Simultaneous Neural Machine Translation

Ashkan Alinejad

Supervisor : Dr. Anoop Sarkar

Simon Fraser University

July 2017

Neural Machine Translation

RNN structure

- Encoder

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

- Decoder

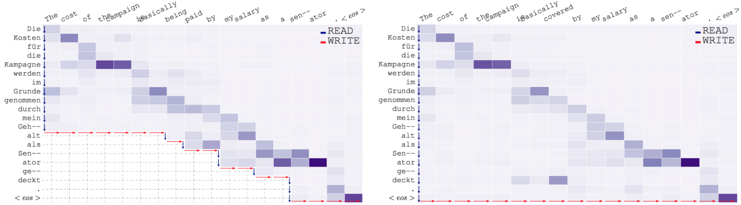
$$p(y) = \prod_{t=1}^{T} p(y_t | \{y_1, \ldots, y_{t-1}\}, c)$$

With an RNN, each conditional probability is modeled as:

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

Simultaneous Machine Translation

- Simultaneous Machine Translation is a challenging task of reading from the source language and at the same time, producing the target translation.
- The objective of translation system is defined as a combination of quality and delay.



(a) Simultaneous Neural Machine Translation (b) Neural Machine Translation

Previous works

- Most of the works in this direction are done in the context of speech translation. incoming speech is transcribed and segmented into a translation unit largely based on acoustic and linguistic cues.
- Each of these segments is then translated largely independent from each other

Neural SMT

- Sequentially making two interleaved decisions:
  1. READ
  2. WRITE
-

$$\begin{aligned} \text{Input sequence} \quad & X = \{x_1, \ldots, x_{T_s}\} \\ \text{Decoded Output} \quad & Y = \{y_1, \ldots, y_{T_t}\} \\ \text{Action sequence} \quad & A = \{a_1, \ldots, a_T\} \end{aligned}$$

$$T = T_s + T_t$$

- The model structure is an attention-based neural network

$$
\begin{aligned}
\text{Encoder} \quad &: \quad h_\eta = \phi_{\textsf{UNI-ENC}}(h_{\eta-1}, x_\eta) \\
\text{Decoder} \quad &: \quad c_\tau^\eta = \phi_{\textsf{ATT}}(z_{\tau-1}, y_{\tau-1}, H^\eta) \\
&\quad\;\; z_\tau^\eta = \phi_{\textsf{DEC}}(z_{\tau-1}, y_{\tau-1}, c_\tau^\eta) \\
\text{Output} \quad &: \quad p(y|y_{<\tau}, H^\eta) \propto \exp[\phi_{\textsf{OUT}}(z_\tau^\eta)] \\
&\quad\;\; y_\tau^\eta = \arg\max_y p(y|y_{<\tau}, H^\eta)
\end{aligned}
$$

Algorithm 1 Simultaneous Greedy Decoding

Require: δ, s_0, Input Pipe X, Output Pipe Y

Initialize $s \leftarrow s_0$, $C \leftarrow \text{READ}(X, s)$, $C' \leftarrow \{\}$
Initialize the decoder's state $\mathbf{z}_0$ based on C
while true do
    $\hat{y}_t = \arg\max_{y_t} \log p(y_t | y_{<t}, C)$
    if $s \geq T_X$ **then**
        $\text{WRITE}(Y, \hat{y}_t)$, $t \leftarrow t + 1$
    else
        $C' \leftarrow \text{READ}(X, \delta)$ **if** $|C'| = 0$.
        if $\Lambda(C, C \cup C')$ **then**
            $C \leftarrow C \cup C'$, $s \leftarrow s + \delta$, $C' \leftarrow \{\}$
            continue
        else
            $\text{WRITE}(Y, \hat{y}_t)$, $t \leftarrow t + 1$
        end if
    end if
    if $\hat{y}_t = \langle \text{eos} \rangle$ **then**
        break
    end if
end while

- **Wait-If-Worse**

$$\begin{aligned}\Lambda(C, C \cup C') \ &: (\log p(\hat{y}|\hat{y}_{<t}, C) \\ &> \log p(\hat{y}|\hat{y}_{<t}, C \cup C')),\end{aligned}$$

where $\hat{y} = \arg\max_{y} p(y|\hat{y}_{<t}, C)$

- **Wait-If-Diff**

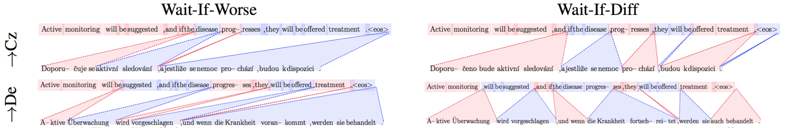
$$\Lambda(C, C \cup C') : (\hat{y} \neq \hat{y}'),$$

where $\hat{y}' = \arg\max_{y} \log p(y|\hat{y}_{<t}, C \cup C').$

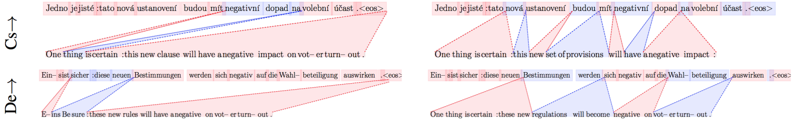
Metrics

- **Quality** The metrics for evaluating quality of the translation is the BLEU score.
- **Delay** $s(t) =$ In each time step for the decoded target symbol $\hat{y}_t$, how many source symbols were required. delay in translation (T):

$$0 < T(X, \hat{Y}) = \frac{1}{|X||\hat{Y}|} \sum_{t=1}^{|\hat{y}|} s(t) \leq 1.$$



(a) From English: "*Active monitoring will be suggested , and if the disease progresses , they will be offered treatment .*"



(b) To English: "*One thing is certain : these new provisions will have a negative impact on voter turn-out .*"

		Cs	De	Ru
En→	Ours	15.2	19.5	17.77
	⋆	13.84	21.75	19.54
→En	Ours	20.47	23.96	22.27
	⋆	20.32	24	22.44

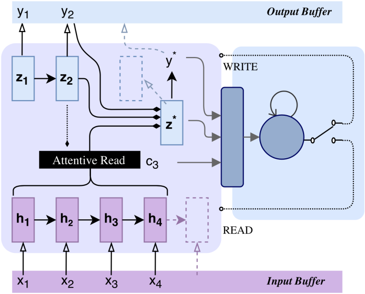
Figure: BLEU scores on the test set (newstest-2015) obtained by the models used in the paper and (⋆) from (Firat et al., 2016). Although our models use a unidirectional recurrent net as an encoder, the translation qualities are comparable.

Discussion

1. They do not have good BLEU score compared to previous works.
2. the waiting criteria proposed in this paper are both manually designed and does not exploit rich information embedded in the hidden representation learned by the recurrent neural networks.
3. The objective of the network is to improve translation quality and do not consider delay during training.

Trainable Agent

- The idea is to have a separate trainable agent
- The framework can be trained using reinforcement learning and it considers both Quality and Delay during training.



Algorithm 1 Simultaneous Greedy Decoding

Require: NMT system ϕ, policy π_θ, τ_{MAX}, input buffer X, output buffer Y, state buffer S.

Init $x_1 \Leftarrow X, h_1 \leftarrow \phi_{\text{ENC}}(x_1), H^1 \leftarrow \{h_1\}$

  $z_0 \leftarrow \phi_{\text{INIT}}(H^1), y_0 \leftarrow \langle s \rangle$

  $\tau \leftarrow 0, \eta \leftarrow 1$

while $\tau < \tau_{\text{MAX}}$ **do**

  $t \leftarrow \tau + \eta$

  $y_\tau^\eta, z_\tau^\eta, o_t \leftarrow \phi(z_{\tau-1}, y_{\tau-1}, H^\eta)$

  $a_t \sim \pi_\theta(a_t; a_{<t}, o_{<t}), S \Leftarrow (o_t, a_t)$

  if $a_t = \text{READ}$ and $x_\eta \neq \langle /s \rangle$ **then**

    $x_{\eta+1} \Leftarrow X, h_{\eta+1} \leftarrow \phi_{\text{ENC}}(h_\eta, x_{\eta+1})$

    $H^{\eta+1} \leftarrow H^\eta \cup \{h_{\eta+1}\}, \eta \leftarrow \eta + 1$

    if $|Y| = 0$ **then** $z_0 \leftarrow \phi_{\text{INIT}}(H^\eta)$

  else if $a_t = \text{WRITE}$ **then**

    $z_\tau \leftarrow z_\tau^\eta, y_\tau \leftarrow y_\tau^\eta$

    $Y \Leftarrow y_\tau, \tau \leftarrow \tau + 1$

    if $y_\tau = \langle /s \rangle$ **then break**

Agent

A trainable agent is designed to make decisions $A = \{a_1, \ldots, a_T\}$, $a_t \in \mathcal{A}$ sequentially based on observations $O = \{o_1, \ldots, o_T\}$, $o_t \in \mathcal{O}$.

- **Observation:** $o_{\tau+\eta} = [c_\tau^\eta; z_\tau^\eta; E(y_\tau^\eta)]$
- **Action:**
  - READ: waits to encode the next word
  - WRITE: accepts the candidate and emits it as the prediction
- **Policy:** a stochastic policy π_θ parameterized by a recurrent neural network

$$s_t = f_\theta(s_{t-1}, o_t),$$
$$\pi_\theta(a_t|a_{<t}, o_{\leq t}) \propto g_\theta(s_t)$$

Reward Function

At each step the agent will receive a reward signal r_t based on (o_t, a_t).

- **Quality** r_t^Q = smoothed BLEU
- **Delay** r_t^D
  1. **Average Proportion**
  2. **Consecutive Wait Length**

The total reward will be computed as:

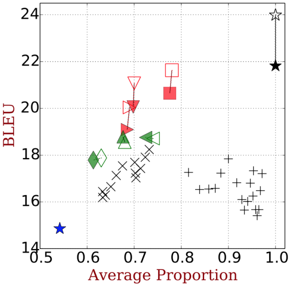
$$r_t = r_t^Q + r_t^D$$

Algorithm 2 Learning with Policy Gradient

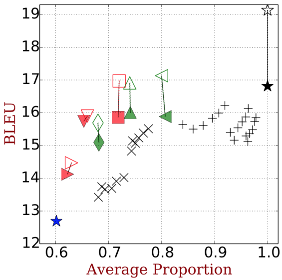
Require: NMT system ϕ, agent θ, baseline φ

```
Pretrain the NMT system φ using MLE;
Initialize the agent θ;
while stopping criterion fails do
    Obtain a translation pairs: {(X, Y*)};
    for (Y, S) ~ Simultaneous Decoding do
        for (o_t, a_t) in S do
            Compute the quality: r_t^Q;
            Compute the delay: r_t^D;
            Compute the baseline: b_φ(o_t);
    Collect the future rewards: {R_t};
    Perform variance reduction: {R̃_t};
    Update: θ ← θ + λ_1 ∇_θ [J − κH(π_θ)]
    Update: φ ← φ − λ_2 ∇_φ L
```

Results



(d) DE→EN



(c) EN→DE

(◀ ◁: CW=8, ▲△: CW=5, ◆◇: CW=2, ▶ ▷: AP=0.3, ▼▽: AP=0.5, ■□: AP=0.7). For each target, we select the model

Discussion

Yajie Miao, Mohammad Gowayyed, and Florian Metze.
Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding.
07 2015.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng.
Deep speech: Scaling up end-to-end speech recognition, 12 2014.

Alex Graves.
Supervised Sequence Labelling with Recurrent Neural Networks.
PhD thesis, Technische Universität München, July 2008.

Vinod Nair and Geoffrey E. Hinton.
Rectified linear units improve restricted boltzmann machines.
In Johannes Fürnkranz and Thorsten Joachims, editors,

Thank You !