Text Generation

Zhiyuan Liu

liuzy@Tsinghua.edu.cn

THUNLP

Outline

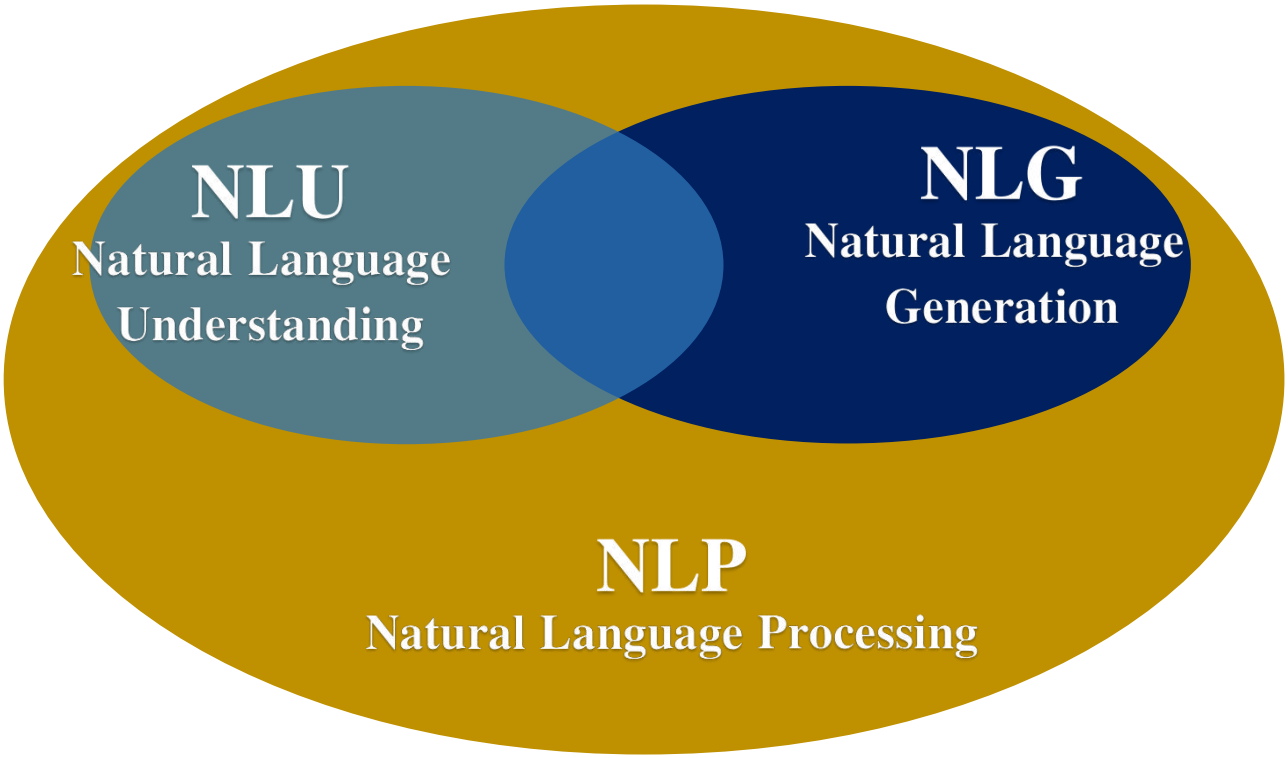
- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
- Text Generation Tasks and Challenges
- Current Trends, and the Future

Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
- Text Generation Tasks and Challenges
- Current Trends, and the Future

Background and Definition

- What is Text Generation?
- The difference between NLU, NLP, NLG or (TG)



[Survey of the State of the Art in Natural Language Generation- Core tasks, applications and evaluation]

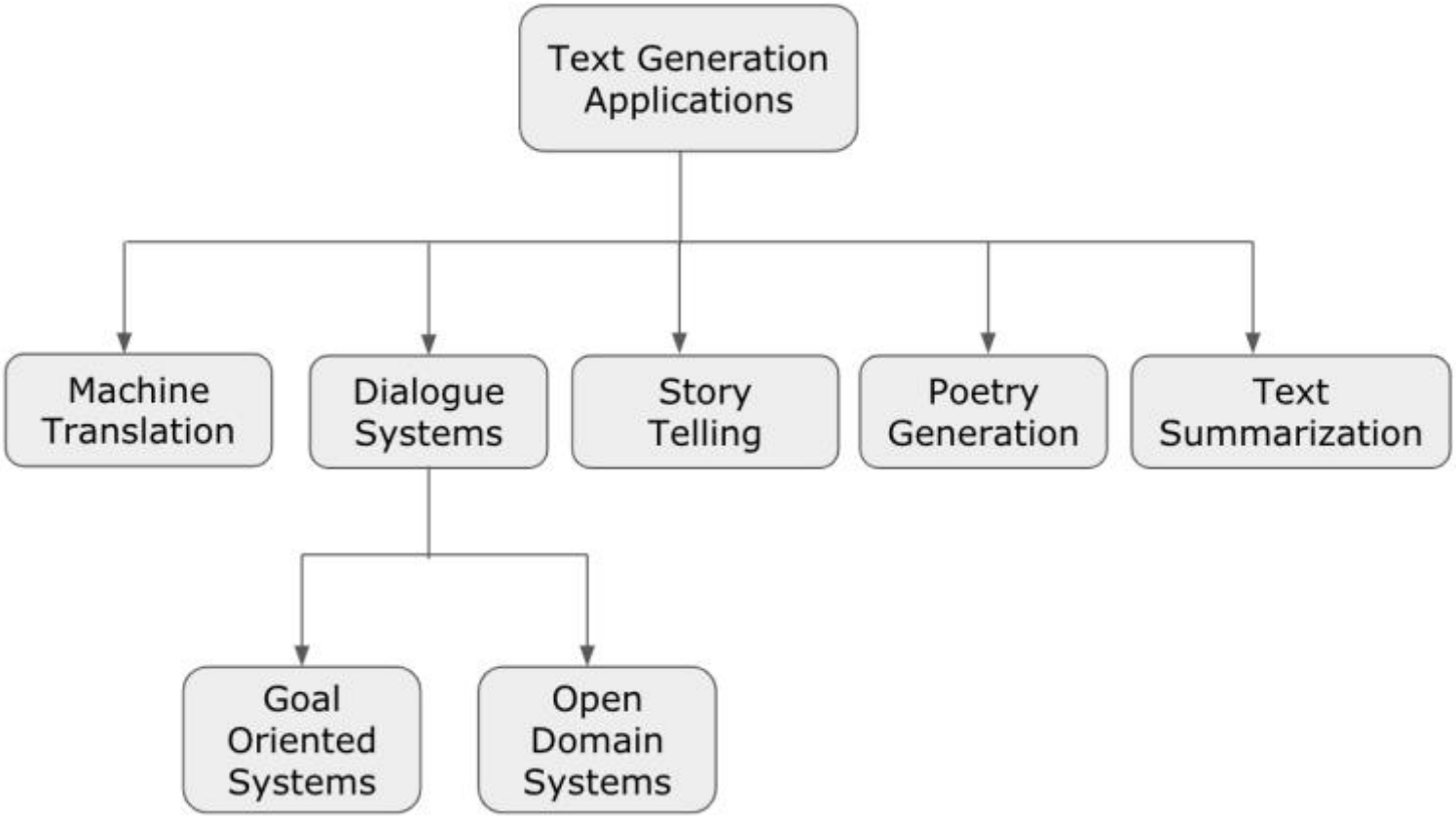
Background and Definition

- What is Text Generation?
- Produce understandable texts in human languages from some underlying non-linguistic representation of information. [Reiter el., 1997]
- Both text-to-text generation and data-to-text generation are instances of Text Generation [Reiter at el., 1997]

[Gatt at el., 2017, Survey of the State of the Art in Natural Language Generation- Core tasks, applications and evaluati

Background and Definition

- An overview of the applications that can be categorized under the umbrella of language ge



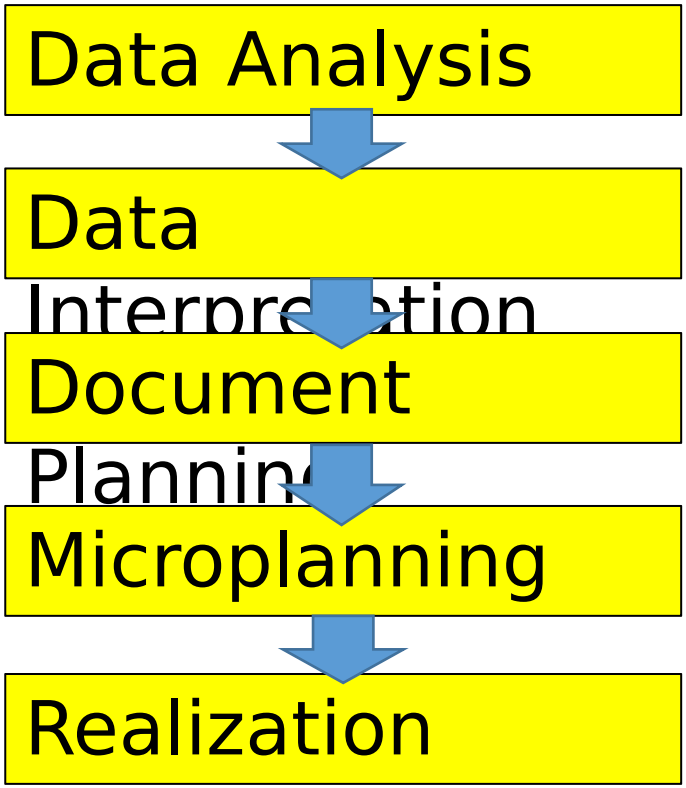
[Gatt at el., 2017, Survey of the State of the Art in Natural Language Generation- Core tasks, applications and evaluati

Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
- Text Generation Tasks and Challenges
- Current Trends, and the Future

Background and Definition

- TG system: Pipeline



Patterns, threads

Analytical insights, which patterns are important

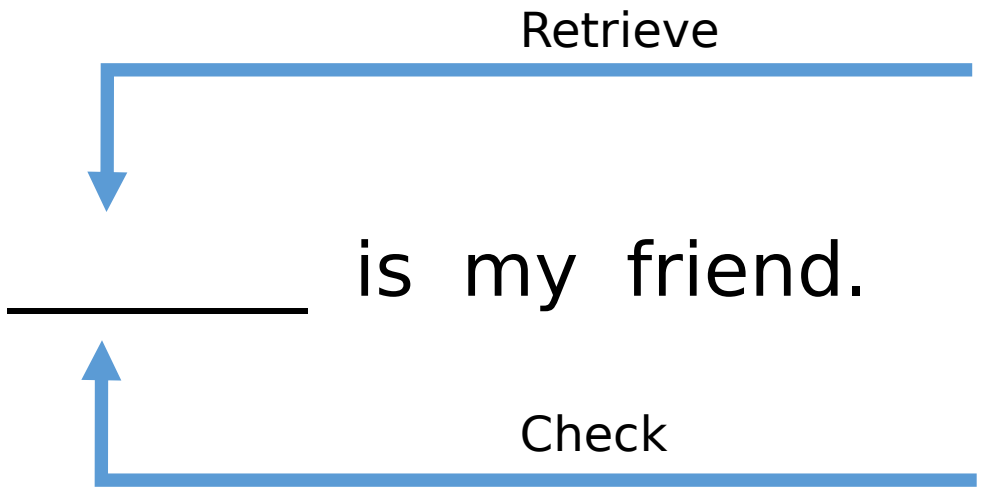
Select events, structures them into narrative

Choose words, syntactic structures, referring experience

Generate actual text

Rule/Template-based

- Rule-based
- Template-based

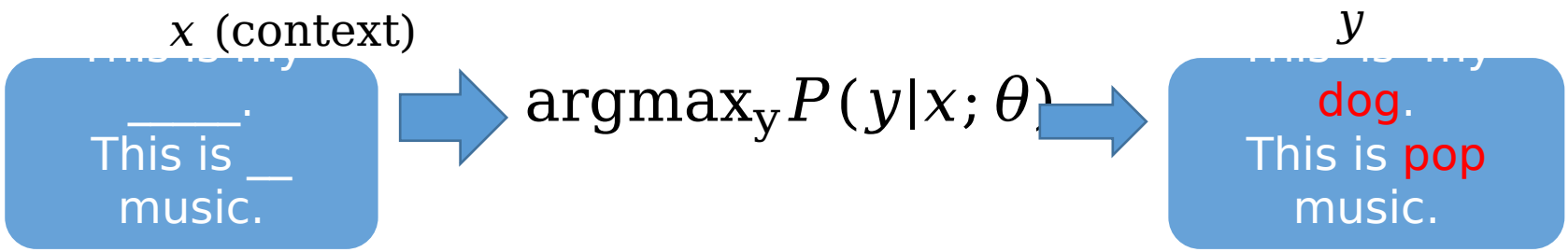


- If we have lots of hand-drafted templates, we can ...

[Sciforce, 2019, A Comprehensive Guide to Natural Language

Statistical Methods

- Core idea: Build a statistical model from data

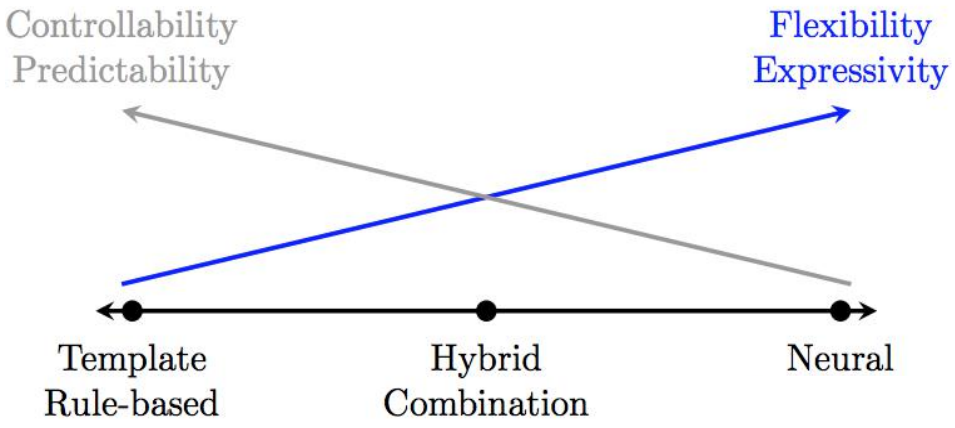


- Similar to statistical machine translation methodology
  - Obtain $P(x|y)$ and $P(y)$ from data
  - Compute the argmax with various algorithm
- Systems extremely complex
  - Separately-designed sub-components
- Lots of human efforts

[Kondadadi at el., 2013, A Statistical NLG Framework for Aggregated Planning

Traditional and Neural Text Generation

- Rule/Template-based and statistical models:
  - Fairly interpretable and well-behaved
  - Lots of hand-engineering and feature engineering
- Neural network models :
  - Demonstrate better performance in many tasks
  - Poorly understood and sometimes poorly behaved as well



Trade-offs between template/rule-based
and neural network models

[Xie at el., 2017, Neural Text Generation: A Practical Guide]

Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
  - Autoregressive
  - Non-Autoregressive
- Text Generation Tasks and Challenges
- Current Trends, and the Future

Language Modeling

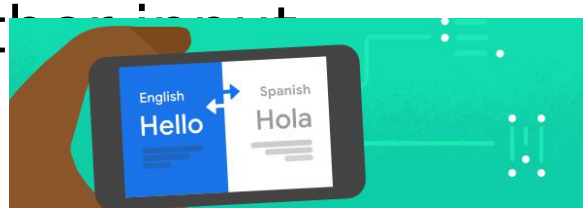
- **Language Modeling:** $P(y_t|y_1,\ y_2,\ \ldots,\ y_{t-1})$
  - The task of predicting the next word, given the words so far

- A system that produces this probability distribution is called a **Language Model**

- RNNs for language modeling

Conditional Language Modeling

- **Conditional Language Modeling:** $P(y_t|y_1, y_2, \cdots, y_{t-1}, x)$
  - The task of predicting the next word, given the words so far, and also some other input x
  - x input/source
  - y output/target sequence



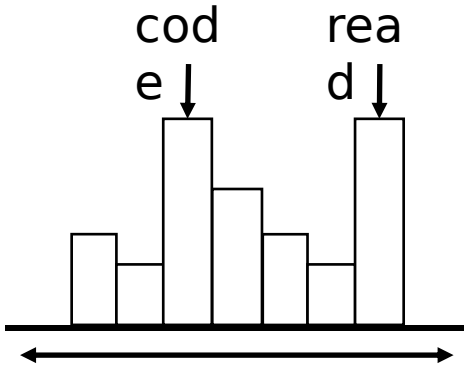
Task	X (example)	Y (example)
language modeling	none (empty sequence)	tokens from news corpus
machine translation	source sequence in English	target sequence in French
grammar correction	noisy, ungrammatical sentence	corrected sentence
summarization	body of news article	headline of article
dialogue	conversation history	next response in turn
Related tasks (may be outside scope of this guide)		
speech transcription	audio / speech features	text transcript
image captioning	image	caption describing image
question answering	supporting text + knowledge base + question	answer

[Xie at el., 2017, Neural Text Generation: A Practical Guide]

RNNs for Language Modeling

output distribution

$$y_4 = \text{softmax}(U h_4 + b_2) \in \mathbb{R}^{|V|}$$



hidden states

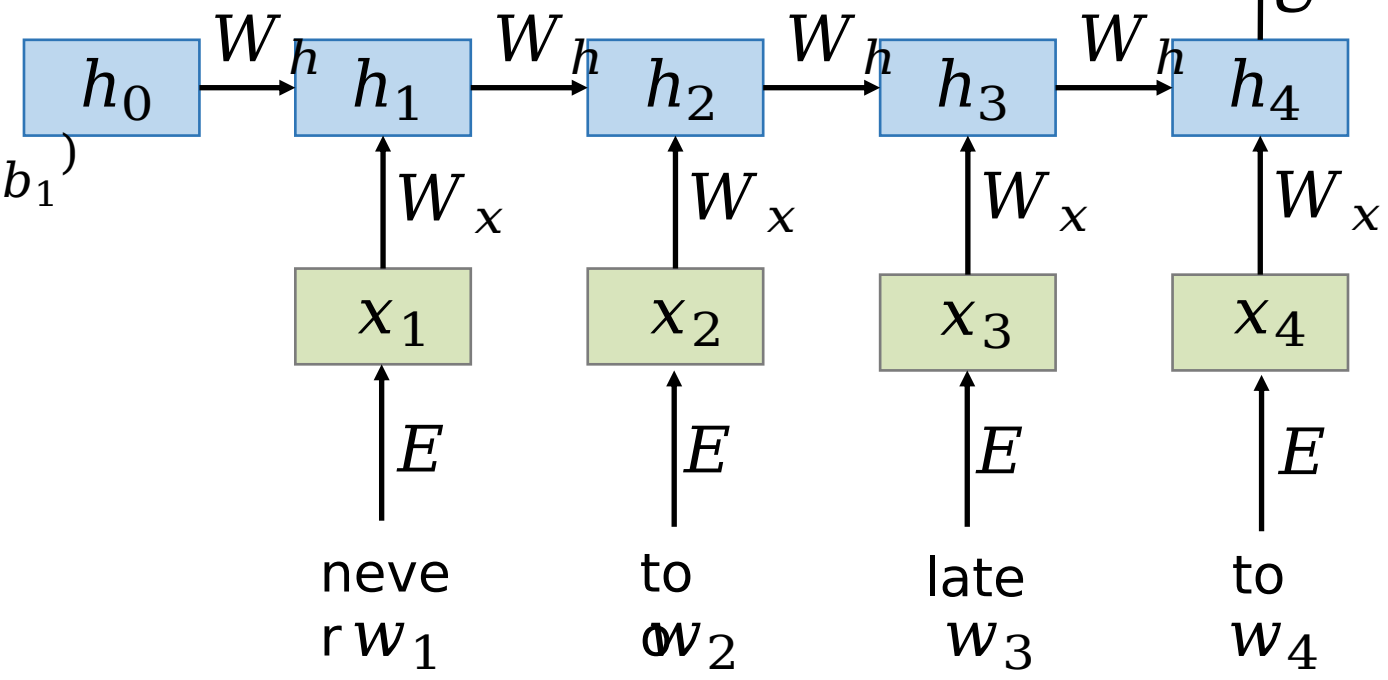
$$h_i = \tanh\left(\begin{matrix} W_x x_i + \\ W_h h_{i-1} + b_1 \end{matrix}\right)$$

word embeddings

$$x_i = E w_i$$

one-hot vectors

$w_i \in \mathbb{R}^{|V|}$



RNNs for Language Modeling

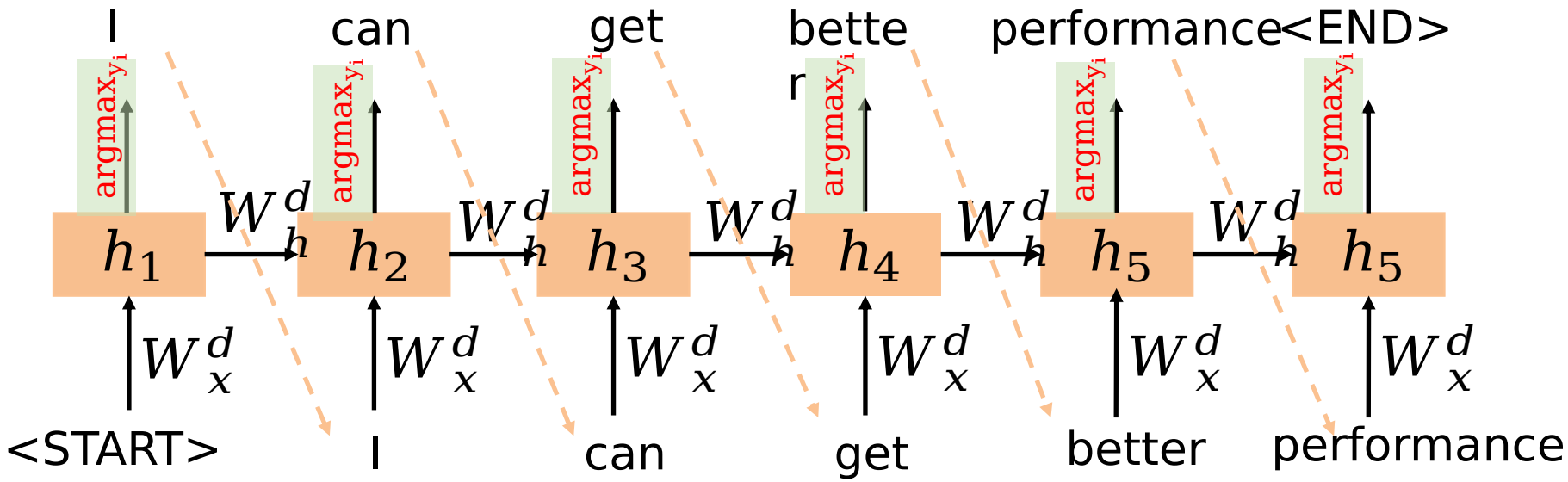
- RNNs are good at modeling sequential information
- General idea: Use RNN as an encoder for building the semantic representation of the sentence

Decoding

- Recap: decoding algorithms
  - Greedy decoding
  - Beam search

Greedy Decoding

- Generate the target sentence by taking argmax on each step of the decoder
  - $\text{argmax}_{y_i} P(y_i | y_1, \ldots, y_{i-1}, x)$



- Due to lack of backtracking, output can be poor

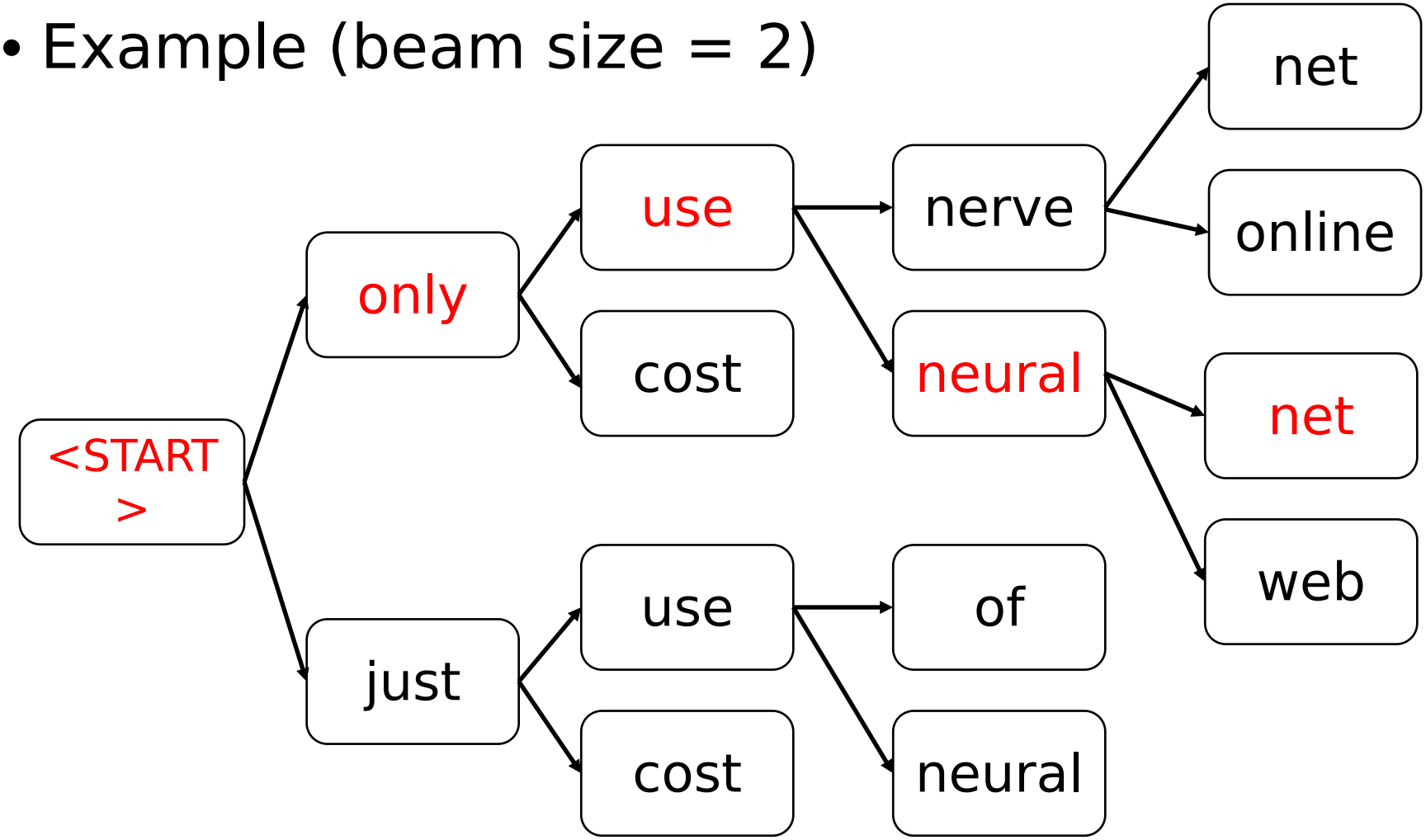
Beam Search Decoding

- We want to find y that maximizes
  - $P(y|x) = P(y_1|x)P(y_2|y_1, x)...P(y_T|y_1, ..., y_{T-1}, x)$
  - Find a high-probability sequence

- Beam search
  - On each step of decoder, keep track of the k most probable partial sequences
  - After you reach some stopping criterion, choose the sequence with the highest probability
  - Not necessarily the optimal sequence

Beam Search Decoding

- Example (beam size = 2)



What's the Effect of Changing Beam Size k?

- What's the effect of changing beam size k?
  - Small k has similar problems to greedy decoding (k=1)
    - Ungrammatical, unnatural, nonsensical, incorrect
  - Larger k means you consider more hypotheses
    - Reduces some of the problems above
    - More computationally expensive
  - But increasing k can introduce other problems
    - For neural machine translation (NMT): Increasing k too much decreases BLEU score (Tu et al., Koehn et al.)
    - clike chit-chat dialogue: Large k can make output more generic (see next slide)

[Tu et al., 2017, Neural Machine Translation with Reconstruction]

Decoding Effect of Beam Size in Chitchat Dialogue

- Effect of beam size in chitchat dialogue

Decoding Effect of Beam Size in Chitchat Dialogue

- Effect of beam size in chitchat dialogue

Beam size	Model response
1	*I love to eat healthy and eat healthy*
2	*That is a good thing to have*
3	*I am a nurse so I do not eat raw food*
4	*I am a nurse so I am a nurse*
5	*Do you have any hobbies?*
6	*What do you do for a living?*
7	*What do you do for a living?*
8	*What do you do for a living?*

Decoding Effect of Beam Size in Chitchat Dialogue

- Effect of beam size in chitchat dialogue

I mostly eat a fresh and raw diet, so I save on groceries

Human chit-chat partner

Beam size	Model response
1	*I love to eat healthy and eat healthy*
2	*That is a good thing to have*
3	*I am a nurse so I do not eat raw food*
4	*I am a nurse so I am a nurse*
5	*Do you have any hobbies?*
6	*What do you do for a living?*
7	*What do you do for a living?*
8	*What do you do for a living?*

Low beam size:
More on-topic but nonsensical; bad English

High beam size:
Converges to safe, "correct" response, but it's generic and less relevant

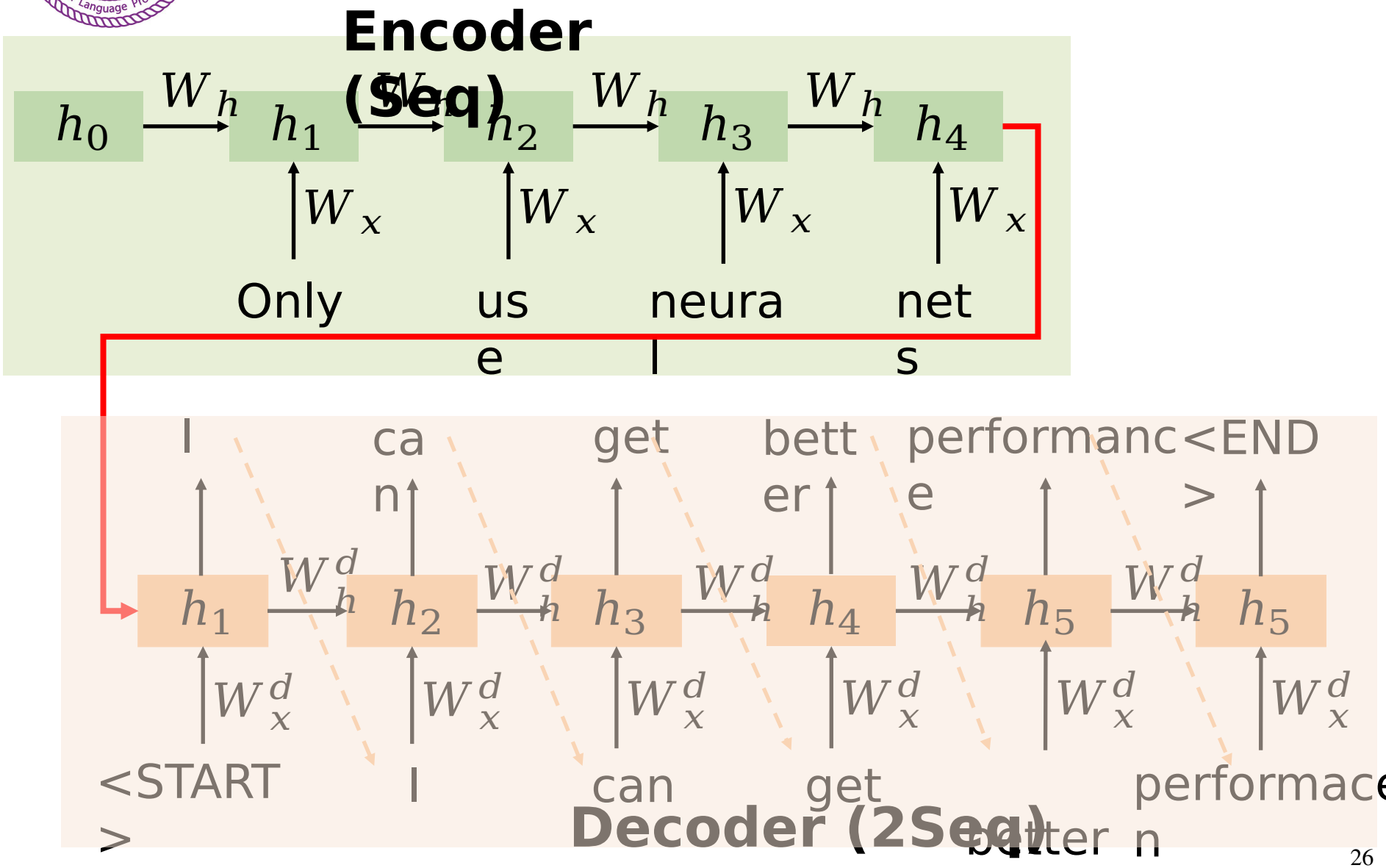
Decoding

- Sampling-based decoding
  - Pure sampling
    - On each step t, randomly sample from the probability distribution P_t to obtain your next word
  - Top-n sampling
    - On each step t, randomly sample from P_t, restricted to just the top-n most probable words
    - $n = 1$ is greedy search, $n = V$ is pure sampling
    - Increase n to get more diverse/risky output
    - Decrease n to get more generic/safe output
  - Both of these are more efficient than Beam search

Decoding

- In summary
  - Greedy decoding
    - A simple method
    - Gives low quality output
  - Beam search
    - Delivers better quality than greedy
    - If beam size is too high, it will return unsuitable output (e.g. Generic, short)
  - Sampling methods
    - Get more diversity and randomness
    - Good for open-ended / creative generation (poetry, stories)
    - Top-n sampling allows you to control diversity

Seq2seq

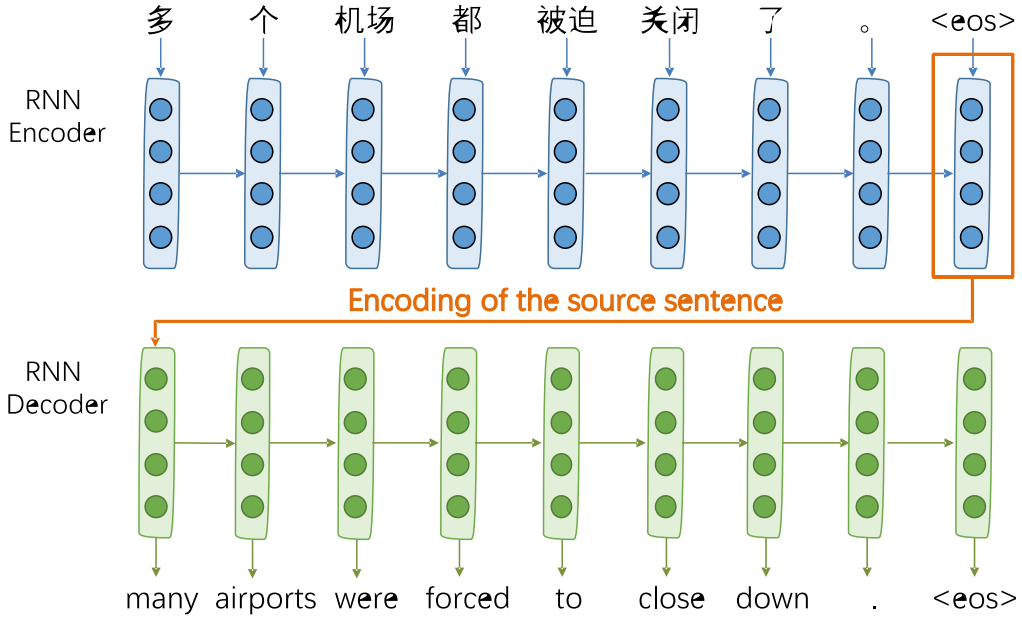

Encoder (Seq)
h_0
W_h
h_1
h_2
h_3
h_4
W_x
Only
us
e
neura
l
net
s
I
ca
n
get
bett
er
performanc
e
<END
>
h_1
h_2
h_3
h_4
h_5
h_5
W_h^d
W_x^d
<START
>
I
can
get
performace
n
Decoder (2Seq)

Seq2seq

- Pipeline methods have some problems
  - It's hard to propagate the supervision signal to each component
  - Once the task is changed, all components need to be trained from scratch
- Sequence-to-sequence model
  - $P(y|x) = P(y_1|x)P(y_2|y_1, x)...P(y_T|y_1, ..., y_{T-1}, x)$
    - It can be easily modeled by a single neural network and trained in an end-to-end fashion (differentiable)
  - The seq2seq model is an example of a conditional language model
  - Encoder RNN produces a representation of the source sentence
  - Decoder RNN is a language model that generates

Attention

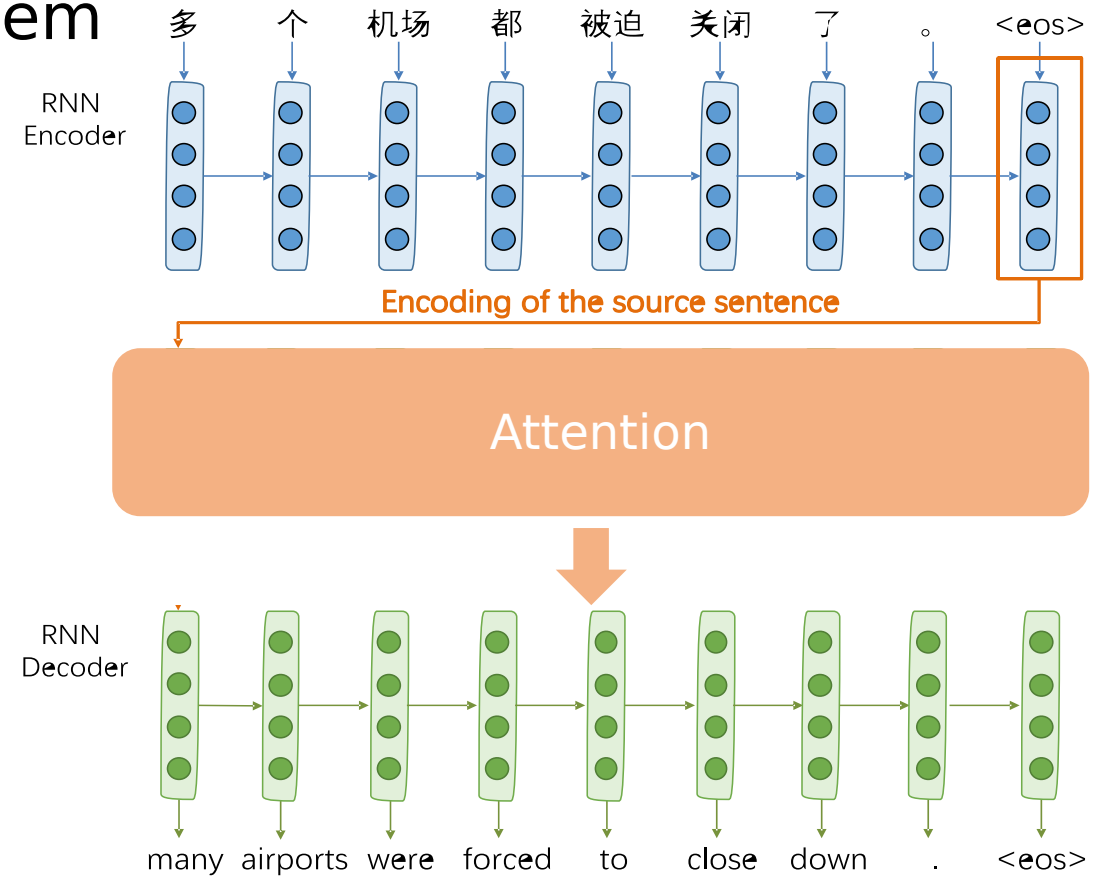
- Seq2seq: the bottleneck problem



  - The single vector of source sentence encoding needs to capture all information about the source sentence.
  - The single vector limits the representation capacity of the encoder, which is the information

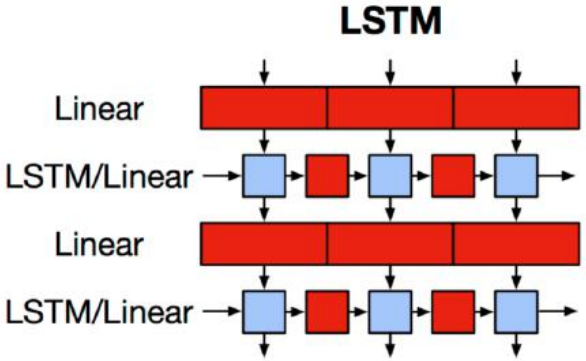
Attention

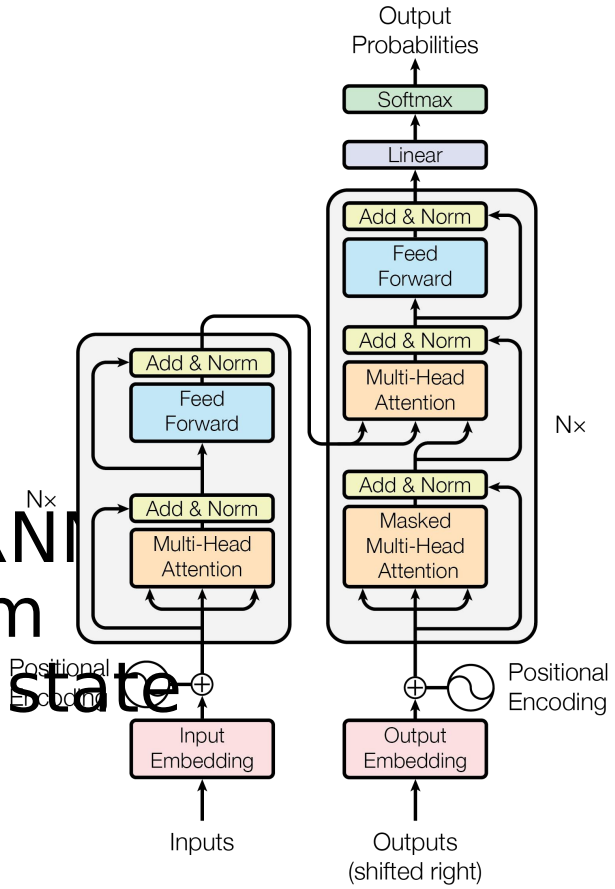
- Architecture
  - Attention provides a solution to the bottleneck problem



Transformer

- Motivations
  - Sequential computation in RNNs prevents





  - Despite using GRU or LSTM, RNN still need attention mechanism which provides access to any state
  - Maybe we do not need RNN?
    - Attention is all you need (2017)

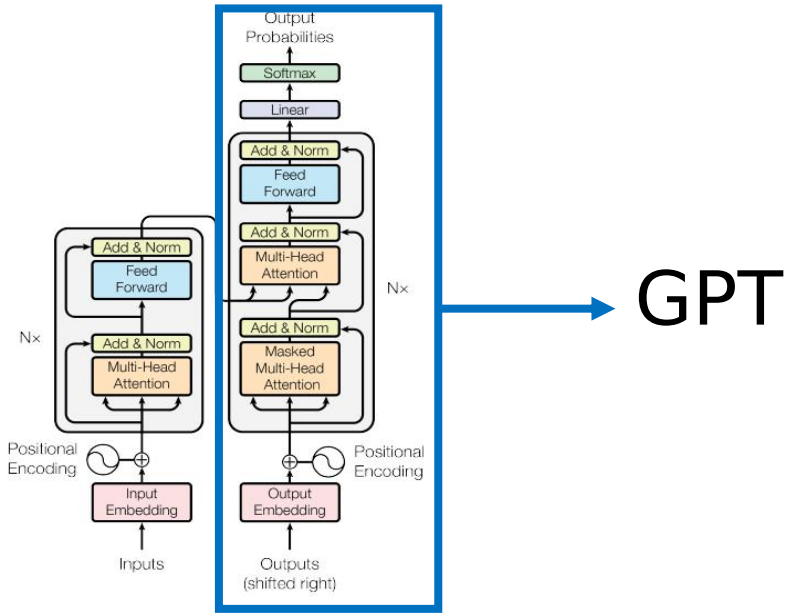
[Vaswani at el. 2017. Attention is all you need]

Transformer

- Replace seq2seq+attention with Transformer
- The effectiveness of the attention mechanism
- The Transformer is powerful and proven to be effective in many text generation tasks

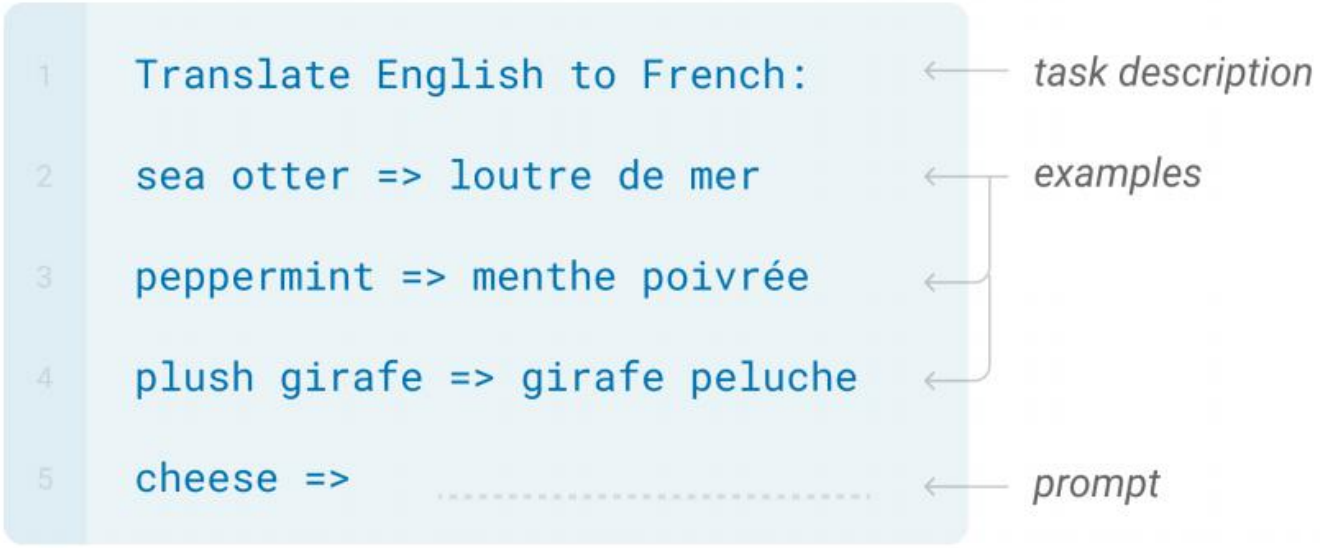
Generative Pre-Training (GPT)

- GPT is a decoder part of a transformer
  - GPT-1: Improving Language Understanding by Generative Pre-training
  - GPT-2: Language Models are unsupervised multitask learners



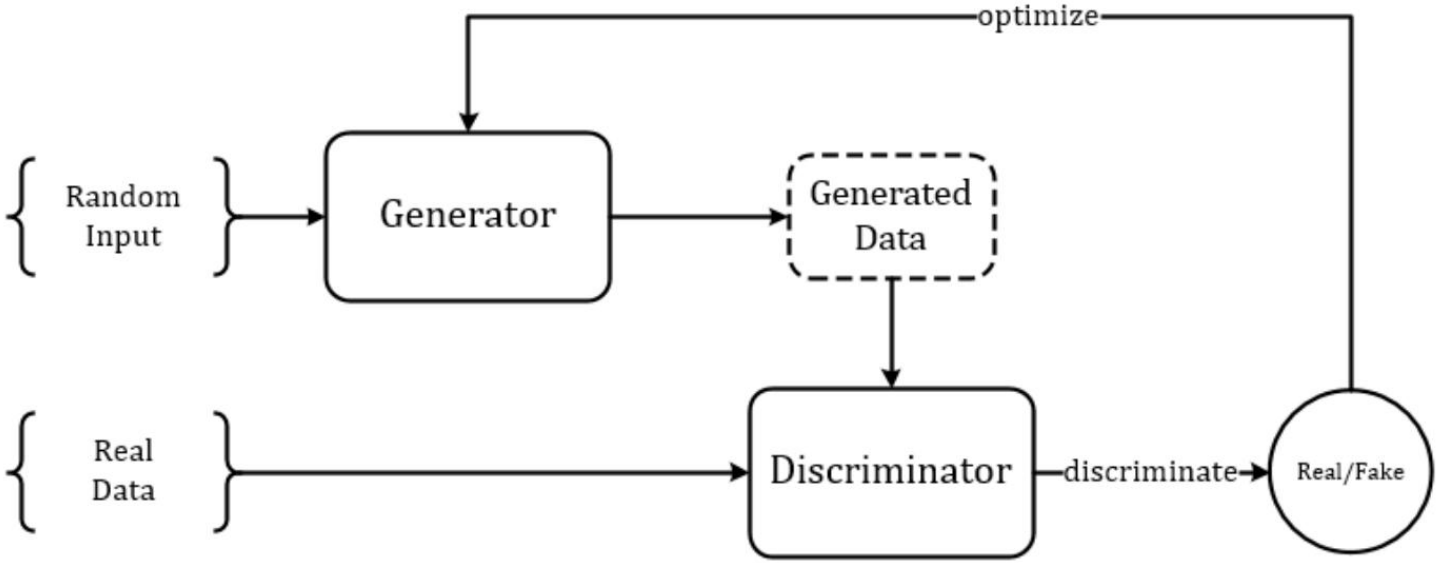
Generative Pre-Training (GPT)

- GPT is a decoder part of a transformer
  - GPT-3: Language models are few shot learners



GAN-based

- GAN architecture:

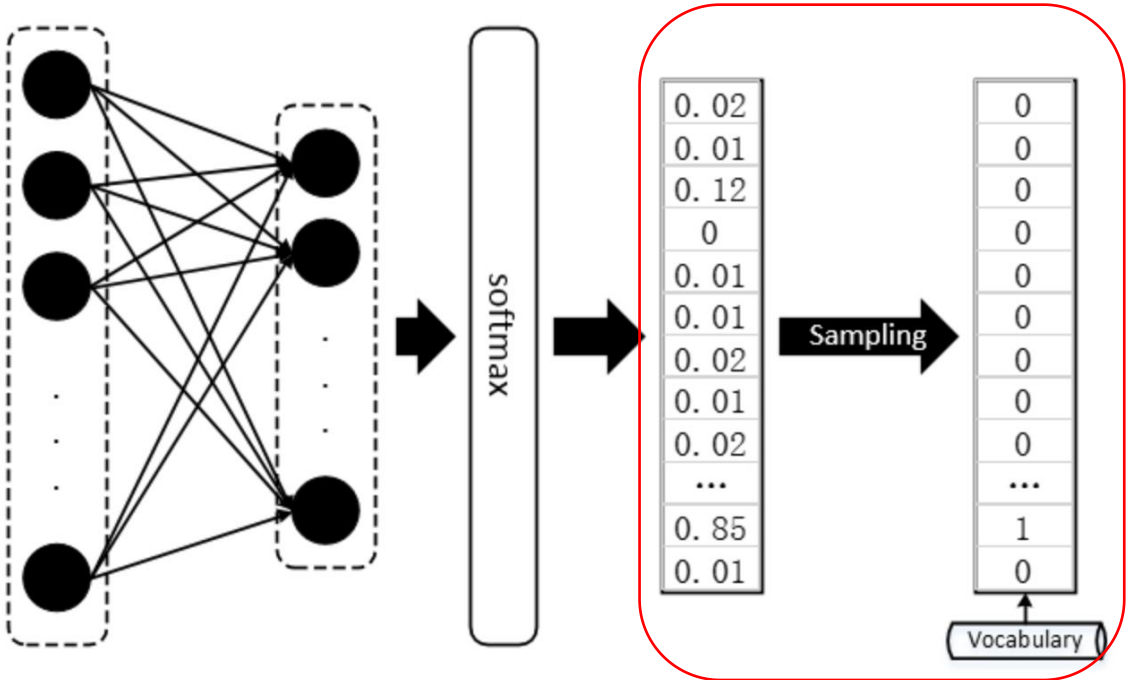


$$argmin_G max_D = V(G, D)$$

- Question: Can GANs be applied to text generation tasks ?
- Answer: Yes. But …

GAN-based

- Challenges
  - The discrete outputs are difficult to pass gradient update to the generative model
  - The discriminative model can only assess a complete sequence



GAN-based

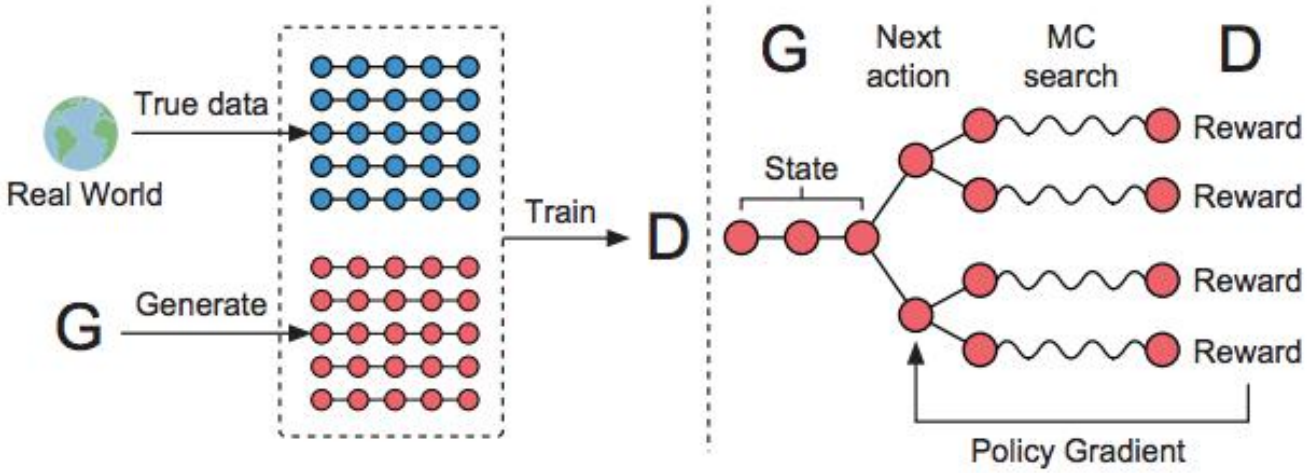
- Solution:
  - Gumbel-softmax
  - Wasserstein GAN (WGAN)
    - Replace KL-Divergence with Wasserstein-Divergence
  - WGAN-GP
    - Add gradient penalty (GP) to WGAN

WGAN with gradient penalty	
Busino game camperate spent odea	Solice Norkedin pring in since
In the bankaway of smarling the	ThiS record (31.) UBS) and Ch
SingersMay , who kill that imvic	It was not the annuas were plogr
Keray Pents of the same Reagun D	This will be us , the ect of DAN
Manging include a tudancs shat "	These leaded as most-worsd p2 a0
His Zuith Dudget , the Denmbern	The time I paidOa South Cubry i
In during the Uitational questio	Dour Fraps higs it was these del
Divos from The ’ noth ronkies of	This year out howneed allowed lo
She like Monday , of macunsuer S	Kaulna Seto consficutes to repor
The investor used ty the present	A can teal , he was schoon news
A papees are cointry congress oo	In th 200. Pesish picriers rega
A few year inom the group that s	Konney Panice rimimber the teami
He said this syenn said they wan	The new centuct cut Denester of
As a world 1 88 ,for Autouries	The near , had been one injostie
Foand , th Word people car , Il	The incestion to week to shorted
High of the upseader homing pull	The company the high product of
The guipe is worly move dogsfor	20 - The time of accomplete , wh
The 1874 incidested he could be	John WVuderenson seqiivic spends
The allo tooks to security and c	A ceetens in indestredly the Wat
Standard GAN objective	
ddddddddddddddddddddddddddddddd	ddddddddddddddddddddddddddddddd
ddddddddddddddddddddddddddddddd	ddddddddddddddddddddddddddddddd

[Gulrajani at el., 2017, Improved Training of Wasserstein GANs]

GAN-based

- Solution:
  - SeqGAN
    - RL+GAN

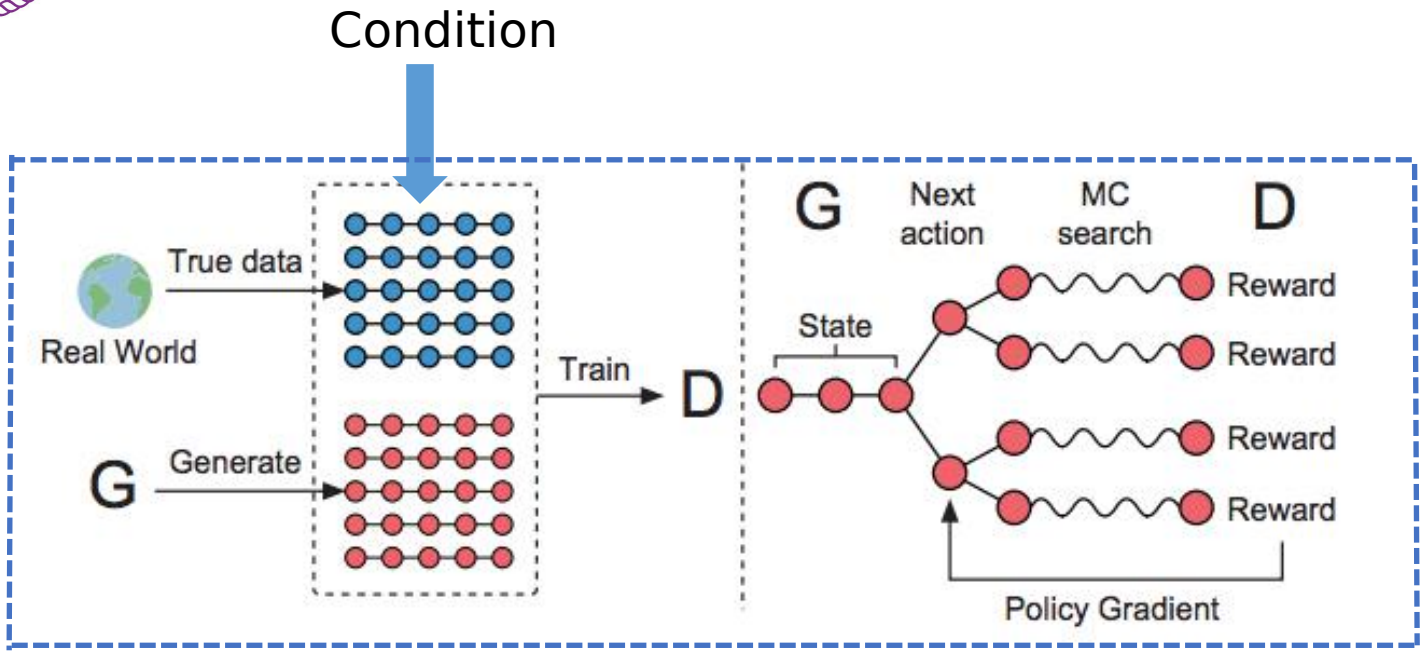


The left D is trained over the real data and the generated data by G.

The right G is trained by policy gradient where the final reward signal is provided by D and is passed back to the intermediate action value via Monte Carlo search.

[Yu at el., 2016, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient]

Conditional SeqGAN



Input	tell me ... how long have you had this falling sickness ?
Vanilla-MLE	i 'm not a doctor .
Vanilla-Sample	well everything you did was totally untrue .
REINFORCE	i don 't know how long it 's been .
REGS Monte Carlo	A few months, I guess .
Input	so i had the doctors test sammy 's response to conditioning .
Vanilla-MLE	sammy wrote the test sammy wrote the test .
Vanilla-Sample	objects pick objects up objects objects objects objects objects
REINFORCE	i 'm not sure that 's relevant .
REGS Monte Carlo	so he took the pills .

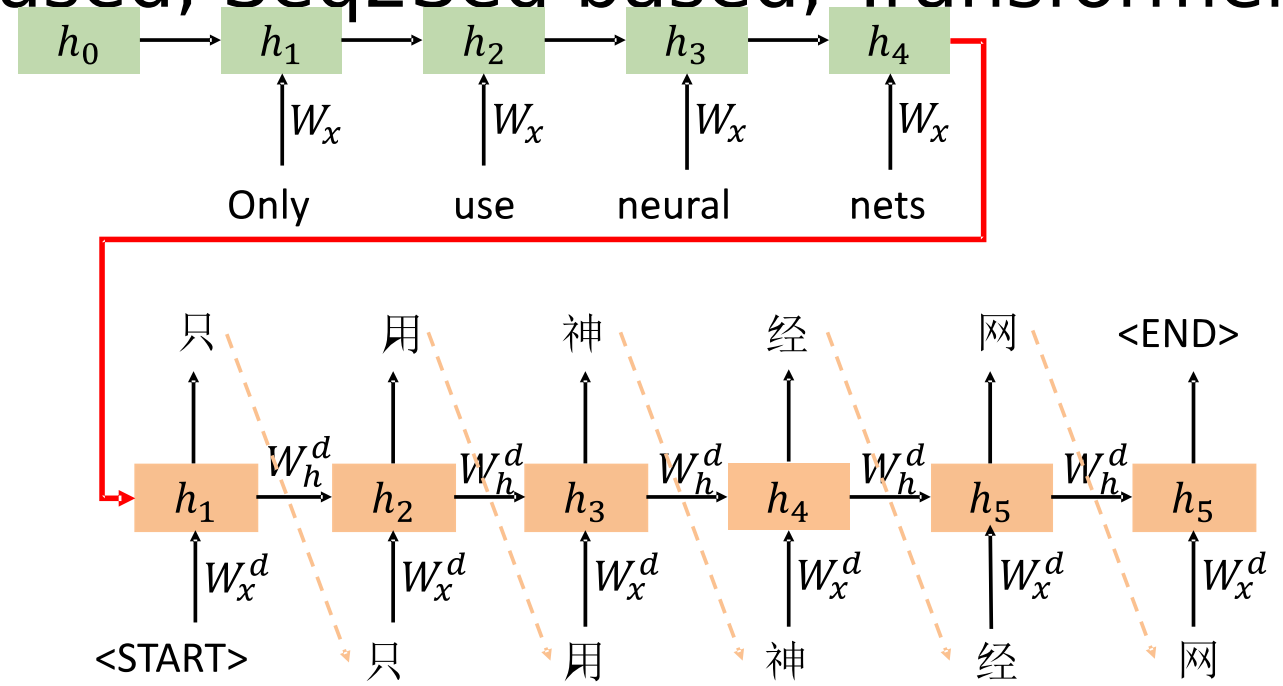
[Li at el., 2017, Adver-sarial Learning for Neural Dialogue Generation]

Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
  - Autoregressive
  - Non-Autoregressive
- Text Generation Tasks and Challenges
- Current Trends, and the Future

Autoregressive

- Given a source $x = (x_1, x_2, \ldots, x_n)$ and target $y = (y_1, y_2, \ldots, y_m)$
- $P(y|x) = \prod_{t=1}^{m} P(y_t|y_{<t}, x, \theta_{enc}, \theta_{dec})$
- RNN based, Seq2Seq based, Transformer based

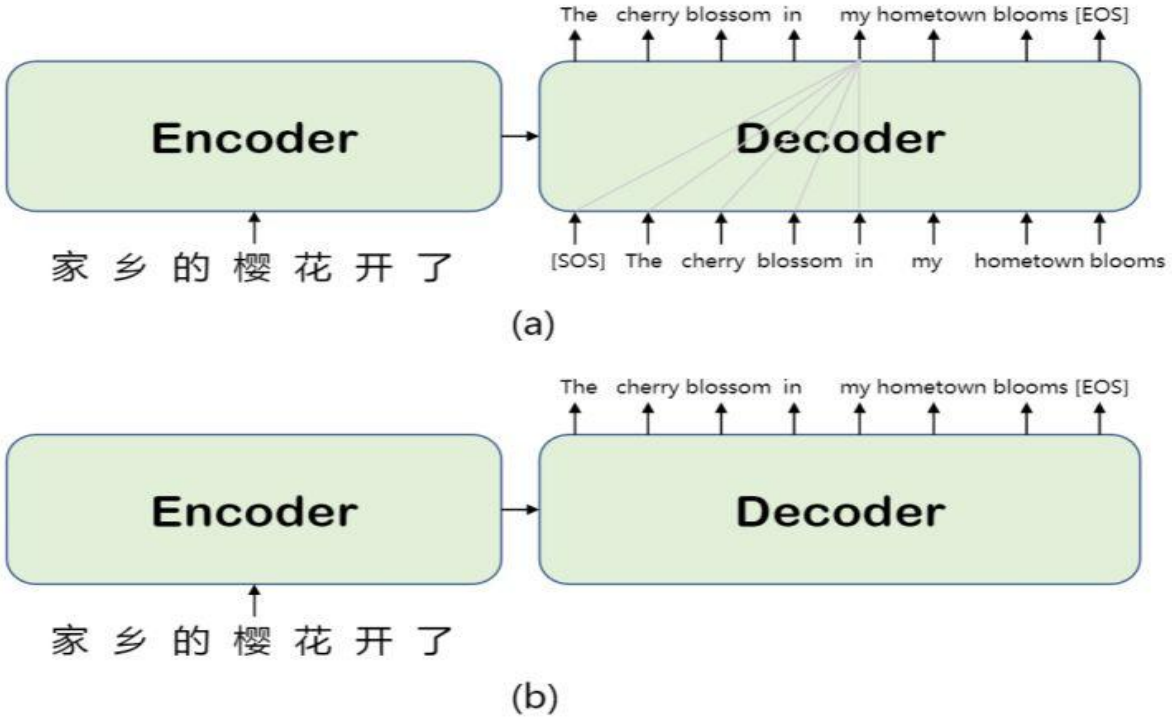


Autoregressive

- The individual steps of the decoder must be run sequentially rather than in parallel
- Time complexity
- Lack of global information
  - Decoding and Transformer based models try to solve this issue

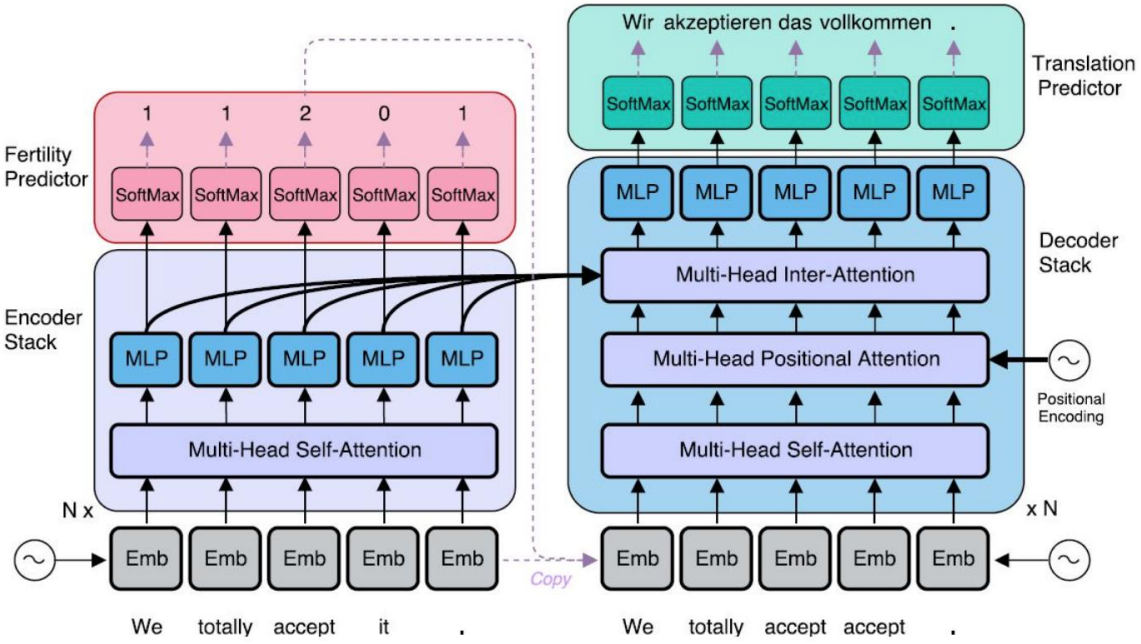
Non-Autoregressive

- Given a source $x = (x_1, x_2, \ldots, x_n)$ and target $y = (y_1, y_2, \ldots, y_m)$
- $P(y|x) = P(m|x) \prod_{t=1}^{m} P(y_t|z, x)$



Non-Autoregressive

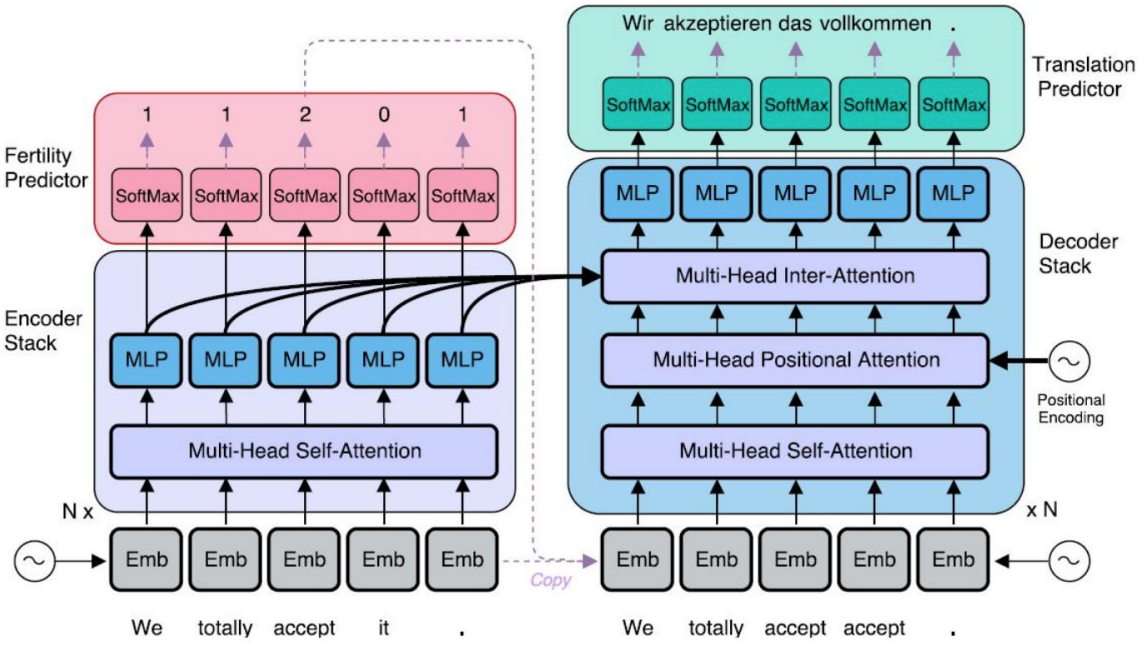
- $P(y|x) = P(m|x) \prod_{t=1}^{m} P(y_t|z, x)$
- Decide the length of the target sequence
- The encoder are the same as non-autoregressive encoder;
- Input $z = f(x; \theta_e$
- Generate the tar
  sequence paralle



[Gu† at el., 2018, NON-AUTOREGRESSIVE NEURAL MACHINE TRANSLATION]

Non-Autoregressive

- Decode target output at the same time
- Fast (20x than Autoregressive)
- Can keep context information well during decoding
- Similar to BERT masked LM deco

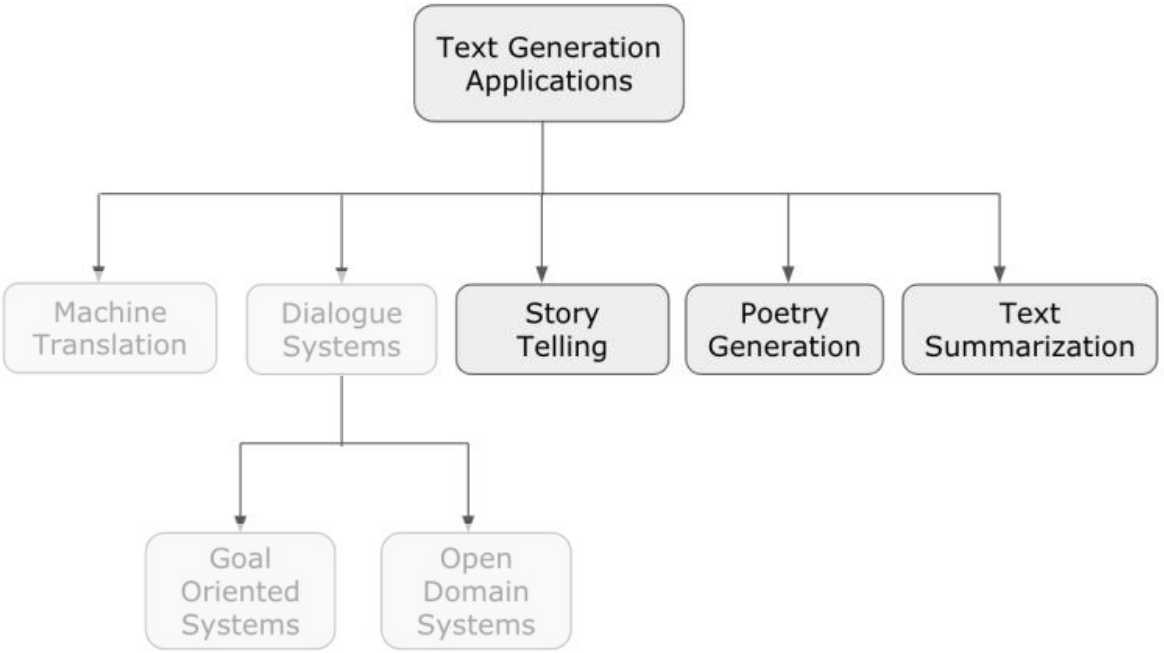


Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
- Text Generation Tasks and Challenges
  - Text Generation Tasks
  - Control Text Generation
  - Knowledge-guided Text Generation
- Current Trends, and the Future

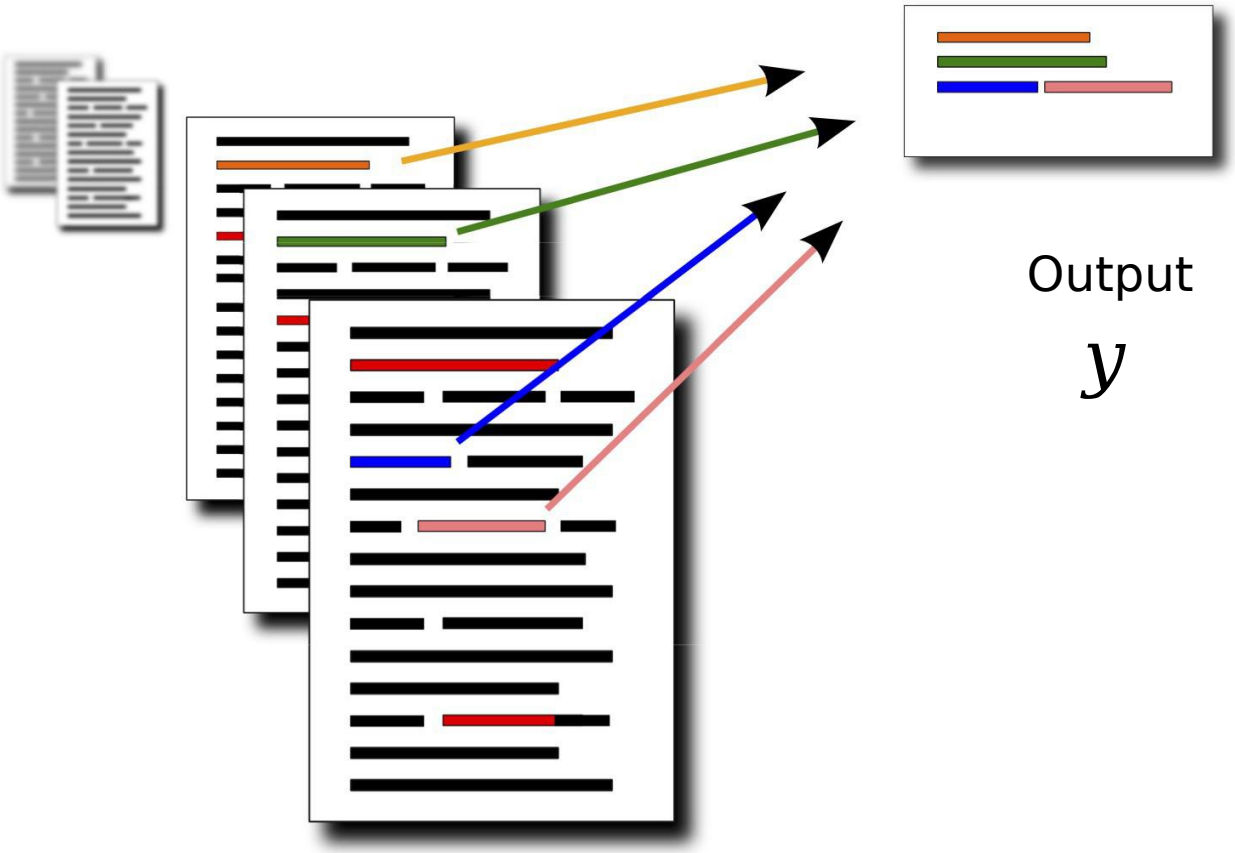
Text Generation Tasks

- Focus on: Text Summarization, Storytelling, and Poetry Generation
- Scenario: Text-to-Text, Data-To-Text (**variety of data**)
- Challeng



[Gatt at el., 2017, Survey of the State of the Art in Natural Language Generation- Core tasks, applications and evaluati

Summarization



Summarization

Extractive summarization

Select parts (typically sentences) of the original text to form a summary.

- Easier
- Restrictive (no paraphrasing)

Abstractive summarization

Generate new text using natural language generation techniques.

- More difficult
- More flexible (more human)

Summarization: Pre-neural

- Pre-neural summarization systems are mostly extractive
- They typically had a pipeline:
  - **Content selection**: Sentence scoring functions, Graph-based algorithms
  - Information ordering
  - Sentence realization

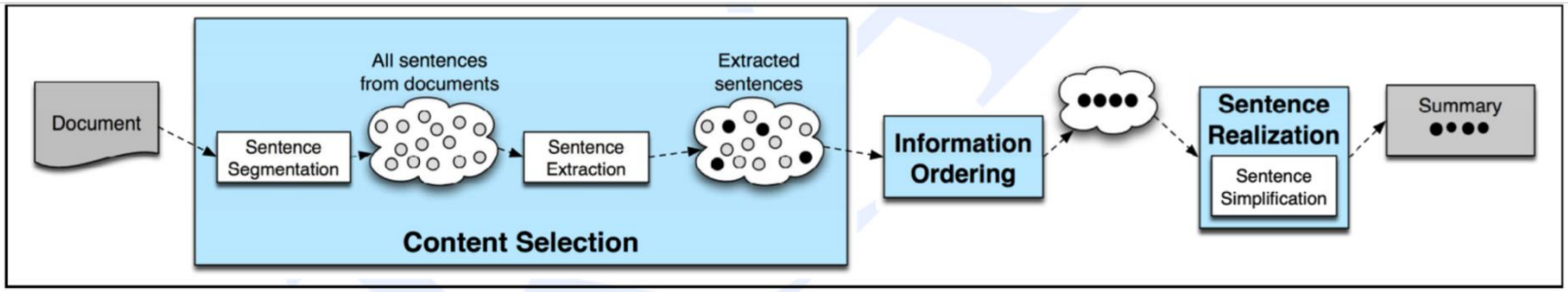
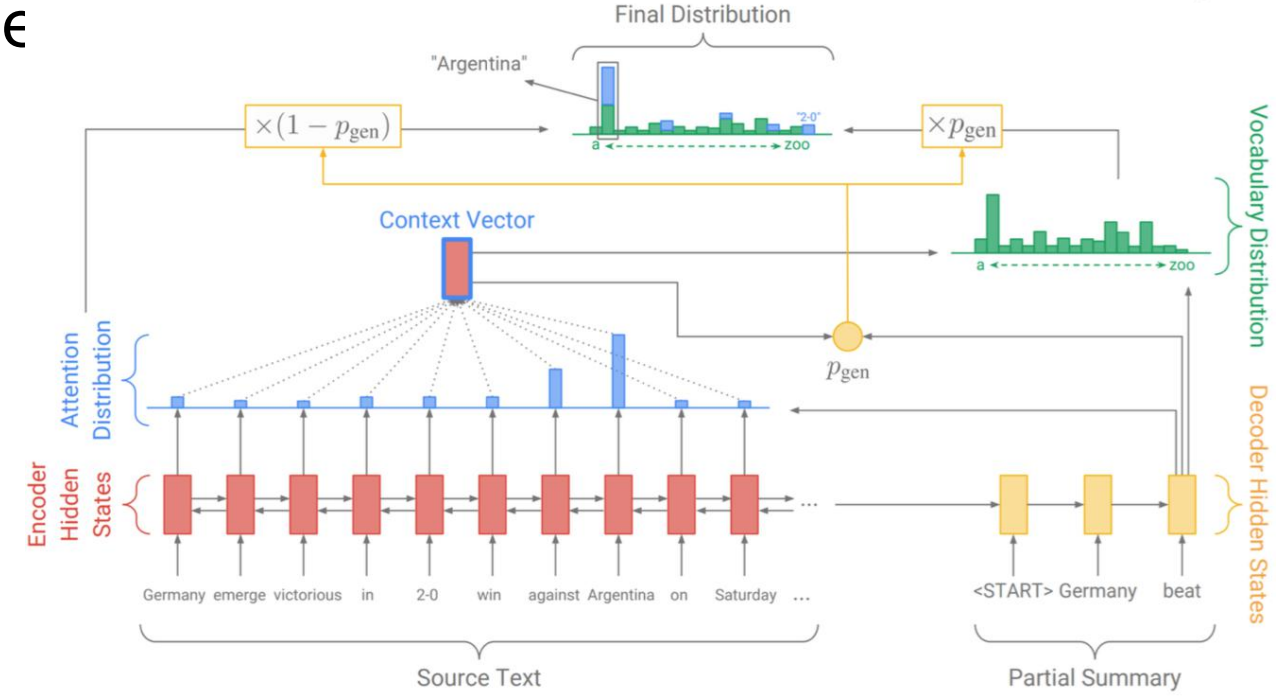


Figure 23.14 The basic architecture of a generic single document summarizer.

Summarization: Neural

- Neural Summarization: Copy mechanisms
  - Calculate p_{gen}, the probability of generating the next word.
  - The final distribu$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t$ion, and the



See et al, 2017, Get To The Point: Summarization with Pointer-

Summarization: Challenges

- Copy mechanisms copy too much
- An abstractive system collapses to a mostly extractive system
- Bad at overall content selection, especially if the input document is long
- No overall strategy for selecting content

Summarization: Better Content Selection

- Better content selection
  - Pre-neural summarization had separate stages for **content selection** and **surface realization** (i.e. text generation)
  - In a standard End2End (seq2seq+attention) summarization system, these two stages are mixed in together
  - One solution: bottom-up summarization

Summarization: Better Content Selection

- Solution: Bottom-up summarization
  - **Content selection stage**
  - **Bottom-up attention stage**

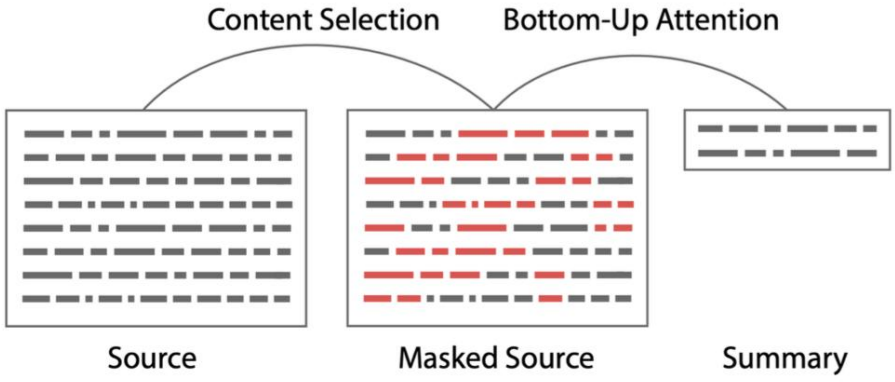


Figure 2: Overview of the selection and generation processes described throughout Section 4.

Simple but effective!

- Better overall content selection strategy
- Less copying of long sequences (i.e. more abstractive output)

Storytelling

- Use the given data to generate text
  - (**Image-to-Text**) Generate a story-like paragraph given an image
  - (**Prompt-to-Text**) Generate a story given a brief writing prompt
  - (**Event-to-Text**) Generate a story given events, entities, state, etc.
  - (**LongDocument-to-Text**) Generate the next sentence of a story, given the story so far (story continuation)
    - This is different to the previous one, because we are not concerned with the system's performance over several generated sentences

Storytelling: Prompt-to-Text

- Generating a story from a writing prompt
  - Each story has an associated brief writing prompt.

Input:

Output:

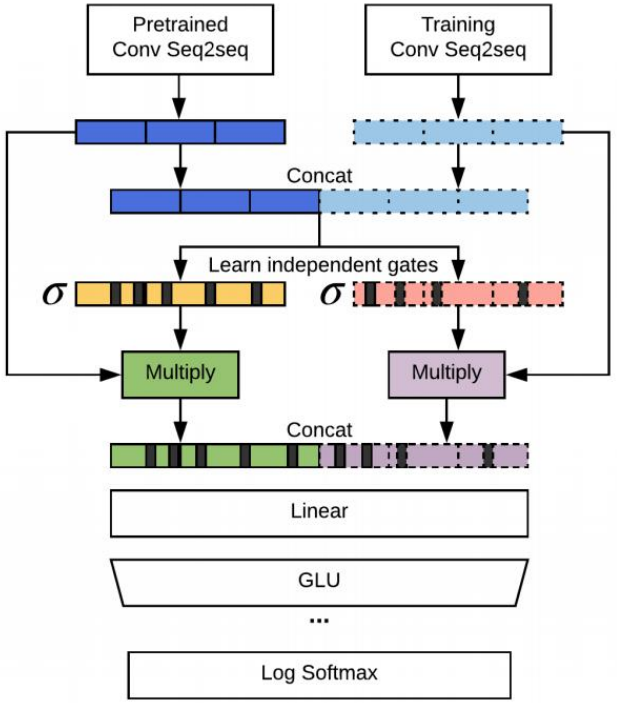
Prompt: The Mage, the Warrior, and the Priest

Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

[Fan et al ,2018, Hierarchical Neural Story Generation]

Storytelling: Prompt-to-Text

- A complex seq2seq prompt-to-story model
  - It's convolutional-based
  - Gated multi-head multi-scale self-attention
  - Model fusion



[Fan et al ,2018, Hierarchical Neural Story Generation]

Storytelling: Prompt-to-Text

- The results are impressive!
  - Related to prompt
  - Diverse; non-generic
  - Stylistically dramatic

Example Prompt 1: Aliens start abducting humans

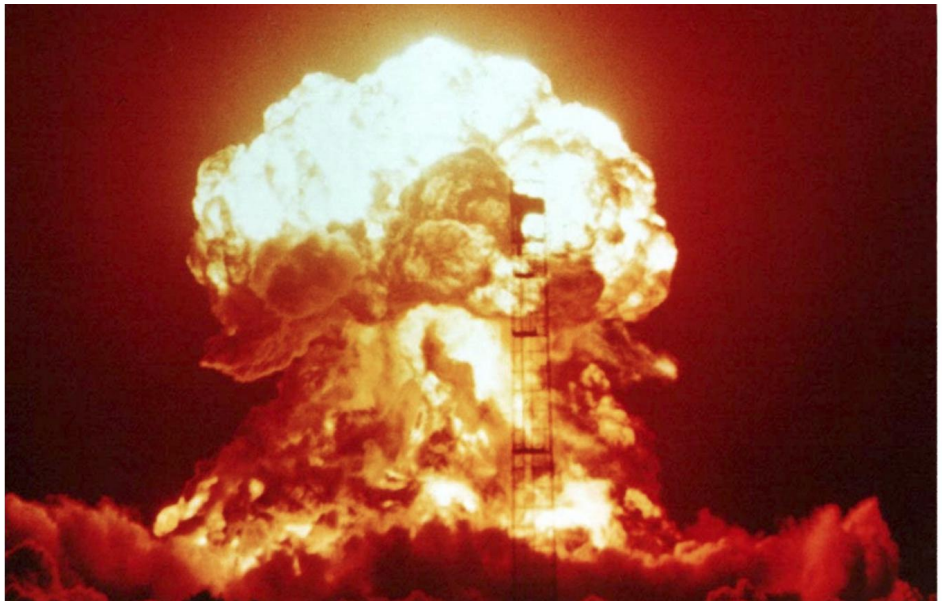
It has been two weeks, and the last of my kind has gone. It is only a matter of time until there will be nothing left. I'm not sure what the hell is going on... I can't think. I can hear a distant scream. I think of a strange, alien sound. I try to speak, but am interrupted by something, something that feels like a drum, I ca not tell. I mean I'm just a little bit older than an average human. But I can, and I can feel the vibrations . I hear the sound of a ship approaching. The ground quakes at the force of the impact, and a loud explosion shatters the silence.

- However:
  - Mostly atmospheric/descriptive/scene-setting; less events/plot
  - When generating for longer, mostly stays on the same idea without moving forward to new ideas – coherence issues

Storytelling: Image-to-Text

- Generating a story from an image

Input



Output

Generated story about image
Model: Taylor Swift Lyrics

"You have to be the only light bulb in the night sky, I thought, Oh, God, it 's so dark out of me that I missed you , I promise."

What's interesting here is that this isn't straightforward supervised image-captioning. There was no paired data to learn from.

Storytelling: Event-to-Text

- Event2event Story Generation

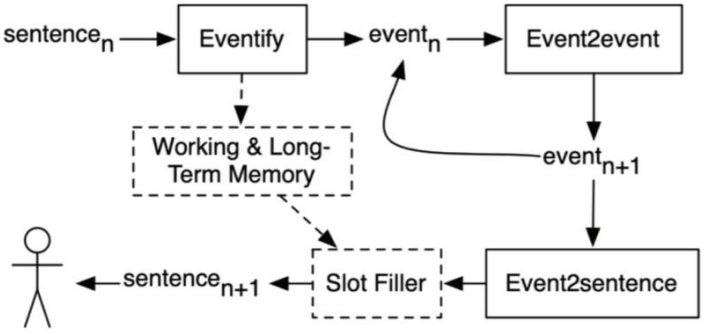


Figure 1: Our automated story generation pipeline. Dashed boxes and arrows represent future work.

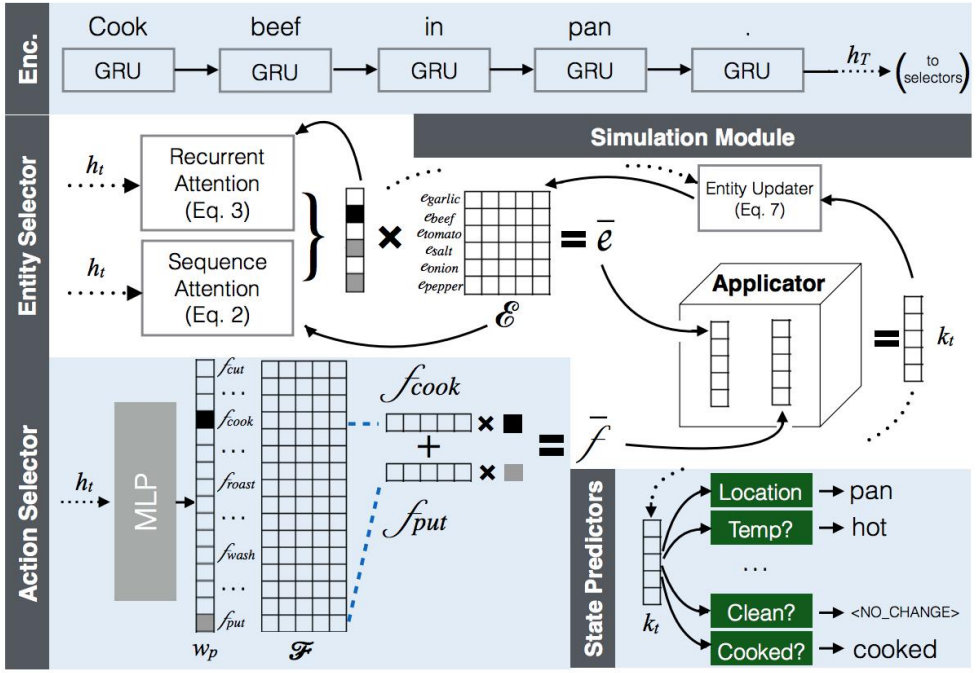
Input

Output

Experiment	Input	Extracted Event(s)	Generated Next Event(s)	Generated Next Sentence
All Generalized Events & Generalized Sentence	He reaches out to Remus Lupin, a Defence Against the Dark Arts teacher who is eventually revealed to be a werewolf.	⟨male.n.02, get-13.5.1, ∅, <CHAR>0⟩ ⟨ORGANIZATION, say-37.7-1, monster.n.01, ∅⟩	⟨monster.n.01, amuse-31.1, sarge, ∅⟩ ⟨monster.n.01, amuse-31.1, realize, ∅⟩ ⟨monster.n.01, conjecture-29.5-1, ∅, ∅⟩ ⟨male.n.02, conduit.n.01, entity.n.01, ∅⟩ ⟨male.n.02, free-80-1, ∅, penal_institution.n.01⟩	When monster.n.01 nemesis.n.01 describes who finally realizes male.n.02 can not, dangerous entity.n.01 male.n.02 is released from penal_institution.n.01.

Storytelling: Event-to-Text

- There's been lots of work on tracking events/ entities/state in neural NLU (natural language understanding)
- However, applying these methods to NLG is even more difficult
- Research: Yejin Choi group



Storytelling: Table-to-Text

- Case: BabyTalk

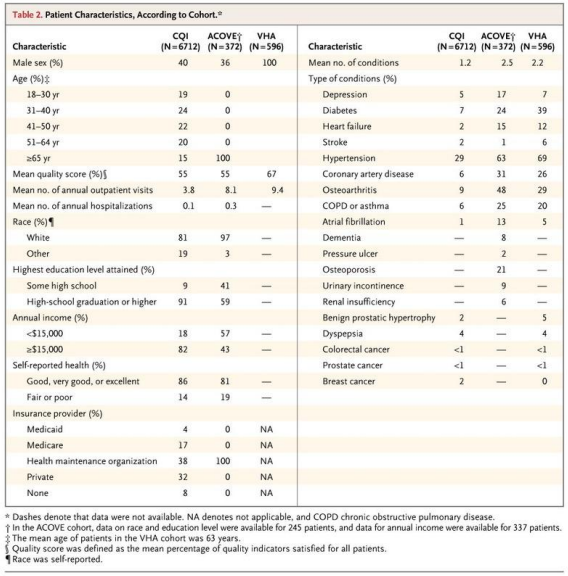
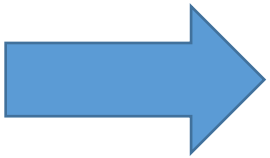


Table 2. Patient Characteristics, According to Cohort.*

Characteristic	CQI (N=6712)	ACOVE† (N=372)	VHA (N=596)	Characteristic	CQI (N=6712)	ACOVE† (N=372)	VHA (N=596)
Male sex (%)	40	36	100	Mean no. of conditions	1.2	2.5	2.2
Age (%)‡				Type of conditions (%)			
18–30 yr	19	0		Depression	5	17	7
31–40 yr	24	0		Diabetes	7	24	39
41–50 yr	22	0		Heart failure	2	15	12
51–64 yr	20	0		Stroke	2	1	6
≥65 yr	15	100		Hypertension	29	63	69
Mean quality score (%)§	55	55	67	Coronary artery disease	6	31	26
Mean no. of annual outpatient visits	3.8	8.1	9.4	Osteoarthritis	9	48	29
Mean no. of annual hospitalizations	0.1	0.3	—	COPD or asthma	6	25	20
Race (%)¶				Atrial fibrillation	1	13	5
White	81	97	—	Dementia	—	8	—
Other	19	3	—	Pressure ulcer	—	2	—
Highest education level attained (%)				Osteoporosis	—	21	—
Some high school	9	41	—	Urinary incontinence	—	9	—
High-school graduation or higher	91	59	—	Renal insufficiency	—	6	—
Annual income (%)				Benign prostatic hypertrophy	2	—	5
<$15,000	18	57	—	Dyspepsia	4	—	4
≥$15,000	82	43	—	Colorectal cancer	<1	—	<1
Self-reported health (%)				Prostate cancer	<1	—	<1
Good, very good, or excellent	86	81	—	Breast cancer	2	—	0
Fair or poor	14	19	—				
Insurance provider (%)							
Medicaid	4	0	NA				
Medicare	17	0	NA				
Health maintenance organization	38	100	NA				
Private	32	0	NA				
None	8	0	NA				

* Dashes denote that data were not available. NA denotes not applicable, and COPD chronic obstructive pulmonary disease.
† In the ACOVE cohort, data on race and education level were available for 245 patients, and data for annual income were available for 337 patients.
‡ The mean age of patients in the VHA cohort was 63 years.
§ Quality score was defined as the mean percentage of quality indicators satisfied for all patients.
¶ Race was self-reported.

Table



[*for real-time clinical decision support*]

By 11:00 the baby had been hand-bagged a number of times causing 2 successive bradycardias. She was successfully re-intubated after 2 attempts. The baby was sucked out twice.

At 11:02 FIO2 was raised to 79%.

Doctor

BT-Nurse text (extract)

[*for nursing shift handover*]

Respiratory Support
Current Status
Currently, the baby is on CMV in 27 % O2. Vent RR is 55 breaths per minute. Pressures are 20/4 cms H2O. Tidal volume is 1.5.

SaO2 is variable within the acceptable range and there have been some desaturations.

...

Events During the Shift
A blood gas was taken at around 19.45. Parameters were acceptable. pH was 7.18. CO2 was 7.71 kPa. BE was -4.8 mmol/L.

Nurse

BT-Family text (extract)

[*to keep parents informed, especially if not at hospital*]

John was in intensive care. He was stable during the day and night. Since last week, his weight increased from 860 grams (1 lb 14 oz) to 1113 grams (2 lb 7 oz). He was nursed in an incubator.

Yesterday, John was on a ventilator. The mode of ventilation is Bilevel Positive Airway Pressure (BiPAP) Ventilation. This machine helps to provide the support that enables him to breathe more comfortably. Since last week, his inspired Oxygen (FiO2) was lowered from 56 % to 21 % (which is the same as normal air). This is a positive development for your child.

During the day, Nurse Johnson looked after your baby. Nurse Stevens cared for your baby during the night.

Family

Storytelling: Challenges

- Challenges in storytelling
  - Stories generated by neural LMs can sound fluent... but are meandering, nonsensical, with no coherent plot

What's missing?

- LMs model sequences of words. Stories are sequences of events.
- To tell a story, we need to understand and model:
  - Events and the causality structure between them
  - Characters, their personalities, motivations, histories, and relationships to other characters
  - State of the world (who and what is where and why)
  - Narrative structure (e.g. exposition → conflict → resolution)
  - Good storytelling principles (don't introduce a story element then never use it)

Storytelling: Challenges

- **Challenges in storytelling**
  Stories generated by neural LMs can sound fl
  but are meandering, nonsensical, with no

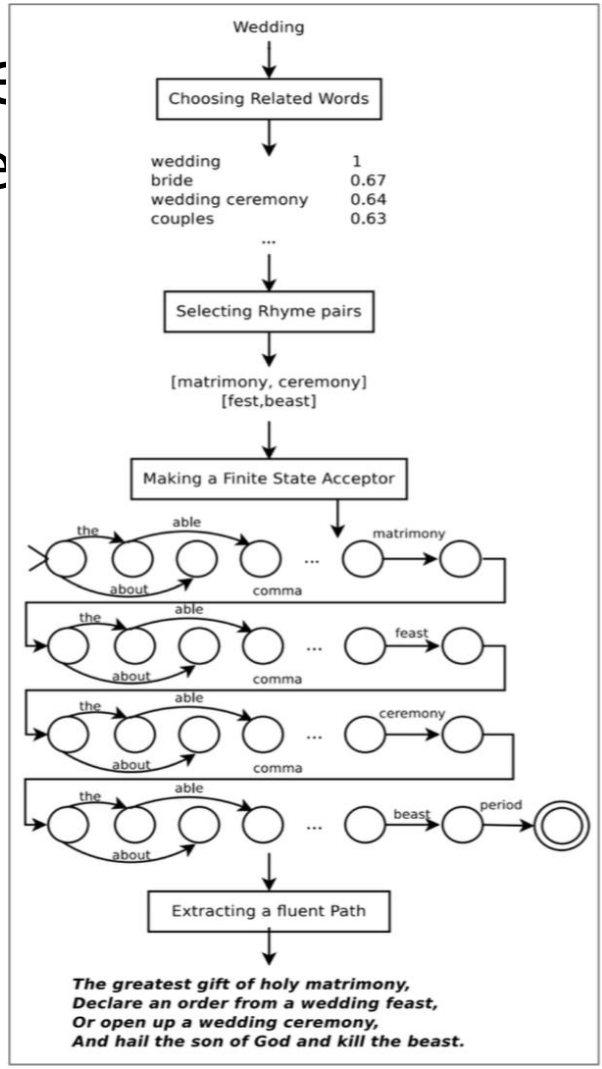
What's missing?

- LMs model sequenc s are sequences of e
- To tell a st understand and model:
  - E y structure between them
    rsonalities, motivations, histories,
    ships to other characters
    te of the world (who and what is where and why)
  - Narrative structure (e.g. exposition → conflict → resolution)
  - Good storytelling principles (don't introduce a story element then never use it)

This is Incredible Difficult

Poetry Generation: Text-to-Text

- Hafez: a poetry generation syste
  - User provides topic words and the will return a poetry for you.



Ghazvininejad et al, 2016, Generating Topical Poetry

Poetry Generation: Text-to-Text

- In a follow-up paper, the authors made the sy interactive and user-controllable
- The control method i simple: during beam search, upweight the scores of words t have the desired features

(a) Poem generated with default style settings

Ghazvininejad et al, 2017, Hafez: an Interactive Poetry Generation System

Use case

- Automated reporting: Financial reports, robot journalism, regulatory reports
- Better dialogue: Chatbots, virtual assistants, computer games
- Decision support: Customer service
- Assistive technology: Summarize graphs for visually impaired

Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
- Text Generation Tasks and Challenges
  - Text Generation Tasks
  - Control Text Generation
  - Knowledge-guided Text Generation
- Current Trends, and the Future

Controlling Text Generation

- Control text generation:
  - Fine-tuned
  - Conditional
  - Plug and play Language Model (PPLM)

Model type	Form of model	Samples	Example models and number of trainable params
Fine-tuned Language Model	$p(x)$	Uncond.	Fine-tuned GPT-2 medium: 345M (Ziegler et al., 2019)
Conditional Language Model	$p(x\|a)$	Cond.	CTRL: 1.6B (Keskar et al., 2019)
Plug and Play Language Model (PPLM)	$p(x\|a) \propto p(x)p(a\|x)$	Cond.	PPLM-BoW: 0 (curated word list) PPLM-Discrim: 4K (not counting pretrained $p(x)$)

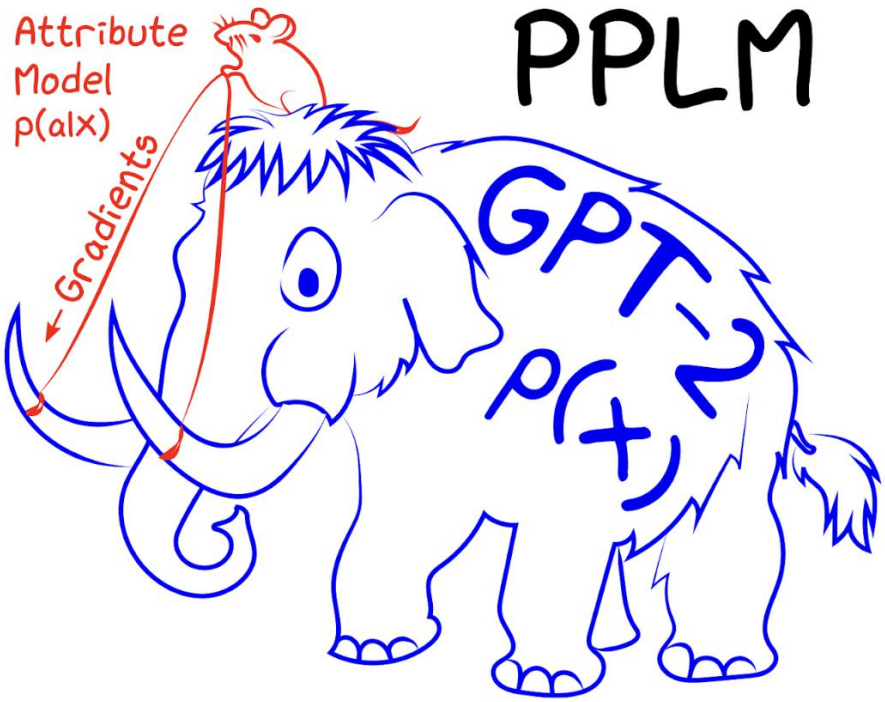
Controlling Text Generation

- Example: PGT-2 + PPLM
  - Control the generation

[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though…

[Negative] *The potato* is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you…

[Positive] *The potato* chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them…



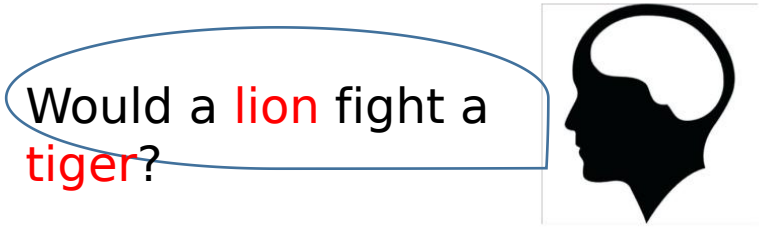
[Uber AI, 2019, Plug and Play Language Models: a Simple Approach to Controlled Text Generation]

Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
- Text Generation Tasks and Challenges
  - Text Generation Tasks
  - Control Text Generation
  - Knowledge-guided Text Generation
- Current Trends, and the Future

Knowledge-based

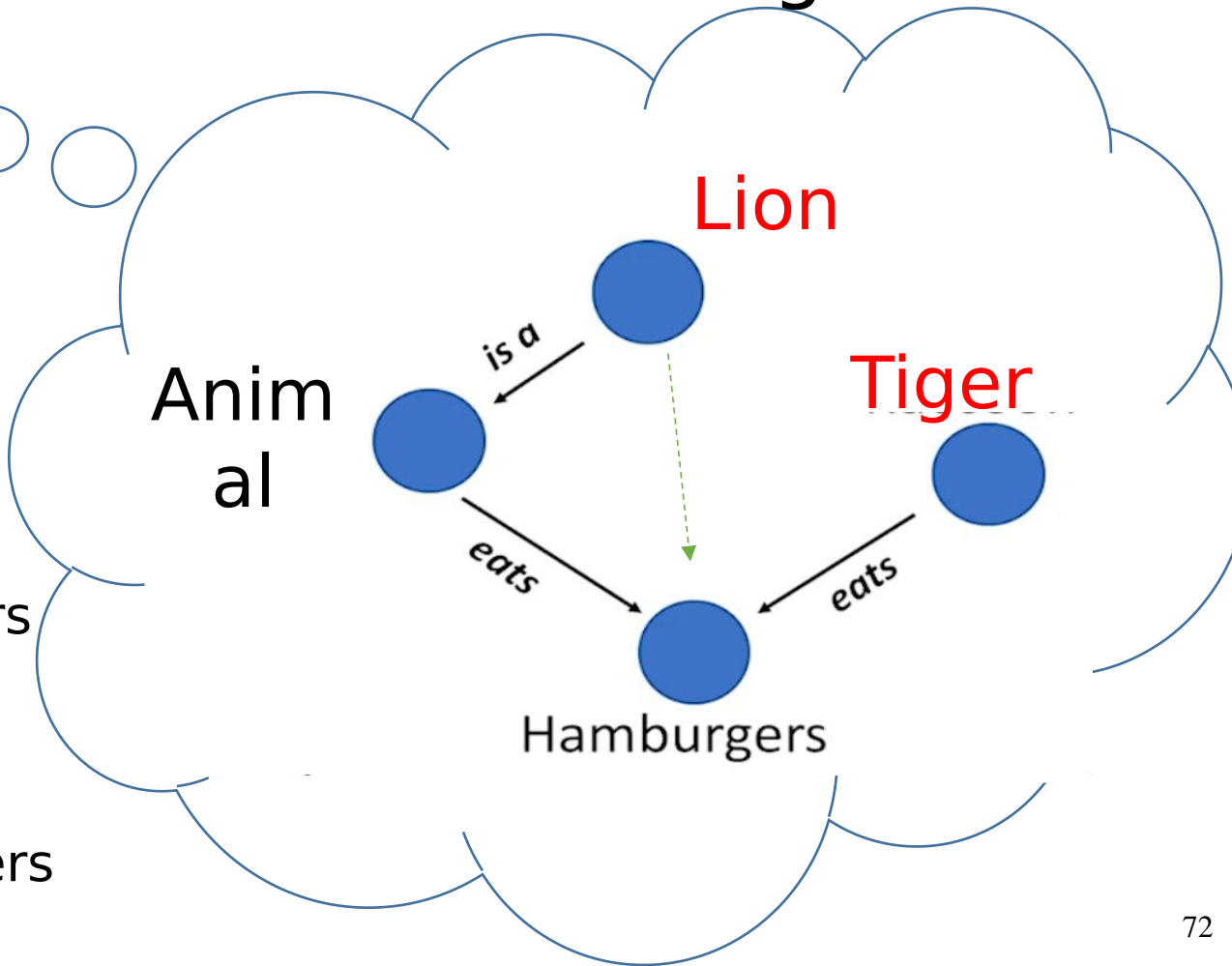
- In the real world: …



Knowledge-based

- In the real world: …

Knowledge



Lion eats Hamburgers

competitor

Tiger eats Hamburgers

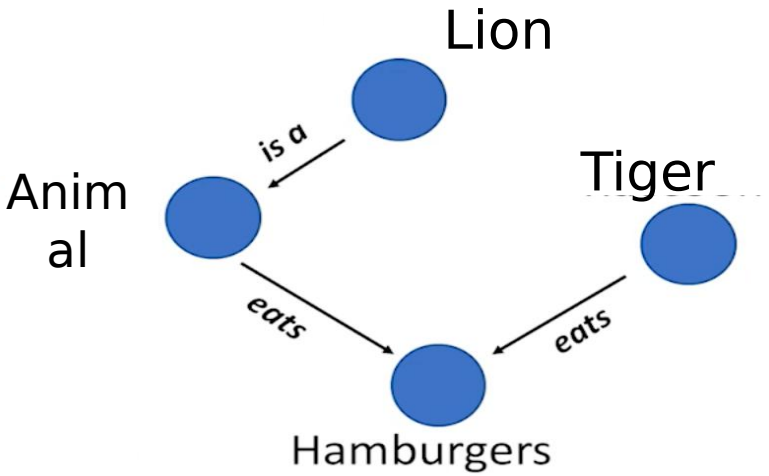
Knowledge-based

- In the real world: …

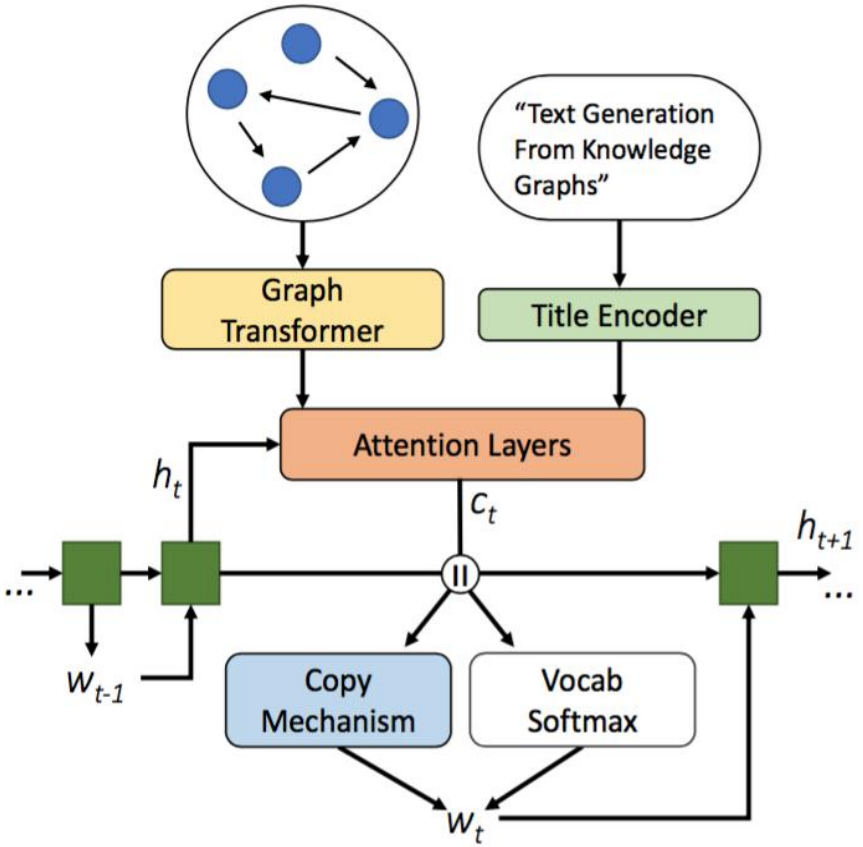
Yes, Lion is animal. Both Lions and tigers eat hamburgers, and so may fight over a hamburger

Knowledge-based

- Incorporate knowledge



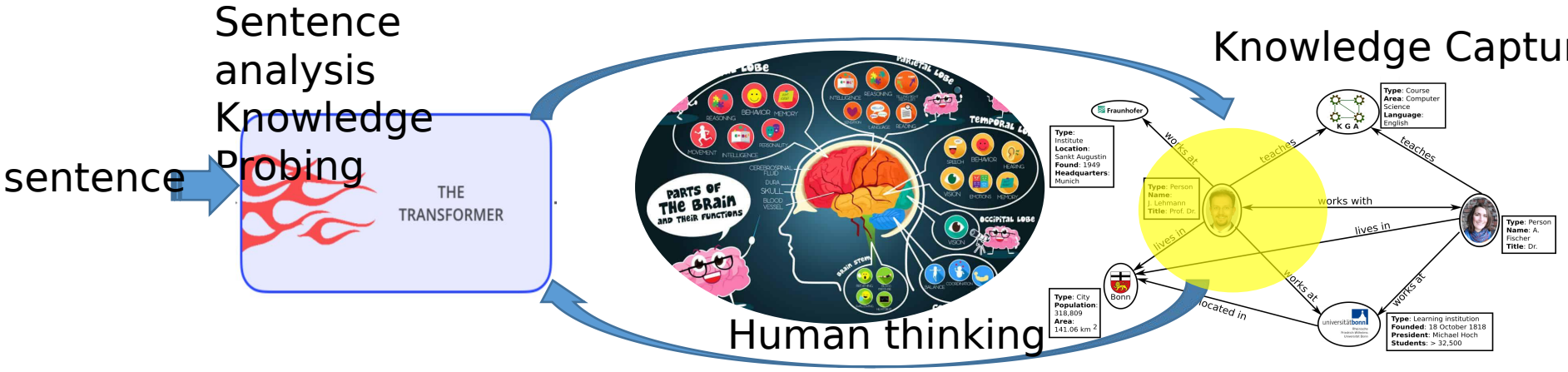
Yes, Lion is animal. Both Lions and tigers eat hamburgers, and so may fight over a Discarded hamburger



[Koncel-Kedziorski at el., 2019, Text Generation from Knowledge Graphs]

Knowledge-based

- Human thinking process
  - Language models capture knowledge and generate language
    - Text Generation from Knowledge Graphs with Graph Transformers



Outline

- Introduction to Text Generation
- Traditional Text Generation
- Neural Text Generation
- Text Generation Tasks and Challenges
- Current Trends, and the Future

Current Trends, and the Future

- Incorporating knowledge into text generation
  - May help with modeling knowledge in tasks that really need it, like storytelling, task-oriented dialogue, etc
- Alternatives to strict left-to-right generation
  - Parallel generation, iterative refinement, top-down generation for longer pieces of text
- Alternatives to maximum likelihood training with teacher forcing
  - More holistic sentence-level (rather than

Current Trends, and the Future

- Text Generation research: Where are we? Where are we going?
  - 5 years ago, NLP + Deep Learning research was a wild west

  - Now (2021), it’s a lot less wild
  - Text Generation seems still like one of the wildest parts remaining

Current Trends, and the Future

- **Neural TG community is rapidly maturing**
  - During the early years of NLP + Deep Learning, the community was mostly transferring successful neural machine translation (NMT) methods to text generation (TG) tasks.
  - Increasingly more (neural) text generation workshops and competitions, especially focusing on open-ended TG
  - It would be particularly useful to organize the community, increase reproducibility, standardize evaluation, etc.

Reading Material

a. Survey
Tutorial on variational autoencoders [link]
Neural text generation: A practical guide [link]
Survey of the state of the art in natural language generation: Core tasks, applications and evaluation [link]4. **Neural Text Generation: Past, Present and Beyond** [link] 5. **Survey of the state of the art in natural language generation: Core tasks, applications and evaluation** [link]
b. Classic
A neural probabilistic language model [link] (NNLM)
Recurrent neural network based language model [link] (RNNLM)
Sequence to sequence learning with neural networks [link] (seq2seq)
c. VAE based
Generating Sentences from a Continuous Space [link]
Long and Diverse Text Generation with Planning-based Hierarchical Variational Model [link]

For reading material recommendation of this course, please refer to our github

Reading Material

d. GAN based

Adversarial feature matching for text generation [link] (TextGAN)

e. Knowledge based

Text Generation from Knowledge Graphs with Graph Transformers [link]

Neural Text Generation from Rich Semantic Representations [link]

For reading material recommendation of this course, please refer to our github


TSINGHUA UNIVERSITY
-1911-


Tsinghua University
NLP
Natural Language Processing

THUNLP