Improving Unsupervised Word-by-Word Translation Using Language Model and Denoising Autoencoder

Yunsu Kim, Jiahui Geng, Hermann Ney

kim@cs.rwth-aachen.de

EMNLP 2018
November 2, 2018

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Unsupervised Machine Translation

Machine translation (MT) requires lots of parallel data

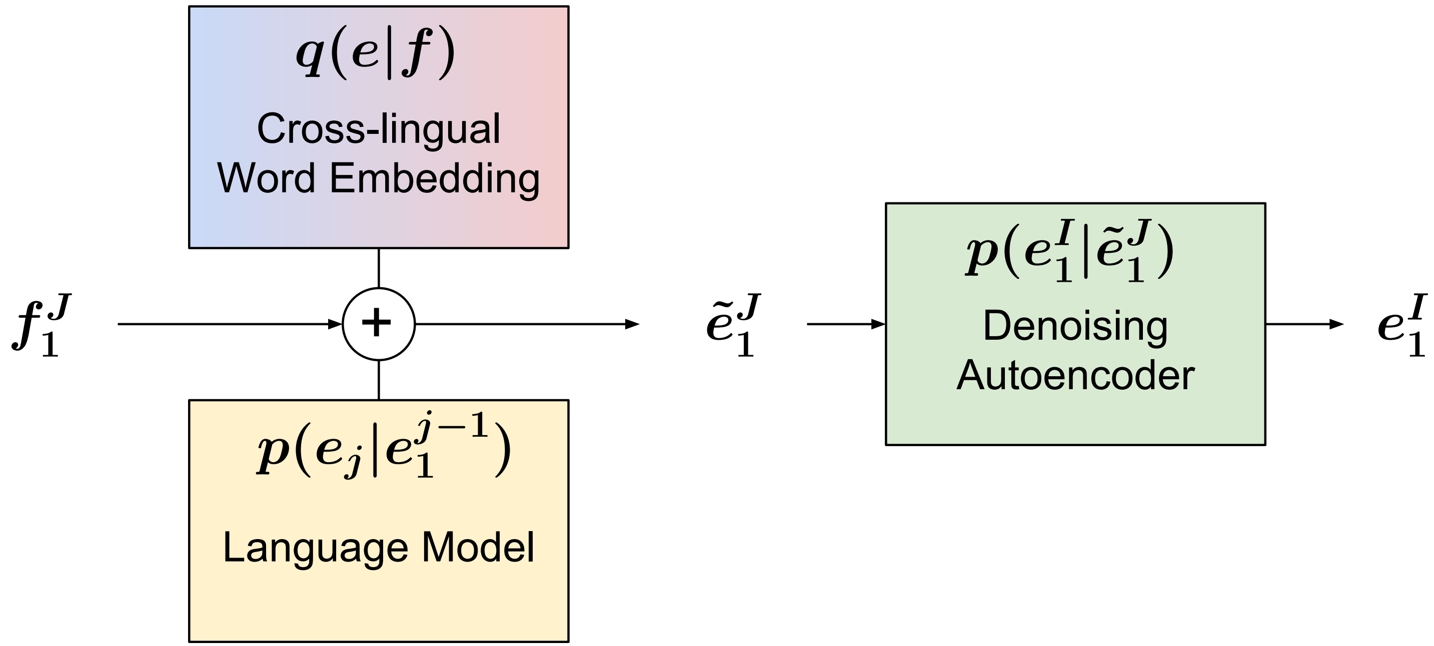
- **Especially for neural models [Koehn & Knowles 17]**
- **Small or no parallel data for many language pairs**

Unsupervised MT: Train only with monolingual data

- **[Artetxe & Labaka$^+$ 18], [Lample & Denoyer$^+$ 18]**
- **Iterative back-translation of both translation directions**
  - **Long training time (e.g. 1-3 weeks)**
- **Model shared for both translation directions but separate training data**
  - **Considerable effort to implement**

Can we build an unsupervised machine translation system quickly & simply?

Our Unsupervised MT System



Combine the ideas from

- **Classic word-based models**
- **Modern neural sequence-to-sequence model**

Minimal implementation & Quick training (1-2 days)

- **Outperforms [Artetxe & Labaka$^+$ 18], [Lample & Denoyer$^+$ 18]**

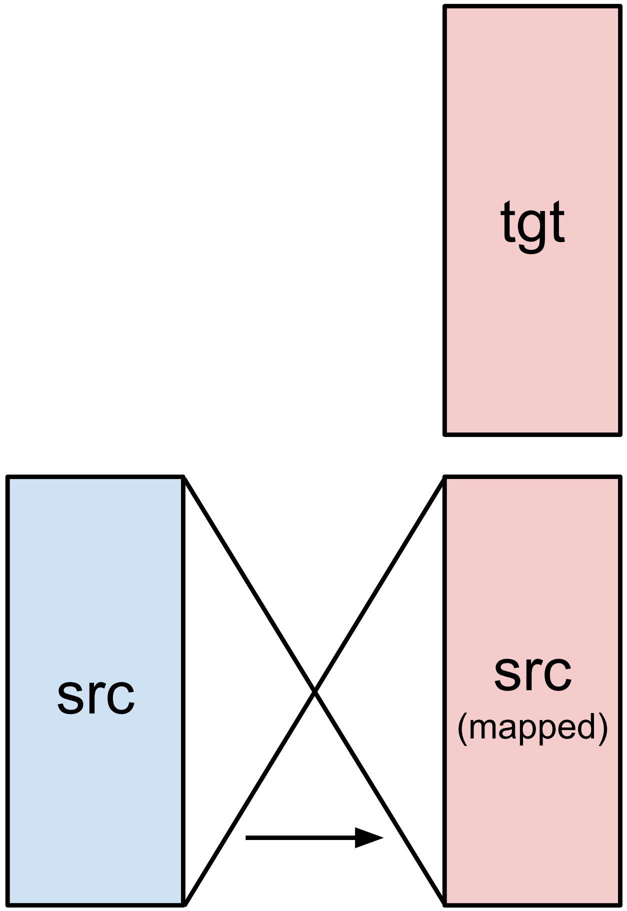
Word Lexicon: Cross-lingual Word Embedding

Monolingual word embedding

- Skip-gram, CBOW
- Individually learned for source and target

Cross-lingual word embedding

- Linear mapping: source $\rightarrow$ target
- Shared embedding space
- Arithmetic operations possible between source and target words



Word Lexicon: Cross-lingual Word Embedding

Unsupervised learning of cross-lingual mapping

1. Initialization: adversarial training [Conneau & Lample$^+$ 18]
2. Training: minimum squared error (MSE)

$$\hat{W} = \operatorname*{argmin}_{W} \left\{ \sum_{(f,e) \in D} \| W f^{\text{emb}} - e^{\text{emb}} \| \right\}$$

   - Dictionary D: mutual nearest neighbors
3. Repeat dictionary induction and MSE training [Artetxe & Labaka$^+$ 17]

Word translation = Nearest neighbor search

$$\hat{e}(f) = \operatorname*{argmin}_{e} \{ d(f, e) \}$$

- $d(f, e)$: cosine similarity with hub penalty [Conneau & Lample$^+$ 18]

Beam Search with Language Model

Word-by-word translation does not consider context

- **And most literature on cross-lingual word embedding evaluate only on word translations!**
- **Ignored so far: behavior of cross-lingual neighbor words within a context**

Beam search with language model (LM)

$$S(e; f, h) = \lambda_{\text{emb}} \log q(f, e) + \lambda_{\text{LM}} \log p(e|h)$$

- $q(f, e) \in [0, 1]$: **linearly scaled cosine similarity**
- e **=** k**-nearest neighbors**
- **Context-aware lexical choices**

Denoising Autoencoder

Cross-lingual word embedding + LM = $f_1^J \rightarrow \tilde{e}_1^J$

- Still one target word per source word
- Reordering is not considered

Denoising: noisy target sentence $\rightarrow$ clean target sentence

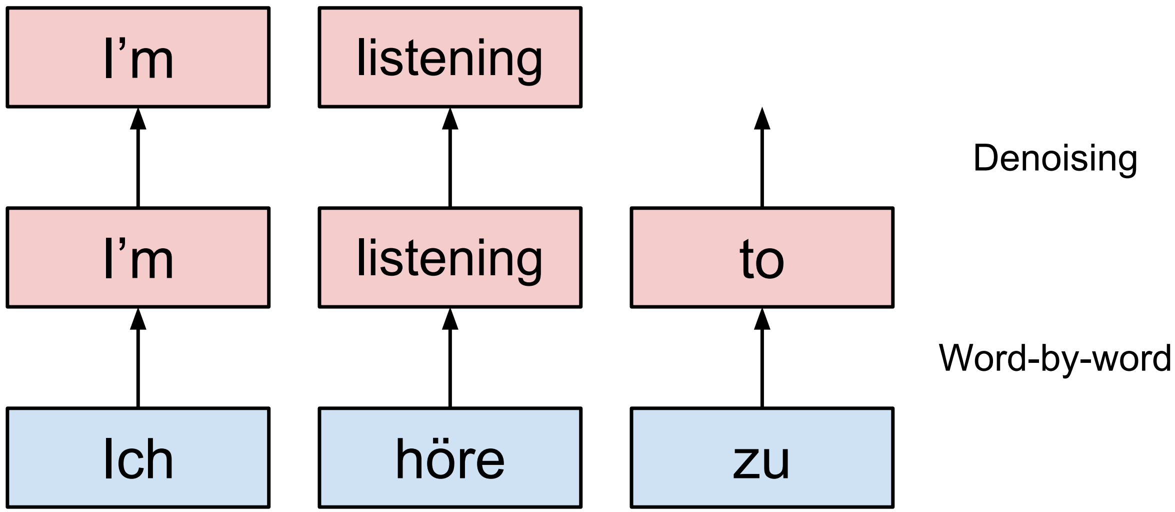
- Neural sequence-to-sequence autoencoder
- Can be trained only with target monolingual data

$$L(E) = -\sum_{e_1^I \in E} \log p(e_1^I \mid \textbf{noise}(e_1^I))$$

- Input noise(e_1^I): target sentence with artificial noise
  - Simulate errors in word-by-word translations
- Output e_1^I: target sentence (original)

Insertion Noise

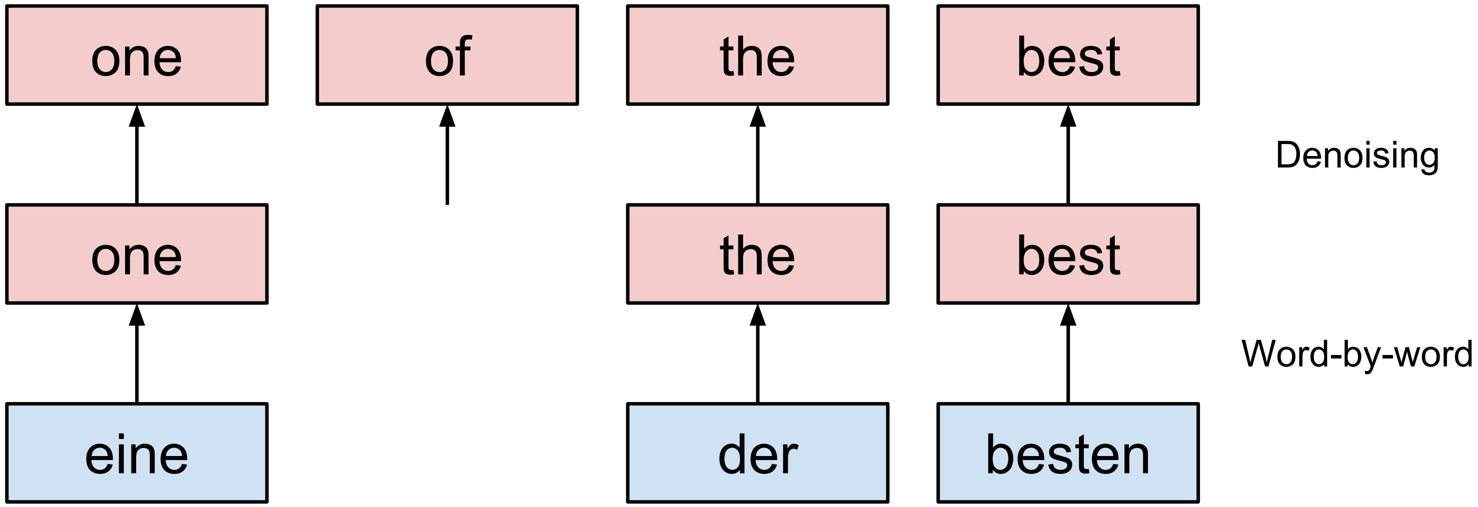
Case 1: multiple source words $\rightarrow$ a single target word



- **Insertion noise: insert a word between original words [This work]**
  - **Randomly with a probability p_{ins} at each position**
  - **Only V_{ins} frequent words are inserted, e.g. articles, prepositions**
- **Denoiser learns to delete such words**

Deletion Noise

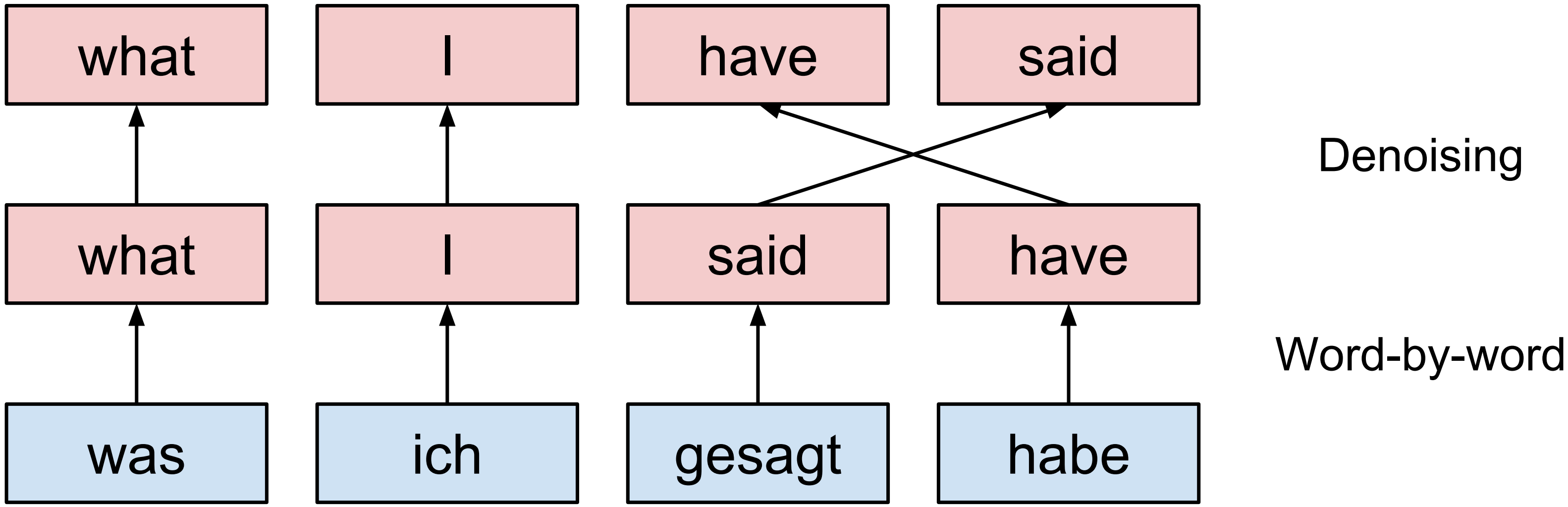
Case 2: a single source word $\rightarrow$ multiple target words



- **Deletion noise: delete words from the original sentence [Hill & Cho^{+} 16]**
  - **randomly with a probability p_{del} at each position**
- **Denoiser learns to insert such words**

Permutation Noise

Case 3: target hypothesis words should be reordered



- Permutation noise: permute original word positions [Hill & Cho^{+} 16]
  - randomly within a limited distance d_{per}: maintain general monotonicity
- Denoiser learns to reorder such words

Experimental Setup

Training data: WMT News Crawl monolingual data

- **English:** $100\mathrm{M}$ **sentences**
- **German:** $100\mathrm{M}$ **sentences**
- **French:** $42\mathrm{M}$ **sentences**

Test sets: WMT News translation task

- **German$\leftrightarrow$English:** **`newstest2016`**
- **French$\leftrightarrow$English:** **`newstest2014`**

Experimental Setup

Cross-lingual word embedding

- **Discriminator input and dictionary induction:** 100k **frequent words**

LM: 5-gram with modified Kneser-Ney smoothing

Denoising autoencoder: 6-layer Transformer encoder/decoder

- 50k **frequent words +** `<unk>`

Search parameters

- **Number of nearest neighbors (k) =** 100
- **Beam size =** 10
- $\lambda_{\text{emb}} = 1.0$, $\lambda_{\text{LM}} = 0.1$

Results

BLEU [%] scores on WMT tasks

	newstest2016		newstest2014	
System	de-en	en-de	fr-en	en-fr
Word-by-Word	11.1	6.7	10.6	7.8
+ LM	14.5	9.9	13.6	10.9
+ Denoising	17.2	11.0	16.5	13.9
[Lample & Denoyer^{+} 18]	13.3	9.6	14.3	15.1
[Artetxe & Labaka^{+} 18]	-	-	15.6	15.1

Conclusion

Fully unsupervised MT system with cross-lingual word embedding

- **Beam search with LM for context-aware lexicon choice**
- **Denoising autoencoder for insertion/deletion/local reordering**
- **Simple to implement and fast to train**
- **Outperforms unsupervised neural MT with iterative back-translations**

Future work

- **Our method to initialize unsupervised neural MT [Lample & Denoyer$^+$ 18, Artetxe & Labaka$^+$ 18]**
- **Artificial noises to regularize neural MT**

Codes available at `https://github.com/yunsukim86/wbw-lm/` $\Rightarrow$

Thank you for your attention

Yunsu Kim

kim@cs.rwth-aachen.de

http://www.hltpr.rwth-aachen.de/

Ablation Study: Denoising

- d_{per}: local reordering range / p_{del}: deletion probability / p_{ins}: insertion vocabulary size

d_{per}	p_{del}	V_{ins}	BLEU [%]
2			14.7
3			14.9
5			14.9
3	0.1		15.7
3	0.3		15.1
3	0.1	10	16.8
3	0.1	50	17.2
3	0.1	500	16.8
3	0.1	5000	16.5

Ablation Study: Vocabulary

	Vocabulary	BLEU [%]
	Merges	
BPE	20k	10.4
	50k	12.5
	100k	13.0
	Cross-lingual training	
Word	20k	14.4
	50k	14.4
	100k	14.5
	200k	14.4

- Word embedding performs better than BPE embedding
- Embedding trained on 20k similar to 200k $\Rightarrow$ Frequent words matter

References

[Artetxe & Labaka^{+} 17] M. Artetxe, G. Labaka, E. Agirre: Learning Bilingual Word Embeddings with (Almost) No Bilingual Data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vol. 1, pp. 451–462, 2017.

[Artetxe & Labaka^{+} 18] M. Artetxe, G. Labaka, E. Agirre, K. Cho: Unsupervised Neural Machine Translation. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*, 2018.

[Conneau & Lample^{+} 18] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou: Word Translation Without Parallel Data. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*, 2018.

[Hill & Cho^{+} 16] F. Hill, K. Cho, A. Korhonen: Learning Distributed Representations of Sentences from Unlabelled Data. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pp. 1367–1377, 2016.

[Koehn & Knowles 17] P. Koehn, R. Knowles: Six Challenges for Neural Machine Translation. In *Proceedings of the 1st ACL Workshop on Neural Machine*

Translation (WNMT 2017), pp. 28–39, 2017.

[Lample & Denoyer^{+} 18] G. Lample, L. Denoyer, M. Ranzato: Unsupervised Machine Translation Using Monolingual Corpora Only. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*, 2018.