---|---|
| 2014-2017 | 2008-2018 | 1.7M | 5.4 | 15.1 |
| | 1995-2008 | 28.6M | 1.5 | 1.9 |

- Degenerates if topics, styles, periods do not match!

# Performance Factor: Initialization

Initialization vs. Back-translation: Which is more important?




Good initialization

- $>$ 20% BLEU already with 1M sentences in back-translation training

Bad initialization

- $<$ 5% BLEU even with 100M sentences in back-translation training

**Important:** decent initialization

- Back-translation training relies on the quality of the initial model

# Why fail?

| Task | Source input | System output | Reference output |
|---|---|---|---|
| **zh-en** | ... 调整要兼顾生产需要和消费需求。 | ... 调整要兼顾生产需要 and 消费需求. | ... adjustment must balance production needs with consumer demands. |

- Input copying (wrong language)
- Reason: trained on copied back-translations

| Task | Source input | System output | Reference output |
|---|---|---|---|
| **de-en** | *München* 1856: *Vier* Karten, die Ihren Blick auf die *Stadt* verändern | *Austrailia* 1856: *Eight* things that can keep your way to the *UK* | Munich 1856: Four maps that will change your view of the city |

- *Vier* (*Four* in English) $\rightarrow$ *Eight*
- Cannot distinguish words that appear in the same context (1856, things)

# Remarks

Unsupervised NMT fails when...

1. source and target languages are linguistically dissimilar
2. source and target monolingual data are from different domains

These conditions are **very common** in low-resource language pairs!

- In practice: if you have ~**50k** bilingual sentence pairs, just do semi-supervised

You can also find in our paper:

- Why does the copied back-translations occur in training?

**When and Why is the Unsupervised Neural Machine Translation Useless?**

Yunsu Kim, Miguel Graça, Hermann Ney

`https://arxiv.org/abs/2004.10581`

# References

[Artetxe & Labaka+ 18a] M. Artetxe, G. Labaka, E. Agirre.
Unsupervised statistical machine translation.
In EMNLP, 3632–3642, 2018.

[Artetxe & Labaka+ 18b] M. Artetxe, G. Labaka, E. Agirre, K. Cho.
Unsupervised neural machine translation.
In ICLR, 2018.

[Artetxe & Labaka+ 19] M. Artetxe, G. Labaka, E. Agirre.
An effective approach to unsupervised machine translation.
In ACL, pp. 194–203, 2019.

[Conneau & Lample 19] A. Conneau, G. Lample.
Cross-lingual language model pretraining.
In NeurIPS, pp. 7057–7067, 2019.

[Kim & Geng+ 18] Y. Kim, J. Geng, H. Ney.
Improving unsupervised word-by-word translation with language model and denoising autoencoder.
In EMNLP, pp. 862–868, 2018.

[Lample & Denoyer+ 18] G. Lample, L. Denoyer, M. Ranzato.
Unsupervised machine translation using monolingual corpora only.
In ICLR, 2018.

[Lample & Ott+ 18] G. Lample, M. Ott, A. Conneau, L. Denoyer, M. Ranzato.
Phrase-based & neural unsupervised machine translation.
In EMNLP, pp. 5039–5049, 2018.

[Liu & Gu+ 20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer.
Multilingual denoising pre-training for neural machine translation, 2020.

# References

[Pourdamghani & Aldarrab+ 19] N. Pourdamghani, N. Aldarrab, M. Ghazvininejad, K. Knight, J. May.
Translating translationese: A two-step approach to unsupervised machine translation.
In ACL, pp. 3057–3062, 2019.

[Ren & Zhang+ 19] S. Ren, Z. Zhang, S. Liu, M. Zhou, S. Ma.
Unsupervised neural machine translation with smt as posterior regularization, 2019.

[Sen & Gupta+ 19] S. Sen, K. K. Gupta, A. Ekbal, P. Bhattacharyya.
Multilingual unsupervised NMT using shared encoder and language-specific decoders.
In ACL, pp. 3083–3089, 2019.

[Song & Tan+ 19] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu.
Mass: Masked sequence to sequence pre-training for language generation.
In ICML, pp. 5926–5936, 2019.

[Sun & Wang+ 19] H. Sun, R. Wang, K. Chen, M. Utiyama, E. Sumita, T. Zhao.
Unsupervised bilingual word embedding agreement for unsupervised neural machine translation.
In ACL, pp. 1235–1245, 2019.

[Yang & Chen+ 18] Z. Yang, W. Chen, F. Wang, B. Xu.
Unsupervised neural machine translation with weight sharing.
In ACL, pp. 46–55, 2018.