자연어처리 논문 리뷰

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

(https://arxiv.org/abs/1409.0473)

발표자: 박규봉

목차

ABSTRACT

- Neural Machine Translation(NMT) 기계어 번역으로 급부상
- 최신 제시된 NMT 모델은 fixed-length vector(fixed size representation)을 가진 encoder-decoder 형태
- fixed-length vector는 encoder-decoder 구조에서 bottleneck으로 판단되어 새로운 구조 제시
- 제시하는 아이디어는 source sentenc로부터 target word와 연관된 부분을 자동으로 찾아오는 것(soft search)
- 제시된 방법으로 영어-프랑스어 번역을 통해 실험 진행
- 질적 분석을 통해 제시된 아이디어를 통해 분석된 모델이 우리 직감과 일치한다는 것을 보여줌

1. INTRODUCTION + 2.RNN ENCODER-DECODER

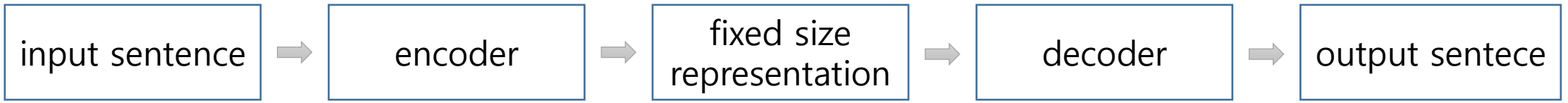
Word to Word translation

source sentence → target sentence
I → nan(난)
love → saranghey(사랑해)
you → nul(널)
I love you → nan saranghey nul

예외) How are you? = Jal jiney?
Word to Word로 번역할 시 입력값과 출력값의 수는 동일하지 않음

seq to seq = encoder-decoders

Encoder Decoder Architecture or
Sequence to Sequence model



Context 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.
- 소스 문장이 주어지면 올바른 번역의 확률을 최대화하기 위해 공동으로 훈련

1. INTRODUCTION + 2. RNN ENCODER-DECODER

y: target sentence
x: source sentence
y: 조건부 확률

$$argmax_y p(y|x)$$

NMT 전형적으로 두 부분으로 분리
- sentence x를 encod하는 부분
- target sentence y를 decode하는 부분

전형적으로 RNN을 통해 이와 같은 부분들을 수행

RNN Encoder-Decoder

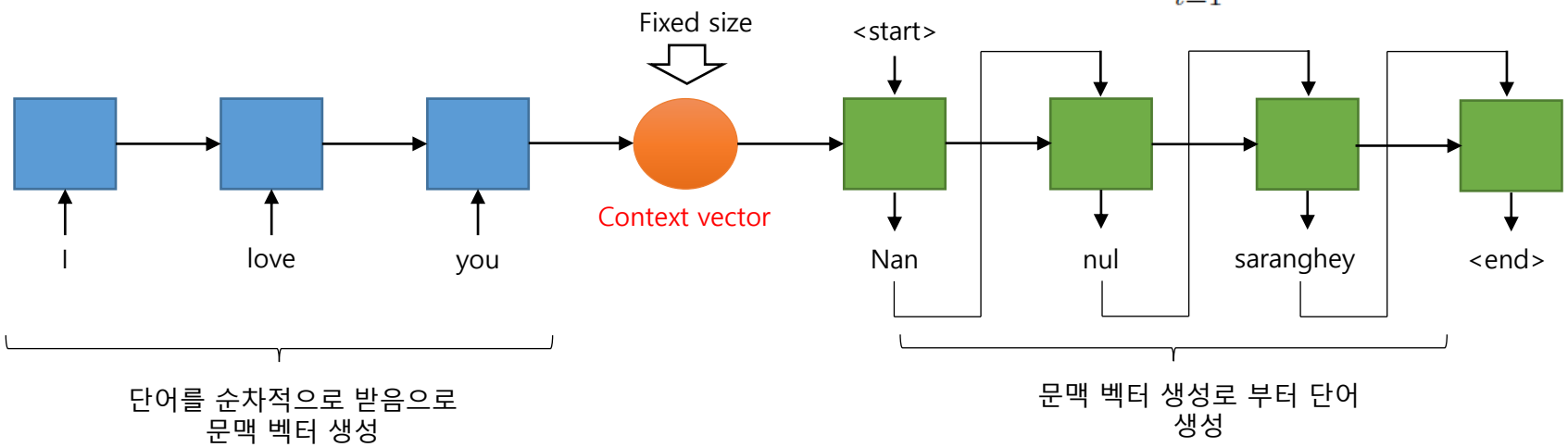
input sentence → x = (x_1, x_2, … , x_Tx) → vector c

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

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

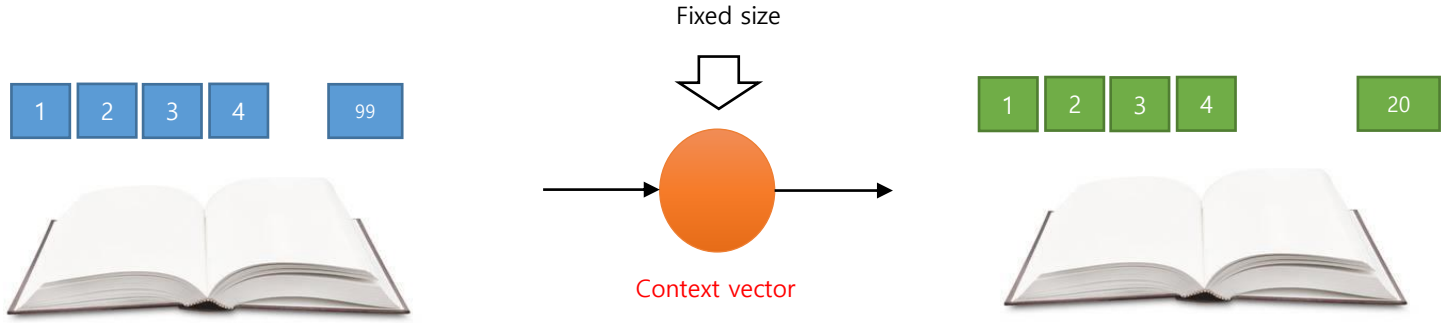
decoder는 벡터 c와
이전 y들을 가지고 다음 word를 추측

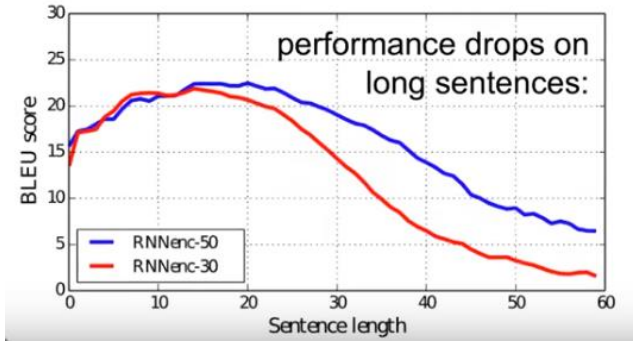
$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c),$$



1. INTRODUCTION + 2. RNN ENCODER-DECODER

RNN Encoder-Decoder의 단점
문맥 벡터가 고정된 사이즈로 문장의 길이가 길어지면 성능이 떨어짐





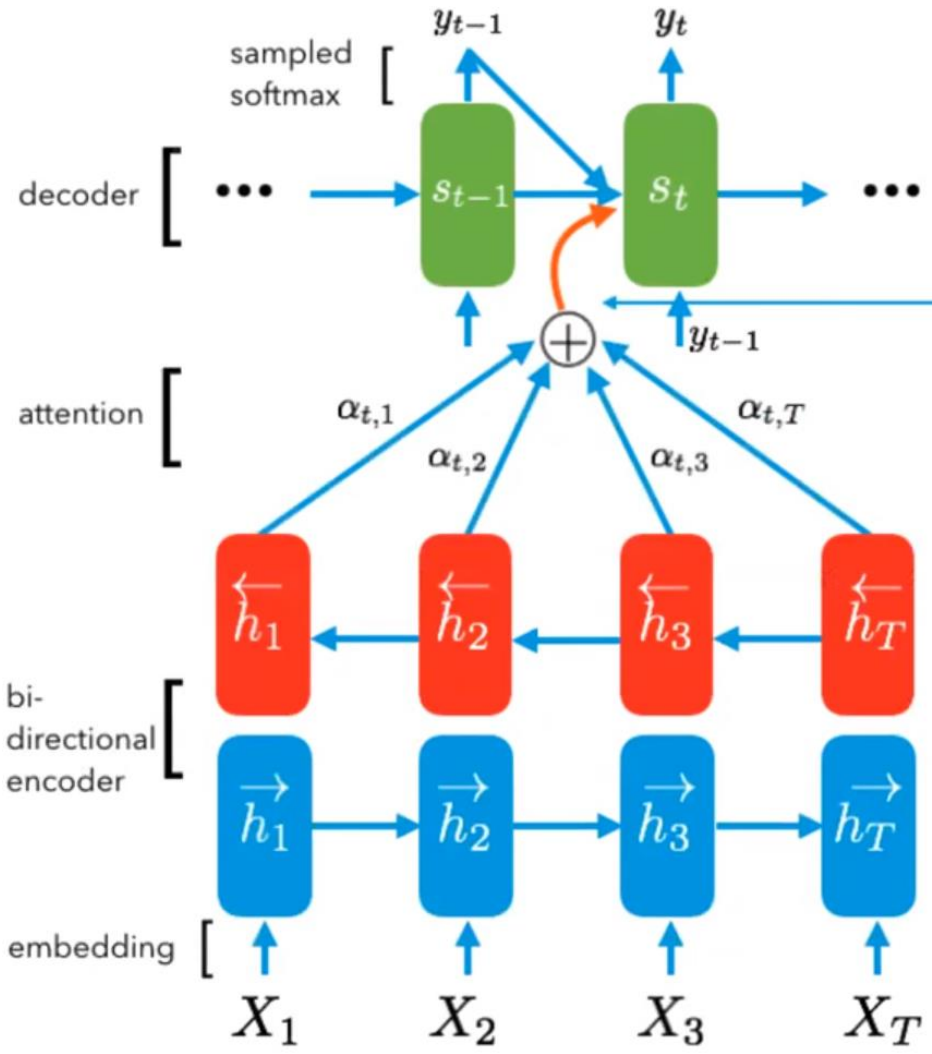
기계번역을 하다보면 충분치 않은 번역이 이루어 진다

새로 제시되는 모델

- encoder-decoder 모델의 확장
- 번역+문장배열 함께 학습
- source sentence에서 가장 관련성 높은 정보가 집중되어 있는 일련의 위치를 찾음(soft-)search
- 소스 위치 및 이전에 생성 된 모든 대상 단어(target word)와 연관된 컨텍스트 벡터를 기반으로 대상 단어를 예측
- 기존 모델과 가장 큰 차이점은 **하나의 single fixed-length vector로 encode하지 않는 다는 것**
- 제시된 모델은 input sencence들을 sequence of vector들로 encod한 후, decoding하는 동시에 이 vector들의 일부를 동적으로 고른다
- 이렇게 함으로써 원본 문장을 고정 길이 벡터로 만들 필요가 없다

3. LEARNING TO ALIGN AND TRANSLATE
new architecture: bidirectional RNN(as a encoder) + soft-search(decoder)

3.1 Decoder: Genenral Description



$$p(y_i|y_1,...,y_{i-1},X) = g(y_{i-1},s_i,c_i)$$

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

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

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

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

where a is a feed forward neural network.

$$h_j = [\rightarrow h_j; \leftarrow h_j]_{concat}$$

vector c_i는 encoder가 input sentence을 맵핑하는 일련의 annotation에 의존적

각 annotation h_i는 입력 시퀀스의 i 번째 단어를 둘러싼 부분에 집중하여 전체 입력 시퀀스에 대한 정보를 포함

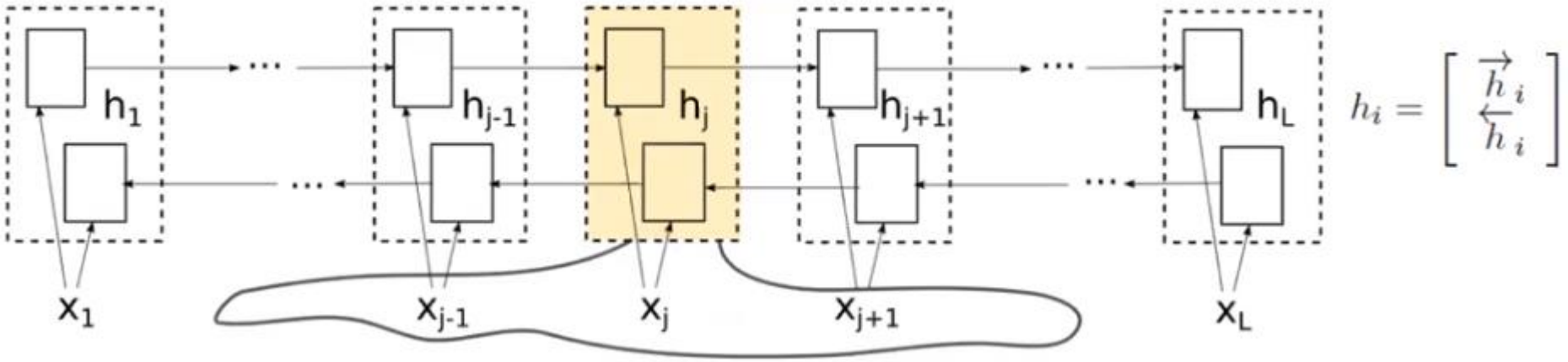
정렬 모델a
j 주변의 입력과 i에서의 출력이 얼마나 잘 일치하는지 평가

the alignment is not considered to be a latent variable. Instead, the alignment model directly computes a soft alignment, which allows the gradient of the cost function to be backpropagated through. This gradient can be used to train the alignment model as well as the whole translation model jointly.

3. LEARNING TO ALIGN AND TRANSLATE
new architecture: bidirectional RNN(as a encoder) + soft-search(decoder)

3.1 Decoder: Genenral Description
The probability α_{ij} , or its associated energy e_{ij} , reflects the importance of the annotation h_j with respect to the previous hidden state s_{i-1} in deciding the next state s_i and generating y_i. **Intuitively, this implements a mechanism of attention in the decoder.**
직관적으로, 이것은 디코더에서 attention 메커니즘을 구현합니다.
이러한 메커니즘 덕분에 고정된 길이의 문맥 벡터를 만들 필요가 없어졌다!!

3.2 ENCODER: BIDIRECTIONAL RNN FOR ANNOTATING SEQUENCES
annotation에서 이전 단어뿐만 아니라 다음에 나올 단오도 요약하길 원함 → bidirectional RNN



4. EXPERIMENT SETTINGS
English-to-French 번역
ACL WMT ‘14
기존 RNN Encoder-Decoder 모델과 비교

4.1 DATASET

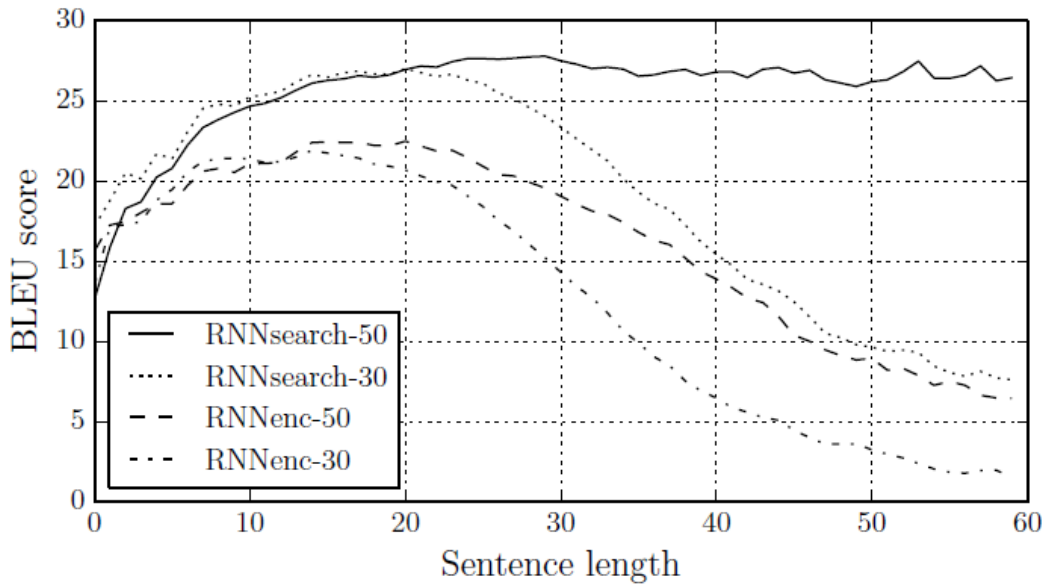
- Europarl (61M words), news commentary (5.5M), UN (421M) and two crawled corpora of 90M and 272.5M words respectively, totaling 850M words → **348M words**
- do not use monolingual data other than the mentioned parallel corpora
- concatenate news-test-2012 and news-test-2013 to make a development (validation) set, and evaluate the models on the test set (news-test-2014) from WMT ‘14
- useal tokenization 후에, 모델을 학습시키기 위해 각 언어로 **30,000 개의 가장 빈번한 단어 목록(shortlist)**을 사용
- shortlist에 포함되지 않는 단어는 special token(UNK)으로 분류됨

4.2 MODELS

- RNN Encoder–Decoder
- RNNsearch(논문이 제시한 모델)
- 문장 길이-30, 문장 길이-50 두 개를 학습

5. RESULTS

5.1 QUANTITATIVE RESULTS



아주 오랫동안 학습하고 모르는 단어를 제외했더니 Moses랑 유사한 성능이 나오더라

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

RNN Enc-Dec 보다 성능이 훨씬 뛰어나다

5. RESULTS

5.2 QUALITATIVE ANALYSIS

5.2.1 ALIGNMENT

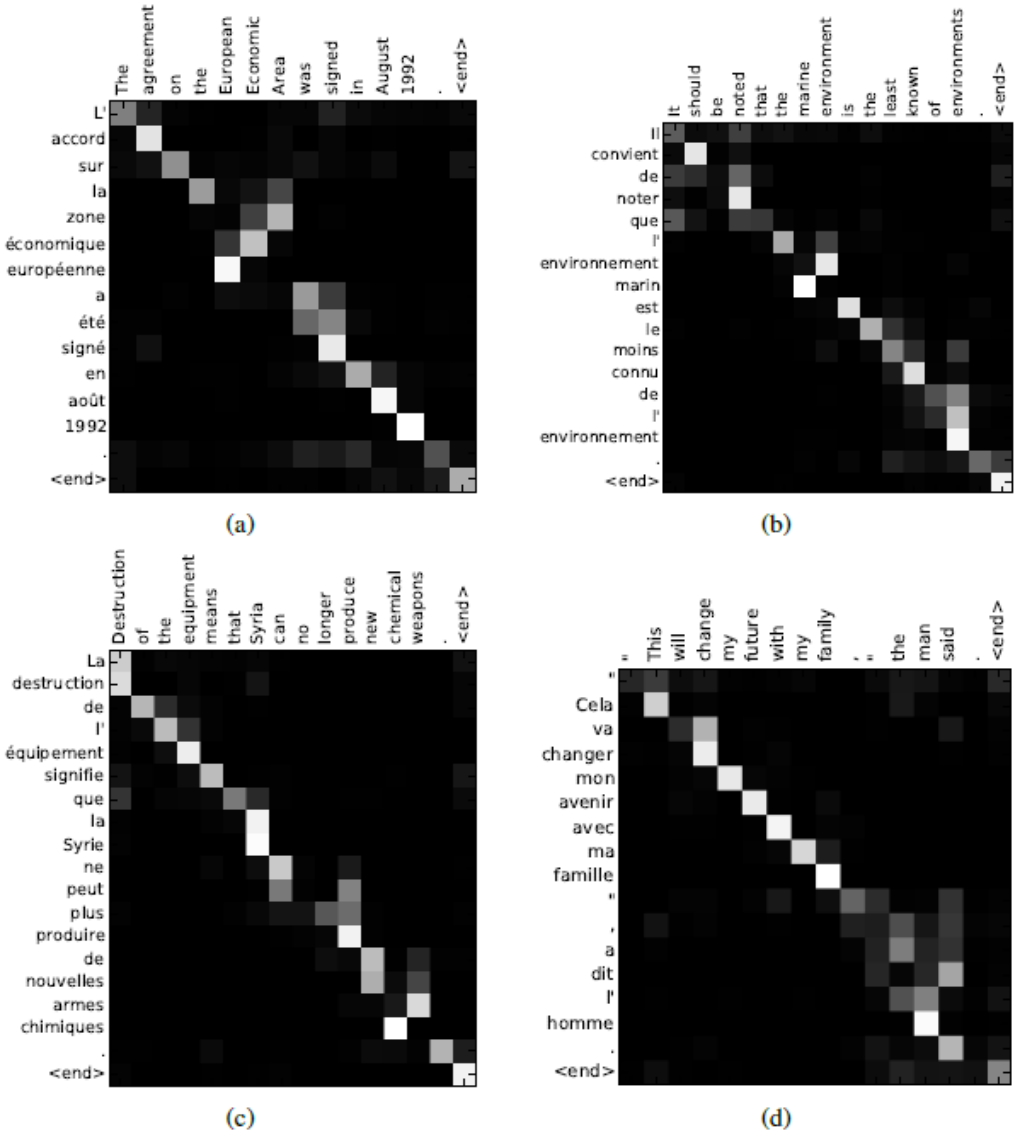


Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word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

(a) 영어와 프랑스어 사이 형용사와 명사의 위치는 전반적으로 반대로 위치

(b) 불어는 정관사가 여러 개라서 다음 단어를 보고 추리해야 되는데 해당 논문은 그것을 잘 찾더라

5. RESULTS

5.2 QUALITATIVE ANALYSIS
5.2.2 LONG SENTENCES
직접 예시 보여줌

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

New Model
RNNsearch-50

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.

Encoder-Decoder
RNNencdec-50

.... d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

6. RELATED WORK

Effective Approaches to Attention-based Neural Machine Translation(Luong, 2015)

- 동일한 biderection RNN + Attention
- https://github.com/tensorflow/nmt

seq to seq with attention code
https://colab.research.google.com/github/tensorflow/tensorflow/blob/master/tensorflow/contrib/eager/python/examples/nmt_with_attention/nmt_with_attention.ipynb

Thank you