Natural Language Generation

CS224n Natural Language Processing with Deep Learning

– Lecture 15 –

UOS STAT NLP Study : Ngel ☺

Changdae Oh

2021. 02. 15

Topics

1. Recap : LMs & Decoding algorithms
2. NLG tasks and neural approaches to them
3. NLG Evaluation and difficulties
4. Trends & Future

Topics

1. Recap : LMs & Decoding algorithms
2. NLG tasks and neural approaches to them
3. NLG Evaluation and difficulties
4. Trends & Future

Recap : LMs & Decoding algorithms

Natural Language Generation

Any setting in which
we generate new text.

subcomponent of :

– Machine Translation
– Dialogue
– Storytelling
– Image captioning
…

Recap

Language Modeling

➡ $P(y_t|y_1,\ldots,y_{t-1})$

Conditional Language Modeling

➡ $P(y_t|y_1,\ldots,y_{t-1},x)$

- Machine Translation
- Summarization
- Dialogue

Recap : LMs & Decoding algorithms

Decoding Algorithms

: Used to determine the text generated from your language model

In lecture 8.

Greedy decoding

- Lack of backtracking

Exhaustive search

- infeasible

Beam search

- Stable and feasible !

Beam search

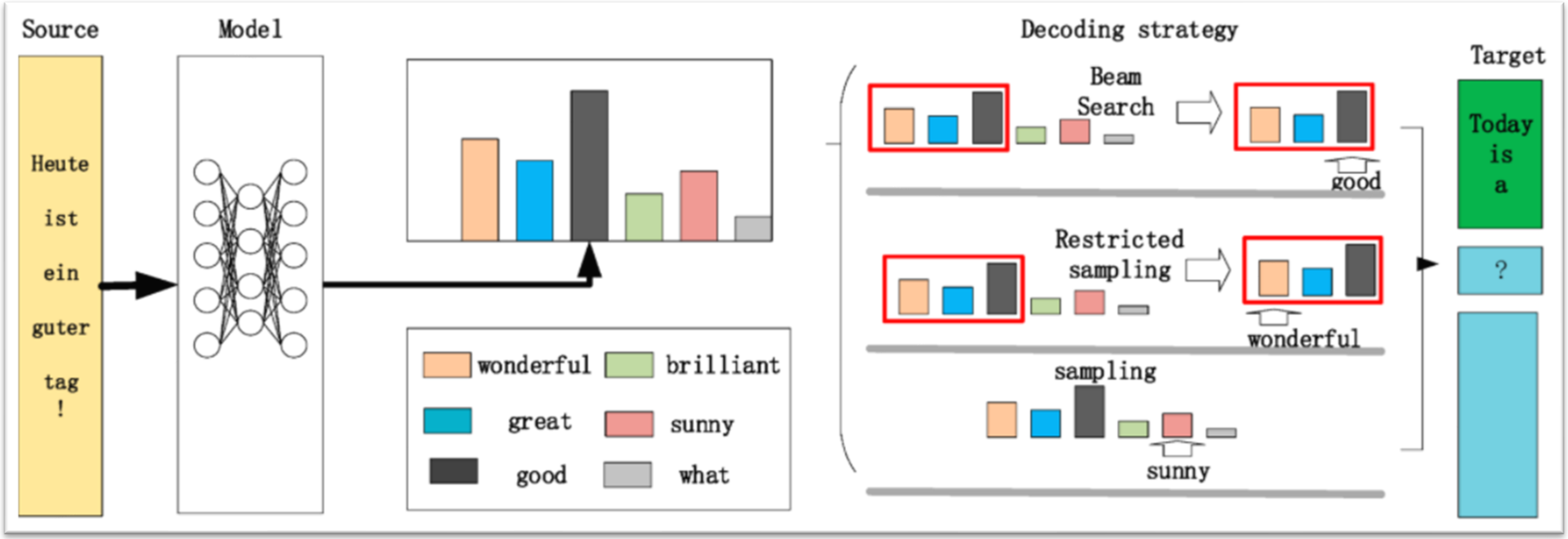
- Aims to find a high-probability
  -> tracking multiple seq. at once

- K(beam size) most probable hypotheses

  - Small k : ungrammatical, incorrect
  - Larger k : reduce errors,
    expensive,
    decreases BLEU score,
    make generic/irrelevant text

Recap : LMs & Decoding algorithms

Sampling-based decoding → More efficient than any other algorithm



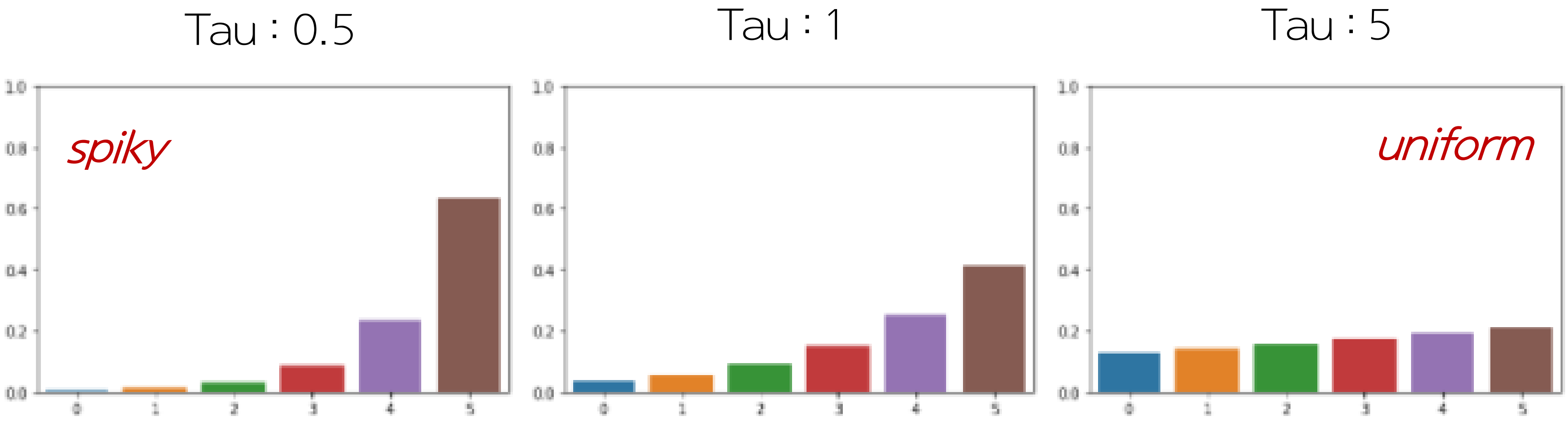
- On each step t, randomly sample from the (truncated)probability distribution
- In Top-n sampling(restricted sampling),

increase n – more diverse/risky
decrease n – more generic/safe

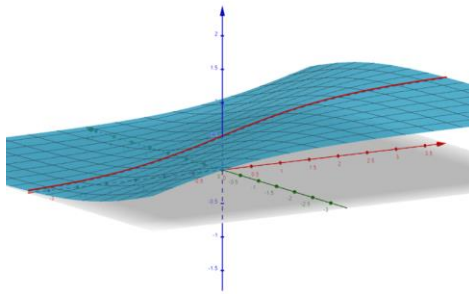
https://www.mdpi.com/2078-2489/11/5/255/htm

Recap : LMs & Decoding algorithms

Softmax temperature



$$P_t(w) = \frac{\exp(s_w)}{\sum_{w' \in V} \exp(s_{w'})} \quad \rightarrow \quad P_t(w) = \frac{\exp(s_w/\tau)}{\sum_{w' \in V} \exp(s_{w'}/\tau)}$$



Similar to
scaled dot-product attention
In transformer

https://3months.tistory.com/491

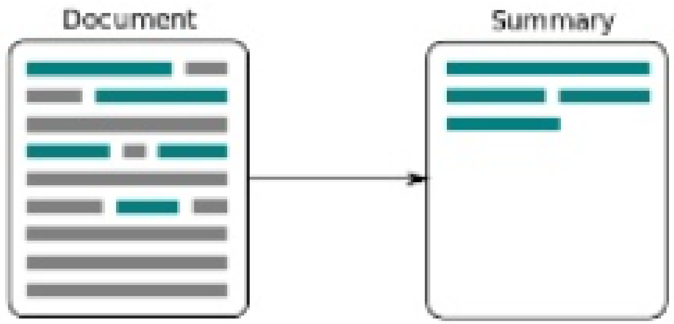
Topics

1. Recap : LMs & Decoding algorithms
2. NLG tasks and neural approaches to them
3. NLG Evaluation and difficulties
4. Trends & Future

NLG tasks and neural approaches to them

Summarization

Given input text x, write a summary y
which is shorter and contains the main information of x.



single-doc

Multi-doc

Few sentence -> headline

News article -> summary

Wiki -> simple wiki

Paragraph -> summary

Manual -> summary

news -> for children

NLG tasks and neural approaches to them

Extractive summarization

- Easy
- Select
- Restrictive

Abstractive summarization

- Difficult
- Generative
- flexible

NLG tasks and neural approaches to them

pre-neural summarization

➢ Mostly extractive

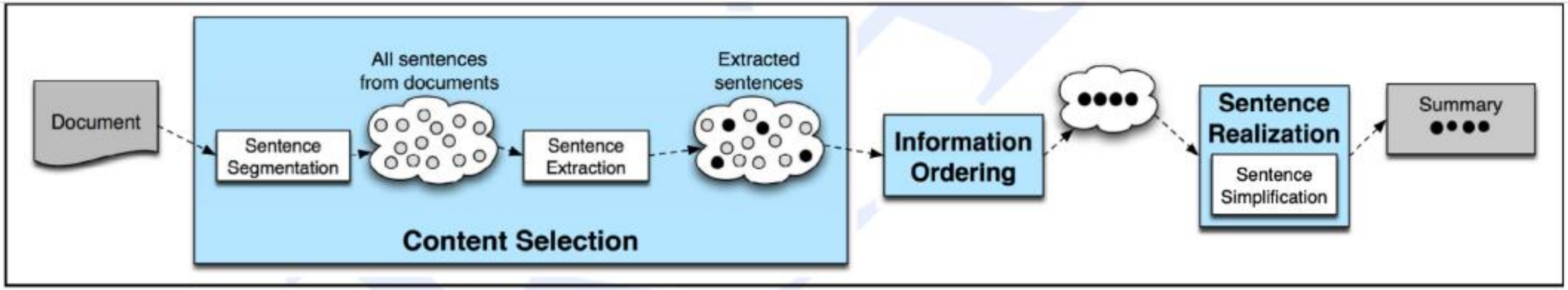


Figure 23.14 The basic architecture of a generic single document summarizer.

Pipeline

* Sentence Scoring
* graph-based algorithms

1. Content selection : Choose important sentences
2. Information ordering : Make sequence
3. Sentence realization : Refine the sequence of sentences

NLG tasks and neural approaches to them

Summarization evaluation : ROUGE

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Count the number of overlapping units such as n-gram, word sequences, and word pairs **between generated texts and reference summaries**

➡ n-gram Co-Occurrence Statistics (n-gram recall between gen. and ref.)

1. No brevity penalty
2. Based on recall , not precision (BLEU)
3. Scoring separately for each n-gram

BLEU

$$\frac{\sum_{n\text{-}gram \in Candidate} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram \in Candidate} Count(n\text{-}gram)}$$

https://www.aclweb.org/anthology/W04-1013/

NLG tasks and neural approaches to them

Neural summarization

- 2015. See "Abstractive Summarization" as translation task
- Standard seq2seq with attention

Lots more developments

1. Copying well
2. Hierarchical / multi-level attention
3. Global / high-level content selection
4. RL approach
5. Resurrecting pre-neural ideas

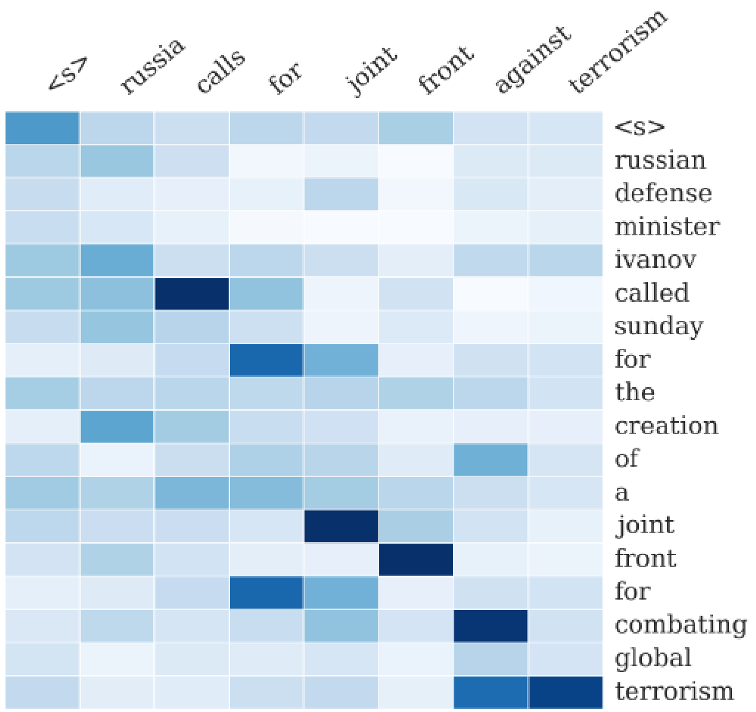
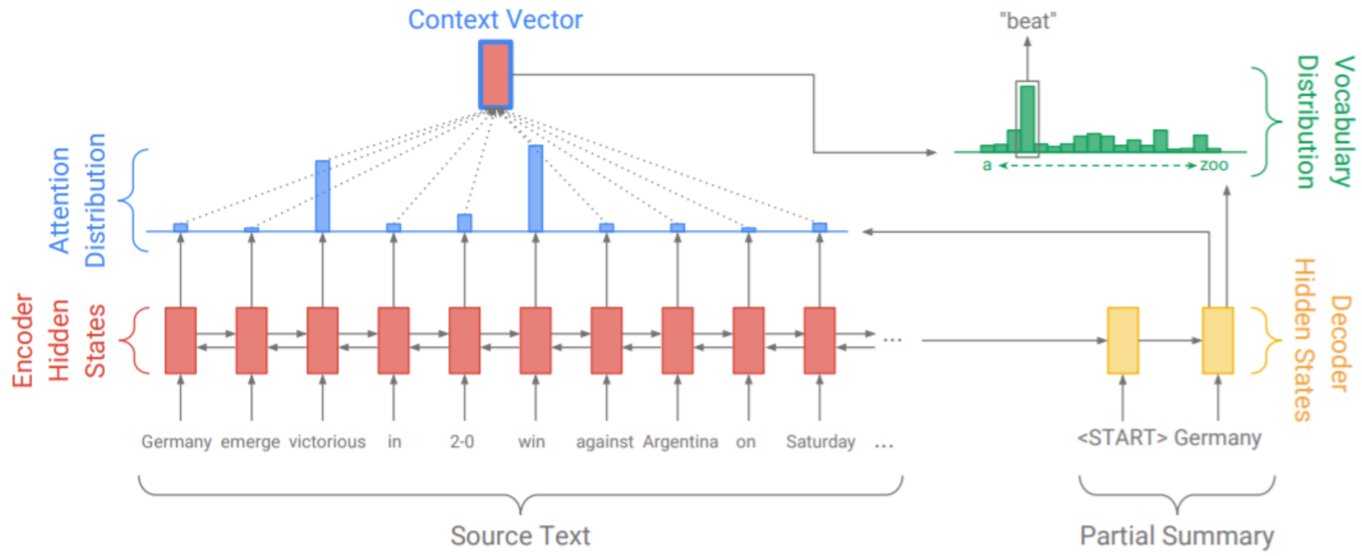


Figure 1: Example output of the attention-based summarization (ABS) system. The heatmap represents a soft alignment between the input (right) and the generated summary (top). The columns represent the distribution over the input after generating each word.

NLG tasks and neural approaches to them

Copy mechanisms

ordinary seq2seq+attention system bad at copying over details(rare words & OOV) correctly



$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}}) \quad (1)$$

$$a^t = \text{softmax}(e^t) \quad (2)$$

$$h_t^* = \sum_i a_i^t h_i \quad (3)$$

$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b') \quad (4)$$

$$P(w) = P_{\text{vocab}}(w) \quad (5)$$

$$\text{loss}_t = -\log P(w_t^*) \quad (6)$$

$$\text{loss} = \frac{1}{T}\sum_{t=0}^{T} \text{loss}_t \quad (7)$$

Original Text (truncated): 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.**

Baseline 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.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

NLG tasks and neural approaches to them

Pointer-Generator Networks

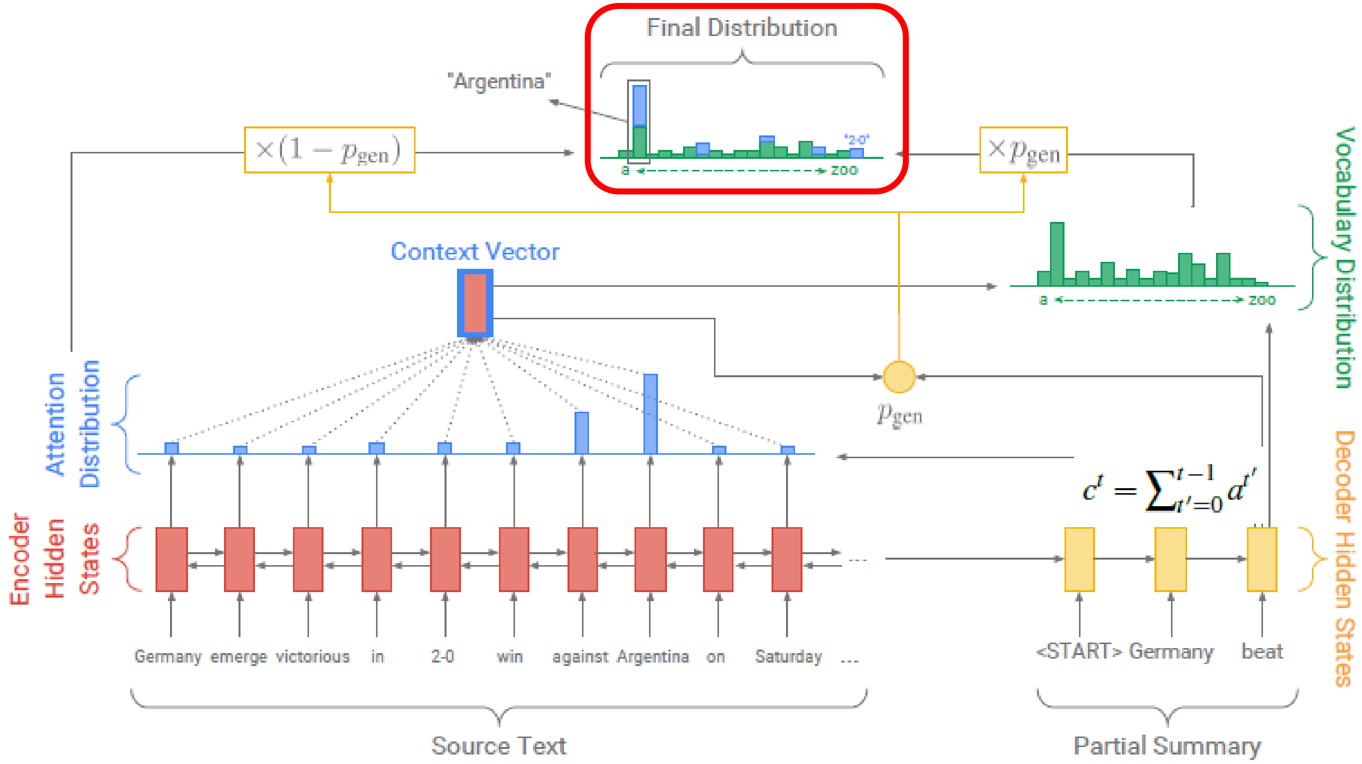
- Pointer-generator network
  : facilitate Copying words from the source text via pointing
  -> improve accuracy and handling of OOV words

- Coverage vector
  : use to track and control coverage of the source document
  -> remarkably effective for eliminating repetition

	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	**39.53**	**17.28**	**36.38**	17.32	18.72

NLG tasks and neural approaches to them

Use attention to both attending and re-phrasing !!



- ✓ The final distribution is a mixture of the generation distribution and the copying distribution
- ✓ Cov vector is a distribution over the source document words that has been covered
- ✓ To avoid repeatedly attending to the same locations, add covloss term

Pointer-generator network

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

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

* Soft switch to choose between generating a word from the vocab or copying a word from the inputs

Coverage mechanism

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

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

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

NLG tasks and neural approaches to them

Problem of copying mechanisms

- ✓ Copy too much
  : what should be an abstractive system collapses to a mostly extractive system.
- ✓ Bad at overall content selection, especially if the input document is long

Note,

- ✓ Pre-neural summarization had separate stages for content selection and surface realization.
- ✓ Standard seq2seq summarization system, two stages are mixed in together

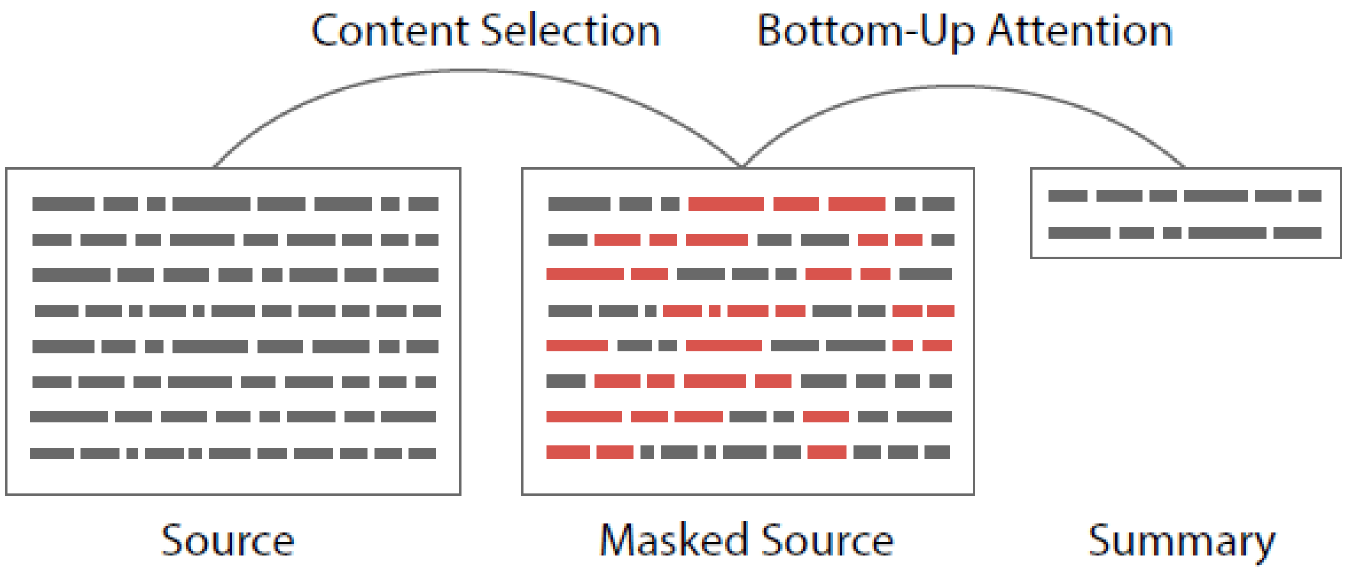
Solution

- ✓ Bottom-up summarization
  - Better overall content selection strategy.
  - Less copying of long sequences (more abstractive output.

NLG tasks and neural approaches to them

Bottom-up summarization

- Content selection stage : neural sequence-tagging model (binary labeling)
- Bottom-up attention stage (apply a mask)



Use attention masks to limit the available selection of the pointer-generator model.

$$p(\tilde{a}_j^i | x, y_{1:j-1}) = \begin{cases} p(a_j^i | x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{ow.} \end{cases}$$

qi is calculated as $\sigma(W_s h_i + b_s)$ by sequence tagging LSTM

NLG tasks and neural approaches to them

Summarization via Reinforcement Learning

- Directly optimize ROUGE via RL
- Only RL model – Higher ROUGE but Lower human judgement score.

Objective

$$L_{ml} = -\sum_{t=1}^{n'} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*, x)$$

$$L_{rl} = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x)$$

$$L_{mixed} = \gamma L_{rl} + (1 - \gamma) L_{ml}$$

Evaluation

Model	ROUGE-1	ROUGE-2	ROUGE-L
ML, no intra-attention	44.26	27.43	40.41
ML, with intra-attention	43.86	27.10	40.11
RL, no intra-attention	**47.22**	30.51	**43.27**
ML+RL, no intra-attention	47.03	**30.72**	43.10

Model	Readability	Relevance
ML	6.76	7.14
RL	4.18	6.32
ML+RL	**7.04**	**7.45**

NLG tasks and neural approaches to them

Dialogue

- Task-oriented dialogue
- Social dialogue

Pre- and post- neural dialogue

✓ Rule based, use predefined templates, or retrieve

⬇

✓ open-ended freeform dialogue system (seq2seq-based)

deficiency

- Genericness
- Irrelevant
- Repetition
- Lack of context
- Lack of consistent persona

NLG tasks and neural approaches to them

Irrelevant response

- Generic
- Unrelated subject

$$\hat{T} = \arg\max_{T} \left\{ \log p(T|S) \right\}$$

$$\hat{T} = \arg\max_{T} \left\{ \log p(T|S) - \log p(T) \right\}$$

Genericness response

✓ Upweight rare words during beam search
✓ Use sampling decoding algorithms
✓ Condition the decoder on additional content
✓ Train a retrieve-and-refine model rather than a generate-from-scratch model

Repetition problem

✓ Block repeating n-grams during beam search
✓ Use coverage mechanism
✓ Define a new objective to discourage repetition

Lack of consistent persona

✓ Persona embeddings
✓ PersonaChat (Dataset)

NLG tasks and neural approaches to them

Storytelling

Given Image / brief writing prompt / story so far

- There was no paired data to learn from.

– Use a **common sentence-encoding** space

1. Using image captioning dataset, learn a mapping from images to the skip-thought encodings of their captions
2. Train a RNN-LM to decode a skip-thought vector to the text
3. Put the two components together

skip-thought vectors

- Highly generic sentence representations
- Sentence level skip-gram method

NLG tasks and neural approaches to them

Challenges in storytelling

- Fluent, but are meandering, with no coherent plot
- Because LMs model sequences of words. Stories are sequences of events.

consider

- Events and the causality structure between them
- characters, personalities, motivations, relationships ···
- State of the world
- Narrative structure
- Good storytelling principles

Event2event story generation

https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17046/15769

Topics

1. Recap : LMs & Decoding algorithms
2. NLG tasks and neural approaches to them
3. NLG Evaluation and difficulties
4. Trends & Future

NLG Evaluation and difficulties

Word overlap based metrics

- BLEU, ROUGE, METEOR, F1, …
  - Not ideal for machine translation
  - Even worse for summarization, dialogue (more open-ended)

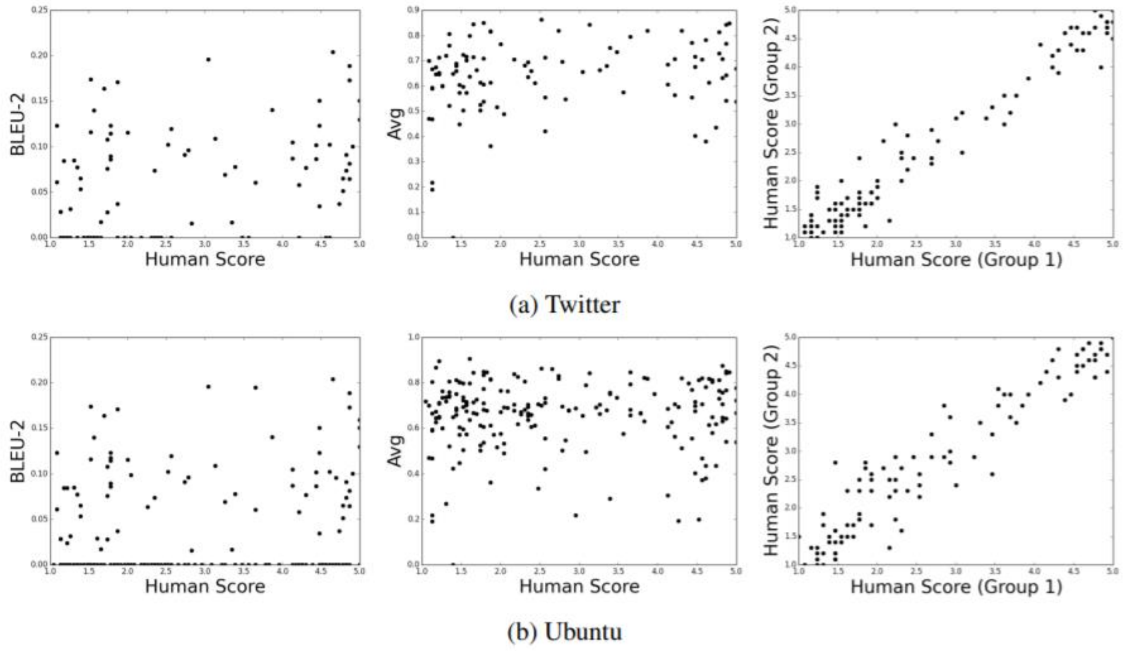


Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

Other automatic metrics

- Perplexity?

  : doesn't tell about generation

- Word embedding based?

  * compare the similarity of the word embeddings !

  doesn't correlate well with human judgments

NLG Evaluation and difficulties

- There are no automatic metrics to adequately capture overall quality.

But we can define more focused(specific) automatic metrics
To capture particular aspects of generated text

- Fluency
- Correct style
- Diversity
- Relevance to input
- Length and repetition
- task-specific metrics (compression rate for summarization)

➡ They can hep us track some important qualities

NLG Evaluation and difficulties

Separates out the important factors that contribute to overall chatbot quality

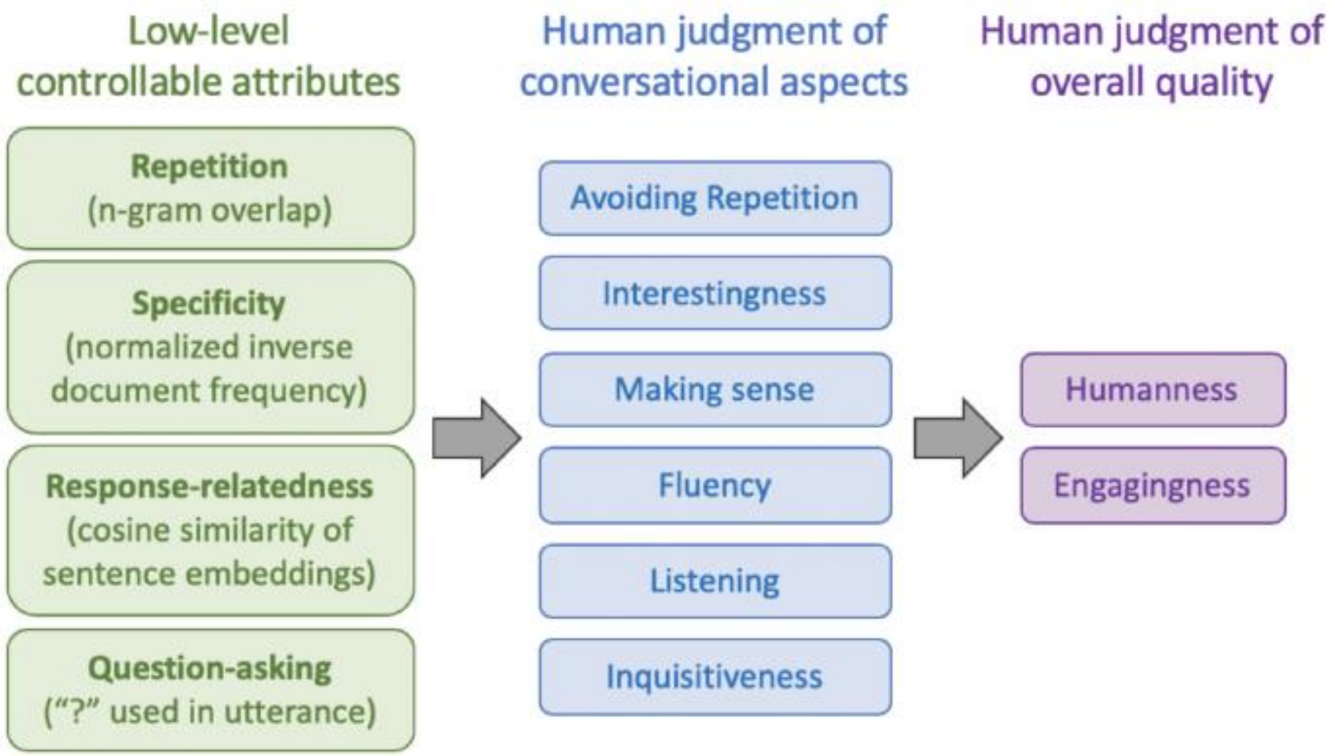


Figure 1: We manipulate four low-level attributes and measure their effect on human judgments of individual conversational aspects, as well as overall quality.

Topics

1. Recap : LMs & Decoding algorithms
2. NLG tasks and neural approaches to them
3. NLG Evaluation and difficulties
4. Trends & Future

Trends & Future

Trends in NLG(2019)

- Incorporating discrete latent variables into NLG
- Alternatives to strict left-to-right generation (Parallel generation)
- Alternative to maximum likelihood training with teacher forcing

NLG seems like one of the wildest parts remaining..

Practical tips

- The more open-ended, the harder –
- Aim Specific improvement!
- Improving the LM -> improve generation quality
- Look at your output
- You need an automatic metric
- Human eval : as focused as possible
- Reproducibility is a huge problem
- can be very frustrating. But also very funny