Key Points of Text Automation Summarisation

Minquan Gao 2017. Sept. 14

Summarisation: Intro

- Definition: Summarisation is the task of condensing a piece of text to a shorter version that contains the main information from the original.
- Relate areas: Machine Translation, Speech Recognition.
- Historical and Pragmatic Approach: Extractive

Historical & Pragmatic: Extractive

- *Extractive models* assembles summaries exclusively from the source text.
- Advantages: 1. Ensuring baselines levels of grammaticality and accuracy; 2. Solve as an optimisation problem, could use many mathematical techniques; 3. Could embedded prior artifical knowledge.
    - e.g: *Text Rank, Sentence Embedding, etc*
- Disadvantage: 1. It's hard to generalise; 2. Could not paraphrasing, incorporation of real-word knowledge.
    - e.g: 1. *The algorithm of a current news could not fit for a storyline news;*
    - *2. '对方做出了不是很积极的应对'（13 words）==> '对方消极应对' (6 words)*

Abstractive framework

- Abstractive framework or generation framework, using the an agent to produce the summarisation by words generation.
- Approaches: 1. Sequence2Sequence Model, *by google machine translation*; 2. Attention, *attention is all you need, by google;* 3. Convolutional. *Convolutional Seq2Seq, Facebook;* 4. Customizable Affective GAN. *By CMU (ACL 2017)*
- Advantage: High-quality summarisation become **possible.**

Abstractive framework

- Main Approaches: Based on Sequence-2-Sequence Model, use an LSTM (or GRU) as encoder, and use another LSTM(or GRU) as decoder, to encode the sequence as an embedding *vector* then to decode a sequence from this embedding *vector*.

- Shortcoming: 1. Liable to reproduce factual details inaccurately; 2. They tend to repeat themselves; 3. Inability to deal with out-of-vocabulary (OOV) words.

Pointer-Generator Network

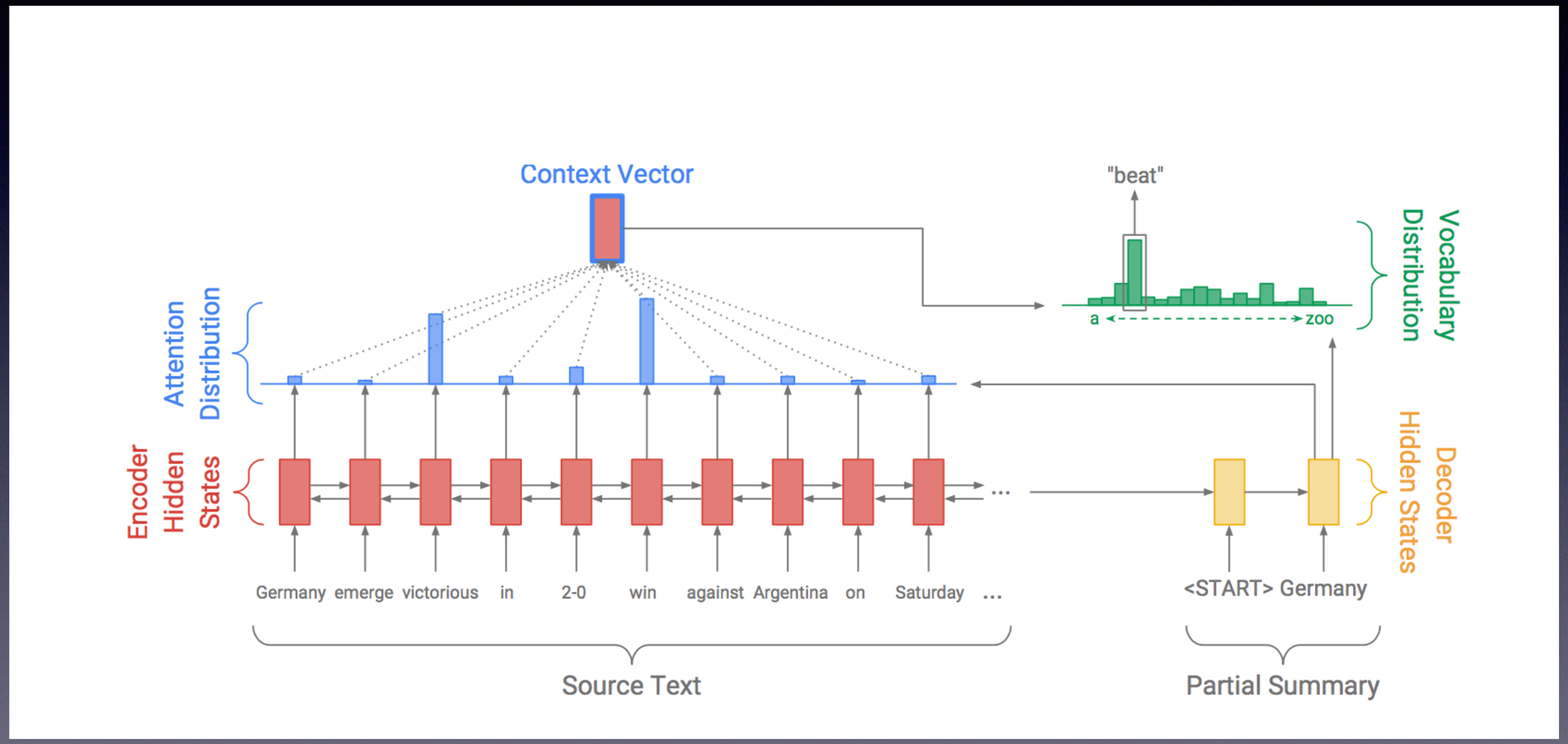
- A hybrid (abstractive-extractive) framework can copy words from the source text *via pointing;* Use *coverage* to keep track of what has been summarized. *Announced by: Stanford NLP Group, Google Brain. ACL 2017.*

- Output performing the current abstractive state-of-art by at least 2 ROUGE points.

  - ROUGE: Recall-Oriented Understudy for Gisting Evaluation, a metric for automatic summarization and machine translation. **FOR RECALL**

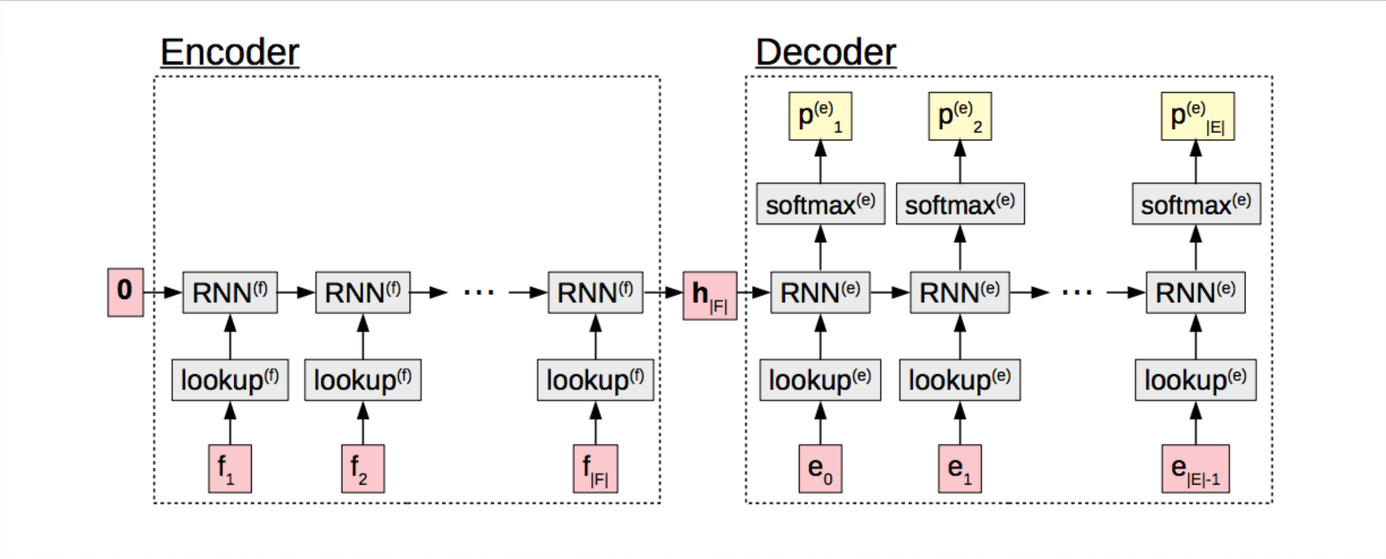
  - BIEU: Bilingual evaluation understudy. Evaluate how similar the candidate text is to the reference texts. **FOR ACCURACY**

Attention based seq2seq model



Attention based seq2seq model

Attention based seq2seq model



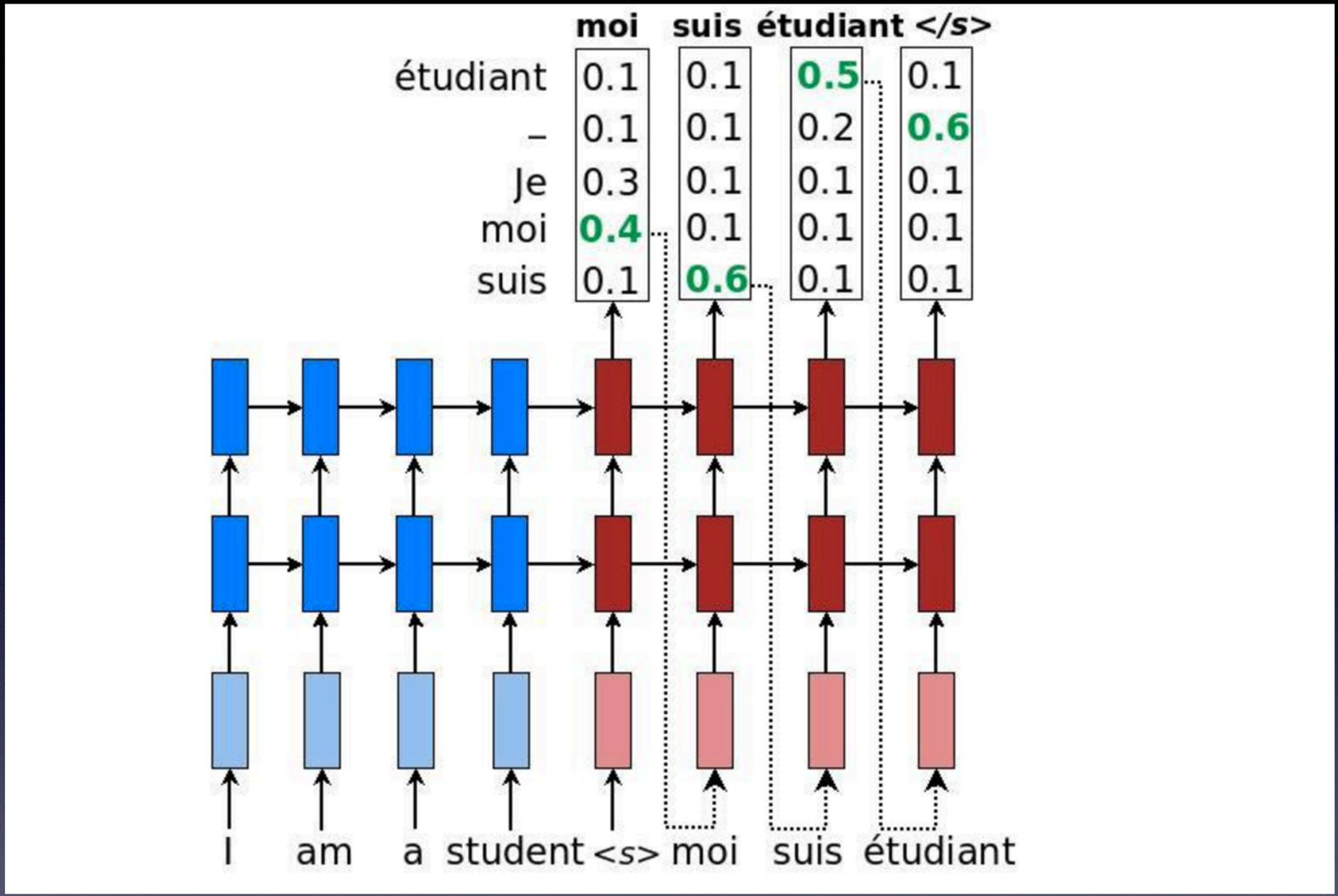
$$\boldsymbol{m}_t^{(f)} = M^{(f)}_{\cdot,f_t}$$

$$\boldsymbol{h}_t^{(f)} = \begin{cases} \mathrm{RNN}^{(f)}(\boldsymbol{m}_t^{(f)}, \boldsymbol{h}_{t-1}^{(f)}) & t \geq 1, \\ \boldsymbol{0} & \text{otherwise.} \end{cases}$$

$$\boldsymbol{m}_t^{(e)} = M^{(e)}_{\cdot,e_{t-1}}$$

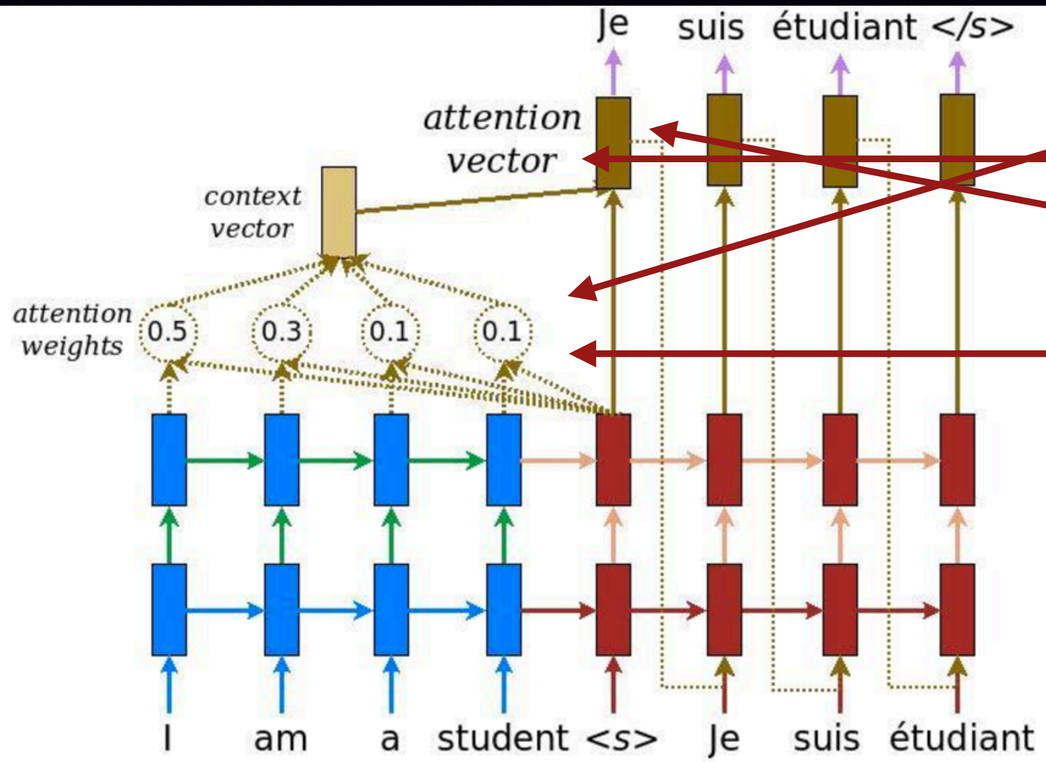
$$\boldsymbol{h}_t^{(e)} = \begin{cases} \mathrm{RNN}^{(e)}(\boldsymbol{m}_t^{(e)}, \boldsymbol{h}_{t-1}^{(e)}) & t \geq 1, \\ \boldsymbol{h}_{|F|}^{(f)} & \text{otherwise.} \end{cases}$$

$$\boldsymbol{p}_t^{(e)} = \mathrm{softmax}(W_{hs}\boldsymbol{h}_t^{(e)} + b_s)$$



1. Random Sampling: Randomly select an output E from the probability distribution P (E | F). This is usually denoted $\hat{E} \sim P(E \mid F)$.
2. 1-best Search: Find the E that maximizes P (E | F), denoted $\hat{E} = \text{argmax } P(E \mid F)$.
3. n-best Search: Find the n outputs with the highest probabilities according to P (E | F).

Attention Model



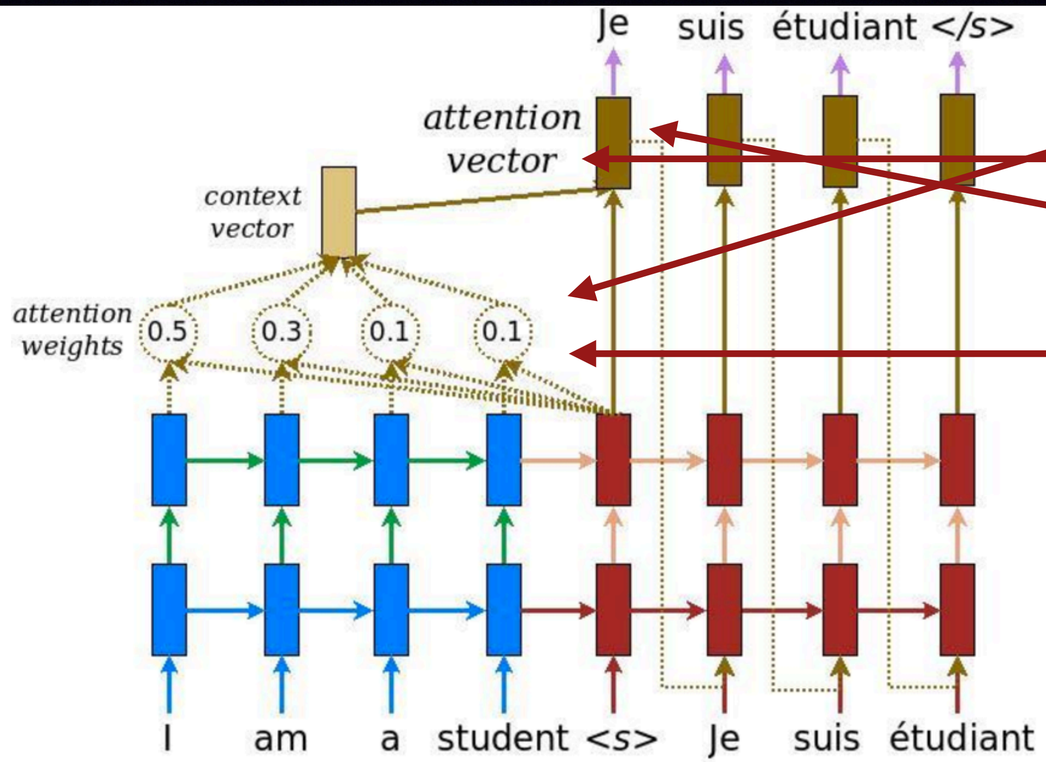
$$\alpha_{ts} = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \quad \text{[Attention weights]} \quad (1)$$

$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \quad \text{[Context vector]} \quad (2)$$

$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W}_{\boldsymbol{c}}[\boldsymbol{c}_t; \boldsymbol{h}_t]) \quad \text{[Attention vector]} \quad (3)$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_{\boldsymbol{1}} \boldsymbol{h}_t + \boldsymbol{W}_{\boldsymbol{2}} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases}$$

Attention Model



$$\alpha_{ts} = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \quad \text{[Attention weights]} \quad (1)$$

$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \quad \text{[Context vector]} \quad (2)$$

$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W}_c[\boldsymbol{c}_t; \boldsymbol{h}_t]) \quad \text{[Attention vector]} \quad (3)$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_1} \boldsymbol{h}_t + \boldsymbol{W_2} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases}$$

Attention Model: Example

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

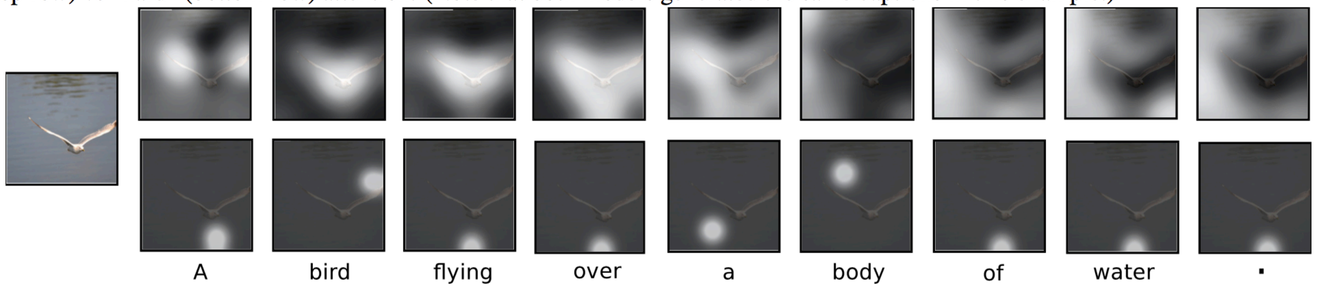
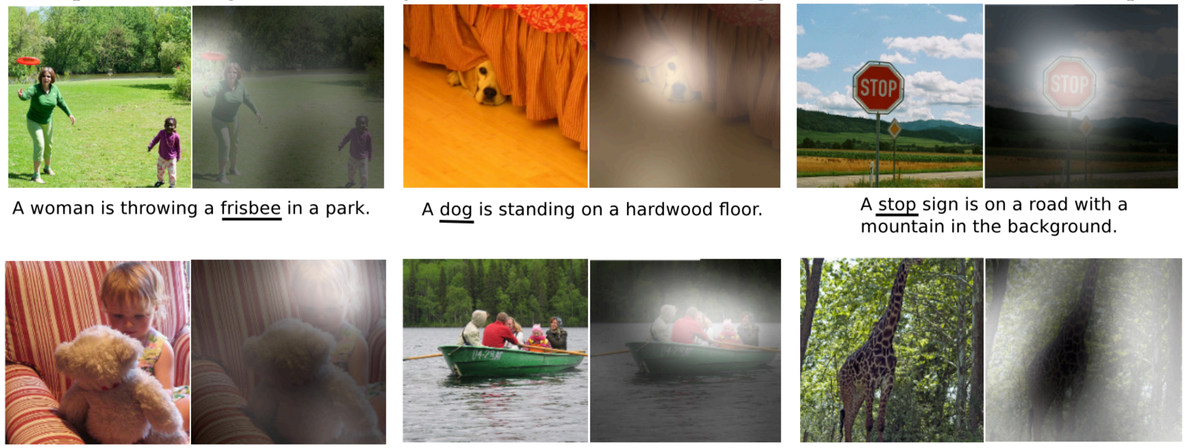
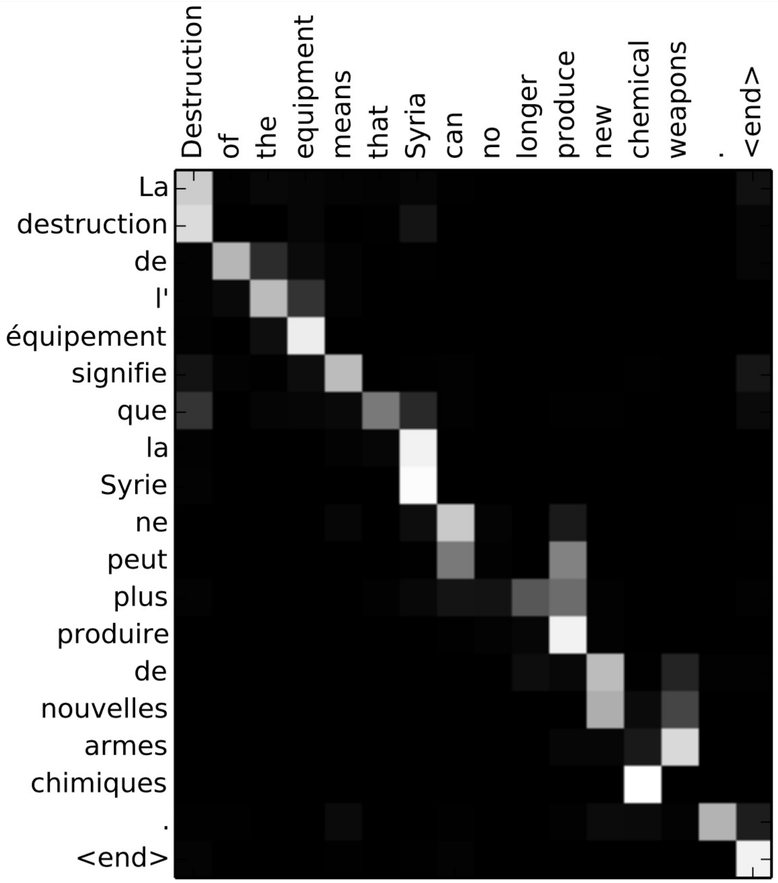


Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)





Pointer-copy

$$loss = \frac{1}{T}\sum_{t=0}^{T} loss_t,\ loss_t = -logP(W_t^*)$$

target_words

$$P(w) = softmax(f(s_t, h_t^*, bias, W))$$

Attention context

The tokens of article *Wi* are fed one-by-one into the encoder, producing a seq of *encoder hidden states h_i.* On the each step t, the decoder receives the word embedding of the previous. And has decoder state *s_t.*

$$h_t^* = \sum_i^T a_i{}^t h_i$$

Hidden state

The attention distribution can be viewed as a probability distribution over the source words, that tells the decoder where to look to produce the next word.

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}})$$
$$a^t = \text{softmax}(e^t)$$

Weight of time t

Pointer-copy

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t$$

p_gen is the generate probability. Such as if p_gen is 1, then the P(w) is definitely equal to P_vocab(W), which produced by decoder.
On the other hand, if p_gen is 0, the P(w) is equal to 0 + sum(1 * a_t^t), in which a_i is attention weight for word w. And the sum is iterated over an *extended vocabulary.* Which is all the words appeared in source document.

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$

p_gen is calculated by context vector h_t*, decoder state s_t, decoder_input x_t. Sigmoid let value to be [0, 1]

Pointer-copy

```
for i, output in enumerate(decoder_outputs):
vocab_scores.append(tf.nn.xw_plus_b(output, w, v))
vocab_dists = [tf.nn.softmax(s) for s in
               vocab_scores]  # The vocabulary distributions. List length max_dec_steps of
(batch_size, vsize) arrays. The words are in the order they appear in the vocabulary file.


final_dists = self._calc_final_dist(vocab_dists, self.attn_dists)

for dec_step, dist in enumerate(final_dists):
    targets = self._target_batch[:,
              dec_step]  # The indices of the target words. shape (batch_size)
    indices = tf.stack((batch_nums, targets), axis=1)  # shape (batch_size, 2)
    gold_probs = tf.gather_nd(dist,
                              indices)  # shape (batch_size). prob of correct words on this step
    losses = -tf.log(gold_probs)
    loss_per_step.append(losses)
```

Coverage Vector

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$

Sum of attention distribution over all previous decoder timesteps.

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}})$$

Change Attention vector

$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$

Penalize repeatedly attending to the same location.

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_i^t, c_i^t)$$

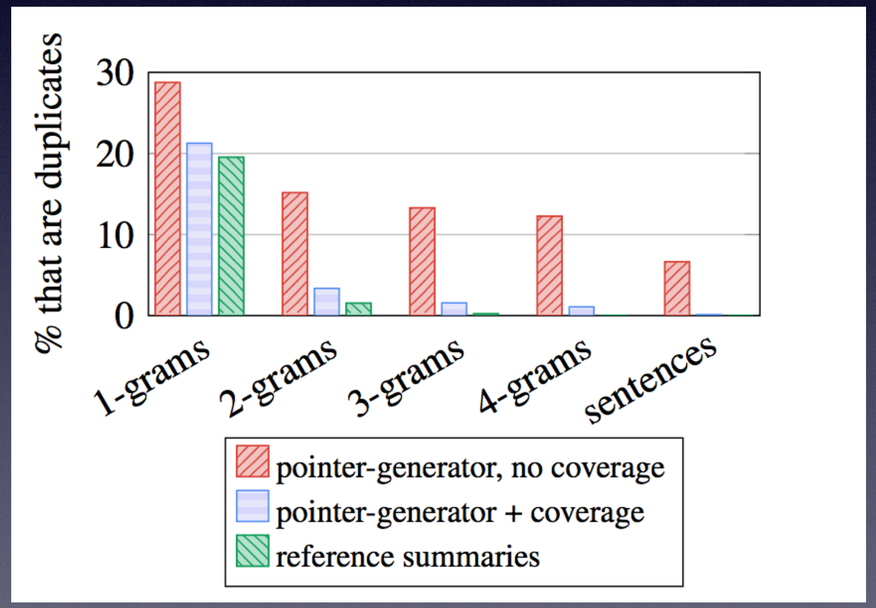
If there is no repeat in time t of word i. The sum(a_i^{0}, a_i^{1}, … a_i^{t-1}) will be very small. And the *covloss* should and would be very small.

Coverage Vector

```
coverage = tf.zeros_like(attn_dists[0])  # shape (batch_size,
attn_length). Initial coverage is zero.
covlosses = []  # Coverage loss per decoder timestep. Will be
list length max_dec_steps containing shape (batch_size).
for a in attn_dists:
    covloss = tf.reduce_sum(tf.minimum(a, coverage), [1])  #
calculate the coverage loss for this step
    covlosses.append(covloss)
    coverage += a  # update the coverage vector
coverage_loss = _mask_and_avg(covlosses, padding_mask)
return coverage_loss
```

Experiment

- 1. Introduce very few additional parameters.
- 2. Use gradient clipping with a maximum gradient norm of 2.
- 3. Use beam search.



Article (truncated): munster have signed new zealand international francis *saili* on a two-year deal . utility back *saili* , who made his all blacks debut against argentina in 2013 , will move to the province later this year after the completion of his 2015 contractual commitments . the 24-year-old currently plays for *auckland-based* super rugby side the blues and was part of the new zealand under-20 side that won the junior world championship in italy in 2011 . *saili* 's signature is something of a coup for munster and head coach anthony foley believes he will be a great addition to their backline . francis *saili* has signed a two-year deal to join munster and will link up with them later this year . ' we are really pleased that francis has committed his future to the province , ' foley told munster 's official website . ' he is a talented centre with an impressive *skill-set* and he possesses the physical attributes to excel in the northern hemisphere . ' i believe he will be a great addition to our backline and we look forward to welcoming him to munster . ' *saili* has been capped twice by new zealand and was part of the under 20 side that won the junior championship in 2011 .

Some Examples

Reference Summary:
utility back francis *saili* will join up with munster later this year .
the new zealand international has signed a two-year contract .
saili made his debut for the all blacks against argentina in 2013 .

Baseline:
dutch international francis UNK has signed a two-year deal to join **irish** UNK super rugby side the blues .
UNK 's signature is something of a coup for munster and his head coach anthony foley believes he will be a great addition to their **respective prospects** .
UNK has been capped twice by new zealand .

Pointer-Generator, No Coverage:
new zealand international francis *saili* will move to the province later this year .
utility back *saili* made his all blacks debut against argentina in 2013 .
utility back *saili* will move to the province later this year .

Pointer-Generator, With Coverage:
francis *saili* has signed a two-year deal to join munster later this year .
the 24-year-old was part of the new zealand under-20 side that won the junior world championship in italy in 2011 .
saili 's signature is something of a coup for munster and head coach anthony foley .

- Thanks !