Neural machine translation by jointly learning to align and translate

Bahdanau, D., Cho, K., & Bengio, Y. (2014)

집현전 논문 리뷰
소현지

Contents

Abstract

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

- NMT(Neural machine translation)의 특징 및 한계
  - 최근 제안된 기계 번역 방법론
    - 기존 기계 번역 방법론: statistical machine translation
  - 번역 성능을 극대화 하기 위하여 공동으로 조정할 수 있는 단일 신경망 구축을 목표로 함
    - 기존 번역 방법론이 문장을 작은 구성요소들로 나누어 번역을 수행한 것과 달리, NMT는 단일 문장 단위의 번역을 수행함
  - 일반적으로 encoder-decoder 형태의 모델을 사용함
    - Source sentence(번역 대상 문장; 번역을 수행하고자 하는 문장)를 fixed-length vector로 인코딩함
    - Fixed-length vector가 해당 모델의 성능 개선을 저해하는 요소임

Abstract

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

- 본 논문에서 제안하는 방법론
  - Target word 예측 시 source sentence 내의 관련된 단어를 자동으로 탐색하는 모델을 제안함
    - Source sentence: 나는 학생이다
    - Target sentence: I am a student
  - 영어-프랑스어 번역 작업에서 기존 sota와 필적하는 번역 성능 달성
  - 모델에 의하여 찾아진 (soft-alignments)를 정성적 분석 결과, 우리의 직관과 일치함
    - Alignments: 원본 텍스트의 단어와 번역 결과의 단어를 매칭시키는 것

Introduction

Neural machine translation is a newly emerging approach to machine translation, recently proposed by Kalchbrenner and Blunsom (2013), Sutskever *et al.* (2014) and Cho *et al.* (2014b). Unlike the traditional phrase-based translation system (see, e.g., Koehn *et al.*, 2003) which consists of many small sub-components that are tuned separately, neural machine translation attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation.

Most of the proposed neural machine translation models belong to a family of *encoder–decoders* (Sutskever *et al.*, 2014; Cho *et al.*, 2014a), with an encoder and a decoder for each language, or involve a language-specific encoder applied to each sentence whose outputs are then compared (Hermann and Blunsom, 2014). An encoder neural network reads and encodes a source sentence into a fixed-length vector. A decoder then outputs a translation from the encoded vector. The whole encoder–decoder system, which consists of the encoder and the decoder for a language pair, is jointly trained to maximize the probability of a correct translation given a source sentence.

A potential issue with this encoder–decoder approach is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector. This may make it difficult for the neural network to cope with long sentences, especially those that are longer than the sentences in the training corpus. Cho *et al.* (2014b) showed that indeed the performance of a basic encoder–decoder deteriorates rapidly as the length of an input sentence increases.

- NMT 모델의 Fixed-length vector
  - 모델이 source sentence의 필요한 모든 정보를 고정 길이 벡터로 압축해야 함
  - 긴 문장 (특히 훈련 말뭉치 문장보다 긴 문장)의 처리가 어려움
    - 저는 학생입니다.
    - 저는 AA 대학교 BB 학과의 CC 교수님 지도 하 자연어처리를 전공하고 있는 학생입니다.
  - 실제 선행 연구를 통하여 입력 문장의 길이 증가에 따라 인코더-디코더의 급격한 성능 저하가 입증됨

Introduction

In order to address this issue, we introduce an extension to the encoder–decoder model which learns to align and translate jointly. Each time the proposed model generates a word in a translation, it (soft-)searches for a set of positions in a source sentence where the most relevant information is concentrated. The model then predicts a target word based on the context vectors associated with these source positions and all the previous generated target words.

The most important distinguishing feature of this approach from the basic encoder–decoder is that it does not attempt to encode a whole input sentence into a single fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation. This frees a neural translation model from having to squash all the information of a source sentence, regardless of its length, into a fixed-length vector. We show this allows a model to cope better with long sentences.

In this paper, we show that the proposed approach of jointly learning to align and translate achieves significantly improved translation performance over the basic encoder–decoder approach. The improvement is more apparent with longer sentences, but can be observed with sentences of any length. On the task of English-to-French translation, the proposed approach achieves, with a single model, a translation performance comparable, or close, to the conventional phrase-based system. Furthermore, qualitative analysis reveals that the proposed model finds a linguistically plausible (soft-)alignment between a source sentence and the corresponding target sentence.

- Fixed-length vector의 한계를 해결하기 위하여 제안된 방법론
  - 번역 모델이 단어를 생성할 때마다, source sentence 내에서 해당 단어와 가장 관련성이 높은 영역(source position)을 탐색함
    - Source position, 앞서 생성된 모든 target words, context vector를 활용하여 target words를 예측함
  - 제안된 방법론의 가장 중요한 특징은 전체 입력 문장을 단일 고정 길이 벡터로 인코딩하지 않는다는 점임
    - 입력 문장을 sequence of vectors로 인코딩하고
    - 번역을 디코딩 하는 동안 해당 벡터의 하위 집합을 적절히 선택함
    - 따라서, 모델이 긴 문장에 더 잘 대처할 수 있음
  - 제안된 방법론은 기존 encoder-decoder 모델에 비해 번역 성능이 크게 향상됨
    - 문장의 길이가 짧아도 성능이 좋으나, 길수록 성능 차이가 뚜렷함
    - 단일 모델로도 기존 encoder-decoder 모델과 유사한 번역 성능을 달성함
    - 정성적 분석 결과, 모델이 타당한 (soft-) alignment를 도출한다는 것을 확인함

Background: Neural Machine Translation

From a probabilistic perspective, translation is equivalent to finding a target sentence y that maximizes the conditional probability of y given a source sentence x, i.e., $\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$. In neural machine translation, we fit a parameterized model to maximize the conditional probability of sentence pairs using a parallel training corpus. Once the conditional distribution is learned by a translation model, given a source sentence a corresponding translation can be generated by searching for the sentence that maximizes the conditional probability.

Recently, a number of papers have proposed the use of neural networks to directly learn this conditional distribution (see, e.g., Kalchbrenner and Blunsom, 2013; Cho *et al.*, 2014a; Sutskever *et al.*, 2014; Cho *et al.*, 2014b; Forcada and Ñeco, 1997). This neural machine translation approach typically consists of two components, the first of which encodes a source sentence x and the second decodes to a target sentence y. For instance, two recurrent neural networks (RNN) were used by (Cho *et al.*, 2014a) and (Sutskever *et al.*, 2014) to encode a variable-length source sentence into a fixed-length vector and to decode the vector into a variable-length target sentence.

Despite being a quite new approach, neural machine translation has already shown promising results. Sutskever *et al.* (2014) reported that the neural machine translation based on RNNs with long short-term memory (LSTM) units achieves close to the state-of-the-art performance of the conventional phrase-based machine translation system on an English-to-French translation task.[1] Adding neural components to existing translation systems, for instance, to score the phrase pairs in the phrase table (Cho *et al.*, 2014a) or to re-rank candidate translations (Sutskever *et al.*, 2014), has allowed to surpass the previous state-of-the-art performance level.

- 기존의 NMT
  - 초기 확률론적 관점에서의 translation task: source sentence (x) 가 주어졌을 때 target sentence (y)의 조건부 확률을 최대화 하는 것

$$\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$$

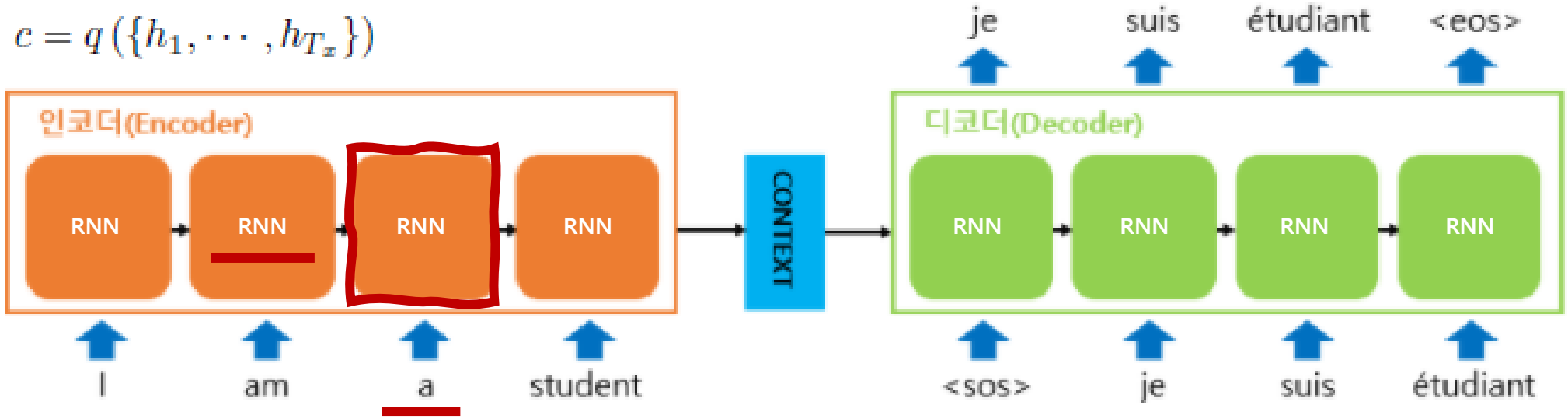
  - 최근에는 딥러닝을 활용하여 이러한 조건부 확률의 분포 자체를 학습하는 방향으로 변화하게 됨
    - 가장 대표적 연구: RNN 기반의 encoder-decoder

Background: Neural Machine Translation

- 2.1 RNN encoder–decoder
  - Encoder: input sentence를 입력 받아, fixed-length context vector 생성
    - Input: $\mathbf{x} = (x_1, \cdots, x_{T_x})$
    - Output: c
  - 계산 과정

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

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

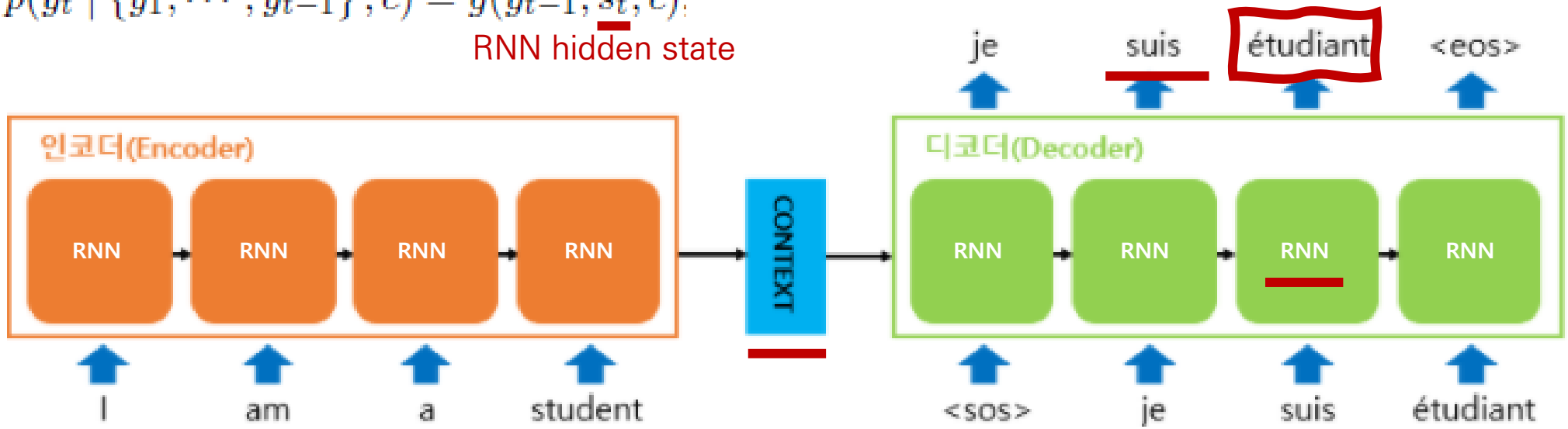


Background: Neural Machine Translation

- 2.1 RNN encoder–decoder
  - Decoder: context vector를 입력 받아, next word y_t,를 예측
    - Input: c, 앞서 예측(생성)된 모든 단어 $\{y_1, \cdots, y_{t'-1}\}$
    - Output: $y_{t'}$
  - 계산 과정

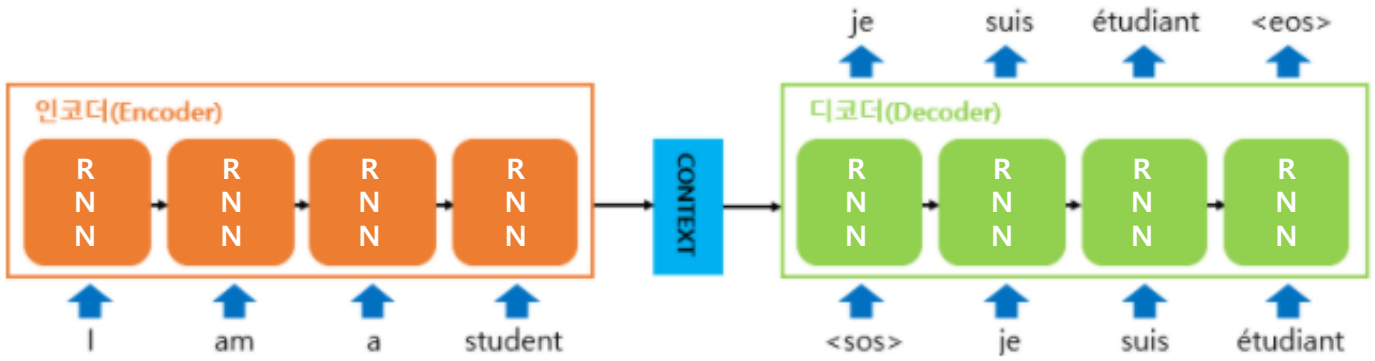
$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c).$$

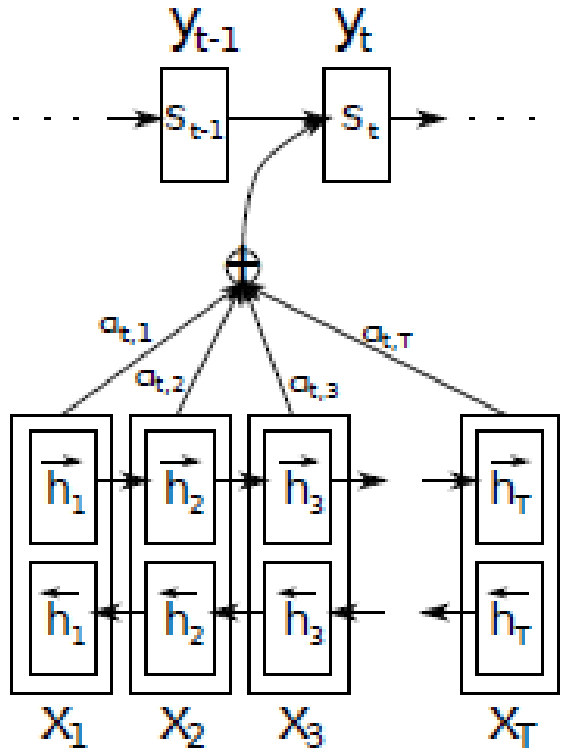
RNN hidden state



Learning to align and translate

- 본 논문에서 제안하는 방법론
  - Neural machine translation by jointly learning to align and translate
  - Align과 translate을 동시에 학습하는 새로운 아키텍쳐





Learning to align and translate

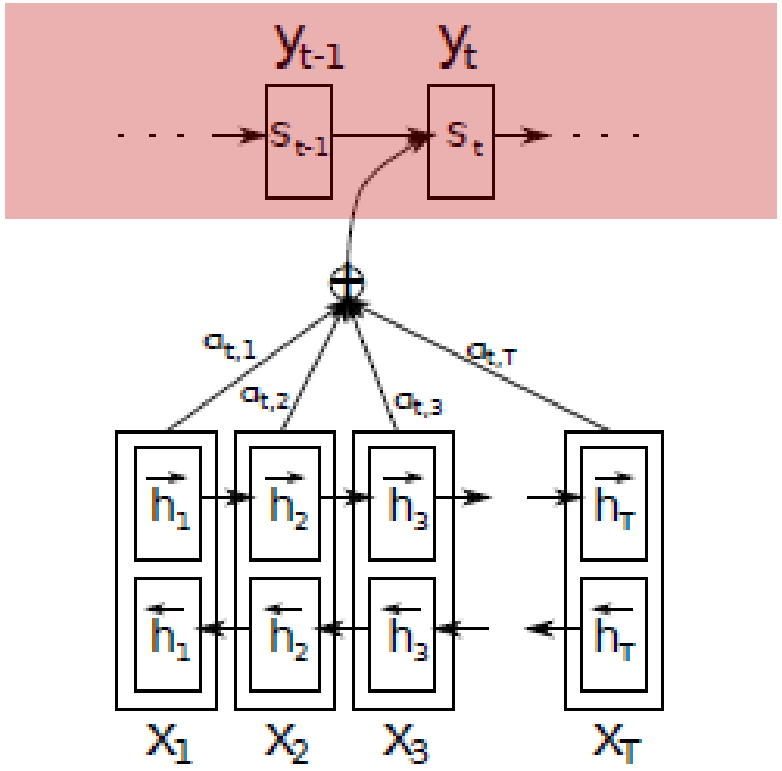


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

- 3.1 Decoder: general description
  - 조건부 확률이 다음과 같이 정의됨

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

  - RNN hidden state (s_i)

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

Learning to align and translate

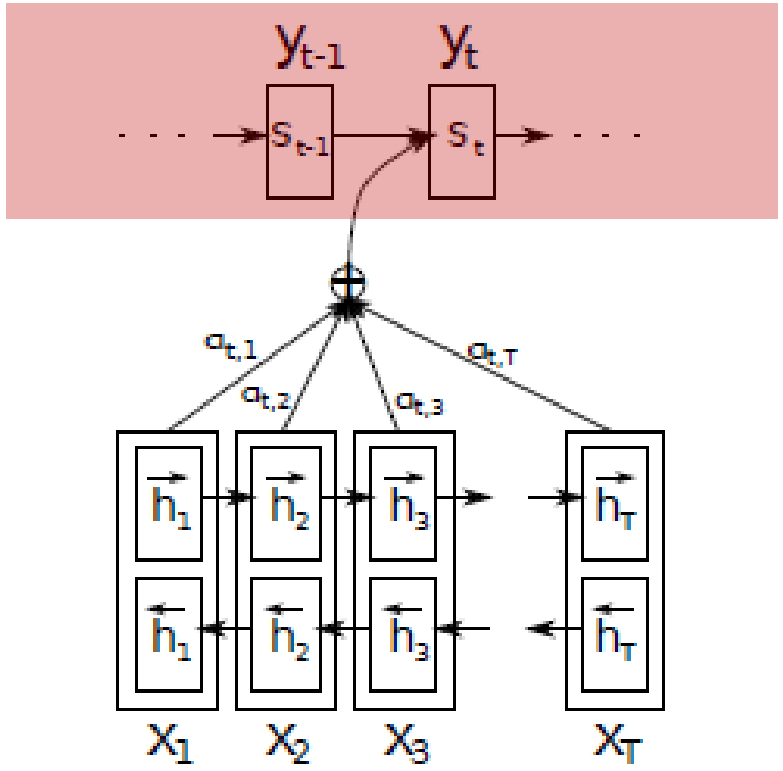


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

- 3.1 Decoder: general description
  - Context vector (c_i)
    - 각각의 target word(y_i)에 대한 개별 context vector (c_i)
    - h_j 의 가중합으로 계산됨
      > h_j는 인풋 시퀀스 x_j 의 annotation을 의미
      > 전체 인풋 시퀀스의 정보를 담고 있으나, x_j 에 대한 정보가 특히 많이 저장됨

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

Learning to align and translate

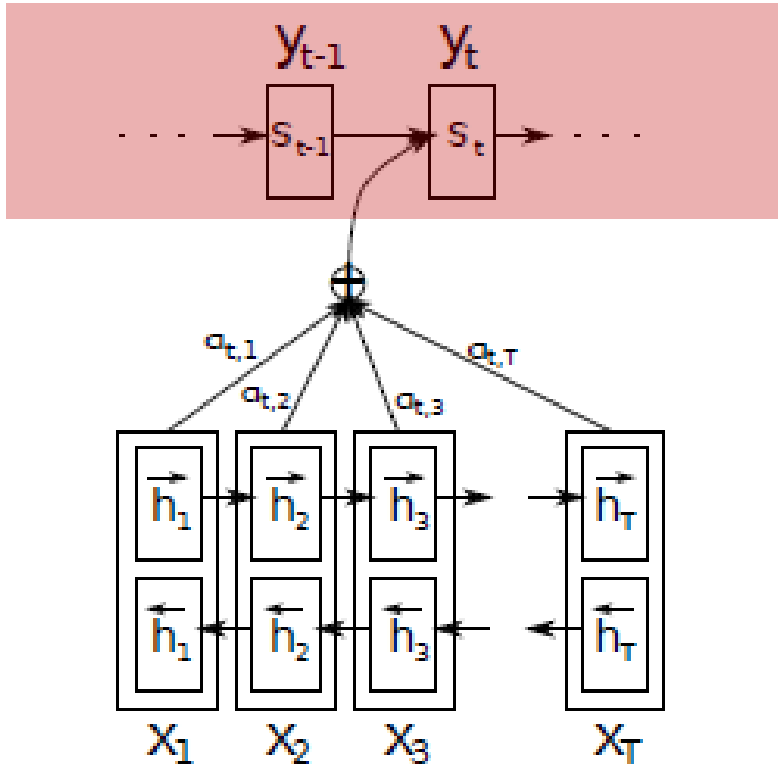


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

- 3.1 Decoder: general description
  - Context vector (c_i)

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

  - 각각의 annotation (h_j)에 대한 가중치 (α_{ij})

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

  - Alignment model
    - 주변의 인풋 단어와 위치의 아웃풋 단어가 일치하는 정도에 대한 점수를 부여함
    - Feedforward neural network를 사용하여 다른 시스템과 함께 학습될 수 있도록 함

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

Learning to align and translate

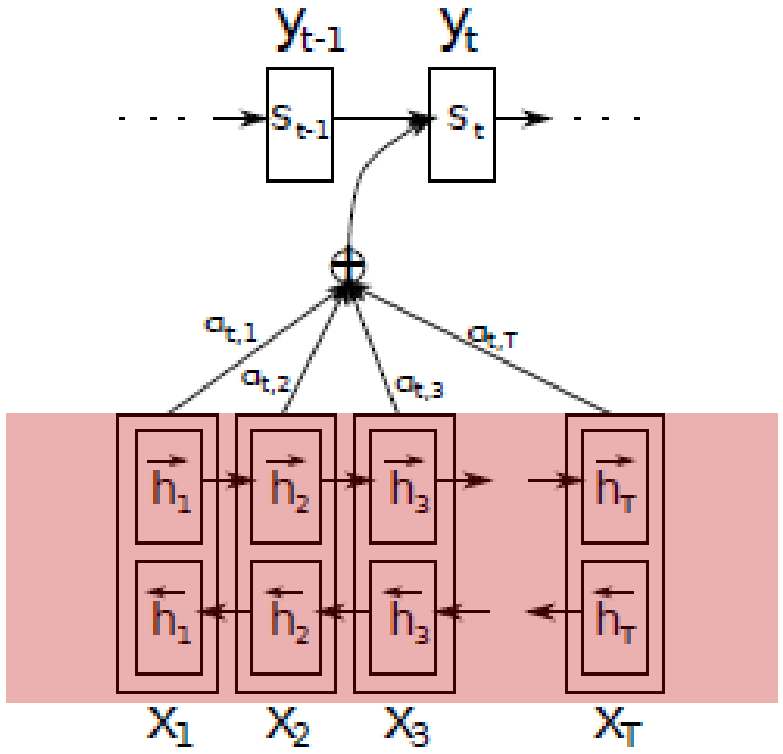


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

- 3.2 Encoder: bidirectional RNN for annotating sequences
  - 양방향 정보를 보유한 annotation 생성을 위하여 Bi-RNN을 사용함
    - 기존 encoder-decoder 모델은 주로 vanilla RNN을 사용함 (단방향 정보만 보유)
    - 각 단어 x_j에 대한 주석은 다음과 같음

      > Forward hidden state 및 backward hidden state의 결합

$$h_j = \left[\overrightarrow{h}_j^\top ; \overleftarrow{h}_j^\top\right]^\top$$

Experiment setting

- 4.1 Dataset
  - English-French parallel corpora data를 사용함
  - 토큰화 이후, 각 언어별 사용 빈도가 높은 어휘 30,000개를 사용하여 모델을 학습함
    - 30,000개에 포함되지 않는 단어는 특수 토큰 [UNK]에 매핑됨

- 4.2 Models
  - 다음 두 종류의 모델을 훈련 시킴
    - 기존 RNN Encoder-decoder 모델
    - 제안된 RNNsearch 모델
  - 각 모델을 문장 길이에 따라 두 번씩 훈련 시킴
    - 최대 문장 길이 30 단어: RNNencdec-30, RNNsearch-30
    - 최대 문장 길이 50 단어: RNNencdec-50, RNNsearch-50
  - Encoder, decoder는 각각 1,000개의 hidden unit으로 구성

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

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).

- 5.1 Quantitative results
  - BLEU scores로 측정된 번역 성능
  - 모든 경우 제안된 방법론이 기존 방법론에 비해 우수한 성능을 보임
  - 고빈도 30,000개 어휘로 구성된 문장만을 고려할 시, Moses에 상응하는 성능을 보임
    - Moses: 본 실험에서 사용한 데이터에 추가로 monolingual corpus 데이터를 학습에 사용함

Results

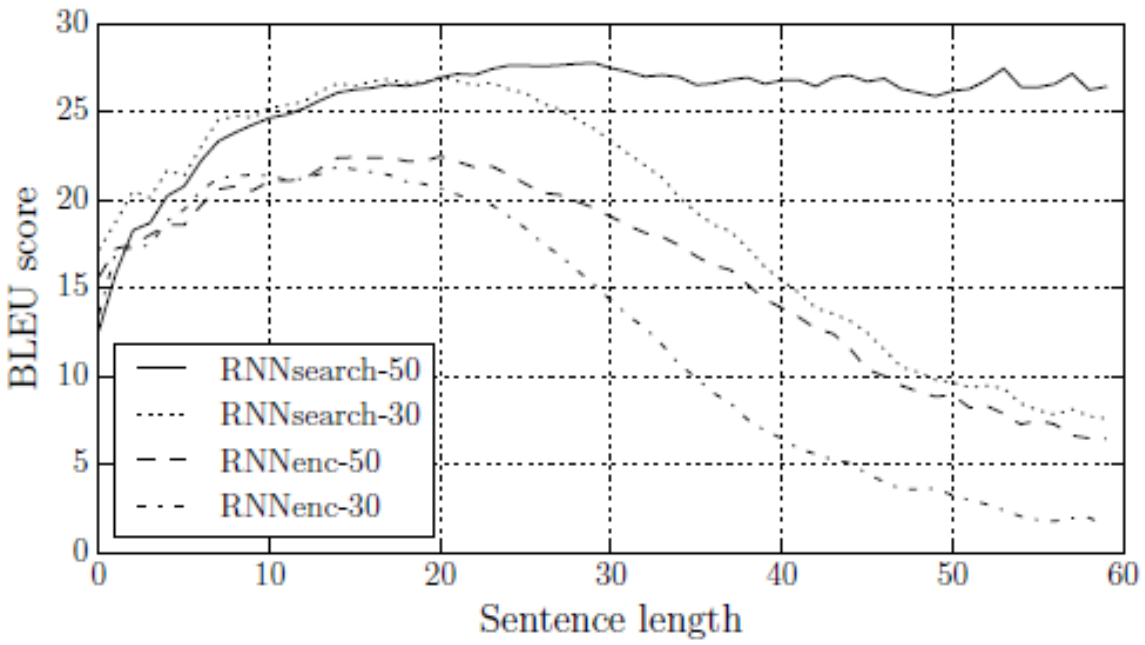


Figure 2: The BLEU scores of the generated translations on the test set with respect to the lengths of the sentences. The results are on the full test set which includes sentences having unknown words to the models.

- 5.1 Quantitative results
  - 문장의 길이 증가에 따라 RNNenc의 성능이 급격히 감소함이 확인됨
  - 제안된 모델은 길이가 50 이상인 문장에서도 성능 저하를 보이지 않음
  - 특히, RNNsearch-30이 RNNenc-50을 능가함

Results

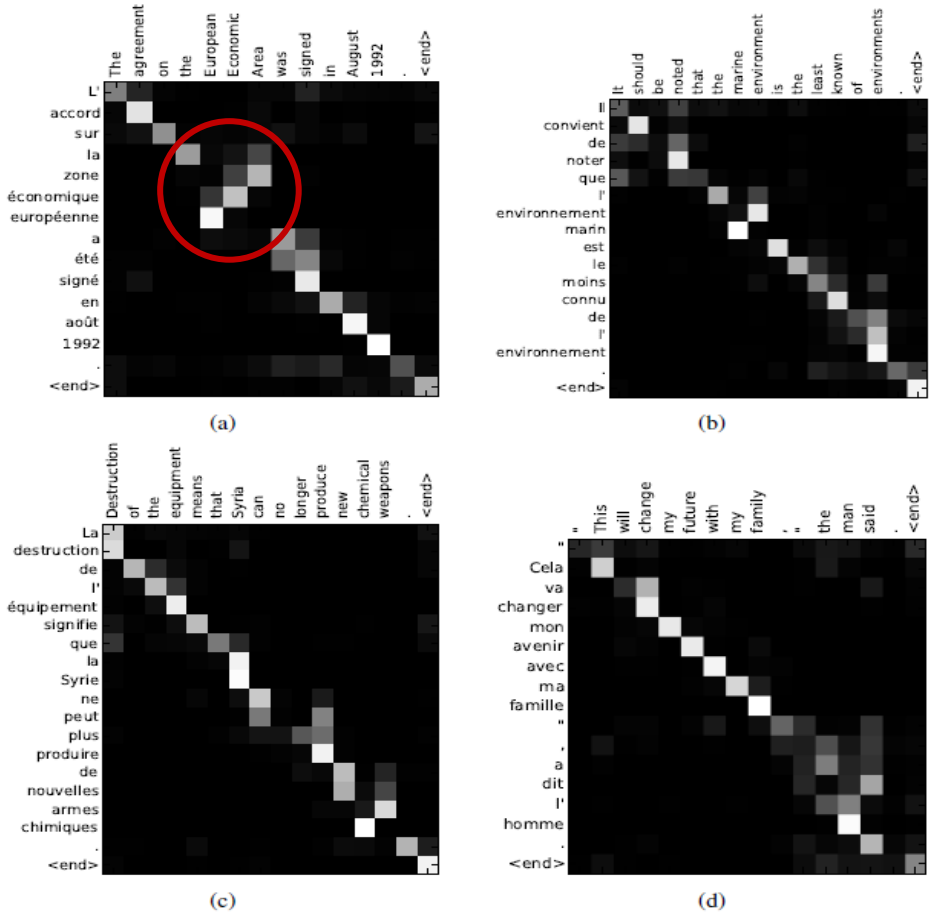


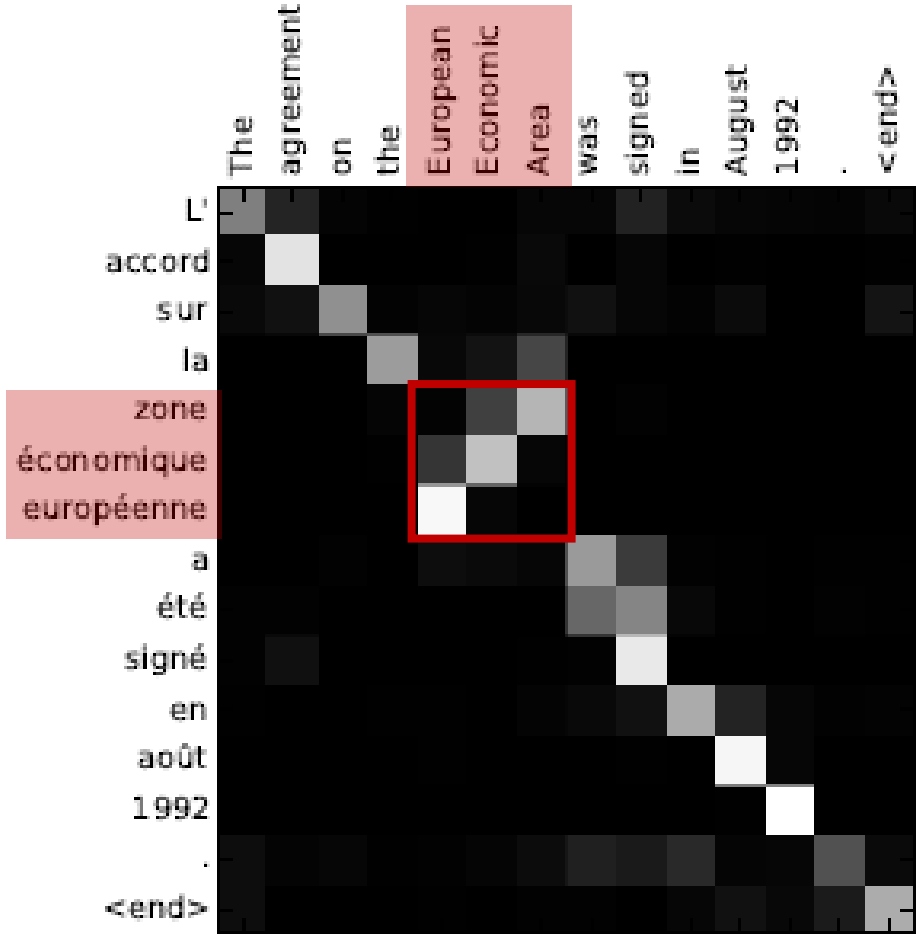
Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word (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
    - Soft alignment의 직관적 확인을 위하여 다음을 시각화 함

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

    - Target words 생성 시 source sentence 내에서 중요하게 판단된 영역을 확인함
      > 밝은 색일수록 가중치가 큰 값을 가지는 것 (단어 간 관련이 높은 것)
      > 영어-불어 간 alignment가 매우 단조로움

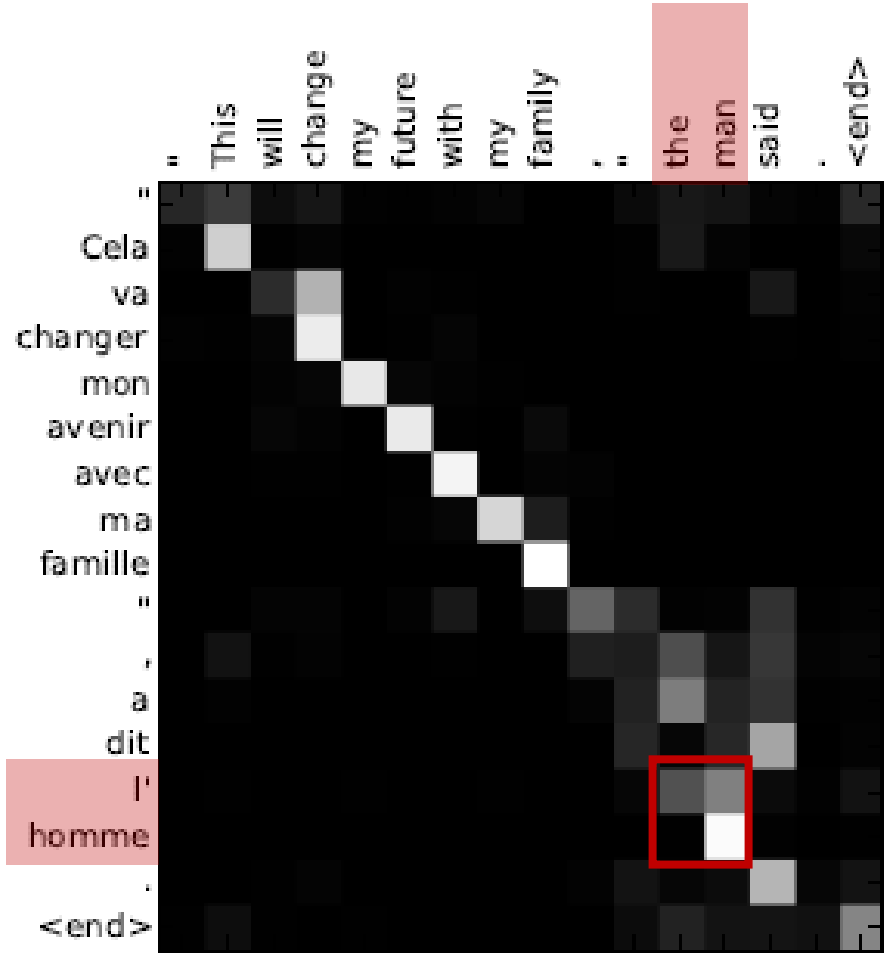
Results



(a)

- 5.2 Qualitative analysis
  - 5.2.1 Alignment
    - 영어: European Economic Area
    - 불어: zone économique européenne
    - 해당 영역의 alignments가 단조롭지 않음에도 불구하고 올바른 번역을 수행함

Results



(d)

- 5.2 Qualitative analysis
  - 5.2.1 Alignment
    - 영어: the man
    - 불어: l' homme
    - 관사 L' 를 생성하기 위하여 모델이 자동으로 the와 man을 모두 고려함

Results

The RNNencdec-50 translated this sentence into:

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

On the other hand, the RNNsearch-50 generated the following correct translation, preserving the whole meaning of the input sentence without omitting any details:

Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.

The translation by the RNNencdec-50 is

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.

As with the previous example, the RNNencdec began deviating from the actual meaning of the source sentence after generating approximately 30 words (see the underlined phrase). After that point, the quality of the translation deteriorates, with basic mistakes such as the lack of a closing quotation mark.

Again, the RNNsearch-50 was able to translate this long sentence correctly:

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.

- 5.2 Qualitative analysis
  - 5.2.2 Long sentences
    - RNNenndec-50 모델은 문장의 후반부 (약 30개 단어 이후) 번역의 질이 악화됨
      > 실제 의미에서 벗어남
      > 마감된 따옴표가 없는 등의 실수
    - RNNsearch-50 모델은 정상적으로 번역함

Related work

A similar approach of aligning an output symbol with an input symbol was proposed recently by Graves (2013) in the context of handwriting synthesis. Handwriting synthesis is a task where the model is asked to generate handwriting of a given sequence of characters. In his work, he used a mixture of Gaussian kernels to compute the weights of the annotations, where the location, width and mixture coefficient of each kernel was predicted from an alignment model. More specifically, his alignment was restricted to predict the location such that the location increases monotonically.

The main difference from our approach is that, in (Graves, 2013), the modes of the weights of the annotations only move in one direction. In the context of machine translation, this is a severe limitation, as (long-distance) reordering is often needed to generate a grammatically correct translation (for instance, English-to-German).

Our approach, on the other hand, requires computing the annotation weight of every word in the source sentence for each word in the translation. This drawback is not severe with the task of translation in which most of input and output sentences are only 15–40 words. However, this may limit the applicability of the proposed scheme to other tasks.

- 6.1 Learning to align
  - Handwriting synthesis로부터 aligning approach 제안됨 (Graves, 2013)
    - 특징: Annotations의 weight mode가 단 방향으로 제한됨
      > 기계 번역에서는 문법적으로 옳은 문장을 생성하기 위하여 (long-distance)reordering이 요구됨
    - 제안된 방법론: source sentence의 모든 단어에 대하여 annotation weight을 계산해야 함
      > 기계번역의 input, output 문장은 주로 15-40개 단어로 구성되어 있으므로 큰 한계는 아님
      > 그러나 다른 태스크에 적용하기에는 어려움이 존재함

Related work

Since Bengio *et al.* (2003) introduced a neural probabilistic language model which uses a neural network to model the conditional probability of a word given a fixed number of the preceding words, neural networks have widely been used in machine translation. However, the role of neural networks has been largely limited to simply providing a single feature to an existing statistical machine translation system or to re-rank a list of candidate translations provided by an existing system.

For instance, Schwenk (2012) proposed using a feedforward neural network to compute the score of a pair of source and target phrases and to use the score as an additional feature in the phrase-based statistical machine translation system. More recently, Kalchbrenner and Blunsom (2013) and Devlin *et al.* (2014) reported the successful use of the neural networks as a sub-component of the existing translation system. Traditionally, a neural network trained as a target-side language model has been used to rescore or rerank a list of candidate translations (see, e.g., Schwenk *et al.*, 2006).

Although the above approaches were shown to improve the translation performance over the state-of-the-art machine translation systems, we are more interested in a more ambitious objective of designing a completely new translation system based on neural networks. The neural machine translation approach we consider in this paper is therefore a radical departure from these earlier works. Rather than using a neural network as a part of the existing system, our model works on its own and generates a translation from a source sentence directly.

- 6.2 Neural network for machine translation
  - 초기 연구
    - 신경망이 기존 통계 기계 번역 시스템 내의 단일 기능 수행
    - 신경망이 기존 시스템에서 도출된 후보 번역 목록 재정렬
  - 제안된 방법론: 신경망 모델이 단일로 작동하며 소스 문장에서 직접 번역을 생성함

Conclusion

The conventional approach to neural machine translation, called an encoder–decoder approach, encodes a whole input sentence into a fixed-length vector from which a translation will be decoded. We conjectured that the use of a fixed-length context vector is problematic for translating long sentences, based on a recent empirical study reported by Cho *et al.* (2014b) and Pouget-Abadie *et al.* (2014).

In this paper, we proposed a novel architecture that addresses this issue. We extended the basic encoder–decoder by letting a model (soft-)search for a set of input words, or their annotations computed by an encoder, when generating each target word. This frees the model from having to encode a whole source sentence into a fixed-length vector, and also lets the model focus only on information relevant to the generation of the next target word. This has a major positive impact on the ability of the neural machine translation system to yield good results on longer sentences. Unlike with the traditional machine translation systems, all of the pieces of the translation system, including the alignment mechanism, are jointly trained towards a better log-probability of producing correct translations.

We tested the proposed model, called RNNsearch, on the task of English-to-French translation. The experiment revealed that the proposed RNNsearch outperforms the conventional encoder–decoder model (RNNencdec) significantly, regardless of the sentence length and that it is much more robust to the length of a source sentence. From the qualitative analysis where we investigated the (soft-)alignment generated by the RNNsearch, we were able to conclude that the model can correctly align each target word with the relevant words, or their annotations, in the source sentence as it generated a correct translation.

Perhaps more importantly, the proposed approach achieved a translation performance comparable to the existing phrase-based statistical machine translation. It is a striking result, considering that the proposed architecture, or the whole family of neural machine translation, has only been proposed as recently as this year. We believe the architecture proposed here is a promising step toward better machine translation and a better understanding of natural languages in general.

One of challenges left for the future is to better handle unknown, or rare words. This will be required for the model to be more widely used and to match the performance of current state-of-the-art machine translation systems in all contexts.

- 본 연구 요약
  - 기존 NMT encoder-decoder의 fixed-length context vector를 긴 문장의 번역 성공률 저하 요인으로 추측함
  - 다음과 같은 새로운 아키텍처를 제안함
    - 기존 encoder-decoder를 확장: 모델이 각 target word 생성 시 입력 단어의 집합 또는 인코더에 의해 계산된 주석을 검색하도록 함
      > 모델이 전체 source sentence를 고정 길이 벡터로 인코딩할 필요가 없음
      > Target word의 생성과 관련된 정보에만 집중 가능함
    - 기존 기계 번역 시스템과 달리 alignment를 포함한 번역 시스템이 전체(단일 모델)가 공동으로 훈련됨
  - 실제로 (특히 긴 문장에서) 우수한 번역 성능을 달성함
  - 향후 Unknown, rare 단어 또한 잘 활용할 수 있어야 함

Neural machine translation by jointly learning to align and translate

Bahdanau, D., Cho, K., & Bengio, Y. (2014)