Neural Machine Translation By Jointly Learning To Align And Translate

(Bahdanau et al. 2015)

- Introduction
- Background
- Attention
- Experiments
- Results
- Related Works

Introduction

- Machine translation uses encoder-decoder models
- The model need to encode all information in a fixed-size vector
- The fixed-size vector is the bottleneck of the model
- The model should learn which part of the input sentence is important for a translation

Background

RNN encoder-decoder

- Encoder turns an input sequence x=(x_1,.., x_{T_x}) into a fixed-size vector c
  - $c = q(\{h_1, h_2, .., h_{T_x}\})$
  - $h_t = f(x_t, h_{t-1})$
- Decoder turns a fixed-size vector c into a sequence y
  - $p(y) = \prod_{t=1}^{T} p(y_t | \{y_1, .., y_{t-1}\}, c)$
  - g(y_{t-1}, s_t,c)
  - $s_t = f(s_{t-1}, y_{t-1}, c)$

Attention

- The Encoder is a bidirectional RNN
  - Reads the input sequence forward and backwards
  - Hidden states are concatenated $h_j = \left[h_{>j}, h_{<j}\right]$

Attention

- Decoder $g(y_{i-1}, s_i, c_i)$
- $s_i = f(s_{i-1}, y_{i-1}, c_i)$
- $c_i = \sum_{j=1}^{T_x} \alpha_{ij}\, h_j$
- $\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$
- $e_{ij} = a(s_{i-1}, h_j)$

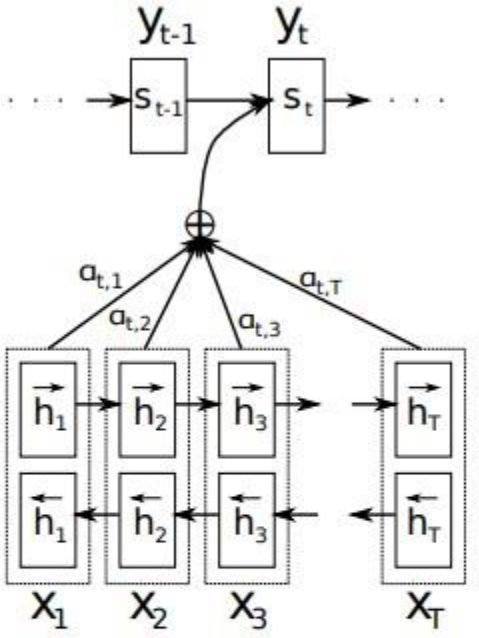


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Dataset

- Bilingual, parallel corpus from ACL WMT 14
- List of 30000 most frequent words in each language as dictionary
- Words not on the list are mapped to an „unknown“ token
- No other preprocessing is used

Experiments

- One traditional encoder-decoder model (RNNencdec)
  - Encoder and decoder have 1000 hidden units
- One encoder-decoder with attention (RNNsearch)
  - Forward and backward encoder have 1000 hidden units each
  - Decoder has 1000 hidden units
  - Attention MLP is single layer
- Models are trained twice, once with sentences of up to 30 words and once with sentences of up to 50 words

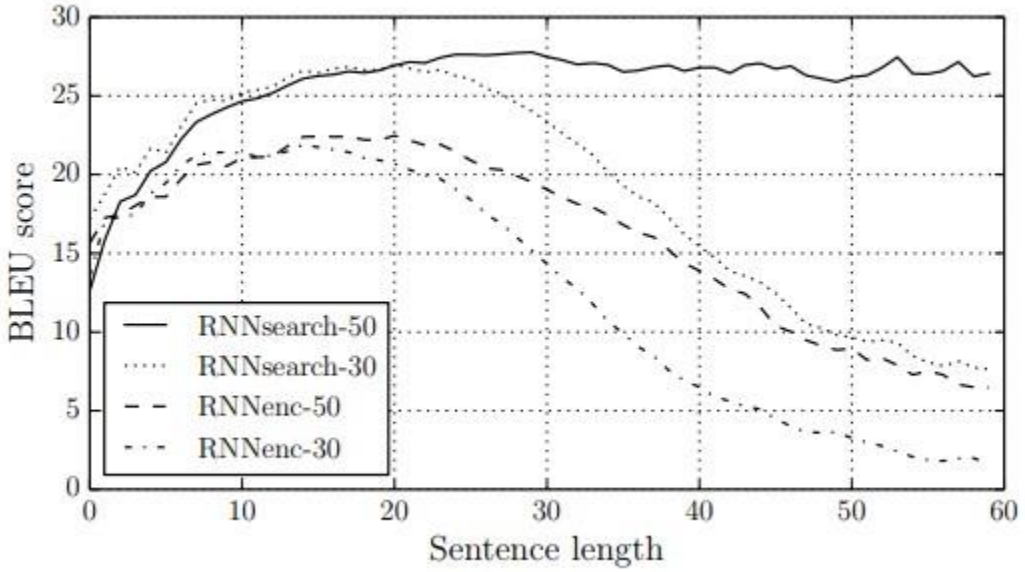
Quantitative Results

- RNNsearch outperforms RNNencdec
- Comparable to phrase-based system
- Better if unknown words are removed from the test set

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

Quantitative Results

- RNNsearch-50 performs consistent for growing sequence lengths
- RNNsearch-30 is similar to RNNencdec



Translations

Source sentence:

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.

RNNencdec50:

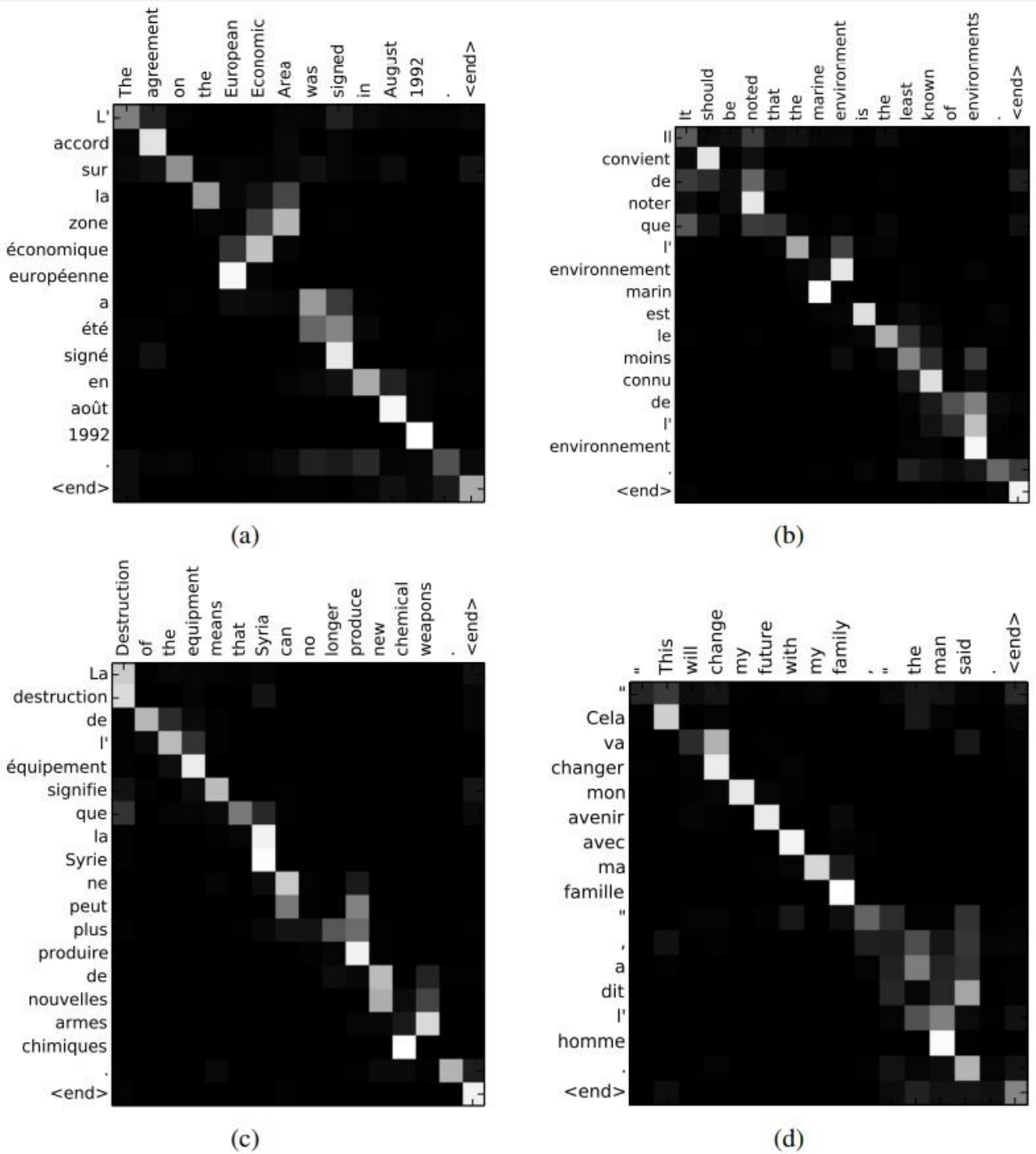
Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

RNNsearch50:

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.

Qualitative Results

- Weights visualize soft alignments
- Words are aligned to several words in the output



Related Works

- Transformer
- Attention in CNNs