Unsupervised Neural Machine Translation

Mikel Artetxe, Gorka Labaka & Eneko Agirre, Kyunghyun Cho

ICLR 2018

정보 및 지능시스템 연구실
이상헌
lawlee1@naver.com

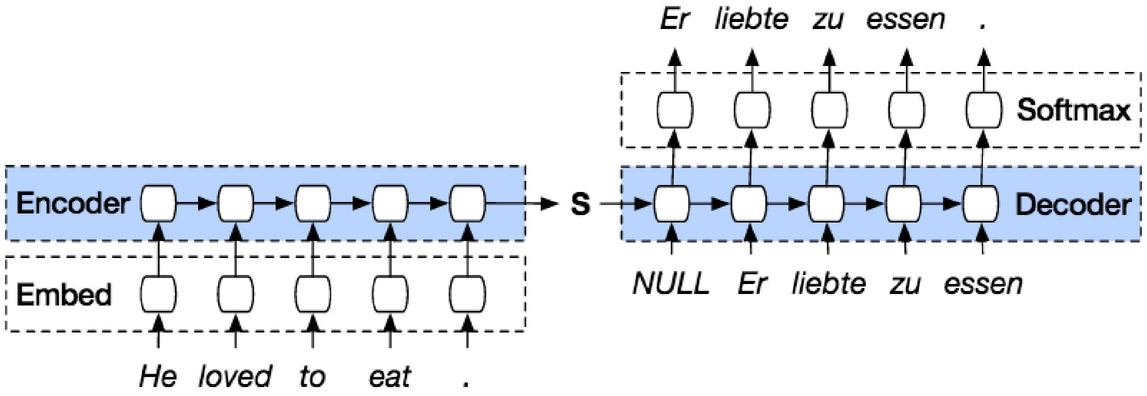
2018/11/15

Contents

- Introduction
- Related Works
- Proposed Method
- Experiments
- Conclusion

Introduction

- Neural Machine Translation
  - Sequence-to-sequence model



- Advantage
  - Better performance, trained end-to-end, less human effort
- Disadvantage (Limitation)
  - Requires a **large parallel corpus**, less interpretable

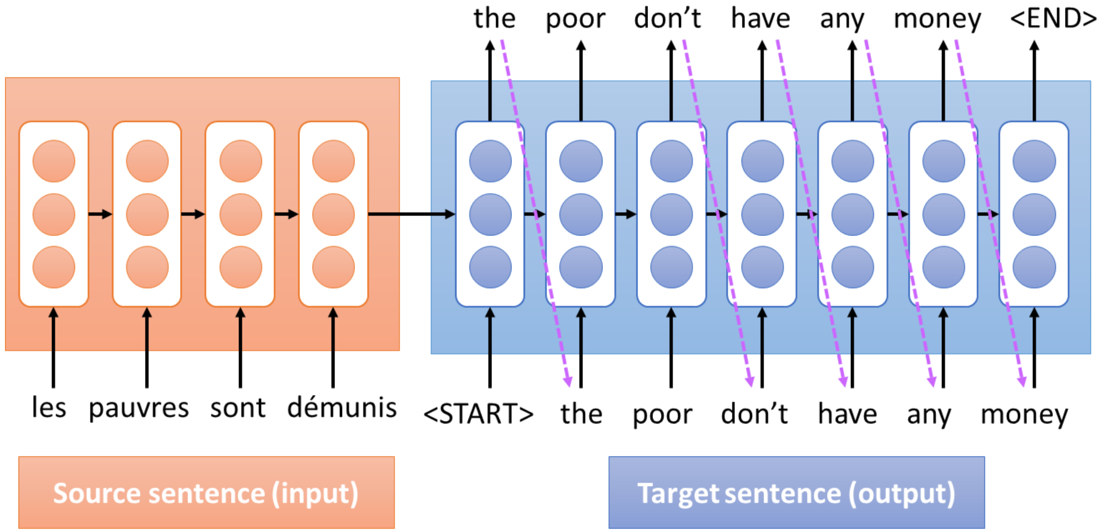
Introduction (Cont'd)

- Lack of large parallel corpora
  - Low-resources languages (e.g. Basque)
  - Many combinations of major languages (e.g. German-Russian)

- Propose NMT in a completely unsupervised manner
  - Relying solely on monolingual corpora
  - Unsupervised cross-lingual embeddings + denoising + backtranslation

Related Works

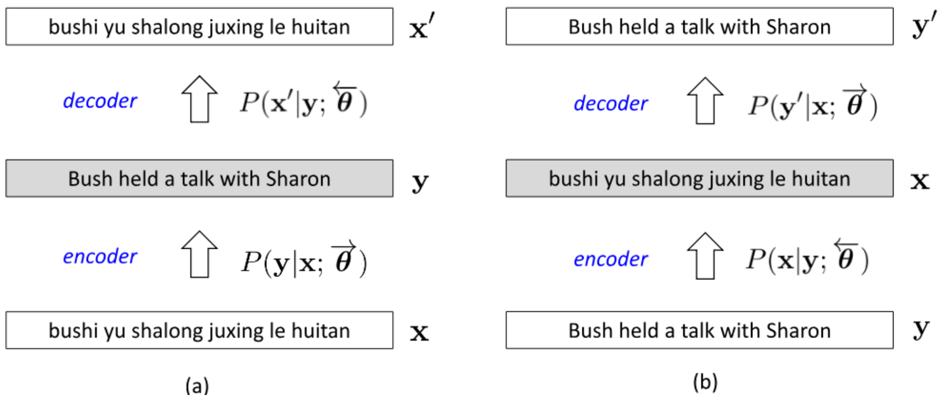
- Neural Machine Translation
  - Sequence-to-sequence model



  - Basically, two RNNs used as encoder and decoder
  - Encoder RNN produces encoding of the source sentence (embedding)
  - Decoder RNN generates target sentence condition on encoding

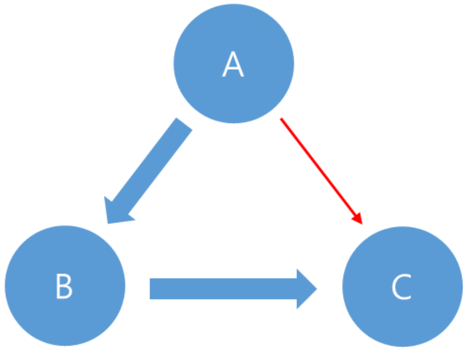
Related Works (Cont'd)

- Low-resource neural machine translation
  - Semi-supervised NMT (Sennrich et al. 2016)
    - monolingual corpora + scarce parallel corpora
    - Back-translate monolingual corpus in the target language
  - Triangulation techniques (Chen et al. 2017)
    - (A, C) little, but (A, B) and (B, C) plenty


bushi yu shalong juxing le huitan
$\mathbf{x}'$
decoder
$P(\mathbf{x}'|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$
Bush held a talk with Sharon
$\mathbf{y}$
encoder
$P(\mathbf{y}|\mathbf{x}; \overrightarrow{\boldsymbol{\theta}})$
bushi yu shalong juxing le huitan
$\mathbf{x}$
Bush held a talk with Sharon
$\mathbf{y}'$
decoder
$P(\mathbf{y}'|\mathbf{x}; \overrightarrow{\boldsymbol{\theta}})$
bushi yu shalong juxing le huitan
$\mathbf{x}$
encoder
$P(\mathbf{x}|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$
Bush held a talk with Sharon
$\mathbf{y}$

(a)

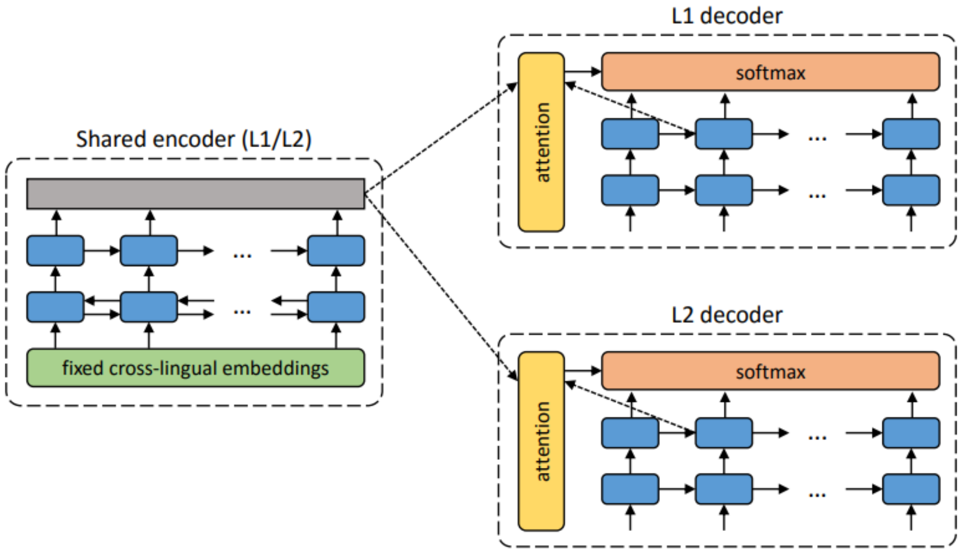
(b)



- Still require a strong cross-lingual signal

Proposed Method

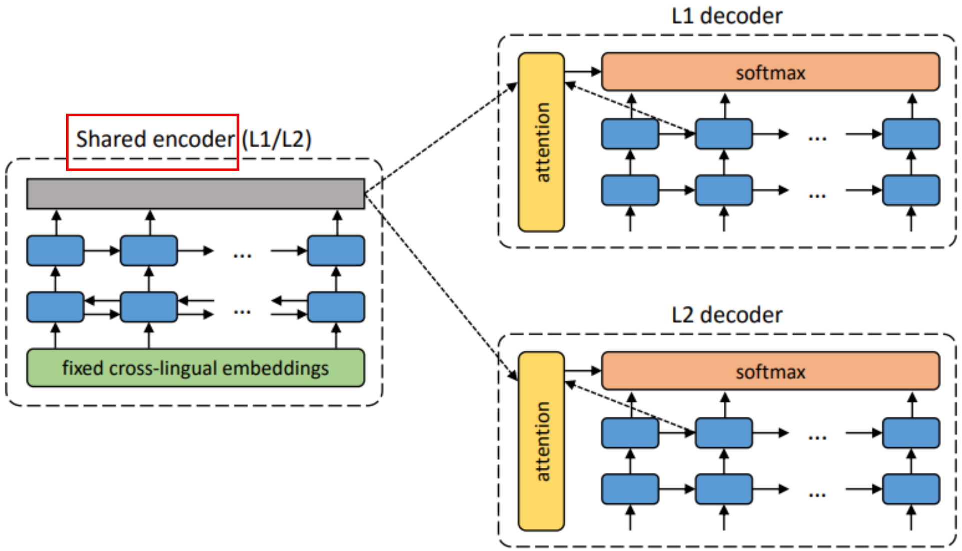
- Model architecture



- Standard encoder-decoder architecture with global attention mechanism
  - Encoder : two-layer bidirectional RNN
  - Decoder : two-layer RNN
  - GRU cells with 600 hidden units, dimension of embeddings = 300

Proposed Method (Cont'd)

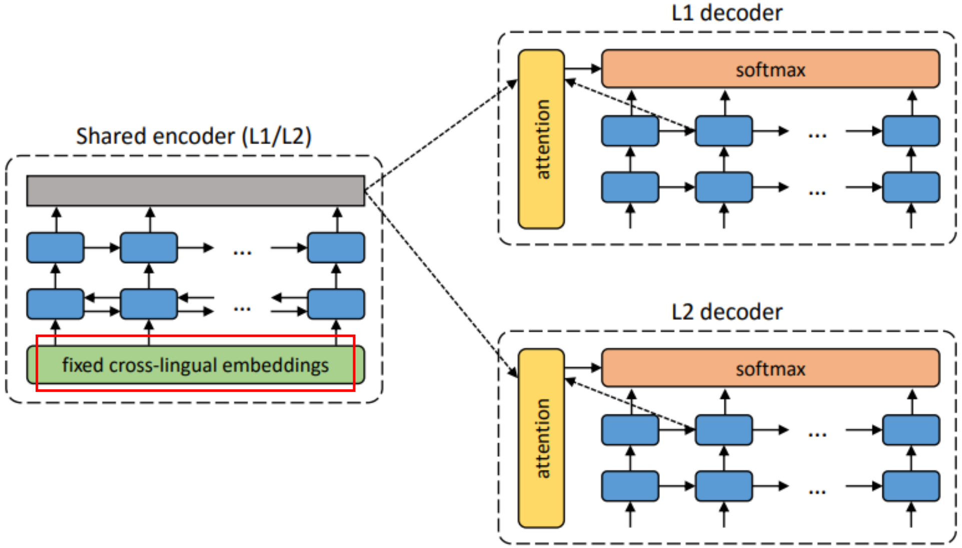
- Model architecture



1. Dual structure
   - Handle both direction (French ↔ English)
2. Shared encoder
   - Using only one encoder for all languages

Proposed Method (Cont'd)

- Model architecture

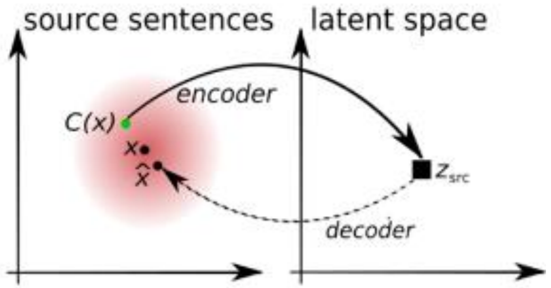


3. Fixed embeddings in the encoder
    - Pre-trained cross-lingual embeddings in the encoder and fixed during training
        - Encoder only needs to learn how to compose larger phrases
    - Cross-lingual embeddings trained with unsupervised way (Artetxe et al. 2017)
        - Linear transformation that mapping each embedding vector to shared space

Proposed Method (Cont'd)

- Unsupervised training

  1. Denoising



  - Idea of **denoising autoencoder** (**DAE**)
    - To avoid simply copying tasks
    - System is trained to reconstruct the original version of **corrupted sentence**
    - Noise : random swaps between contiguous words

  - Input sentence in source language (x)

    → Make random noise in input sentence (C(x))

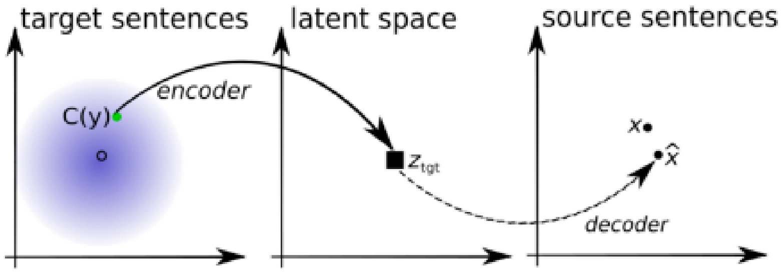
    → Encoding it with shared encoder

    → Reconstruct the original sentence with decoder

Proposed Method (Cont’d)

- Unsupervised training

  2. On-the-fly backtranslation



  - In denoising method, trained single language at each time
  - Adapt **backtranslation** approach proposed by Sennrich et al. (2016)

  - Input sentence in source language (x)

    → Generate target language sentence by greedy decoding (y’)

    → Encode y’

    → Generate source language sentence (x’)

Experiments

- Experiments
  - French-English and German-English dataset from WMT 2014 shared task
  - Three training options
    1. **Unsupervised**
       - Proposed model
       - News Crawl corpus (2007-2013)
    2. **Semi-supervised**
       - Proposed model with small in-domain parallel corpus
       - News Crawl monolingual + 10,000 or 100,000 News Commentary parallel corpus
    3. **Supervised**
       - Proposed model with supervised training
       - All WMT 2014 parallel corpora

  - Test dataset : newstest2014 using tokenized BLEU scores

Experiments (Cont’d)

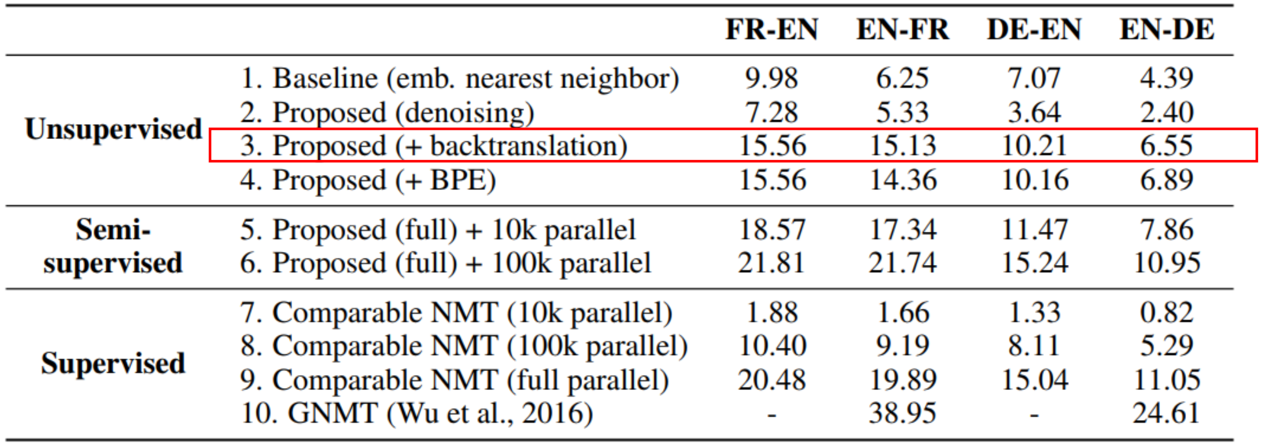
- Experiments
    - Corpus preprocessing using **Byte Pair Encoding** (**BPE**, Sennrich et al. 2016)
        - Most frequent character n-gram as one sub-word (token)
        - Effective way to overcome rare word problem in NMT
        - Vocabulary size : 50,000 tokens, replace others with <UNK>
    - Pre-trained cross-lingual embeddings
        - Train embeddings of each language using Word2Vec
            - Skip-gram model with 10 windows size, 300 dimensions
        - Map each embeddings to a shared space
    - Baseline model : word-by-word translate using nearest neighbor

Experiments (Cont'd)

- Experiments
  - Training
    - cross-entropy loss function, batch size = 50 sentences,
    - Adam optimizer, dropout prob = 0.3, 300,000 iterations
    - 4-5 days on single Titan X GPU : not fully converged
  - Test time inference was done using beam-search
    - beam size = 12

Experiments (Cont'd)

- Quantitative results



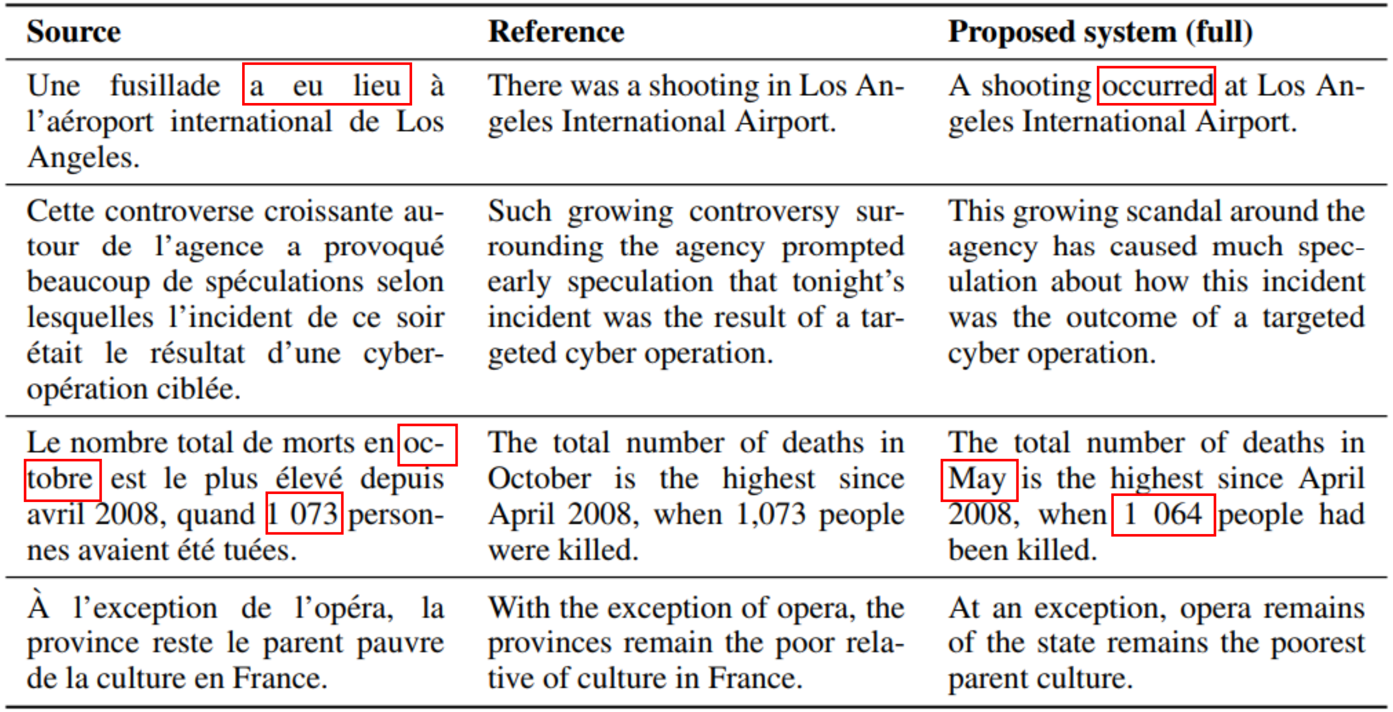
		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 10k parallel	18.57	17.34	11.47	7.86
	6. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	7. Comparable NMT (10k parallel)	1.88	1.66	1.33	0.82
	8. Comparable NMT (100k parallel)	10.40	9.19	8.11	5.29
	9. Comparable NMT (full parallel)	20.48	19.89	15.04	11.05
	10. GNMT (Wu et al., 2016)	-	38.95	-	24.61

- BPE
  - Rare word gets prefixed to a properly translated word
  - Not good for named entities and numerals

- Relatively poor result of Supervised comparable NMT (full parallel)
  - Could be improved in the future

Experiments (Cont'd)

- Qualitative results



Source	Reference	Proposed system (full)
Une fusillade a eu lieu à l'aéroport international de Los Angeles.	There was a shooting in Los Angeles International Airport.	A shooting occurred at Los Angeles International Airport.
Cette controverse croissante autour de l'agence a provoqué beaucoup de spéculations selon lesquelles l'incident de ce soir était le résultat d'une cyber-opération ciblée.	Such growing controversy surrounding the agency prompted early speculation that tonight's incident was the result of a targeted cyber operation.	This growing scandal around the agency has caused much speculation about how this incident was the outcome of a targeted cyber operation.
Le nombre total de morts en octobre est le plus élevé depuis avril 2008, quand 1 073 personnes avaient été tuées.	The total number of deaths in October is the highest since April 2008, when 1,073 people were killed.	The total number of deaths in May is the highest since April 2008, when 1 064 people had been killed.
À l'exception de l'opéra, la province reste le parent pauvre de la culture en France.	With the exception of opera, the provinces remain the poor relative of culture in France.	At an exception, opera remains of the state remains the poorest parent culture.

- Character-level information might help

Conclusion

- Propose train an NMT system in a completely unsupervised manner
  - Unsupervised **fixed cross-lingual embeddings**
  - Attentional encoder-decoder model with **shared encoder**
  - Train through **denoising** and **backtranslation**

- Significant improvements in BLEU score over a baseline system
  - Combine with small parallel corpus can bring further improvements

- Can be improved
  - Comparable supervised NMT is below the state of the art
  - Using character-level information
  - Modification of noise addition in denoising method

References

- NMT, Seq2seq model image : Jongwuk Lee, Text Mining class Chapter 10, 2018.
- Semi-supervised NMT image : Y. Cheng et al., Semi-Supervised Learning for Neural Machine Translation, in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.
- Model training image : G. Lample et al., Unsupervised Machine Translation Using Monolingual Corpora Only, in *Proc. of International Conference on Learning Representations (ICLR)*, 2018.
- BPE : R. Sennrich et al., Neural Machine Translation of Rare Words with Subword Units, *in Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.