Neural Machine Translation by Jointly Learning to Align and Translate

한양대학교 산업공학과
IDSL 석사과정 고경준

HANYANG UNIVERSITY

Contents

1. Introduction

Introduction

- Encoder-decoder approach는 input sequences를 fixed length vector로 encoding
- Encoding 과정에서 bottle-neck problem이 발생하여 성능 악화
- 이를 해결하기 위해 저자들은 extension version을 제안

- Target word를 생성할 때, source sentence에서 관련된 정보에 집중하는 방식
- English-to-French translation에서 기존의 phrase-based system에 준하는 성능을 보임
- Qualitative analysis에서 linguistically plausible alignment를 보여줌

2. Background : Neural Machine Translation

- Probabilistic translation은 argmax p(y|x)와 같음
- Neural machine translation에서는 conditional distributions of sentence pairs를 학습
- 최근 neural machine translation에 관한 많은 연구들은 encoder-decoder를 많이 채택
- 이러한 모델들의 제한적인 활용을 통해 성능을 향상

2.1 RNN Encoder-Decoder

- 일반적인 RNN Encoder-Decoder 방식은 아래와 같음
- Input sentences가 RNN layer output h_t로, 이 h_t 들이 fixed length vector c로 표현됨
- 최종 output sentences를 표현하는 (2)번식에서 conditional probability는 RNN에서 (3)번식 처럼 modeling 됨

$$c = q\left(\{h_1, \cdots, h_{T_x}\}\right), \qquad h_t = f\left(x_t, h_{t-1}\right) \tag{1}$$

$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c), \tag{2}$$

$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c), \tag{3}$$

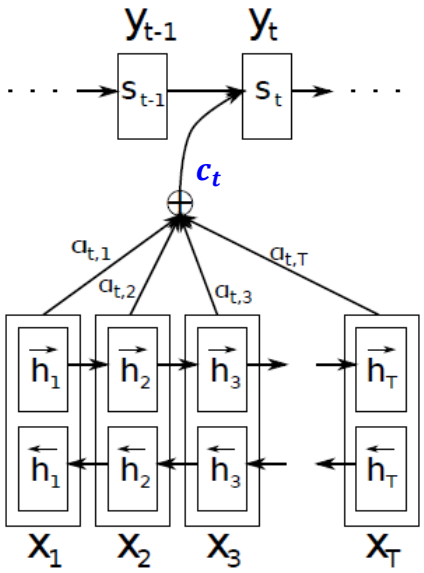
3. Learning to align and translate

3.1 Decoder : general description

$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1},s_i,c_i), \qquad (4)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \qquad (5)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x}\exp(e_{ik})}, \qquad e_{ij} = a(s_{i-1},h_j) \qquad (6)$$



- $s_i = f(s_{i-1},y_{i-1},c_i)$. 는 RNN hidden state for time i
- (4)번 식과 같이 Probability y_i 가 distinct context vector c_i에 영향을 받음
- c_i는 encoder의 모든 hidden state 값이 s_{i-1}에 align될 확률을 가중합하여 계산
- 이는 (6)번 식의 feedforward neural network로 alignment값 생성 후 softmax로 계산

3. Learning to align and translate

3.2 Encoder : bidirectional RNN for annotating sequences

- Alignment에 preceding / following words 모두 고려하기 위해 bidirectional RNN 사용
- Forward RNN의 hidden state를 $\overrightarrow{h}$, backward RNN의 hidden state를 $\overleftarrow{h}$라고 하면

  $h_j = [\overrightarrow{h}_j^{\mathrm{T}}, \overleftarrow{h}_j^{\mathrm{T}}]^{\mathrm{T}}$ (concat $\overrightarrow{h}$ and $\overleftarrow{h}$)
- RNNs의 특징에 따라 h_j는 x_j와 그 주변 word에 focused
- 이 값들은 decoder의 context vector의 계산에 활용됨 ((5), (6)번 식)

4. Experiment setting

4.1 Dataset

- WMT'14 [English to French dataset], total 850M words
  (Europarl 61M, news commentary 5.5M, UN 421M, two crawled 90M & 272.5M words)
- It reduced to 348M words
- Concatenated validation set (news-test-2012 & 2013), test set (news-test-2014)

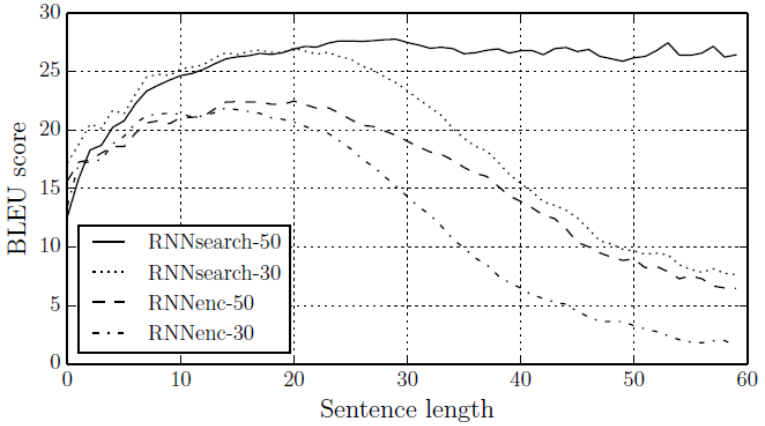
4.2 Models

- RNN Encoder-Decoder (RNNencdec)와 proposed model (RNNsearch)
- 학습하는 sentences의 length에 따라 30이상, 50이상 모델을 각각 학습
- RNNencdec-30, RNNsearch-30, RNNencdec-50, RNNsearch-50
- Adadelta를 이용해 SGD 진행 (minibatch of 80 sentences)
- 대략 5일 정도의 학습 시간이 소모

5. Results

5.1 Quantitative results

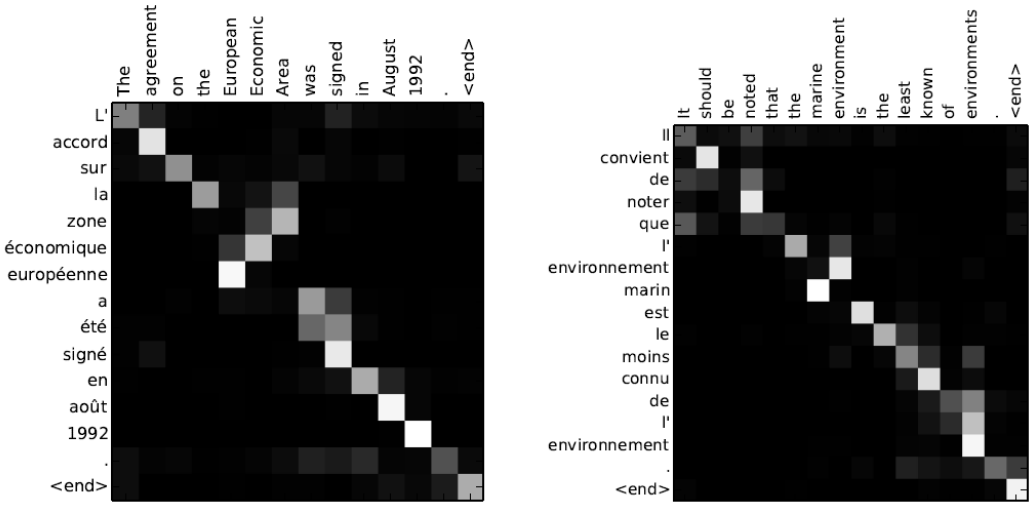
Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63



- Test set에 대한 BLEU scores (No UNK : without unknown words)
- RNNsearch-50*는 성능 향상이 없어질 때 까지 계속 학습시킨 결과
- 각 데이터셋에서 RNNsearch가 RNNencdec를 항상 outperform
- Conventional phrase-based model (Moses)보다 우수한 성능을 보임
- 심지어 RNNsearch-30이 RNNencdec-50보다 더 우수한 모습

5. Results

5.2 Qualitative analysis (alignment)



(a) (b)

- (6)번 식의 annotation weight (α_{ij})를 visualizing 한 모습
- Input sequence와 output sequence의 단어 배열 순서가 적절히 변경됨을 확인 가능
- Soft-alignment를 통해 (beam search) [NULL]로 mapping 되는 단어를 없애고, [the]라는 단어를 주변 단어에 따라 [le], [la], [les], [l'] 등으로 결정될 여지를 부여함
  (위 예시에서는 나타나지 않음. 논문의 Fig. 3 – (d) 참고.)

5. Results

5.2 Qualitative analysis (long sentences)

- RNNencdec model의 경우 long sentences에서 후반부 내용을 적절하게 translate하지 못하는 모습을 보임
- 반면 RNNsearch model의 경우 50 혹은 그 이상의 length sentences도 후반부까지 적절히 translate하는 모습을 보임
- 아래 test set 예시 참고

This kind of experience is part of Disney's efforts to "extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important," he added.

원문

Ce type d'expérience fait partie des initiatives du Disney pour "prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.

RNNencdec

digital readers that become more complex.

Ce genre d'expérience fait partie des efforts de Disney pour "prolonger la durée de vie de ses séries et créer de nouvelles relations avec des publics via des plateformes numériques de plus en plus importantes", a-t-il ajouté.

RNNsearch

via increasingly important digital platforms", he added.

6. Related work & Conclusion

6.1 Learning to align

- 기존에 output symbol과 input symbol 간 aligning 시도가 있었으나 (Graves 2013) weights of the annotations가 one direction인 반면 이 방식은 every word of source sentence를 참고하는 방식

6.2 Neural Networks for machine translation

- Machine translation을 위해 다양한 Neural Network 모형이 기존 phrase-based statistical approach에 접목하는 방식인 반면 이 방식은 단독으로 사용되는 completely new translation system (based on neural networks)

6.3 Conclusion

- 기존 encoder-decoder approach의 fixed length vector로 인한 bottle-neck problem 해결
- 이를 통해 neural network을 활용해 long sentences translation이 가능해짐
- Qualitative analysis를 통해 alignment의 적절성도 확인
- Neural machine translation 연구에 대한 중요한 발판이 됨

감사합니다.