Natural Language Generation

Better Decoding

Claire Gardent

Advanced Language Processing Winter School, 17-22 January 2021

Part 1: Basics of Neural NLG

The Encoder-Decoder Framework

- The Recurrent Encoder-Decoder

Better Decoding

- Attention, Copy and Coverage

Encoders

- Improved Recurrent Neural Network (RNN): LSTM, biLSTM, GRU
- Convolutional Neural Network (CNN)
- Transformer

Evaluation

- BLEU, PARENT, BLEURT, BERTScore
- ROUGE
- Evaluating faithfulness

Some Problems with Neural Generation

Accuracy/Faithfulness

- The output text sometimes contains information not present in the input.

Repetitions

- The output text sometimes contains repetitions

Coverage

- The output text sometimes does not cover all the input

Rare or Unknown Words

T2T Example: Summarisation

Lagos nigeria (CNN). a day after winning nigeria's presidency, muhammadu buhari told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation's unrest. buhari said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. buhari defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

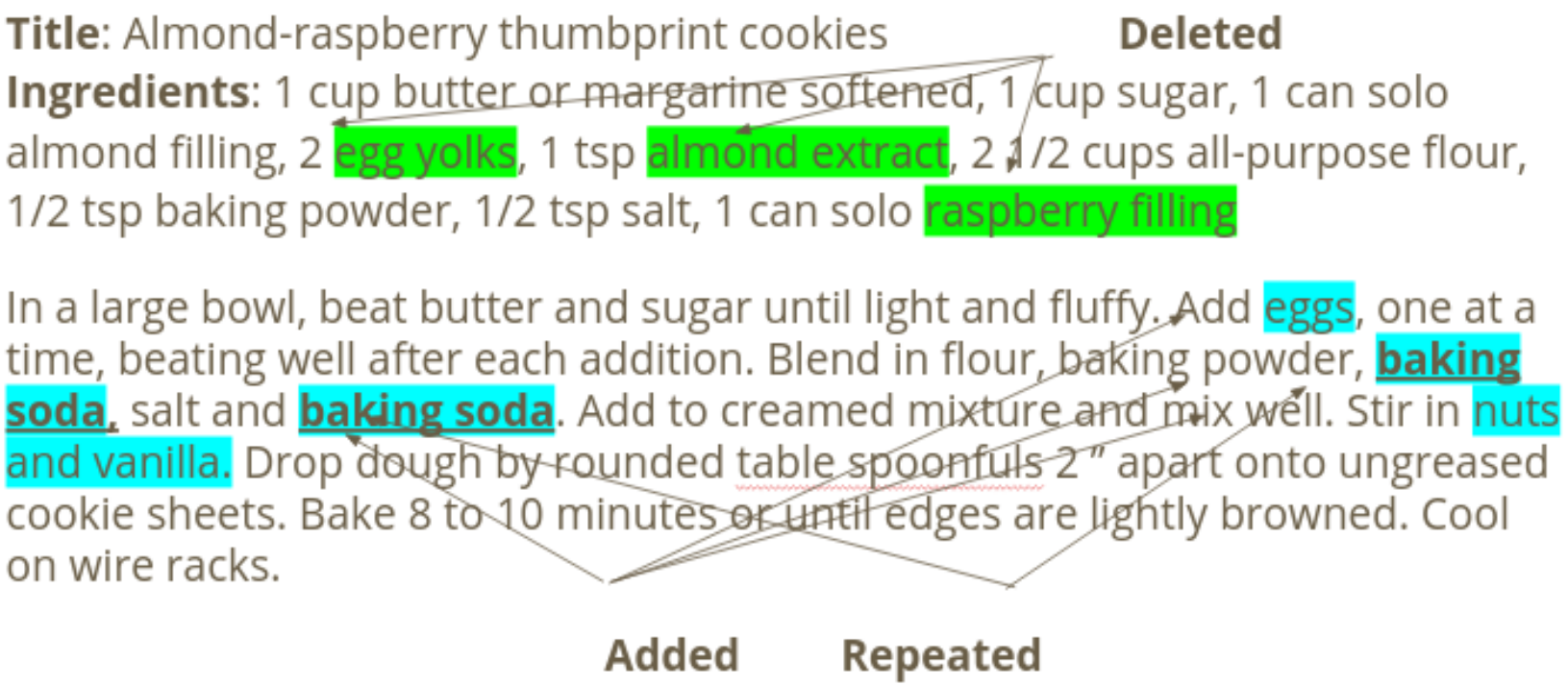
RARE WORD

ADDED

Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize nigeria's economy. UNK says his administration is confident it will be able to thwart criminals and other nigerians. he says the country has long nigeria and nigeria's economy.

Pointer-Gen: muhammadu buhari says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals. See et al. 2017

T2T Example: Generation



T2T Example: MR2T

```
state
 :arg0 ( person
  :arg0-of ( have-org-role
   :arg1 ( committee :mod technical )
   :arg3 ( expert
    :arg1 person
    :arg2 missile
    :mod loc_0 ) ) )
 :arg1 ( evidence
  :arg0 equipment
  :arg1 ( plan :arg1 ( transfer :arg1 ( contrast
   :arg1 ( missile :mod ( just :polarity - ) )
    :arg2 ( capable
     :arg1 thing
     :arg2 ( make :arg1 missile ) ) ) ) )
  :mod ( impeach :polarity - :arg1 thing )
  :mod ( refute :polarity - :arg1 thing ) )
```

DiSFLUENT

ADDED

REF: A technical committee of indian missile experts stated that the equipement was unimpeachable and irrefutable evidence of a plan to transfer not just missiles but missile-making capabilities.

DELETED

SYS: A technical committee expert on the technical committee stated that the equipment is not impeached but it is not refutes.

Three Ways to Improve Decoding

Attention

- To improve accuracy

Copy

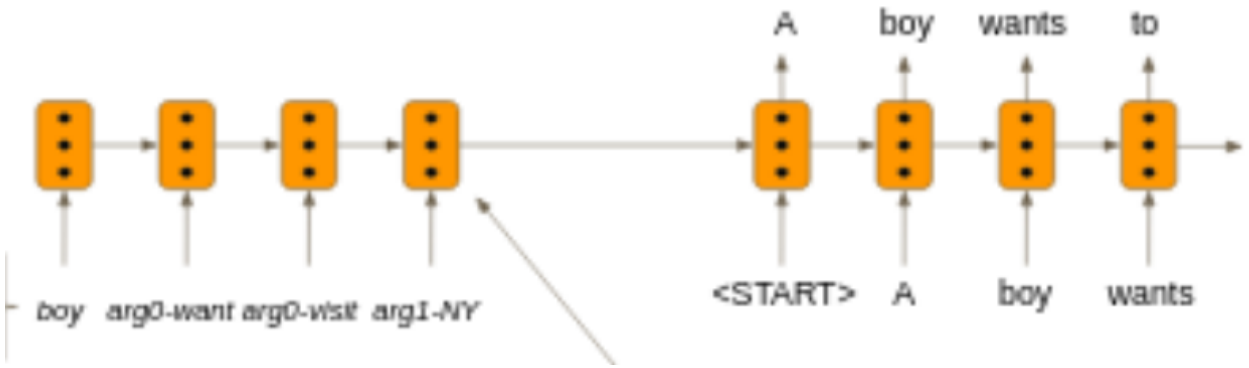
- To copy from the input
- To handle rare or unknown words

Coverage

- To help cover all and only the input
- To avoid repetitions

Attention

Standard RNN Decoding



- The input is compressed into a fixed-length vector
- Performance decreases with the length of the input

Sutskever et al. NIPS 2014

Decoding with Attention

Input

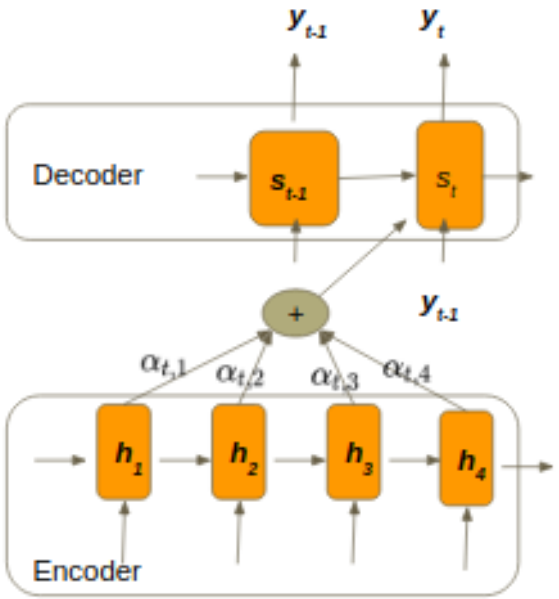
- the previous state s_{t-1}
- the previously generated token y_{t-1} and
- a context vector c_t

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

Context vector

- depends on the previous state and therefore changes at each step
- indicates which part of the input is most relevant to the decoding step

RNN with Attention



$$\alpha_t = softmax(a_t)$$

α can be viewed as a probability distribution over the source words

- A score is computed between each input token encoder state and the current state

$$a_{t,j} = v \top tanh(W_h\ h_j + W_s\ s_t + b)$$

- The context vector is the weighted sum of the encoder states

$$c_t = softmax(\sum_j \alpha_{t,j}.h_j)$$

- The new state is computed taking into account this context vector.

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

- The next predicted token is sampled from the new target vocabulary distribution

$$P_{vocab} = softmax(W s_t)$$

Attention

- Attention is a way to obtain a fixed-size representation
  - of an arbitrary set of representations (the values),
  - dependent on some other representation (the query)

- Encoder-Decoder

  - Query = current decoder state
  - Values = encoder hidden states

- Transformer

  - Query = token embedding
  - Values = surrounding tokens embeddings

Encoder-Decoder Attention

- Values = Encoder hidden states $h_1 \dots h_n$
- Query = Decoder state at time t s_t
- Attention scores at time t $e^t = [s_t \top h_1 \dots s_t \top h_n]$
- Attention distribution $\alpha^t = softmax(e^t)$

- Context vector (weighted sum of the encoder hidden states)

$$c^t = \sum_{i=1}^{n} \alpha_i^t h_i$$

- The new decoder state is computed taking into account this context vector.

$$s^t = f(s^{t-1}, y^{t-1}, c^t)$$

Attention Score Variants

- Dot product

$$e_i = s \top h_i$$

- Multiplicative

$$e_i = s \top W h_i$$

- Additive

$$e_i = v \top tanh(W_1 h_i + W_2 s)$$

Ruder 2017, Britz et al 2017

Copy

Copy

Motivation

- To copy from the input (E.g., in Text Summarisation applications)
- To handle rare or unknown words

Method

- At each time step, the model decides whether to copy from the input or to generate from the target vocabulary.

See et al. ACL 2017

Learn a Copy/Generate Switch

$P_{gen} \in [0, 1]$

Learned soft switch to choose between generating a word from the vocabulary by sampling from P_{vocab}, or copying a word from the input sequence by sampling from the attention distribution α_t.

Probability of generating a word from the vocabulary versus copying a word from the source

$$P_{gen} = \sigma(w_c\ c_t + w_s\ s_t + w_y\ y_{t-1})$$

Final Probability Distribution

Over source and target vocabulary

$$P(w) = p_{gen}\ P_{target}(w) + (1 - p_{gen})\ P_{source}$$

- $P(w)$, probability of generating word w
- P_{target}, decoder probability of generating w
  = 0 if w is not in the target vocabulary (OOV)
- P_{source}, probability of copying word from the source
  Cumulated attention score
  $= \sum_{i:w=w_i} \alpha_{t,i}$
  $= 0$ if w is not in the input

Coverage

Coverage

- Neural models tend to omit or repeat information from the input

Solution

(Tu et al. 2017)

- Coverage: cumulative attention, what has been attended to so far
- Use coverage as extra input to attention mechanism
- Loss: Penalises attending to input that has already been covered

Tu et al. ACL 2016

Coverage

- A coverage vector k_t captures how much attention each input words has received

$$k_t = \sum_{t=0}^{t-1} \alpha_t$$

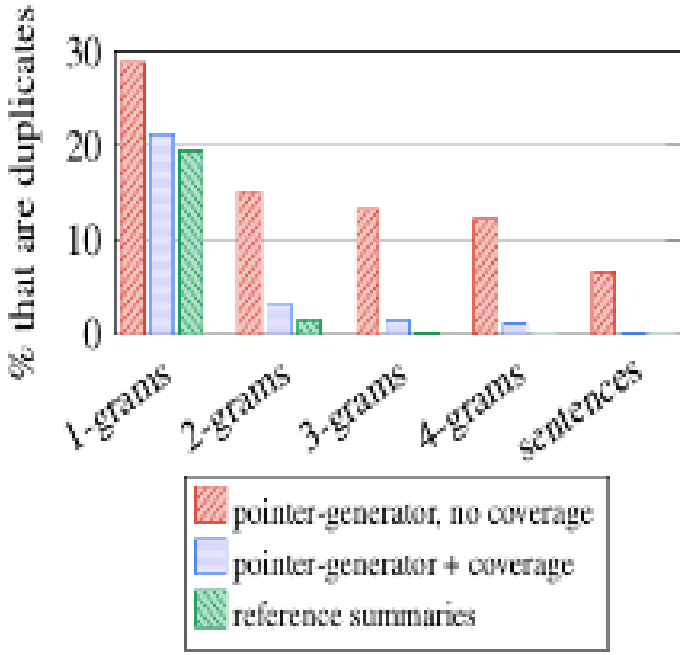
- The attention mechanism takes coverage into account

$$\alpha_{t,j} = v^T tanh(W_h\ h_j + W_s\ s_t + W_k\ k_t + b)$$

- The loss penalizes repeatedly attending to the same location

$$loss_t = -logP(w_t) + \lambda \sum_j min(\alpha_{t,j}, k_{t,j})$$

Impact of Coverage on Duplicate N-Grams



- The proportion of duplicate n-grams is similar in the reference summaries and in the summaries produced by the model with coverage.
- Coverage successfully eliminates repetitions.

See et al. ACL 2017

Rare Words

- Copying
- Delexicalisation
- Character-Based Network
  - smaller vocabulary
  - unknown words handled by copying characters
- WordPieces, Byte Pair Encoding (BPE)

Delexicalisation

- Slot values occurring in training utterances are replaced with a placeholder token representing the slot
- At generation time, these placeholders are then copied over from the input specification to form the final output

inform(restaurant name = Au Midi , neighborhood= midtown , cuisine = french)

Au Midi is in Midtown and serves French food .

inform(restaurant name = **restaurant name**, neighborhood= **neighborhood**, cuisine = **cuisine**)

restaurant name is in **neighborhood** and serves **cuisine** food.

Character-Based Generation

- Much smaller vocabulary
- Word embeddings are composed of character embeddings
- Similar embeddings for words with similar spelling
- Particularly interesting for languages with rich morphology
- Competitive results on
  - E2E benchmark: Generating from dialog moves
  - WebNLG benchmark: Generating from RDF graphs

Jagfeld et al. 2018

Word Pieces

A finite set of word pieces allows for an infinite sets of words

Byte Pair Encoding (BPE)

- Most frequent character pairs $\rightarrow$ new word piece
- Bottom-up clustering of characters
  - Start with unigram vocabulary of (Unicode) characters in the data
  - Add most frequent character pairs to vocabulary
  - and repeat until target vocabulary size is reached
- Segment input words into wordpieces and concatenated output wordpieces into words

Sennrich et al. ACL 2016

BPE Vocabulary

Dictionary

5 l o w
2 l o w e r
6 n e w e s t
3 w i d e s t

Vocabulary

l, o, w, e, r, n, w, s, t, i, d

Start with all characters in vocab

Image credit: Sennrich et al. ACL 2016

BPE Vocabulary

Dictionary

5 l o w
2 l o w e r
6 n e w **es** t
3 w i d **es** t

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, **es**

Add a pair (e, s) with freq 9

BPE Vocabulary

Dictionary

5 l o w
2 l o w e r
6 n e w **est**
3 w i d **est**

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, es, **est**

Add a pair (es, t) with freq 9

BPE Vocabulary

Dictionary

5 **lo** w
2 **lo** w e r
6 n e w est
3 w i d est

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, es, est, **lo**

Add a pair (l, o) with freq 7