NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio
발제_한지윤

Abstract

- 문장 정보 전체를 포함하지 않은 fixed-length vector를 이용하는 것이 성능 향상 저해
- 고정적인 분할(segment)를 하지 않고 번역 대상 단어와 관련된 번역 원문이 어떤 부분인지 자동적으로 (soft-)search
- 영어-프랑스어

**용어 정리:

align ≈ attention
annotation ≈ hidden state

1.Introduction

기존 encoder–decoder 모델의 문제점: 긴 문장 성능 저하

제안 모델: 정렬(align)과 번역(translate)을 함께(joint) 학습하는 확장된 encoder–decoder model

-제안 모델이 번역문에서 단어(word)를 생성할 때마다 원문(source sentence)에서 관련된 정보가 응축된 위치 집합(set of positions)을 탐색(soft-searches)
-모델은 이 원문 위치 정보와 앞서 생성된 대상 단어(target word)와 관련된 맥락(context) vector를 이용하여 대상 단어 예측

2.1. RNN encoder-decoder

Here, we describe briefly the underlying framework, called *RNN Encoder–Decoder*, proposed by Cho *et al.* (2014a) and Sutskever *et al.* (2014) upon which we build a novel architecture that learns to align and translate simultaneously.

In the Encoder–Decoder framework, an encoder reads the input sentence, a sequence of vectors $\mathbf{x} = (x_1, \cdots, x_{T_x})$, into a vector c.[2] The most common approach is to use an RNN such that

$$h_t = f(x_t, h_{t-1}) \tag{1}$$

and

$$c = q(\{h_1, \cdots, h_{T_x}\}),$$

where $h_t \in \mathbb{R}^n$ is a hidden state at time t, and c is a vector generated from the sequence of the hidden states. f and q are some nonlinear functions. Sutskever *et al.* (2014) used an LSTM as f and $q(\{h_1, \cdots, h_T\}) = h_T$, for instance.

The decoder is often trained to predict the next word $y_{t'}$ given the context vector c and all the previously predicted words $\{y_1, \cdots, y_{t'-1}\}$. In other words, the decoder defines a probability over the translation $\mathbf{y}$ by decomposing the joint probability into the ordered conditionals:

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

where $\mathbf{y} = (y_1, \cdots, y_{T_y})$. With an RNN, each conditional probability is modeled as

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

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_t, and s_t is the hidden state of the RNN. It should be noted that other architectures such as a hybrid of an RNN and a de-convolutional neural network can be used (Kalchbrenner and Blunsom, 2013).

Encoder: 유동적인 길이의 source sequence를 fixed-legnth vector로 변환

Decoder: vector representation을 유동적인 target sequence로 변환

주어진 source sequence에서 target sequence의 조건부 확률을 극대화

3.1. Decoder: General Description

In a new model architecture, we define each conditional probability in Eq. (2) as:

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

where s_i is an RNN hidden state for time i, computed by

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

It should be noted that unlike the existing encoder–decoder approach (see Eq. (2)), here the probability is conditioned on a distinct context vector c_i for each target word y_i.

The context vector c_i depends on a sequence of *annotations* $(h_1, \cdots, h_{T_x})$ to which an encoder maps the input sentence. Each annotation h_i contains information about the whole input sequence with a strong focus on the parts surrounding the i-th word of the input sequence. We explain in detail how the annotations are computed in the next section.

The context vector c_i is, then, computed as a weighted sum of these annotations h_i:

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

The weight α_{ij} of each annotation h_j is computed by

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

where

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

s_i: RNN hidden state
c_i: context vector (5)
α_{ij}: 가중치, softmax (6)
e_{ij}: alignment model, j번째 input 주변의 정보들이 얼마나 i번째 output과 적합한지에 대한 점수

- α는 하나의 feedforward neural network로 구성되며 다른 네트워크와 같이 학습. 기존의 NMT에서는 alignment를 latent variable로 간주했으나 이 모델에서는 네트워크의 variable

- 두개의 네트워크가 함께 학습되며, 학습과정의 cost gradient는 역전파를 통해 두 네트워크에 적용됨

-

- target word y_i와 source word x_i에 대한 확률이라 생각하면, i번째 context vector c_i는 annotation의 기대값

- decoder가 source sentence에서 집중해야 할 부분을 결정 = 어텐션

3.2. Encoder: Bidirectional RNN for Annotating Sequences

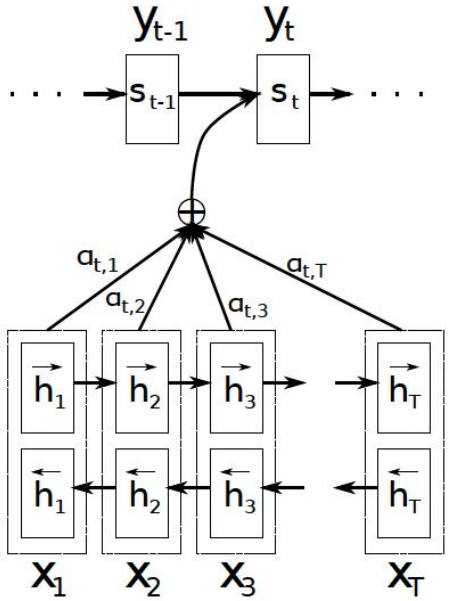

Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

양방향 RNN(bidirection RNN, BiRNN) 차용:
annotation이 앞 단어뿐 아니라 뒤에 오는 정보도 포함할 수 있도록 양방향 RNN 차용.
annotation h_j는 j번째 단어 앞뒤의 정보를 모두 포함

4.1. Dataset

WMT'14 English-French corpora
Data Selection: 데이터 크기 축소
monolingual data는 사용 안함

4.2. Models

기존 모델: RNN Encoder-Decoder모델(RNNencdec)
제안 모델: RNNsearch
학습 방법 1: 문장길이 30 제한
학습 방법 2: 문장길이 50 제한
hidden unit: 두 모델의 Encoder-decoder는 각각 1,000개
Minibatch SGD 알고리즘 (80)
Adadelta

5.1. Quantitative Results

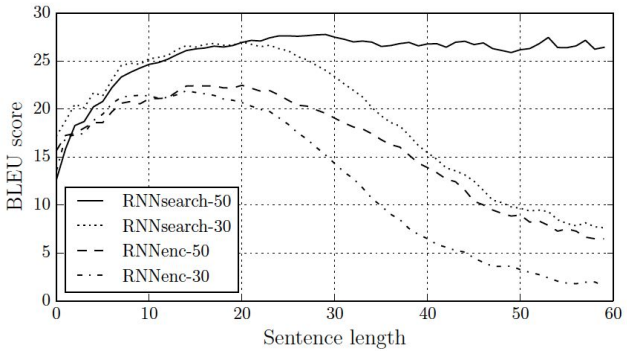

Figure 2: The BLEU scores of the generated translations on the test set with respect to the lengths of the sentences. The results are on the full test set which includes sentences having unknown words to the models.

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

Table 1: BLEU scores of the trained models computed on the test set. The second and third columns show respectively the scores on all the sentences and, on the sentences without any unknown word in themselves and in the reference translations. Note that RNNsearch-50* was trained much longer until the performance on the development set stopped improving. (○) We disallowed the models to generate [UNK] tokens when only the sentences having no unknown words were evaluated (last column).

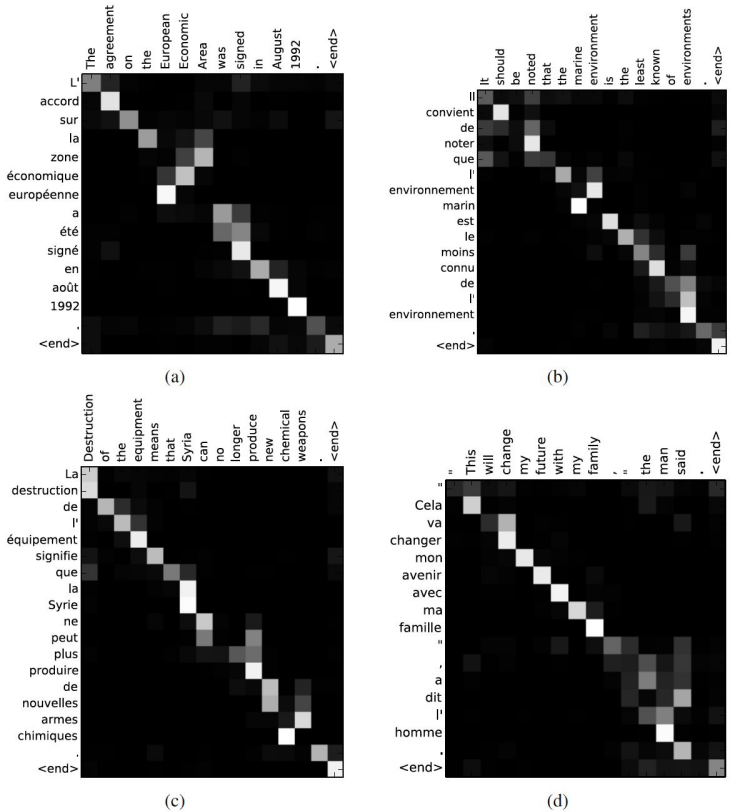

Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word (see Eq. (6)), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b–d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.

5.2. Qualitative Analysis

5.2.1. Alignment

영어와 프랑스어에서 특징적인 차이는 형용사와 명사의 어순인데 이를 잘 정렬함.

soft alignment의 경우는 관사도 잘 잡아내어, 원어와 대상어의 길이가 다른 경우도 잘 잡아냄.

5.2.2. Long Sentences

긴 문장의 경우도 디테일의 누락없이 잘 번역함.

감사합니다:)

6, 7의 경우는 모델 설명이 아니라 생략할게요..