Unsupervised Neural Machine Translation

조재춘 소아람 이찬희 김규경

KOREA UNIVERSITY

Natural Language Processing & Artificial Intelligence

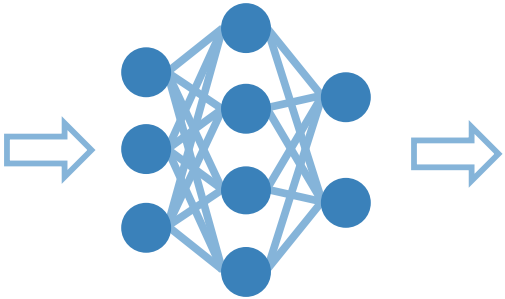
INDEX

Background

Brief Introduction to Sequence to Sequence Models

Neural Networks

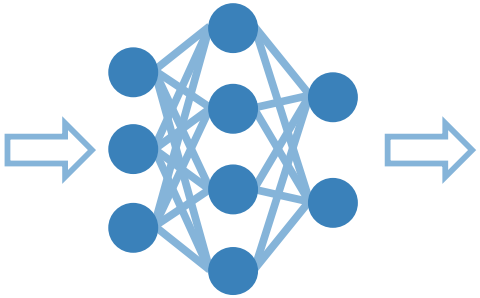
- Transforms input to output
- NN is a “set of weights(parameters)”
- Need to change the weights(train) to make it do what we want



Cat

Sequence to Sequence Models

- Transforms one sequence to another sequence
  - E.g. He loved to eat. -> Er liebte zu essen.

He loved to eat.  Er liebte zu essen.

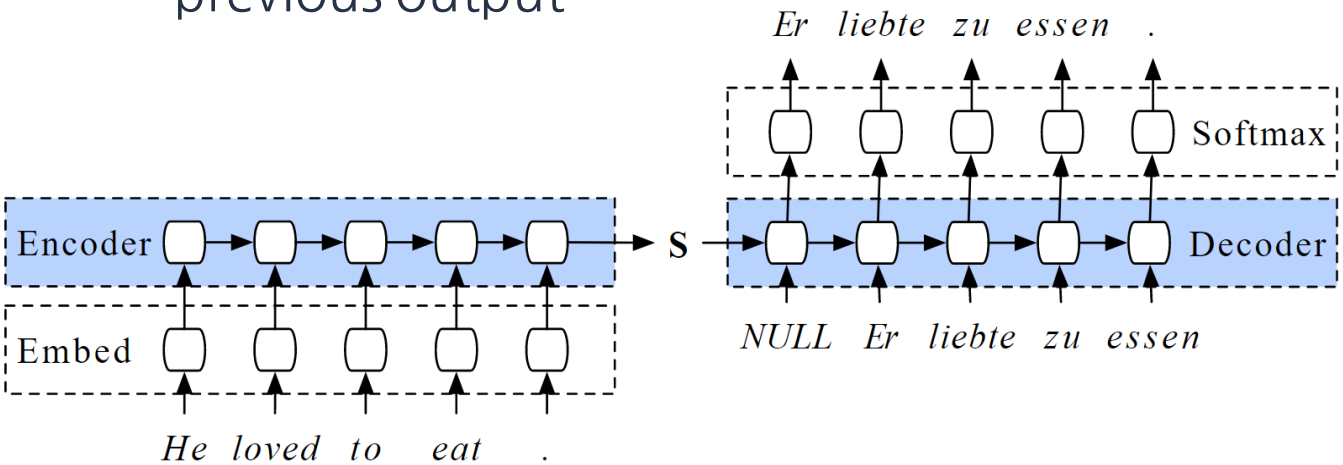
- Use Recurrent Neural Networks(RNN) to handle sequences of varying length

Sequence to Sequence Models

- Sequence to sequence(seq2seq) model has two main components - encoder, decoder
- Encoder - encodes a sequence of tokens(e.g. words) into a sequence of vectors
- Decoder - decodes the output of encoder(vectors) into a sequence of tokens(e.g. words)

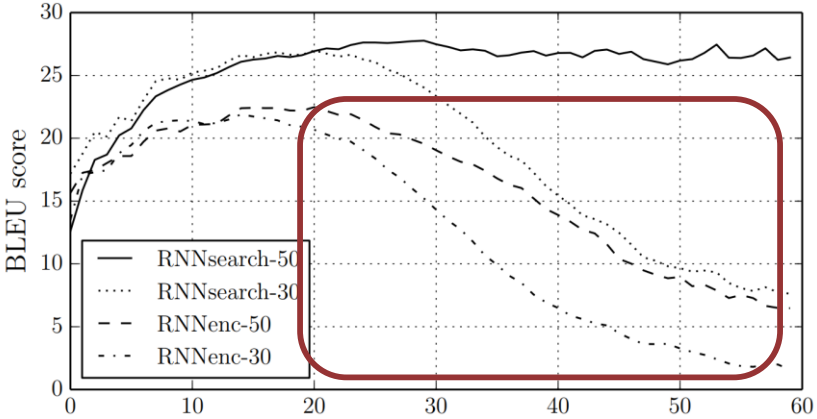
Seq2seq w/o Attention

- Use the final output vector of encoder to initializer decoder state
- Decoder performs greedy decoding using its previous output

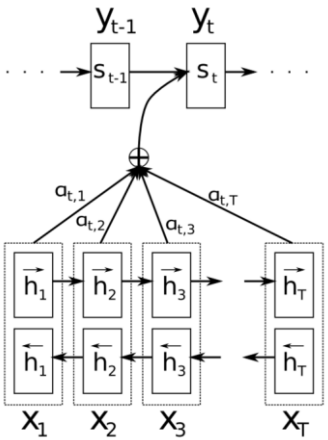


Information Bottleneck

- Input sequence is encoded into a single vector, which becomes a bottleneck
- Translation quality decreases as the input sentence gets longer



Seq2seq with Attention



$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

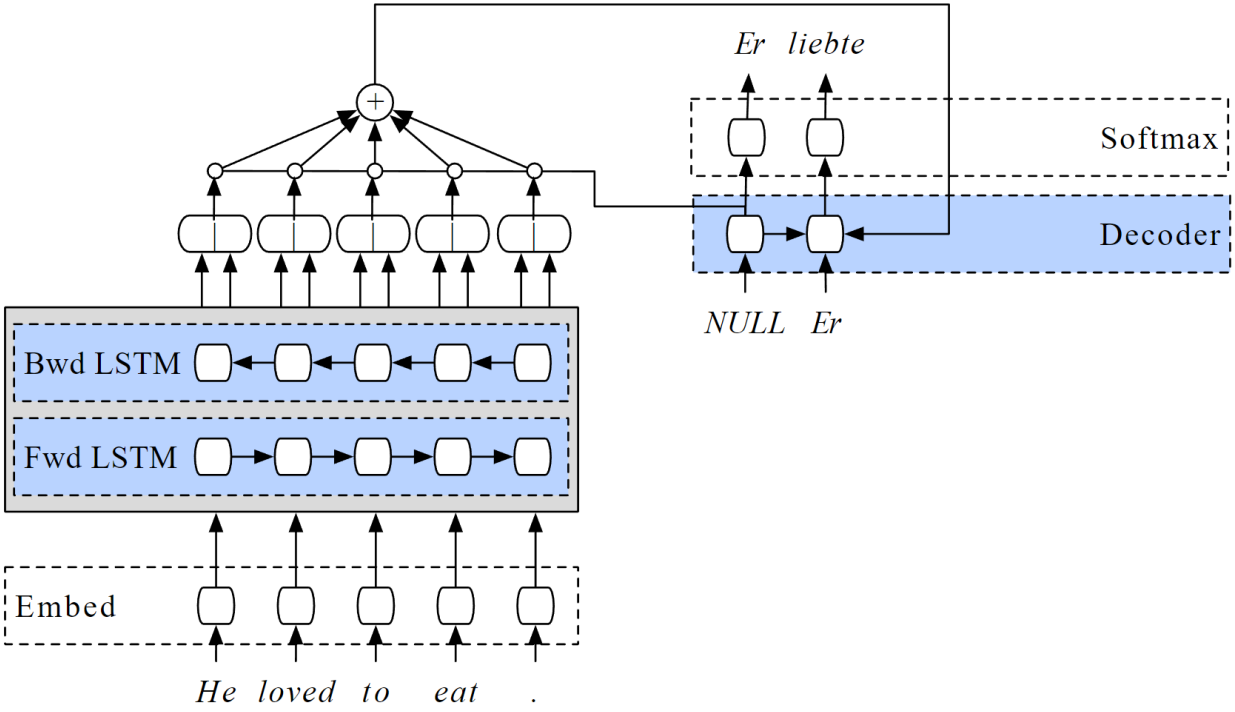
$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

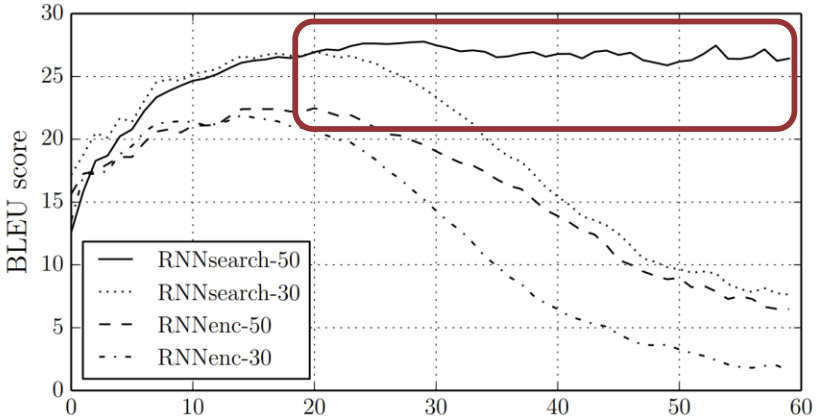
- Consider the weighted combination of all the encoder outputs, not just the last hidden state

Seq2seq with Attention



Seq2seq with Attention

- Seq2seq models with attention can handle long sequences well



Introduction

Unsupervised Neural Machine Translation

Introduction

- Neural Machine Translation(NMT) has become the dominant paradigm to machine translation
- However, NMT requires a large parallel corpus to be effective
- Lack of large parallel corpora is a practical problem for the vast majority of language pairs, including low-resource languages (e.g. Basque) as well as many combinations of major languages (e.g. German-Russian)

Introduction

- This paper proposes a method to train a NMT in a completely unsupervised manner, relying solely on monolingual corpora
- It builds upon the recent work on unsupervised cross-lingual embeddings
- They test the proposed method on unsupervised, semi-supervised, and supervised settings

Related Work & Method

Unsupervised Neural Machine Translation

Unsupervised Cross-Lingual Embeddings

- The proposed method uses a pre-trained cross-lingual embeddings in the encoder
  - The following will be a brief explanation of how it is formed.
  - The premises of the cross-lingual embeddings are based on the previous work from the authors of this paper.
    - Learning bilingual word embeddings with (almost) no bilingual data by Mikel Artetxe, Gorka Labaka, Eneko Agirre

Unsupervised Cross-Lingual Embeddings

- Distinctive Features of the Unsupervised Cross-Lingual embeddings are…
  - Reduced the need of large bilingual dictionaries to much smaller seed dictionaries
  - Dictionary is used to learn the embedding mapping and the embedding mapping is used to induce a new dictionary.
    → Stated as a self learning fashion

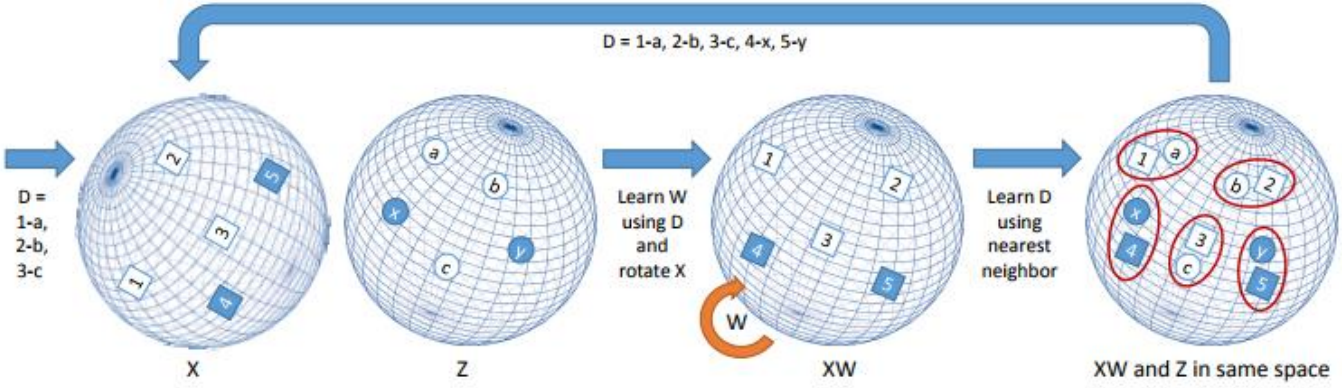
Unsupervised Cross-Lingual Embeddings

- The proclaimed 'Self-learning framework' is as follows:

Algorithm 1 Traditional framework

Input: X (source embeddings)
Input: Z (target embeddings)
Input: D (seed dictionary)

1: $W \leftarrow$ LEARN_MAPPING(X, Z, D)
2: $D \leftarrow$ LEARN_DICTIONARY(X, Z, W)
3: EVALUATE_DICTIONARY(D)



Unsupervised Cross-Lingual Embeddings

- The proclaimed 'Self-learning framework' works as follows:
  - Learn mapping W_1 based on X, Z and seed dictionary D_0.

$$W_1 \leftarrow \text{LEARN\_MAPPING}(X,\ Z,\ D_0)$$

  - Learn dictionary D_0 based on X, Z and W_1.

$$D_1 \leftarrow \text{LEARN\_DICTIONARY}(X,\ Z,\ W_1)$$

  - Assuming that the D_1 is better than the D_0, D_1 should serve to learn a better mapping W_2 and, consequently, an even better dictionary D_2 the second time.
  - The process is to be repeated iteratively to obtain a hopefully better mapping and dictionary each time until some convergence criterion is met.

Unsupervised Cross-Lingual Embeddings

- The proclaimed 'Self-learning framework' works as follows:

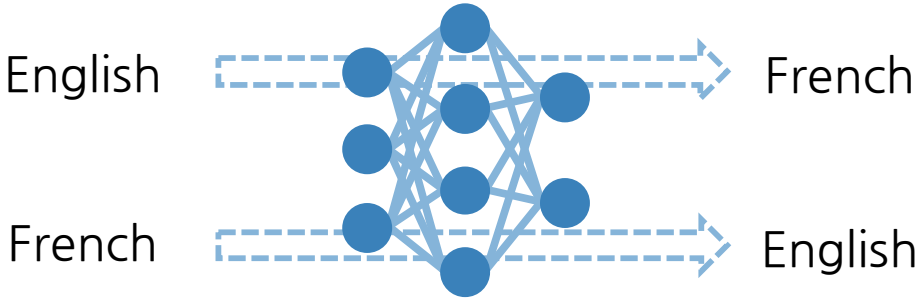
Algorithm 2 Proposed self-learning framework

Input: X (source embeddings)
Input: Z (target embeddings)
Input: D (seed dictionary)

```
repeat
    W ← LEARN_MAPPING(X, Z, D)
    D ← LEARN_DICTIONARY(X, Z, W)
until convergence criterion
EVALUATE_DICTIONARY(D)
```

System Architecture

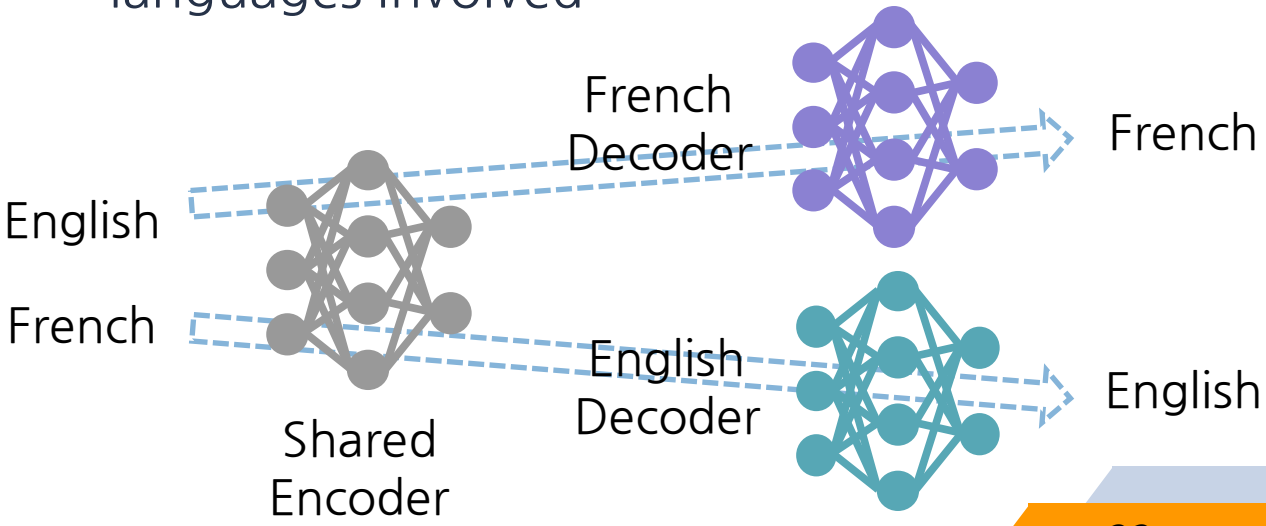
- Dual structure - handle both directions together (e.g. French ↔ English)



Unsupervised NMT

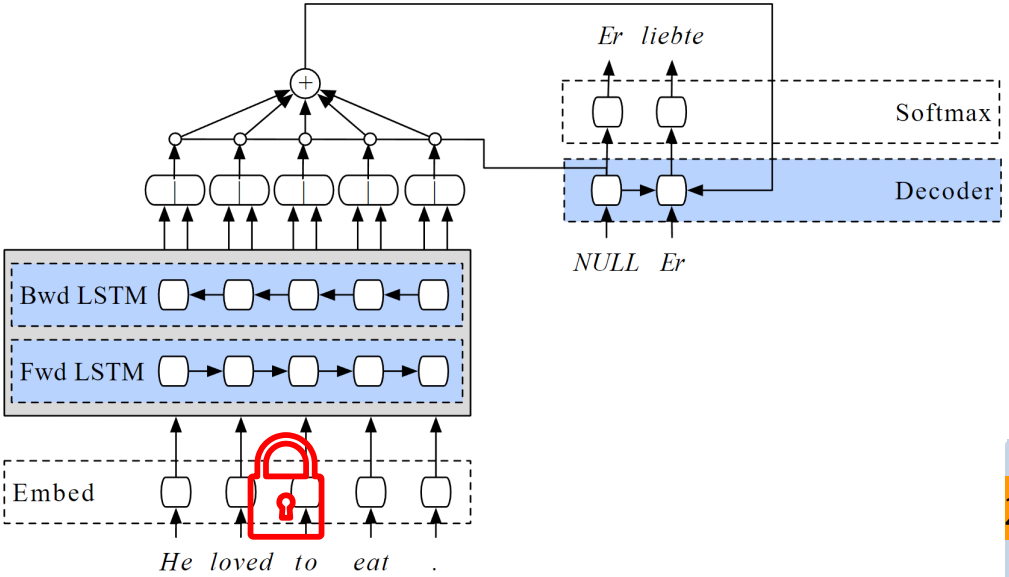
System Architecture

- Shared encoder - makes use of one and only one encoder that is shared by both languages involved



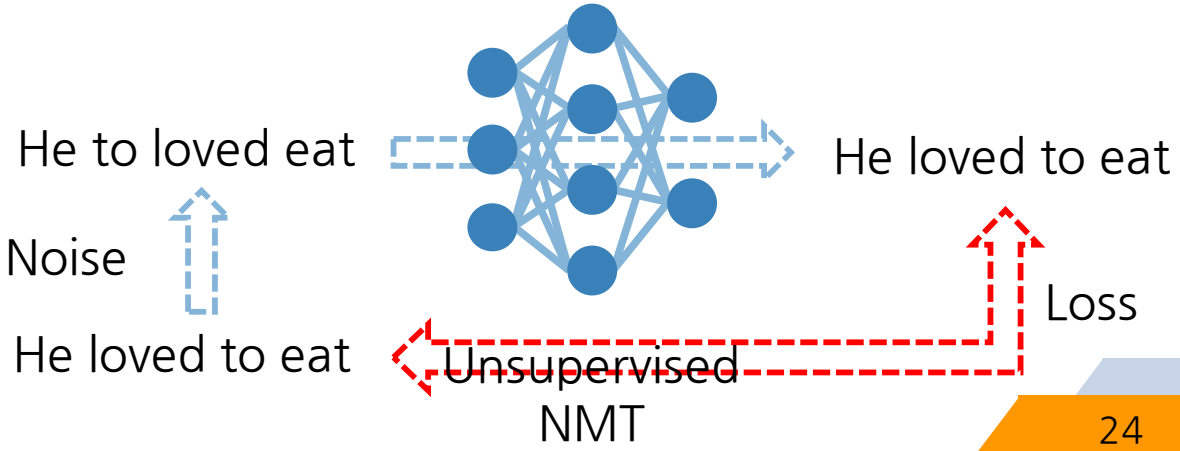
System Architecture

- Fixed embeddings in the encoder - use pre-trained cross-lingual embeddings in the encoder that are kept fixed during training



Unsupervised Training

- Denoising - like denoising autoencoders, alter the word order of the input sentence by making random swaps between contiguous words

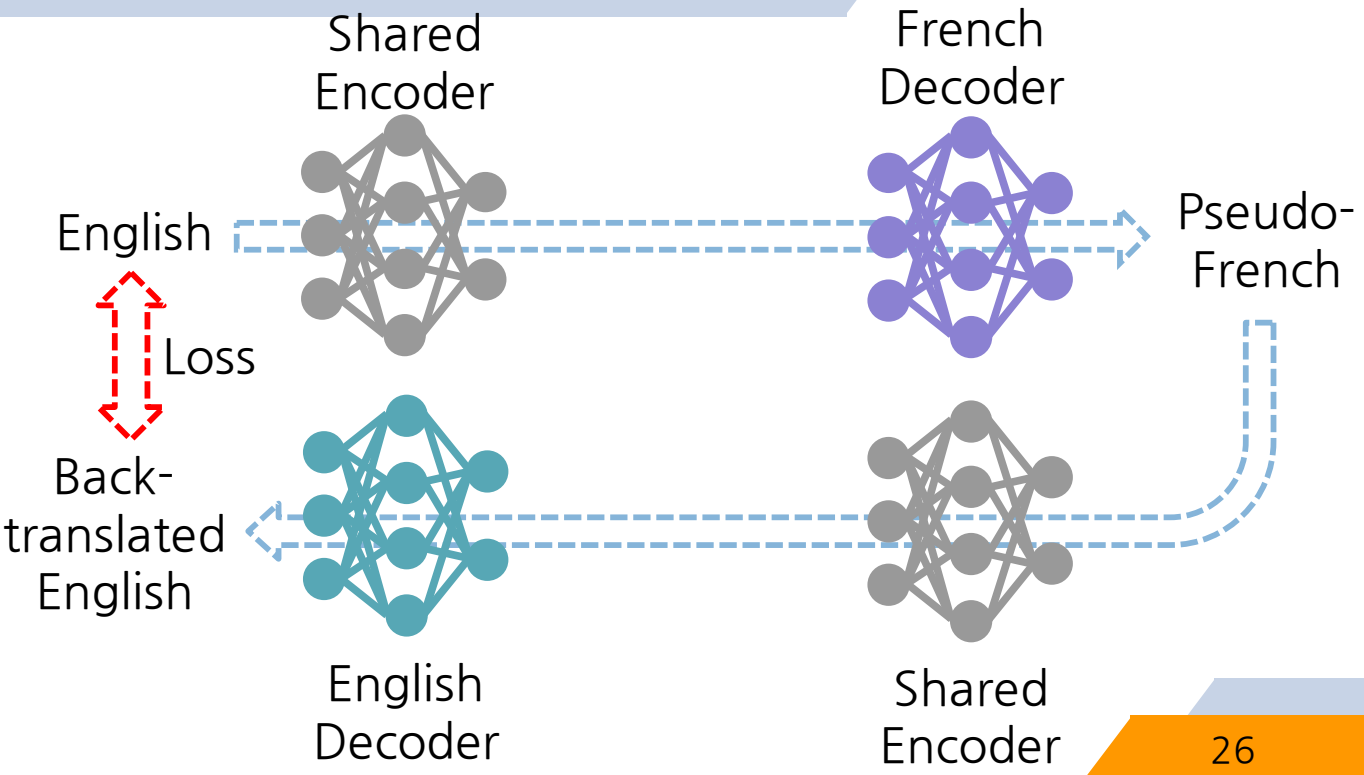


Unsupervised Training

- Backtranslation

1. Obtain a pseudo-parallel corpus: use the system in inference mode with greedy decoding to translate it to the other language
2. Train the system to predict the original sentence from this translation, using the pseudo-parallel corpus

Unsupervised Training



Losses and Training Procedure

- Losses (Languages L1, L2)
  1. Denoising L1
  2. Denoising L2
  3. Backtranslation L1 ⇨ L2 ⇨ L1
  4. Backtranslation L2 ⇨ L1 ⇨ L2
- Alternate these different training objectives from batch to batch

Experiments

Unsupervised Neural Machine Translation

Experimental settings

- **Datasets** *(WMT 2014)*
  - French ↔ English
  - German ↔ English
- **Evaluation**
  - Tokenized BLEU(Bilingual Evaluation Understudy) score

- **BLEU score**
  - N-gram overlap between machine translation output and reference translation
  - Compute precision for n-grams of size 1 to 4

System A: Israeli officials responsibility of airport safety

2-gram match

Reference: Israeli officials are responsible for airport security

System B: airport security Israeli officials are responsibility

4-gram match

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

- **Unsupervised**
  - System has access to nothing but monolingual corpora
  - Data
    - News Crawl Corpus with articles form 2007-2013
- **Semi-supervised**
  - monolingual corpora, small in-domain parallel corpus
  - Data
    - News Crawl Corpus with articles form 2007-2013
    - 100,000 random sentence pairs from News commentary parallel corpus

- **Supervised**
  - Large parallel corpus
  - Data
    - Europarl
    - Common Crawl
    - News Commentary
    - UN corpus
    - Gigaword corpus (French - English)

Corpus preprocessing

- **Tokenization**
- **Truecasing**
- **Byte pair encoding(BPE)**
  - Using 50,000 operations
  - Replacing OOV words with special token *<UNK>*

Quantitative analysis

- BLEU scores

		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) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	6. Comparable NMT	20.48	19.89	15.04	11.05
	7. GNMT (Wu et al., 2016)	-	38.95	-	24.61

Qualitative analysis

- **BPE** (French → English)

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.

Conclusion

- Proposed a novel method to train an NMT system in a completely unsupervised manner, training the system from monolingual corpora alone, combining denoising and backtranslation
- The trained system is able to model complex cross-lingual relations and produce high-quality translations
- Combining the proposed method with a small parallel corpus can bring further improvements

THANKS!

Any questions?