IR for NLP
Learning Structure for Text Generation

Niyati Chhaya

nchhaya@adobe.com

True or False

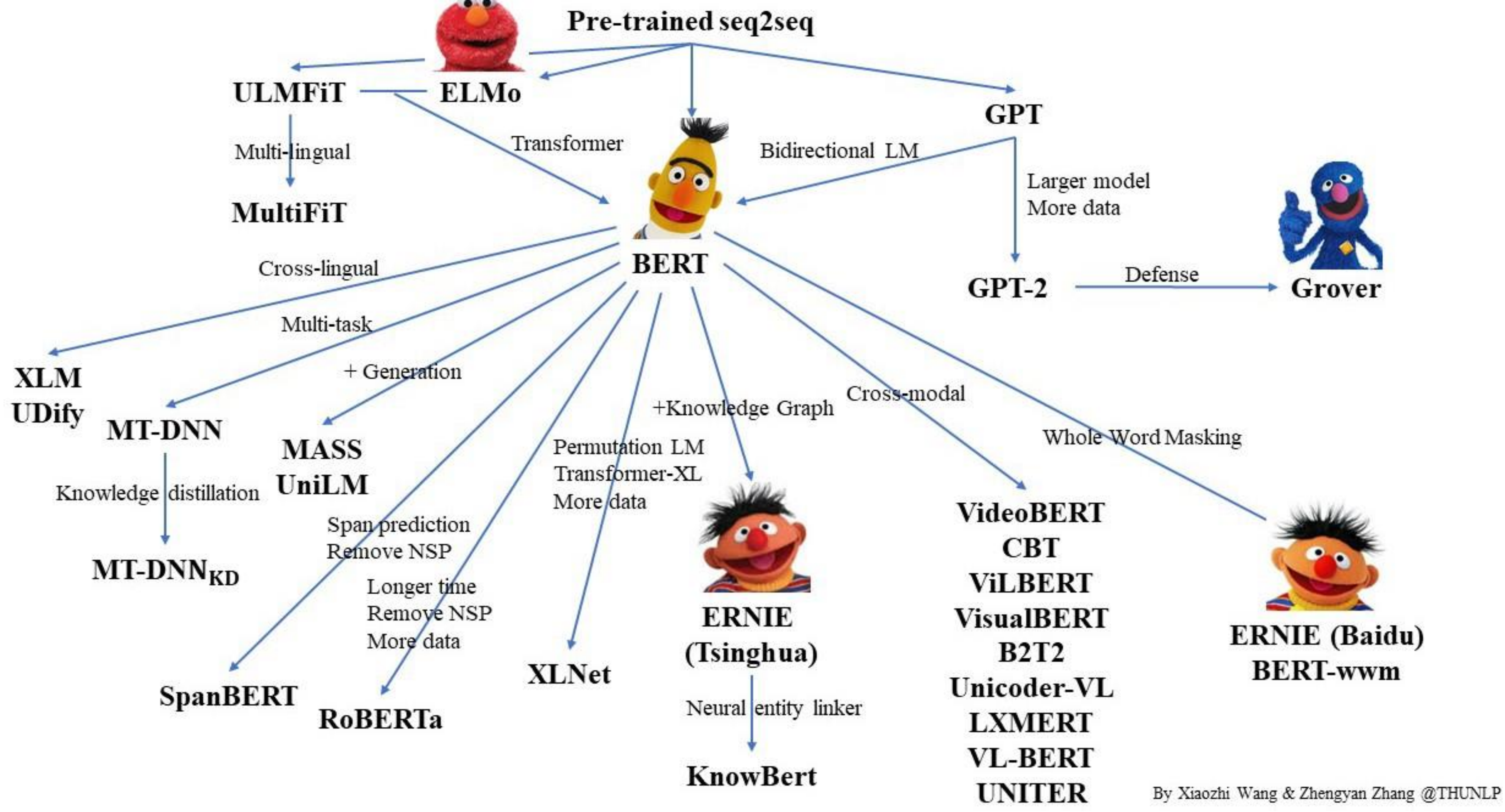
1. Deepweb is indexed by Google
2. You should always use words for indexing
3. Stemming is encouraged for IR applications
4. Syntactic analysis helps understand the meaning of words
5. Word embeddings are a type of semantic analysis
6. IR techniques cannot be applied to NLP tasks.

Pulse Check

1. What is NLP ?
2. Examples of NLP tasks ?
3. What applications need NLP approaches ?

Agenda

- Learning Structure for Text Generation
- AI for Legal Discovery
- Structured Document Retrieval


Pre-trained seq2seq
ULMFiT
ELMo
GPT
Multi-lingual
Transformer
Bidirectional LM
MultiFiT
Larger model
More data
BERT
Cross-lingual
GPT-2
Defense
Grover
Multi-task
XLM
UDify
+ Generation
MT-DNN
MASS
UniLM
Cross-modal
+Knowledge Graph
Whole Word Masking
Permutation LM
Transformer-XL
More data
Knowledge distillation
Span prediction
Remove NSP
MT-DNN$_{KD}$
Longer time
Remove NSP
More data
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
ERNIE
(Tsinghua)
ERNIE (Baidu)
BERT-wwm
SpanBERT
RoBERTa
XLNet
Neural entity linker
KnowBert
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

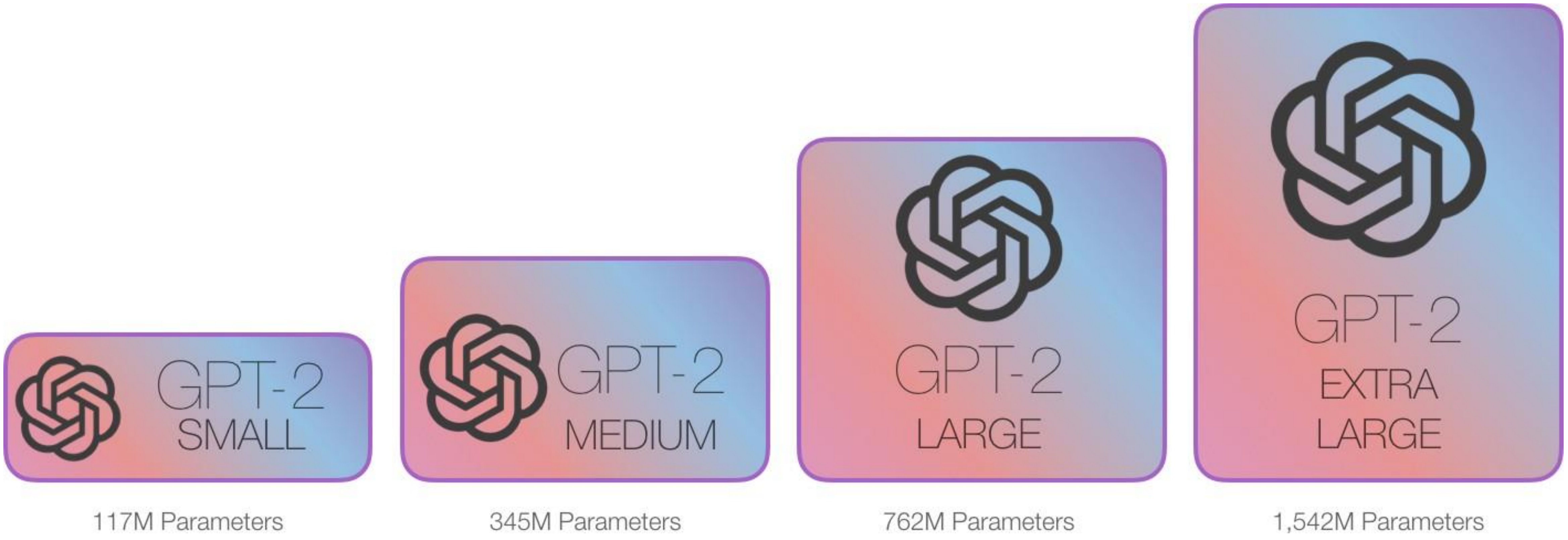
Question set A

1. Note down your name / Role number
2. List NLP approaches that you are aware of for generation ?
3. What is BERT ?
4. What is ELMo ?

Neural Language Models are stronger than ever!

Why do we still do research on ***text generation****?*

How good are today's AI writers !



Context (WebText test) as PROMPT	GPT-2 Completion
Aaron loves mint chocolate cake, but he requires that it be paired with mini chocolate chips, so I threw some of those in between the layers. I also had a few Peppermint Jo Jos on hand so I crushed them up and threw some of those in along with some crushed meringue cookies because, why not? It's a total smorgasbord of minty chocolate chippy cookie crunchy goodness. I didn't measure how much of each topping I used, but after I tasted the finished product, I wish I had added more. You can add anything you want- crushed candy canes, peppermint bark, etc. And don't be afraid to use a heavy hand. Texture = good. If you don't have 7-inch cake pans, you can get 3 shorter 8-inch layers out of this	recipe. If you do, and you know how to make that cake, then your life will never be the same. I have two very simple rules when it comes to cookie crumbs or chocolate chips. 1) When **I buy raw sugar**, I do not use it until it's cooled down and has at **least half a teaspoon of sugar on it**. That's my rule. And 2) **If I have a recipe that calls for something and I want to use it in a recipe,** I do not substitute it with something else on-line. Like this. I found this really cute card with cute little **kittens** on

Radford, *et al.* 2018

Context (WebText test) as PROMPT

Aaron loves mint chocolate cake, but he requires that it be paired with mini chocolate chips, so I threw some of those in between the layers. I also had a few Peppermint Jo Jos on hand so I crushed them up and threw some of those in along with some crushed meringue cookies because, why not? It's a total smorgasbord of minty chocolate chippy cookie crunchy goodness. I didn't measure how much of each topping I used, but after I tasted the finished product, I wish I had added more. You can add anything you want- crushed candy canes, peppermint bark, etc. And don't be afraid to use a heavy hand. Texture = good.

If you don't have 7-inch cake pans, you can get 3 shorter 8-inch layers out of this

GPT-2 Completion

recipe, which will likely yield about the same amount of cake.

Chocolate Buttermilk Cake Print Chunks of chocolate cake, coated in buttermilk and topped with crunchy chopped chocolate, drizzled with chocolate sauce. Simple but a nice change from chocolate cake!
Author: Momma Made
Ingredients For the chocolate buttermilk cake
1 1/2 cup unsalted butter, room temp.
1 1/2 cup granulated sugar
3 large eggs plus 1 egg yolk For the chocolate glaze 1/3 cup cocoa powder
1 3/4 cups powdered sugar 6 ounces…

Radford, *et al.* 2018

Weaknesses of MEGA Language Models for GENERATION!

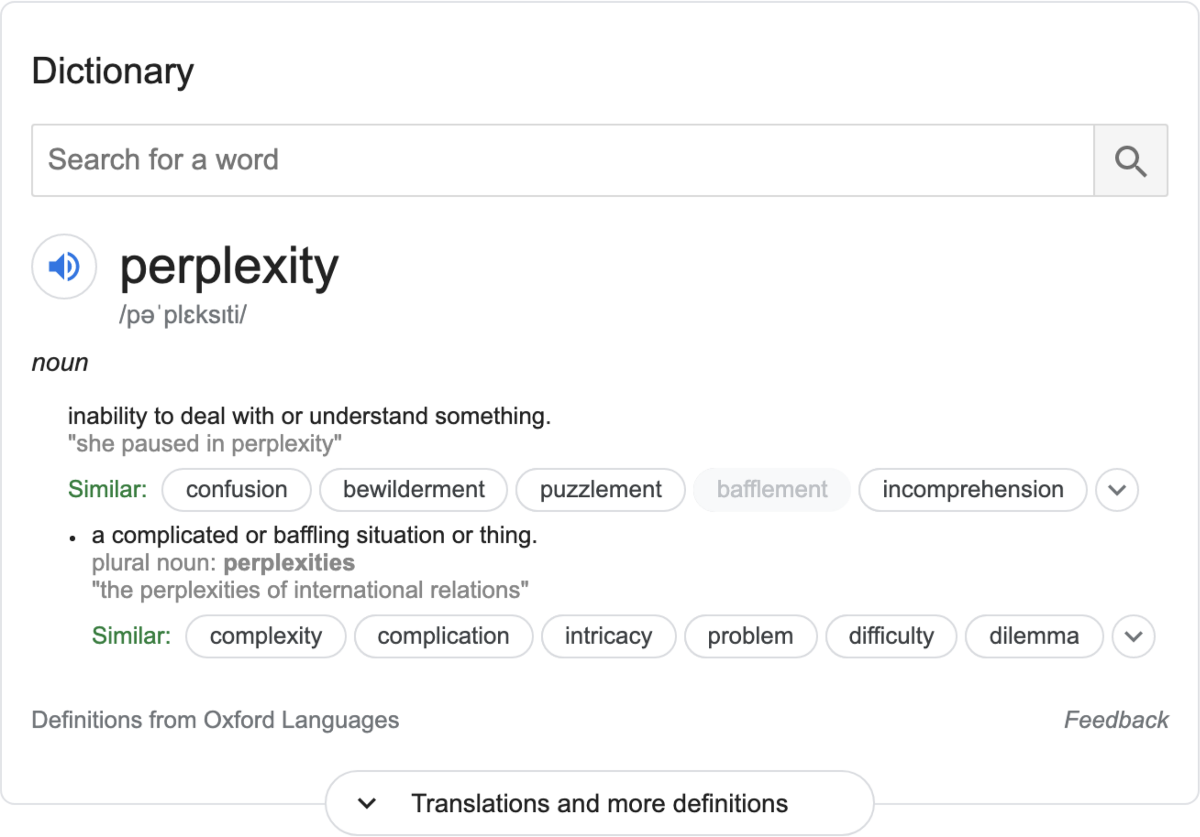
- Inconsistent output
- Crippled by length
- Coreference issues
- Longer strings that are repeated many times in the dataset
- Repeating entities
- MLE!
- We evaluate them with “perplexity”!

unsound, loops

antecedents can go missing

Pulse Check

- What is perplexity ?
- Perplexity is a measure of uncertainty



Open Questions in Long Text Generation

* * Information about what to say next based on probability of observed word sequences (it reads and writes based on that!)

Challenges:

- **common sense reasoning** -> understanding
- **sentence ordering** -> discourse structure
- **relational information** - > entity relations
- **Structure** -> composition, grammar, etc.

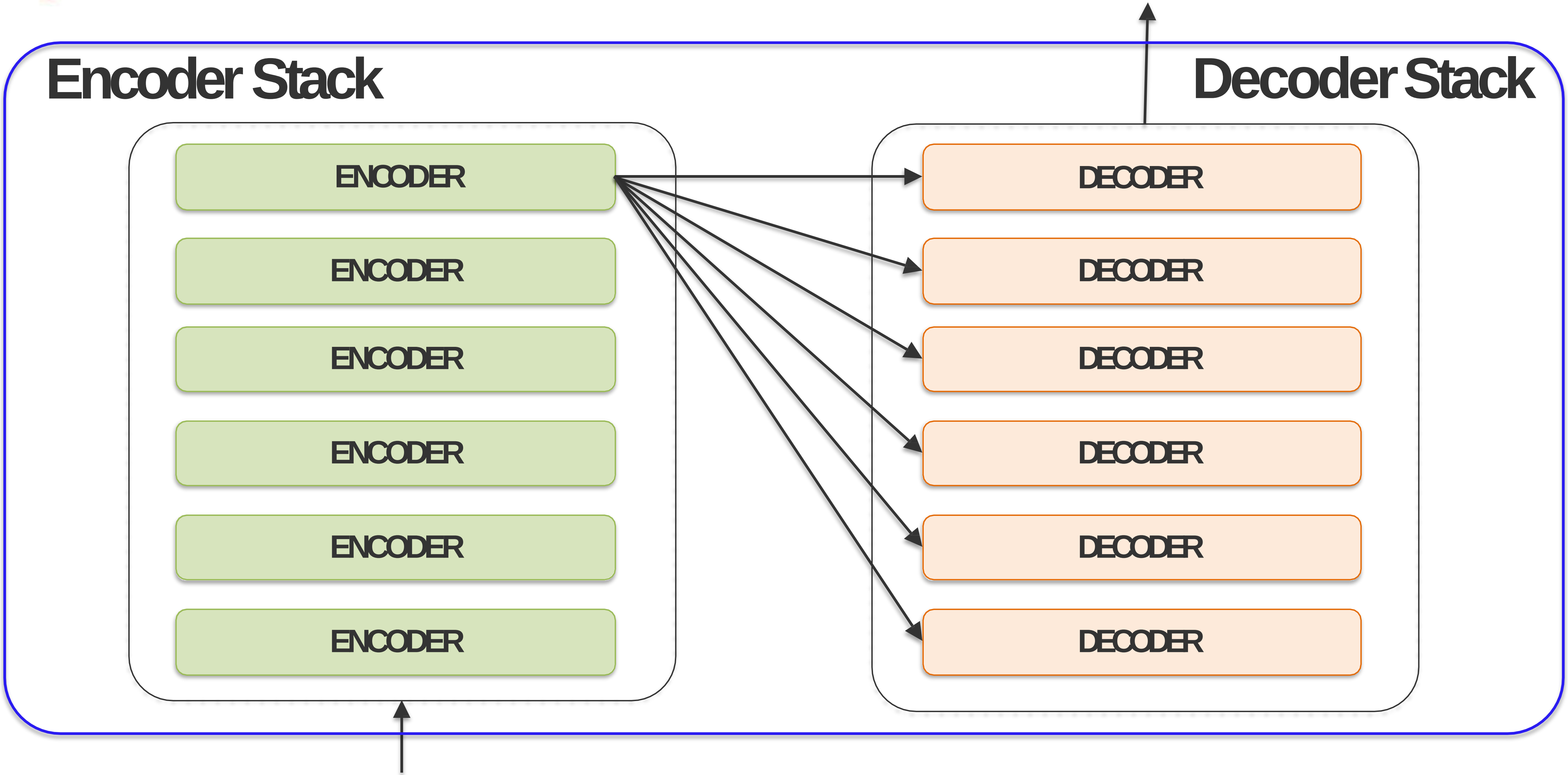
Conditional Generation

- How do we:
  - learn **narrative flow?**
  - **guide** long text generation
  - capture **long range dependencies**?
  - **leverage knowledge** embedded in **pre-trained LMs**?
- Tasks:
  - Summarization
  - Story Generation
  - Knowledge Graph Completion

Background of Transformer Models for Text Input

The Transformer

IRE क्लास Microsoft टीमों का उपयोग करके आयोजित की जाती है



IRE class is conducted using Microsoft Teams

GPT-2

DECODER

. . .

DECODER

DECODER

Open AI

BERT

ENCODER

. . .

ENCODER

ENCODER

Google

TRANSFORMER XL

Recurrent DECODER

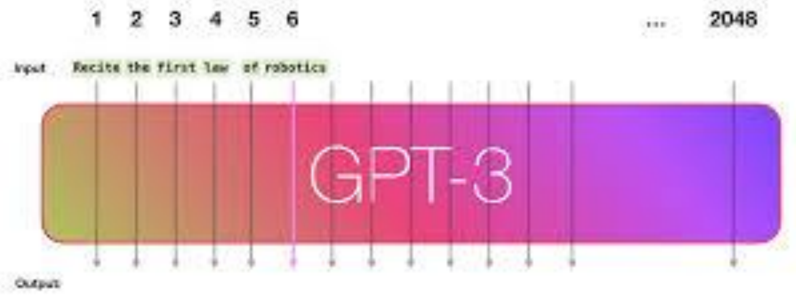
. . .

Recurrent DECODER

Recurrent DECODER

CMU/Google Brain

2020


1 2 3 4 5 6 ... 2048
Input
Recite the first law of robotics
GPT-3
Output

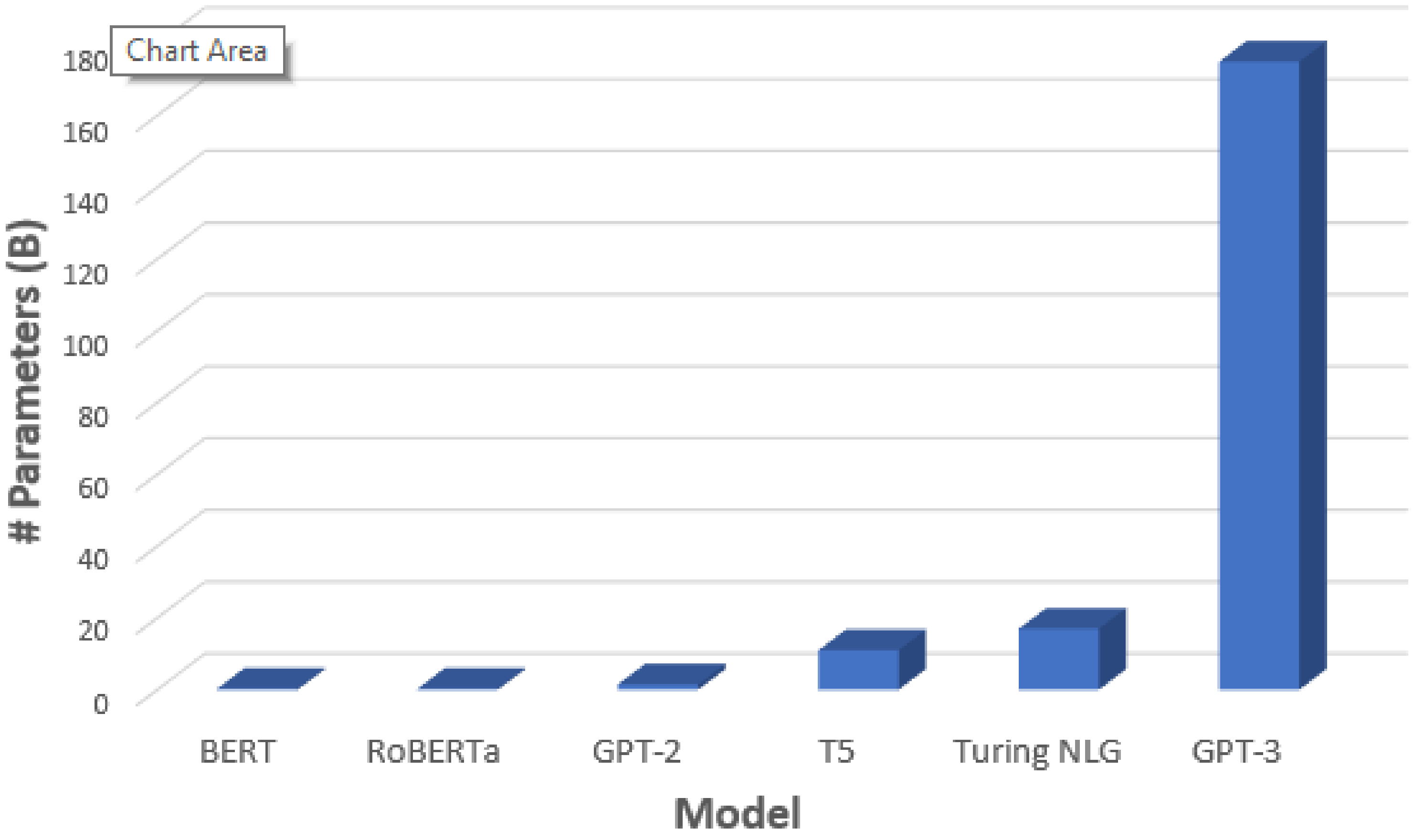
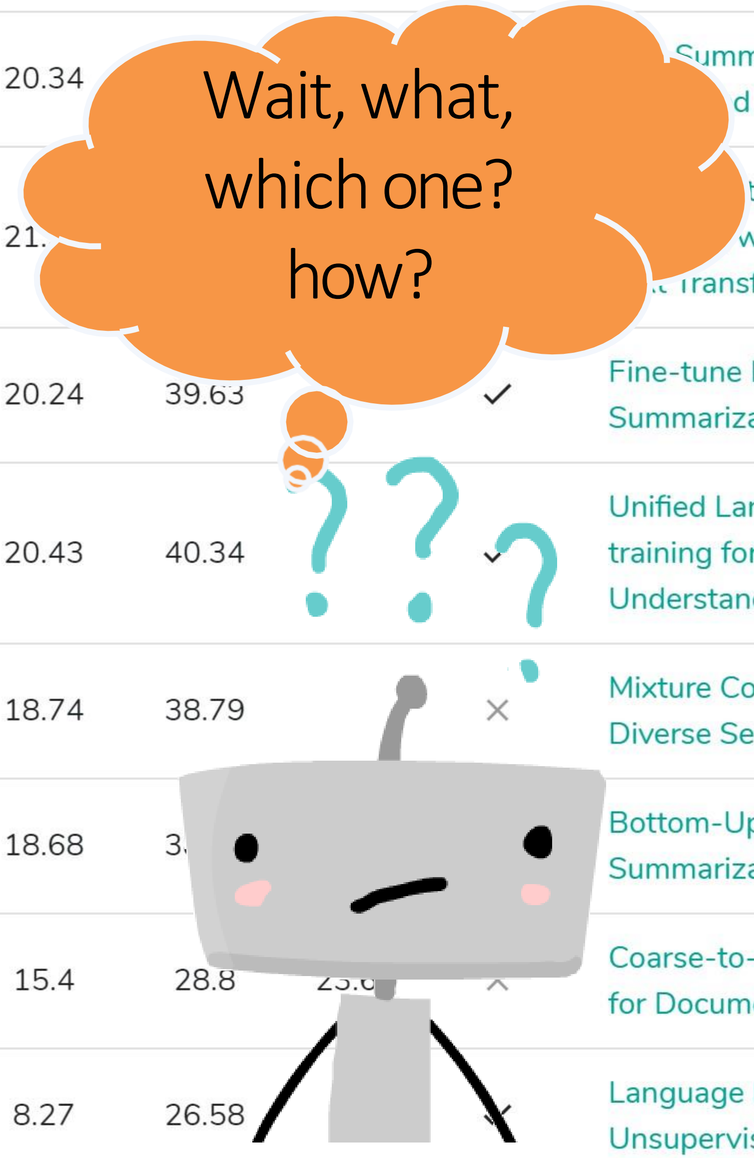


Chart Area
Parameters (B)
0
20
40
60
80
100
120
140
160
180
BERT
RoBERTa
GPT-2
T5
Turing NLG
GPT-3
Model

RANK	METHOD	ROUGE-1	ROUGE-2	ROUGE-L	PPL	EXTRA TRAINING DATA	PAPER TITLE	YEAR	PAPER	CODE
1	BertSumExt	43.85	20.34	[illegible]	[illegible]	[illegible]	[illegible] Summarization with [illegible] d Encoders	2019		
2	T5-11B	43.52	21.[illegible]	[illegible]	[illegible]	[illegible]	[illegible] the Limits of Transfer [illegible] with a Unified Text-to-[illegible] Transformer	2019		
3	BERTSUM+Transformer	43.25	20.24	39.63		✓	Fine-tune BERT for Extractive Summarization	2019		
4	UniLM (Abstractive Summarization)	43.08	20.43	40.34		✓	Unified Language Model Pre-training for Natural Language Understanding and Generation	2019		
5	Selector+Pointer Generator	41.72	18.74	38.79		✗	Mixture Content Selection for Diverse Sequence Generation	2019		
6	Bottom-Up Sum	41.22	18.68	3[illegible]	[illegible]	[illegible]	Bottom-Up Abstractive Summarization	2018		
7	C2F + ALTERNATE	31.1	15.4	28.8	23.[illegible]	✗	Coarse-to-Fine Attention Models for Document Summarization	2017		
8	GPT-2	29.34	8.27	26.58		✓	Language Models are Unsupervised Multitask Learners	2019		



https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail

Generate with *discourse understanding*!

Research Questions:

- ***Length*** *of summaries*
- ***Abstractedness*** of summaries

Corpora for
~~Generate with~~ *discourse understanding*!

Research Questions:

- ***Length*** *of summaries*
- ***Abstractedness*** of summaries

How good are Abstractive Summarization Datasets ? (CNN Example)

Article

CNN - We had no idea how much we would really, really, really, really like tom hanks lip-syncing to a Carly Rae Jepsen song, but we really do. Hanks shows up in the new video for "i really like you," singing Jepsen 's part throughout.

The oscar-winning actor is apparently playing himself, signing autographs for fans, and generally being a very cheery movie star, before he and Jepsen take part in a flash mob. So what exactly is Tom Hanks doing in this video in the first place?

Turns out he is good friends with Scooter Braun, manager for Jepsen, and Justin Bieber, who also appears in the video. He even sang and danced at Braun's wedding.

ABC reported that Hanks suggested himself to play the role, after Jepsen said it would be amusing for a man to lip-sync her song. The result, as you can see, is kind of magical.

Abstract / Headlines / Summary ?

Tom Hanks makes surprise appearance in Carly Rae Jepsen video dancing and lip-syncing .
Hanks is friends with Jepsen's manager, Scooter Braun.
Hanks volunteered to be in the video.

Pulse Check

- What is the CNN/Dailymail Dataset used for ?

Can we evaluate *narrative flow* on existing corpora ?

	Narrative Flow?
XSum	✗
Newsroom	✗
CNN	✗
Daily Mail	✗

Summaries à headlines of the news articles

- don't provide **inductive bias**
- unable to learn or measure **narrative flow**
- tend to be **extractive**

XSum [Narayan et.al, 2018a]
Newsroom [Grusky et.al.m 2019]
CNN [Hermann et.al. 2015]
Daily Mail [Hermann et.al. 2015]

Scientific Dataset (arXiv – CS+BIO)

Introduction

Introduction of task

Discourse relation characterizes the internal structure and logical relation of a coherent text.

Body of the introduction, citations

Automatically identifying these relations not only plans an important role in discourse comprehension and generation, but also obtains wide applications in many other relevant natural language processing tasks, such as text summarization (Yoshida et. al., 2014), conversation (Higashinaka et.al., 2014), question answering (Verberne et al., 2007)…

More detailed description of task

Generally, discourse relations can be divided into two categories: explicit and implicit, which can be illustrated in the following example: The company was disappointed by the ruling …

Abstract

Implicit discourse relation recognition is a crucial component for automatic discourse level analysis and natural language understanding.

In this paper, instead, we explore generative models and propose a variational neural discourse relation recognizer.

Inter sentence narrative flow

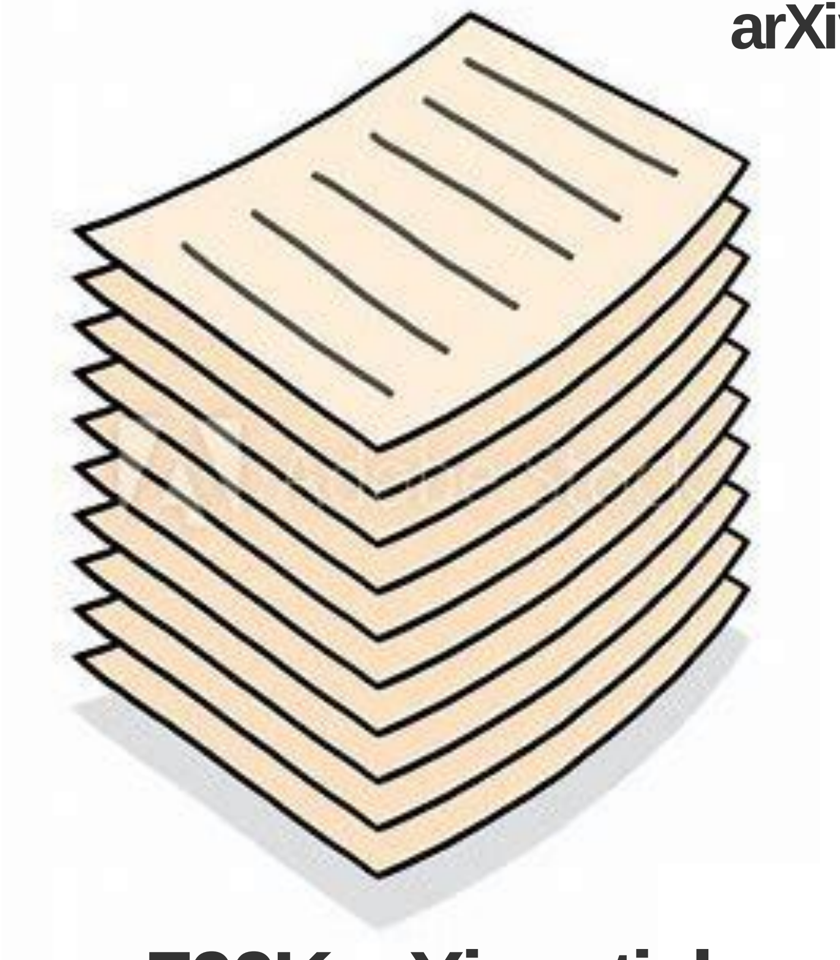
Introductory statements :

- “In this paper”
- “In this work”
- “In the present study”

SAAS : New Abstractive Summarization Dataset

SAAS : Scientific Abstract Summaries
arXiv.org

AAN: ACL Anthology Network

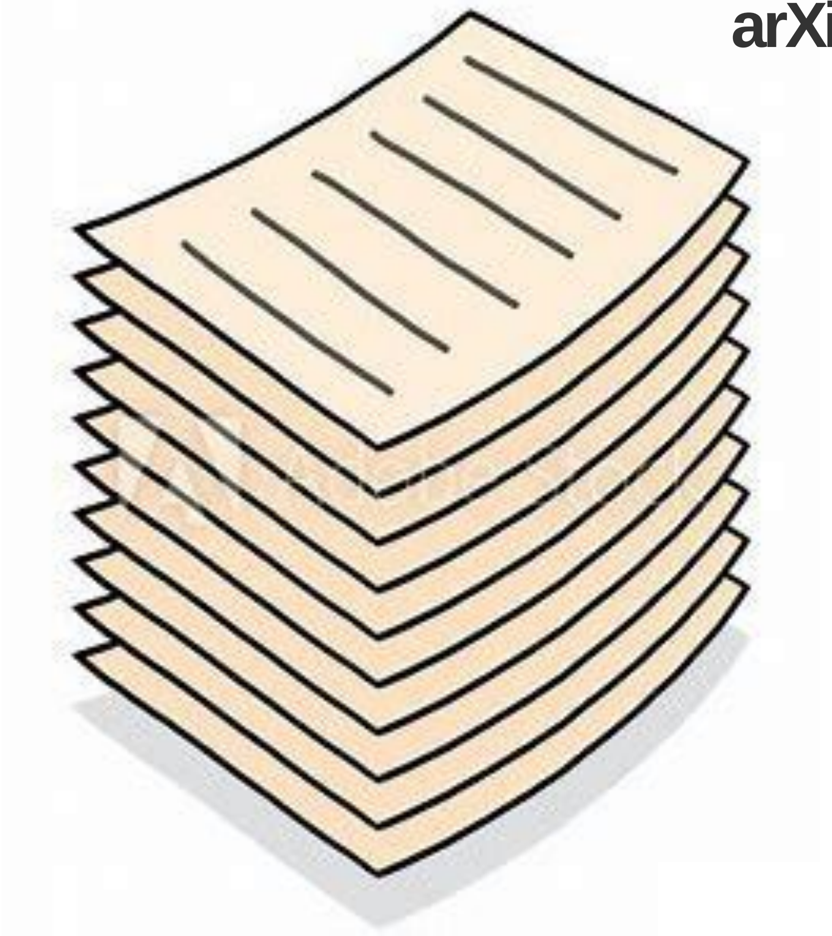


12K NLP articles

700K arXiv articles

SAAS : New Abstractive Summarization Dataset

SAAS : Scientific Abstract Summaries
arXiv.org

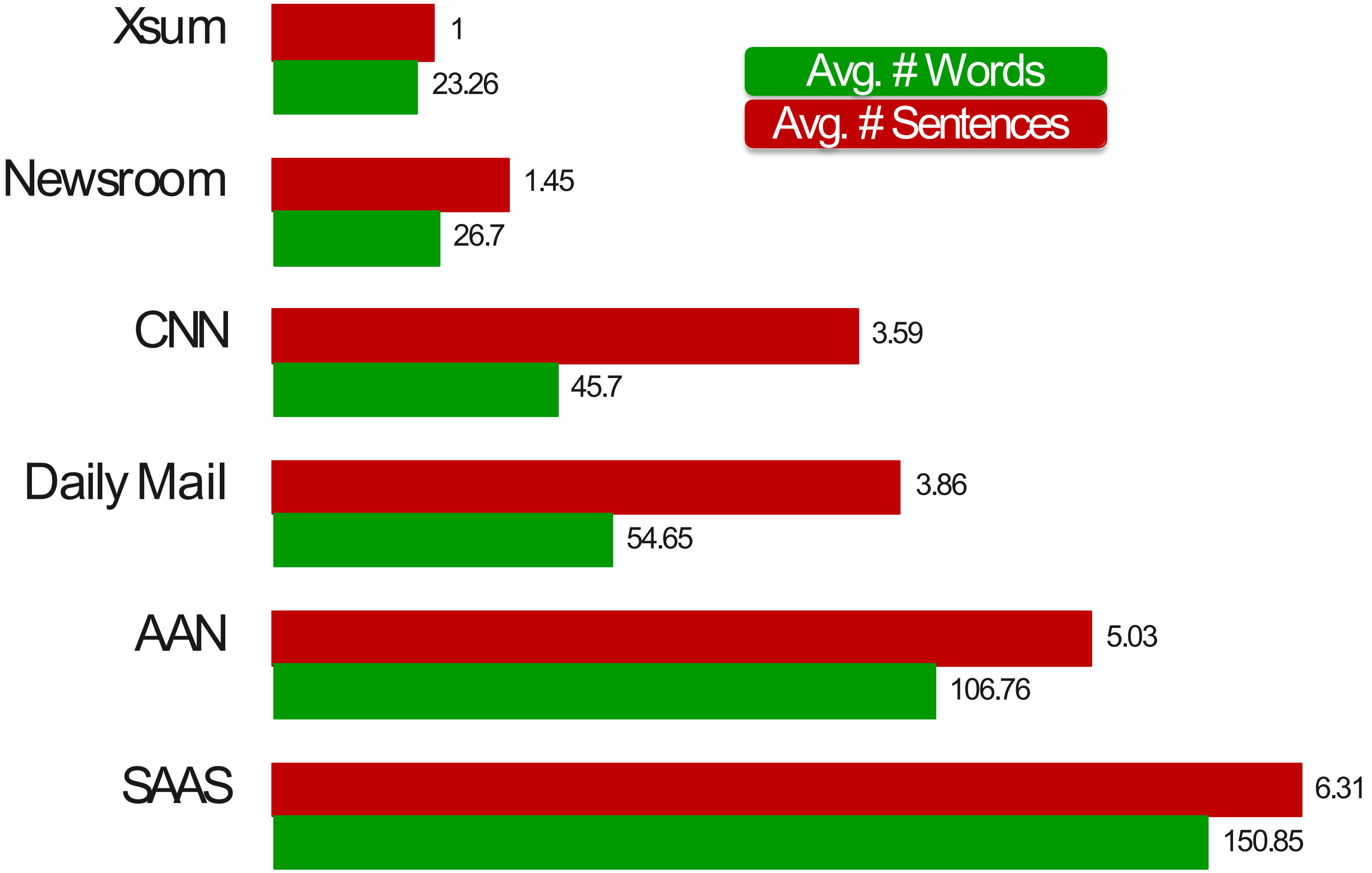


Dataset for three new tasks:

- Abstract - > Title
- Introduction -> Title
- Introduction -> Abstract

700K arXiv articles

How Useful Existing Summarization Corpora



	Narrative Flow?
XSum	✗
Newsroom	✗
CNN	✗
Daily Mail	✗
AAN	✓
SAAS	✓

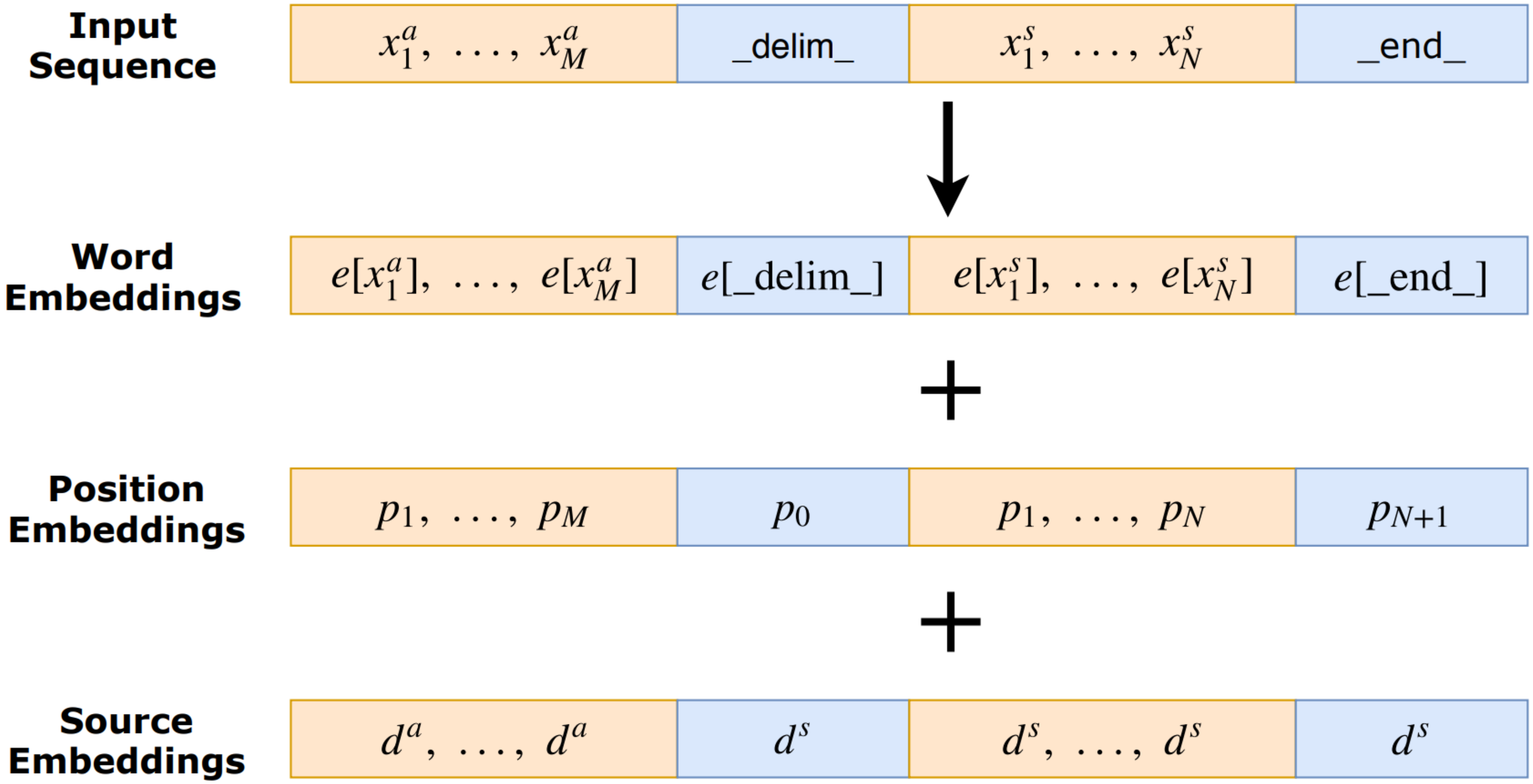
XSum [Narayan et.al, 2018a]
Newsroom [Grusky et.al.m 2019]
CNN [Hermann et.al. 2015]
Daily Mail [Hermann et.al. 2015]

Generate text *discourse understanding*!

Co-opNET: Cooperative Generator Discriminator Networks

Gabriel et al., Cooperative Generator-Discriminator Networks for Abstractive Summarization with Narrative Flow
https://arxiv.org/abs/1907.01272

Co-opNET : Generator Networks



Co-opNET : Generator Networks

Transformer-Decoder

Decoder-Block

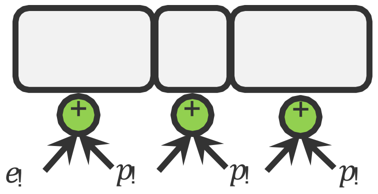
$\times N$

Feed Forward NN

Masked Self-Attention

Gold Abstract

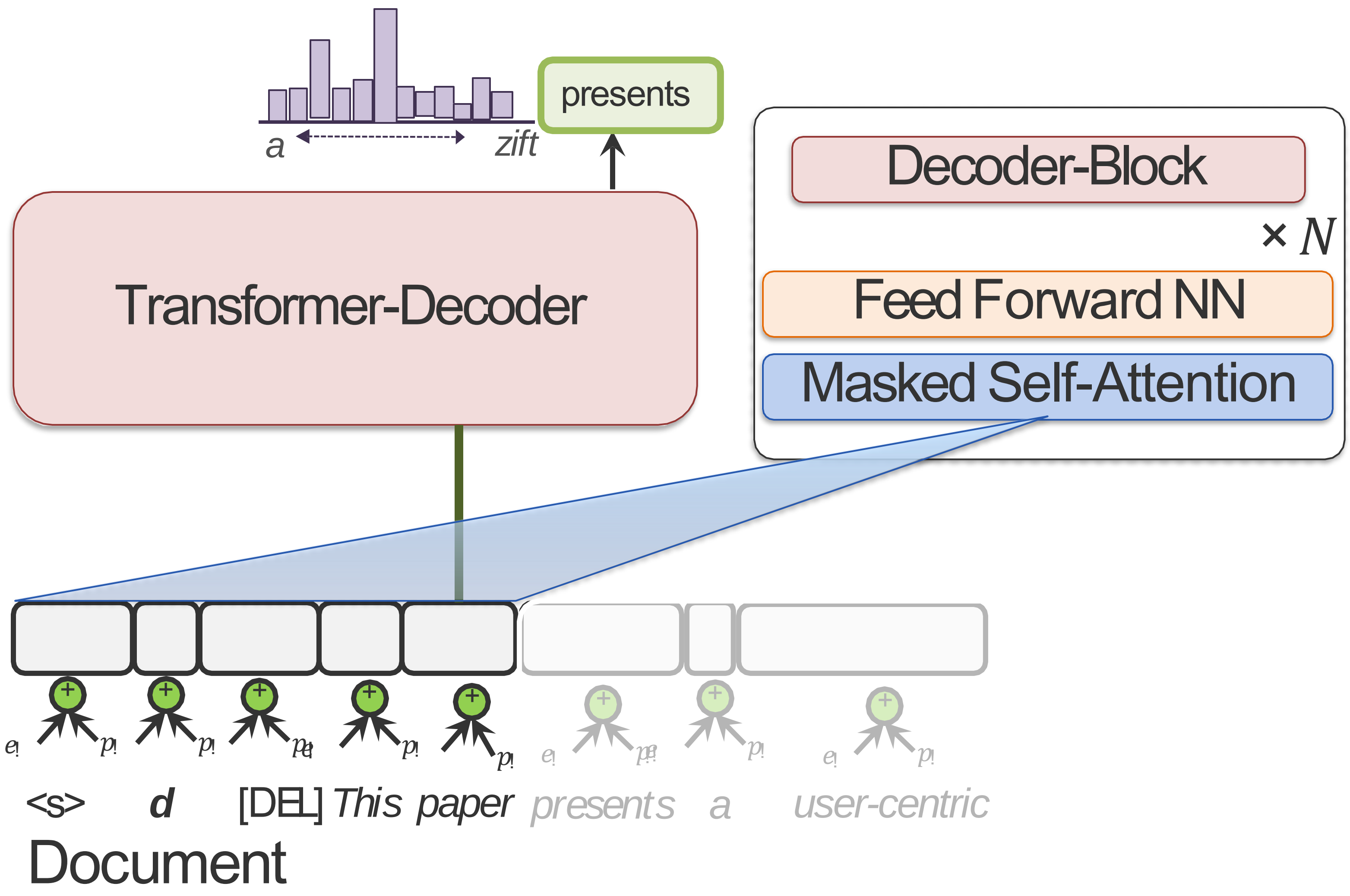
This research is concerned with making recommendations to museum visitors based on their history within the physical environment, and textual information associated with each item in their history. (...)



<s> *d* [DEL]

Document

Co-opNET : Generator Networks



Gold Abstract

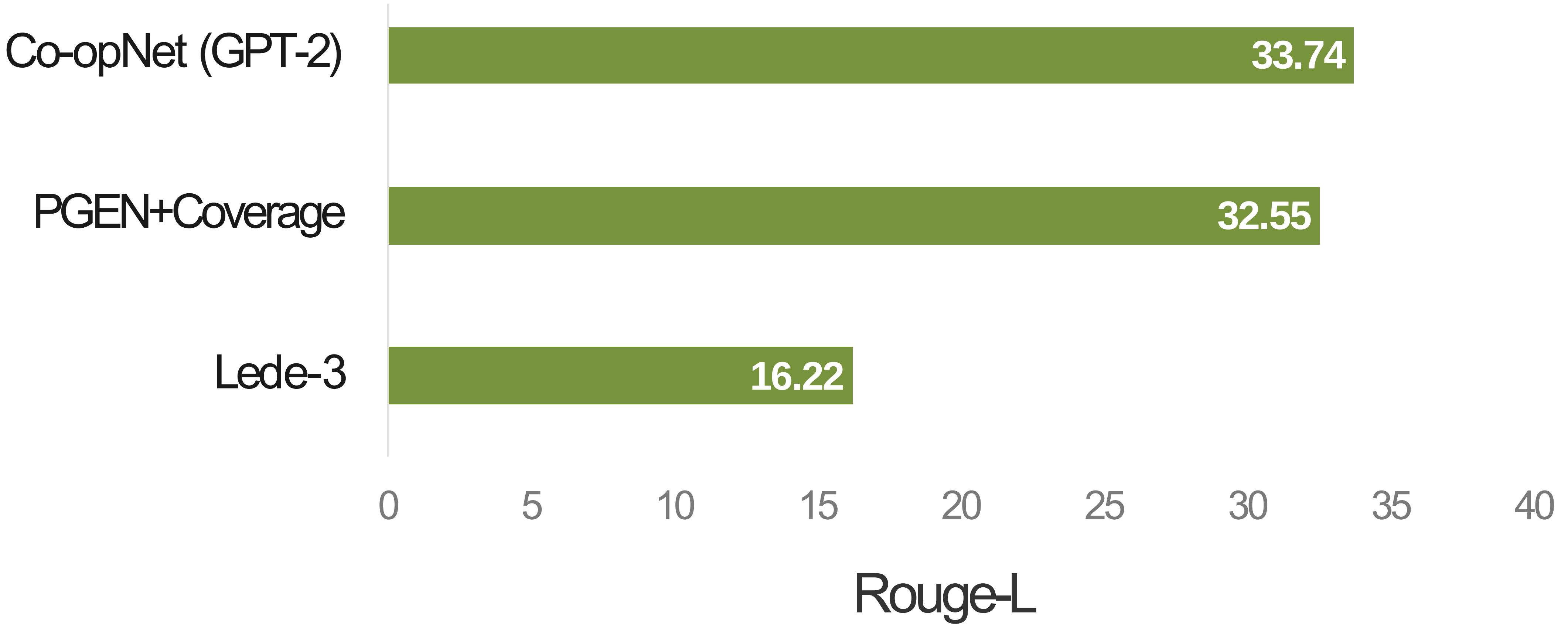
This research is concerned with making recommendations to museum visitors based on their history within the physical environment, and textual information associated with each item in their history. (...)

$$\mathscr{L}_{!"\#} = \sum_{\$\%\&}^{|'|(|*|} \log p(e_{\$} | e_{+}, \ldots e_{\$,\&})$$

Prediction

This paper presents a user-centric perspective on the property of location, focusing on some relevant factors in deciding which exhibit a user intends to visit. (...)

Automatic Metric Evaluations on **SAAS** Dataset



Can '**Generator Only**' Model Improve Coherence ?

Gold Abstract

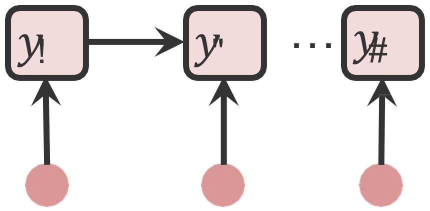
This research is concerned with making recommendations to museum visitors based on their history within the physical environment, and textual information associated with each item in their history. (...) This study compares and analyses different methods of path prediction including an adapted naive Bayes method, document similarity, visitor feedback and measures of lexical similarity.

Co-opNET (Generator Only) Generated Abstract

This paper proposes a novel approach to measuring the success of machine learning methods in a user's selection of a particular exhibit to be produced. An unsupervised framework is used to jointly compute the likelihood of the value of the best exhibit to be produced. (...) *The experiments show* that models produced by supervised methods improve user performance in selecting exhibits over unsupervised methods.

Green: Transitional Phrases

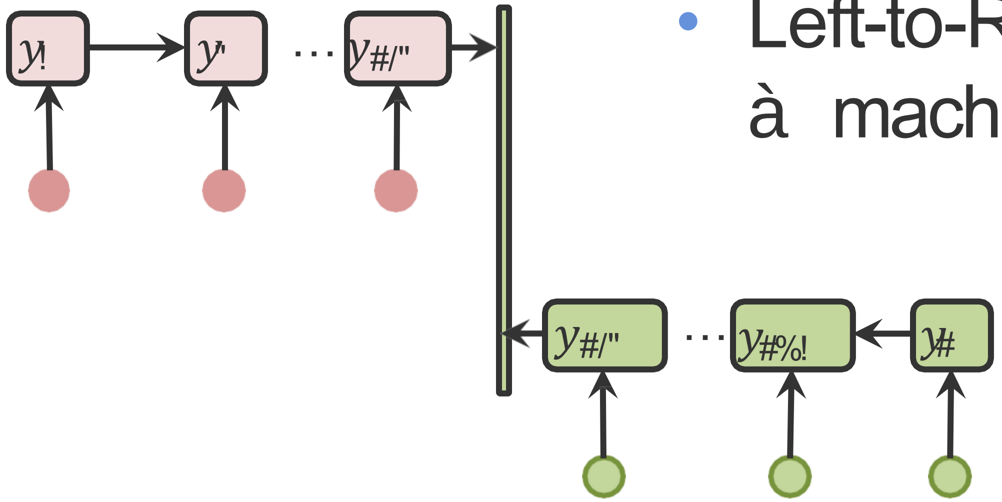
Autoregression issue for Summarization Flow



Left-to-Right decoding

- Cannot achieve narrative flow across multiple sentences
- No way of telling if the sequence of sentences follow a certain discourse structure or narrative flow!

Left-to-Right decoding



Right-to-Left decoding

- Left-to-Right and Right-To-Left decoding [Zhou et.al., 2019]
  à machine translation where alignment matters!

Co-opNET: Cooperative Generator Discriminator Networks

Introduction

Discourse relation characterizes the internal structure and logical relation of a coherent text.

Automatically identifying these relations not only plans an important role in discourse comprehension and generation, but also obtains wide applications in many other relevant natural language processing tasks, such as text summarization (Yoshida et. al., 2014), conversation (Higashinaka et.al., 2014), question answering (Verberne et al., 2007)…

Generally, discourse relations can be divided into two categories: explicit and implicit, which can be illustrated in the following example: The company was disappointed by the ruling …

Abstract

Implicit discourse relation recognition is a crucial component for automatic discourse level analysis and natural language understanding.

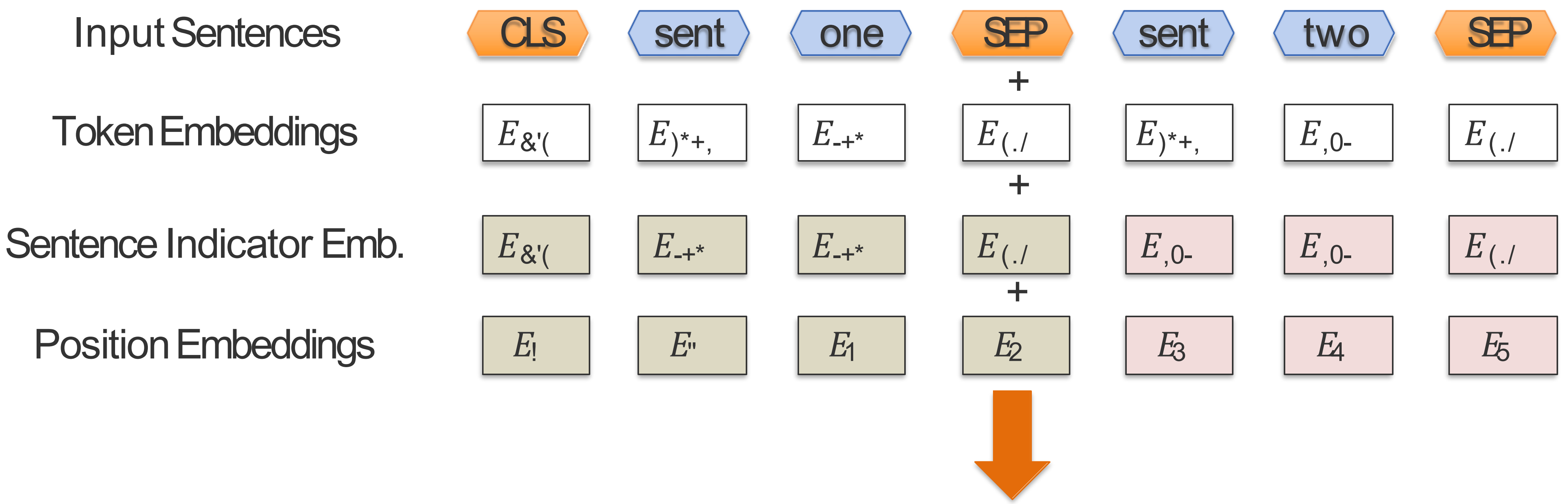
In this paper, instead, we explore generative models and propose a variational neural discourse relation recognizer.

Transformer Generator

BERT Discriminator

Scoring function for likelihood of adjacency

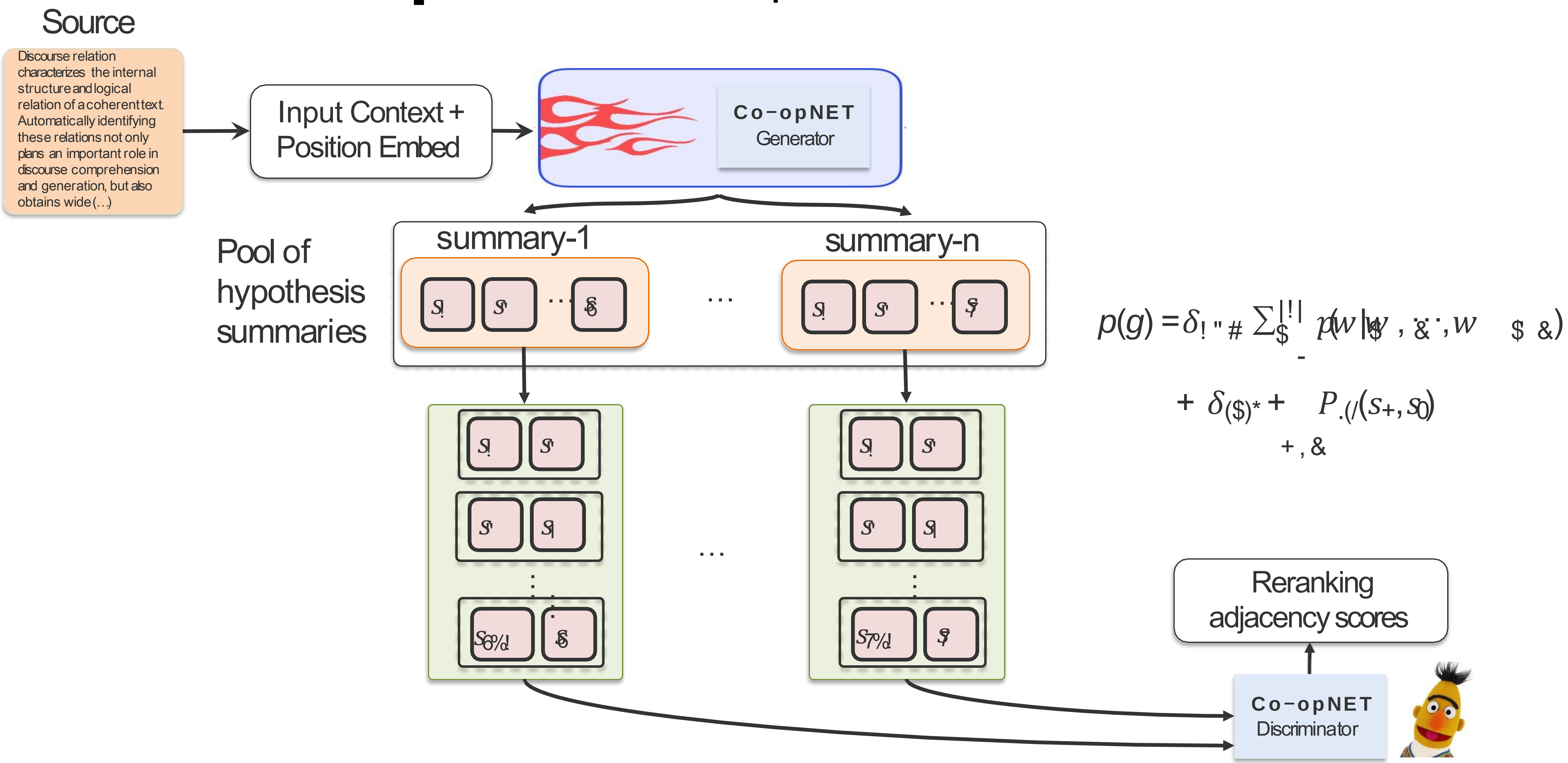
Co-opNET : Discriminator Networks



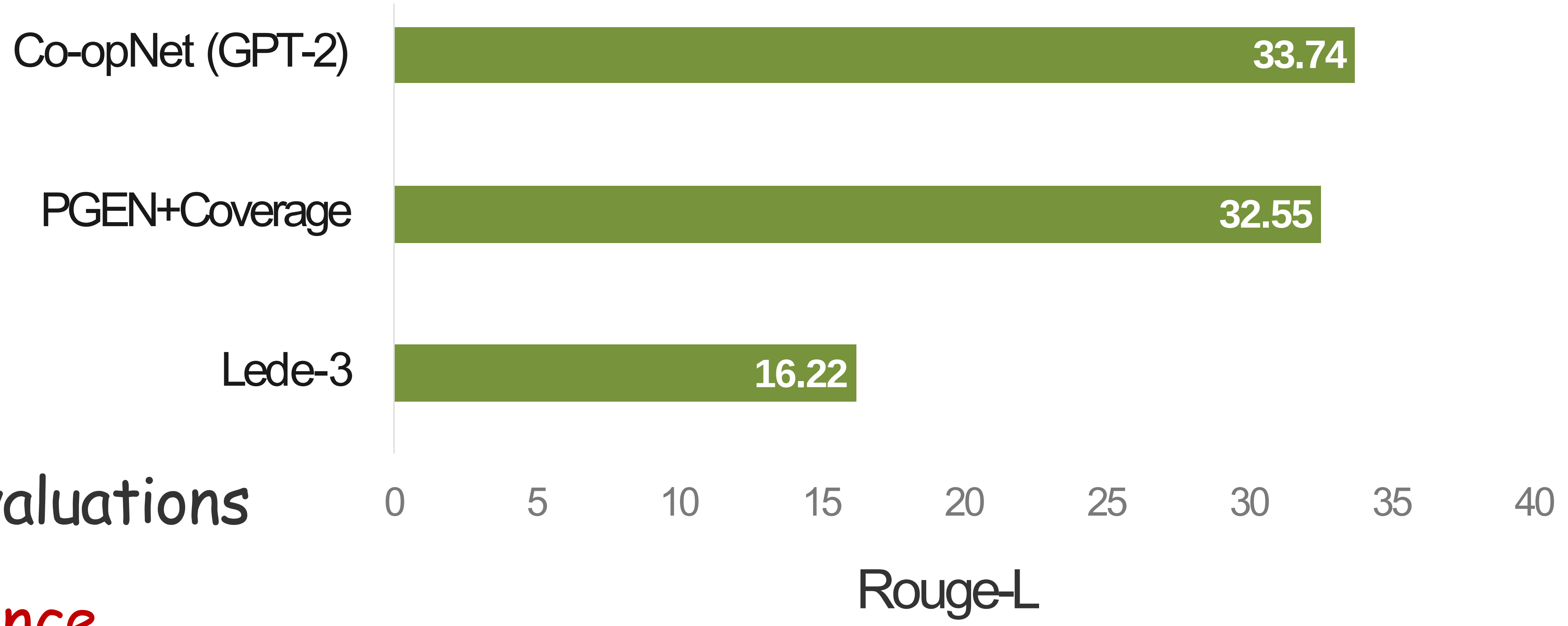
Adjacency Classifier (Probability of adjacency between 2 sentences)

Adjacency Learning. (minimize the likelihood of predicting whether 2 sentences are adjacent or not)

Co-opNET: Cooperative Generation



Automatic Metric Evaluations on **SAAS** Dataset



Human Evaluations

- Flow
- Relevance

Can '*Generator Only*' Model Improve Coherence ?

Gold Abstract

This research is concerned with making recommendations to museum visitors based on their history within the physical environment, and textual information associated with each item in their history. (...) This study compares and analyses different methods of path prediction including an adapted naive Bayes method, document similarity, visitor feedback and measures of lexical similarity.

Co-opNET (Generator +Discriminator) Generated Abstract

This paper presents a user-centric perspective on the property of location, focusing on some relevant factors in deciding which exhibit a user intends to visit. We exploit variation and infrequency in data from the (...) *We make three contributions: (1)* Our experimental system provides empirical evidence for the effectiveness of supervised learning techniques in predicting (…); *(2)* Our structure based method allows unsupervised learning to be applied to multiple sets of related information. *(3)* Our experimental system uses unsupervised model adaptation in a supervised setting.

Question set B

1. What is coherence in text (1/2 lines) ?
2. Co-opNET has a discriminator and a generator (True or False)
3. Co-opNET can be used for summarize hindi text(True or False)

Conditional Generation

- How do we:
  - learn **narrative flow?**
  - **guide** long text generation
  - capture **long range dependencies**?
  - **leverage knowledge** embedded in **pre-trained LMs**?
- Tasks:
  - Summarization
  - Story Generation
  - Knowledge Graph Completion

PLOTMachines: Generate Stories from Outlines

Reference: Raskin et al., **PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking**
https://arxiv.org/abs/2004.14967

How do human’s write a story ?

big bird’s birthday celebration

cookie monster eats

roller skating ring

big birthday cake

PLOTMachines: Outlines for Better Story Generation

Story Outline

- *big bird’s birthday celebration*
- *cookie monster eats*
- *roller skating ring*
- *big birthday cake*

It **is Big Bird's birthday**, and he goes to the **roller skating rink** with his friends . Back at Sesame Street, Maria and Susan take out the **big birthday cake** and leave it on a table . **Cookie Monster** sees the **cake**, but instead of eating it and spoiling the party, he eats a chair and other things all over Sesame Street .

Big Bird and the other skaters return to Sesame Street and are shocked at what **Cookie Monster ate**, though the cake is safe. Gina and Count Von Count presents **the cake** to Big Bird . It has 548 candles even though **Big Bird** is 6 years old . At the end, when Gina announces the sponsors, **Cookie Monster eats** them along with his **cake** .

Story-Outline Dataset

Wikipedia Plots Article [Riedl, 2017]

A criminologist narrates the tale of the newly engaged couple, Brad Majors and Janet Weiss, who find themselves lost and with a flat tire on a cold and rainy late November evening, somewhere near Denton in 1974 . Seeking a telephone, the couple walk to a nearby castle where they discover a group of strange and outlandish people who are holding an Annual Transylvanian Convention . They are soon swept into the world of dr Frank-N-Furter, a self-proclaimed ”sweet transvestite from Transsexual, Transylvania” . The ensemble of convention attendees also includes servants Riff Raff, his sister Magenta, and a groupie named Columbia . In his lab, Frank claims to have discovered the ”secret to life itself” . His creation, Rocky, is brought to life . The ensuing celebration is soon interrupted by Eddie (an ex-delivery boy, both Frank and Columbia's ex-lover, **(…)**

Keypoints using RAKE

- the rocky horror picture show
- convention attendees also includes servants riff raff
- annual transylvanian convention
- old high school science teacher
- frank justifies killing eddie
- enraged rocky gathers frank
- rainy late november evening
- dr scott investigates ufos
- jealous frank kills eddie
- live cabaret floor show

Challenges in Outline Guided Story Generation

- Challenge #1:
  - **outline** only provides **rough elements** of the plot
- Challenge #2:
  - appropriate **beginning**, **setting** and **conclusion** required
- Challenge #3:
  - stories should include **all the key points** in a natural way
  - keep track of what has been written so far

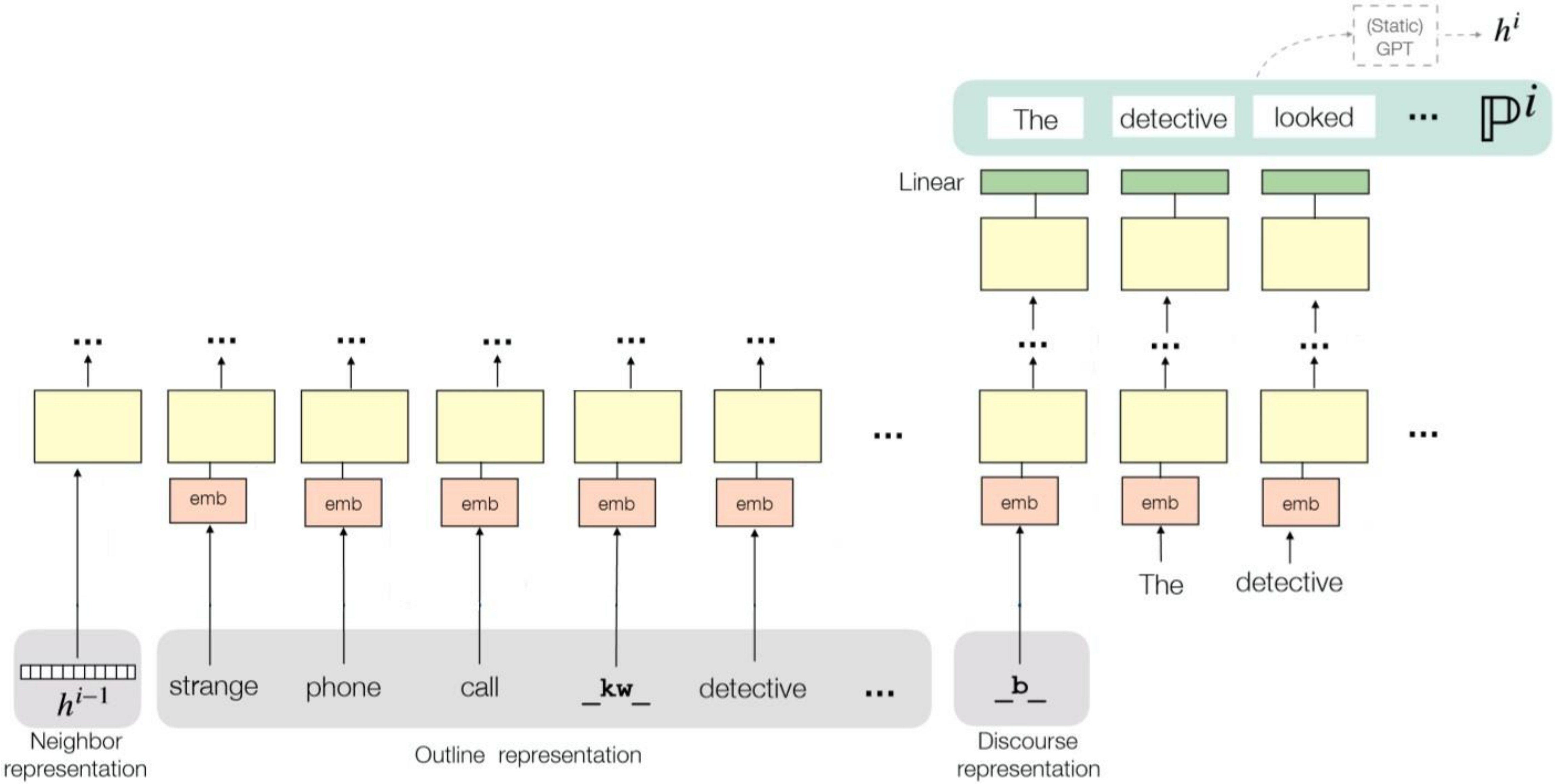
Generate Documents given an Outline

Outline / Keypoints

- the detective
- strange phone call
- detective Leland and another detective Dave
- New York police detective
- Powerful interests in the city
- Leland holds things together
- The incorruptible detective presses
- relationship between man's suicide and murder

Start key one <kw> key two <kw> … endkey

PM: Generate Documents given an Outline



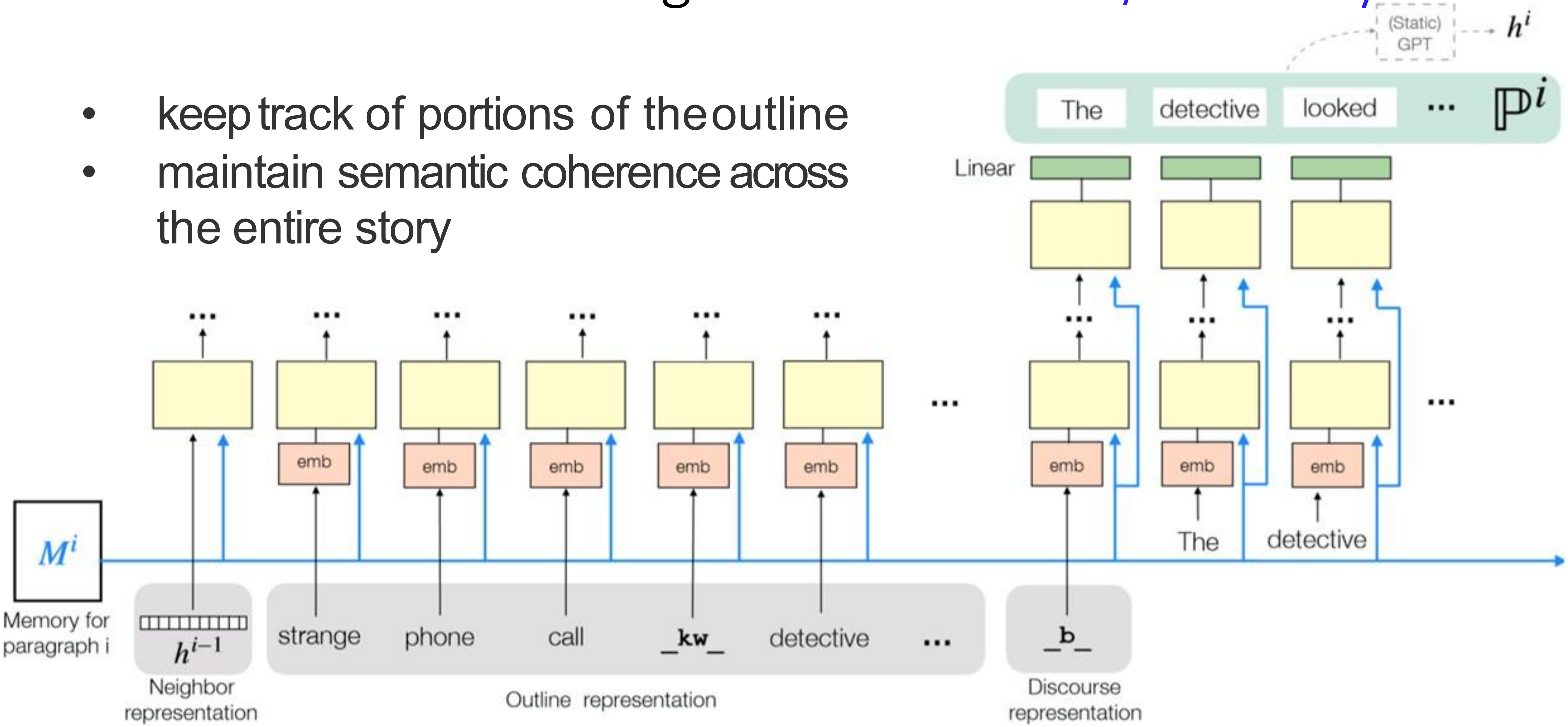
Pulse Check

True / False

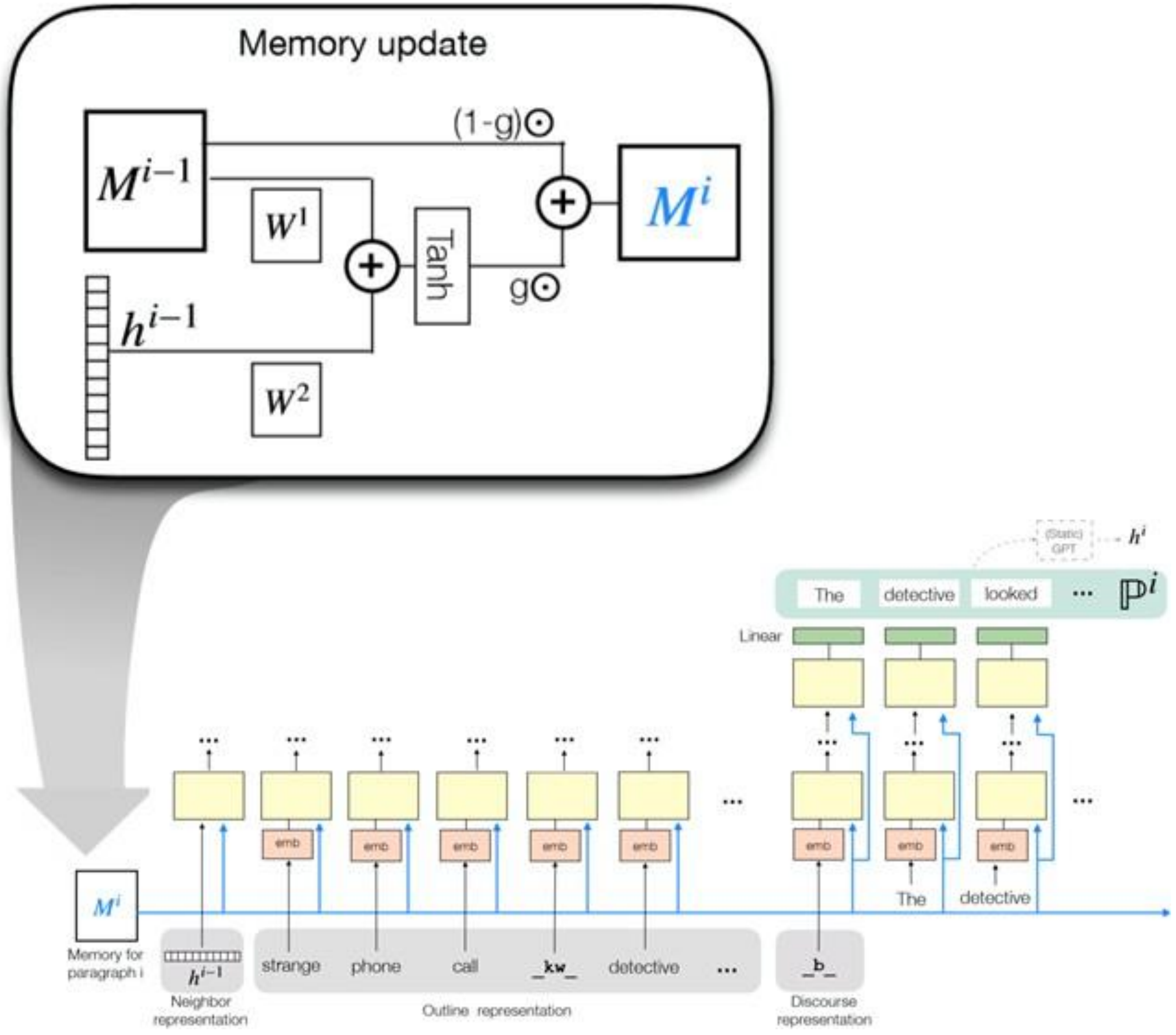
- The model in the previous slide has a memory unit
- The model is based on a seq-seq architecture

PM: Generate Documents given an Outline w/ Memory

- keep track of portions of the outline
- maintain semantic coherence across the entire story

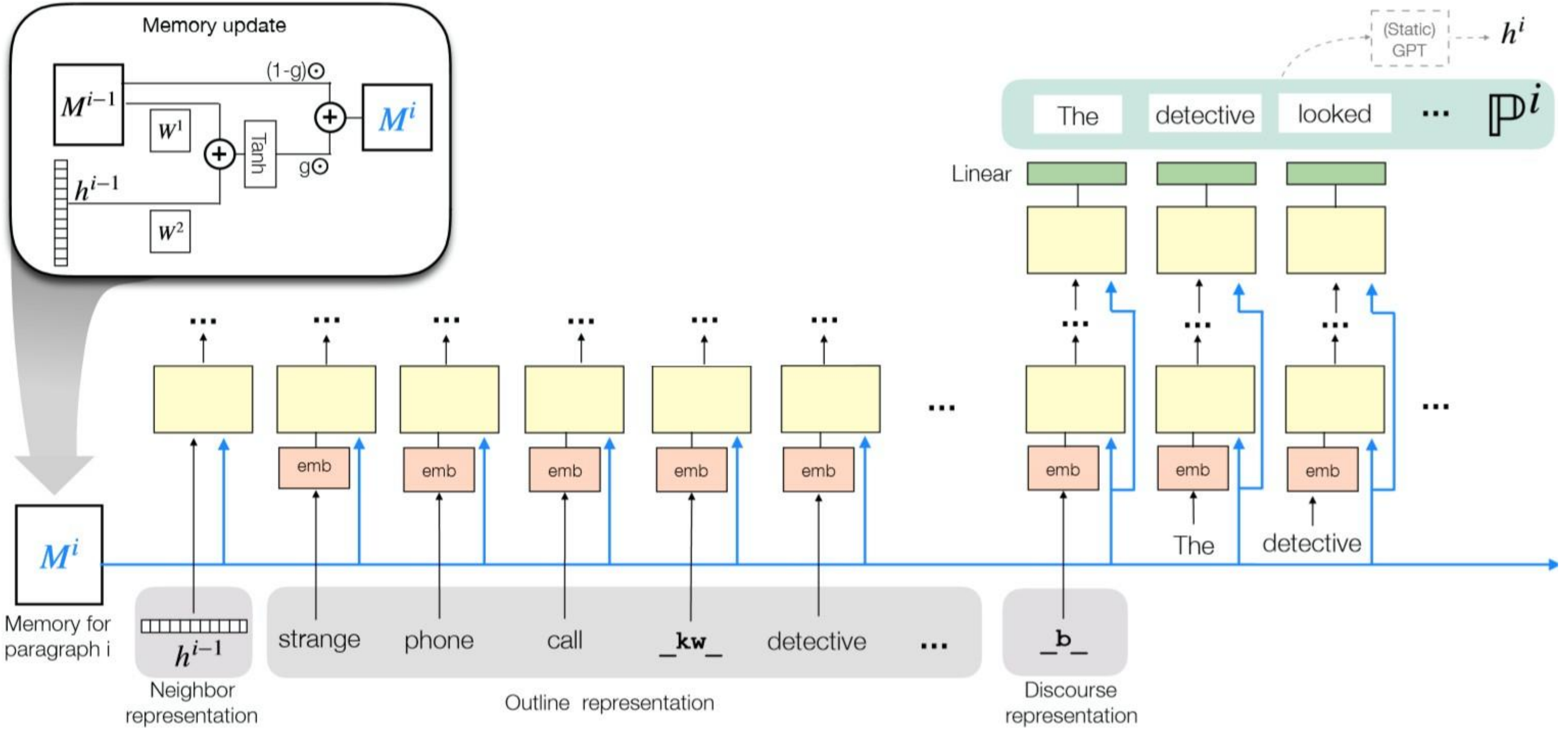


PM: Gated Memory Update Module

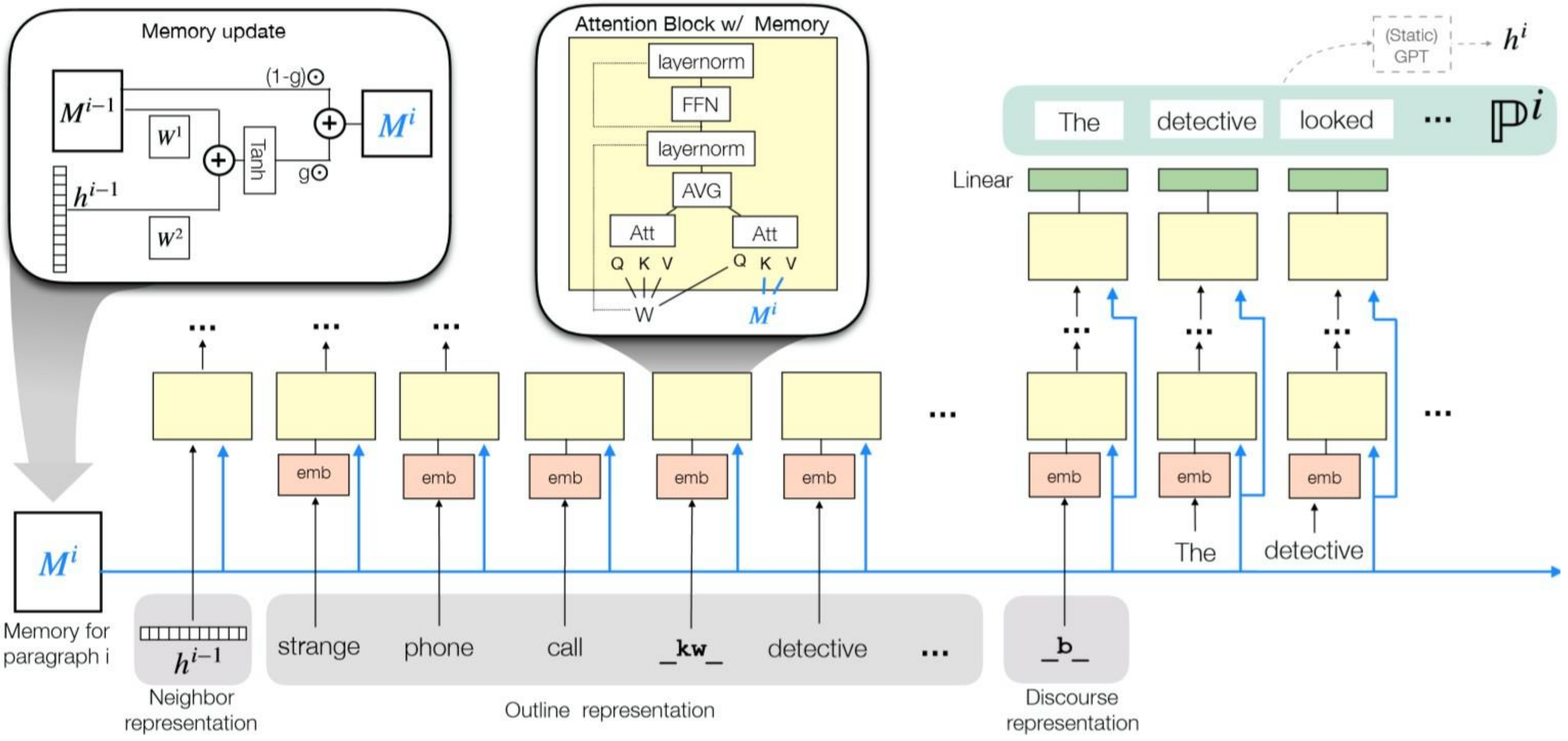


$$M = \begin{bmatrix} K \in \mathscr{R}^{!\times\#} \\ D \in \mathscr{R}^{!\times\#} \end{bmatrix}$$

PM: Generate Documents given an Outline w/ Memory



PM: Generate Documents given an Outline w/ Memory



PLOTMachines : Model Variations

PLOTMachines Full Memory

$$M = \begin{bmatrix} K \in \mathscr{R}^{!\times\#} \\ D \in \mathscr{R}^{!\times\#} \end{bmatrix}$$

PLOTMachines Single Memory

$$M = \_\,[D \in \mathscr{R}^{\$\times\#}]$$

PLOTMachines No Memory

$$M = \_\,[1]$$

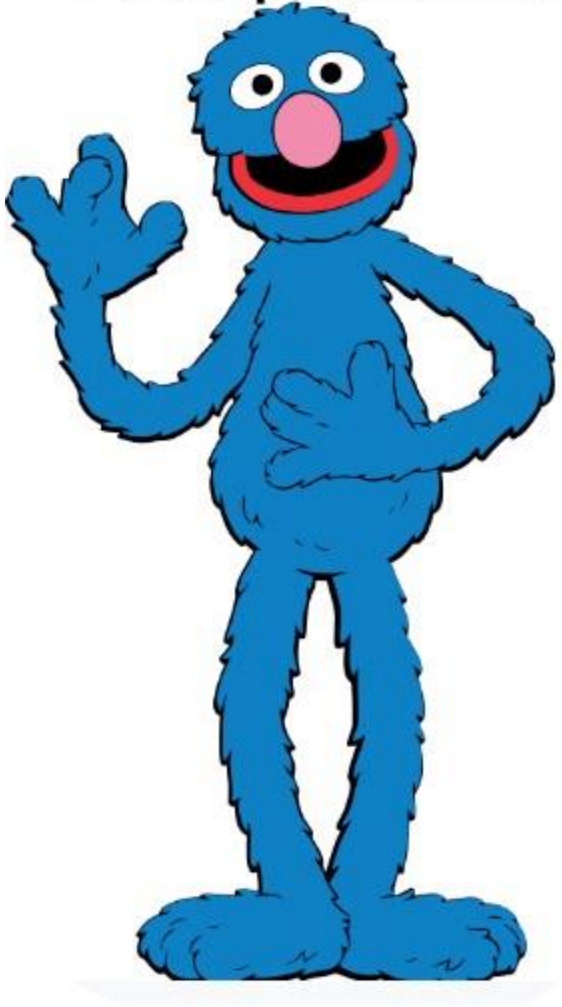
PLOTMachines:

outline conditioned + memory augmented writer

How well does it perform?

Controllable Generation w/ Transformers Baselines

Grover-Large,
345M parameters



salesforce

CTRL-Large

1.6B parameters

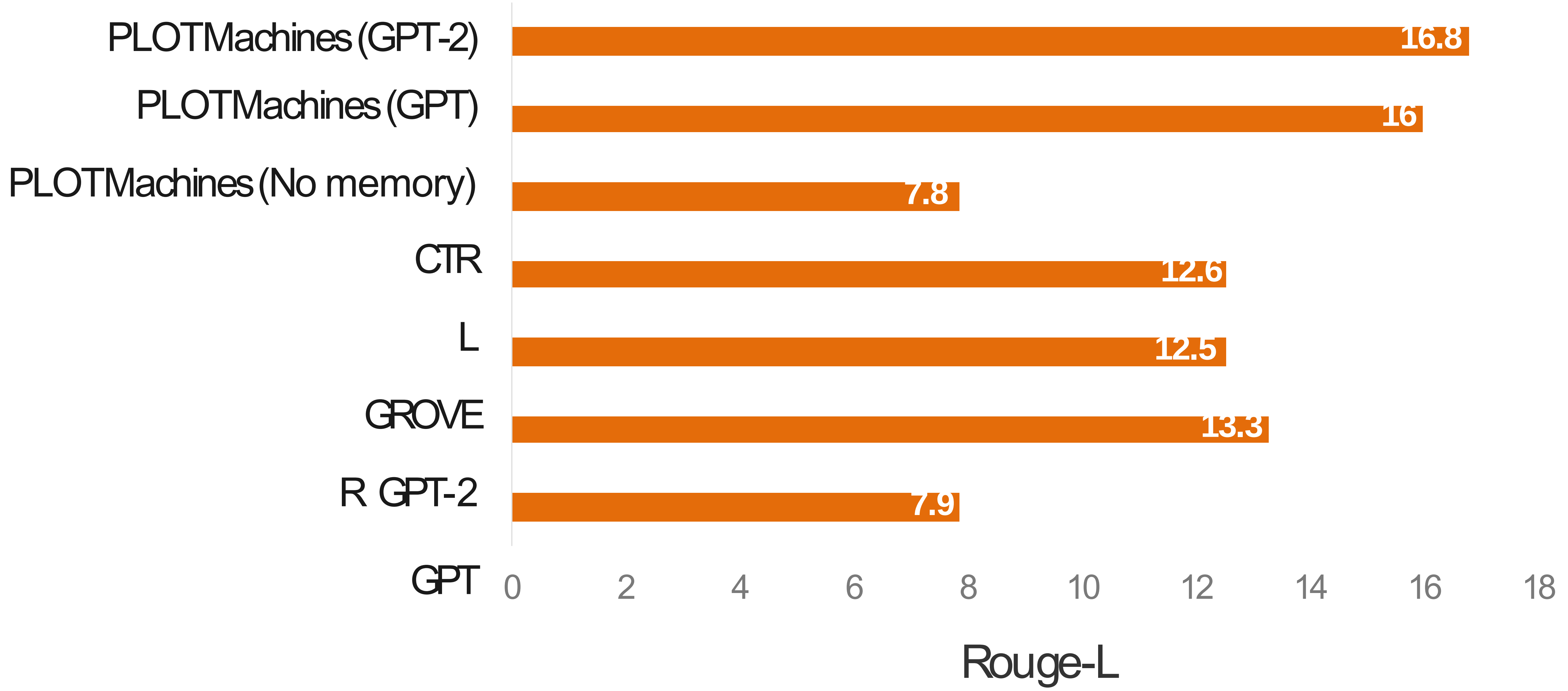
Pulse check

- What is CTRL ?

https://blog.einstein.ai/introducing-a-conditional-transformer-language-model-for-controllable-generation/
Keskar et al., **CTRL: A Conditional Transformer Language Model for Controllable Generation**

- What is PLOTMachine

Automatic Metric Evaluations on WikiPlots Dataset



Human Evaluations: Paragraph base

- **Outline Usage**: (SBS) one is better at utilizing the keywords
- **Narrative Flow**:
  - (SBS)which paragraph contains a single point line?
  - (Single) how smooth is the transition to this paragraph from the previous paragraph?
  - (Single) how repetitive is the information in this paragraph of the information from the previous paragraph?

Human Evaluations: Overall Story

- Which do you think is better at utilizing the keywords ?
- Which do you think is more **repetitive** ?
- Which do you think has better **transitions** ?
- Which do you think is better at following single story line ?
- Which do you think has a **better introduction**?
- Which do you think has a better conclusion ?
- Which do you think has a better order of events ?

Conditional Generation

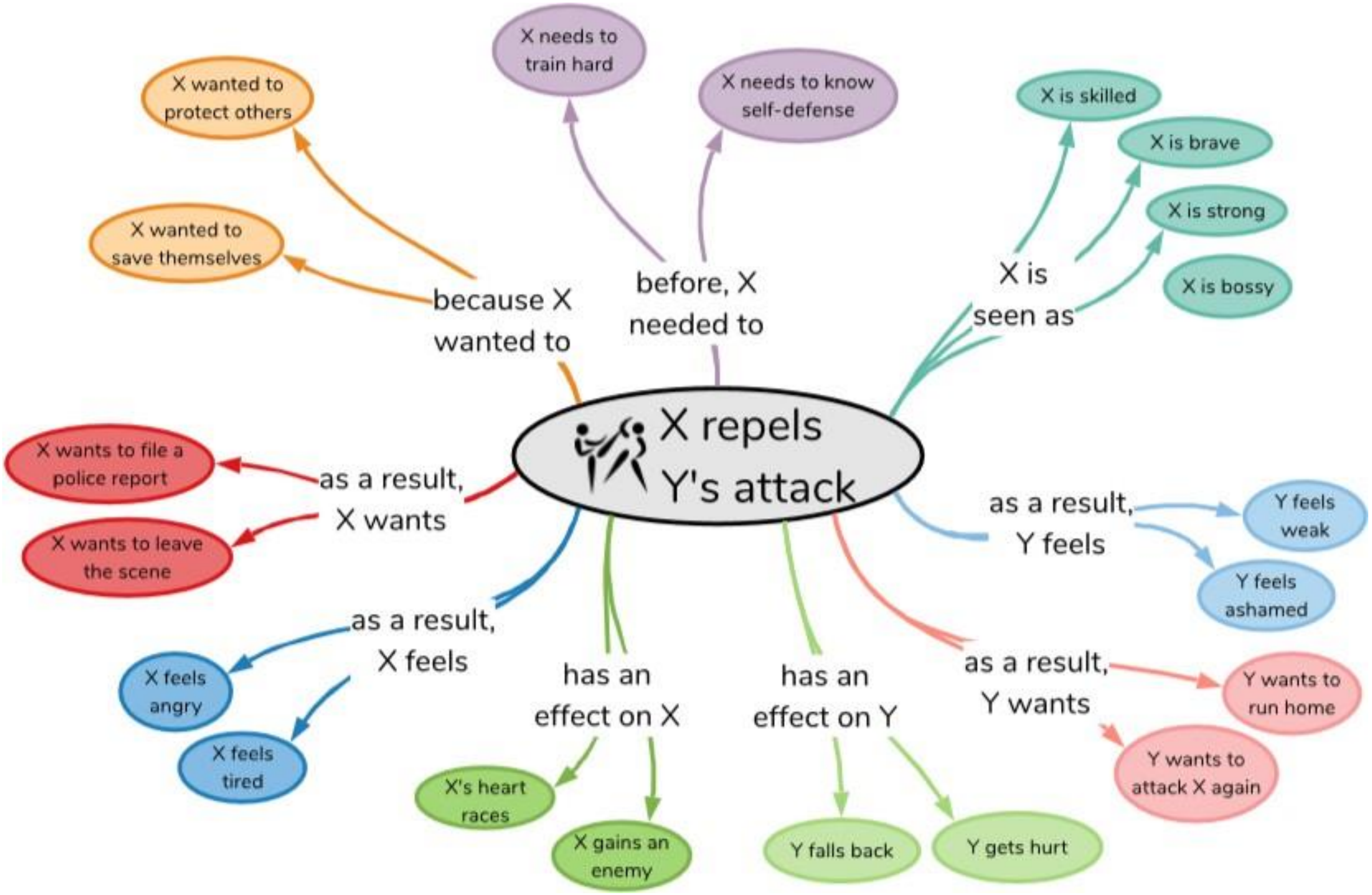
- How do we:
  - learn **narrative flow?**
  - **guide** long text generation
  - capture **long range dependencies**?
  - **leverage knowledge** embedded in **pre-trained LMs**?
- Tasks:
  - Summarization
  - Story Generation
  - Knowledge Graph Completion

COMET: Commonsense Transformers for Automatic Knowledge Graph Construction

Bosselut et al., COMET: Commonsense Transformers for Automatic Knowledge Graph Construction

Generate with common sense knowledge!

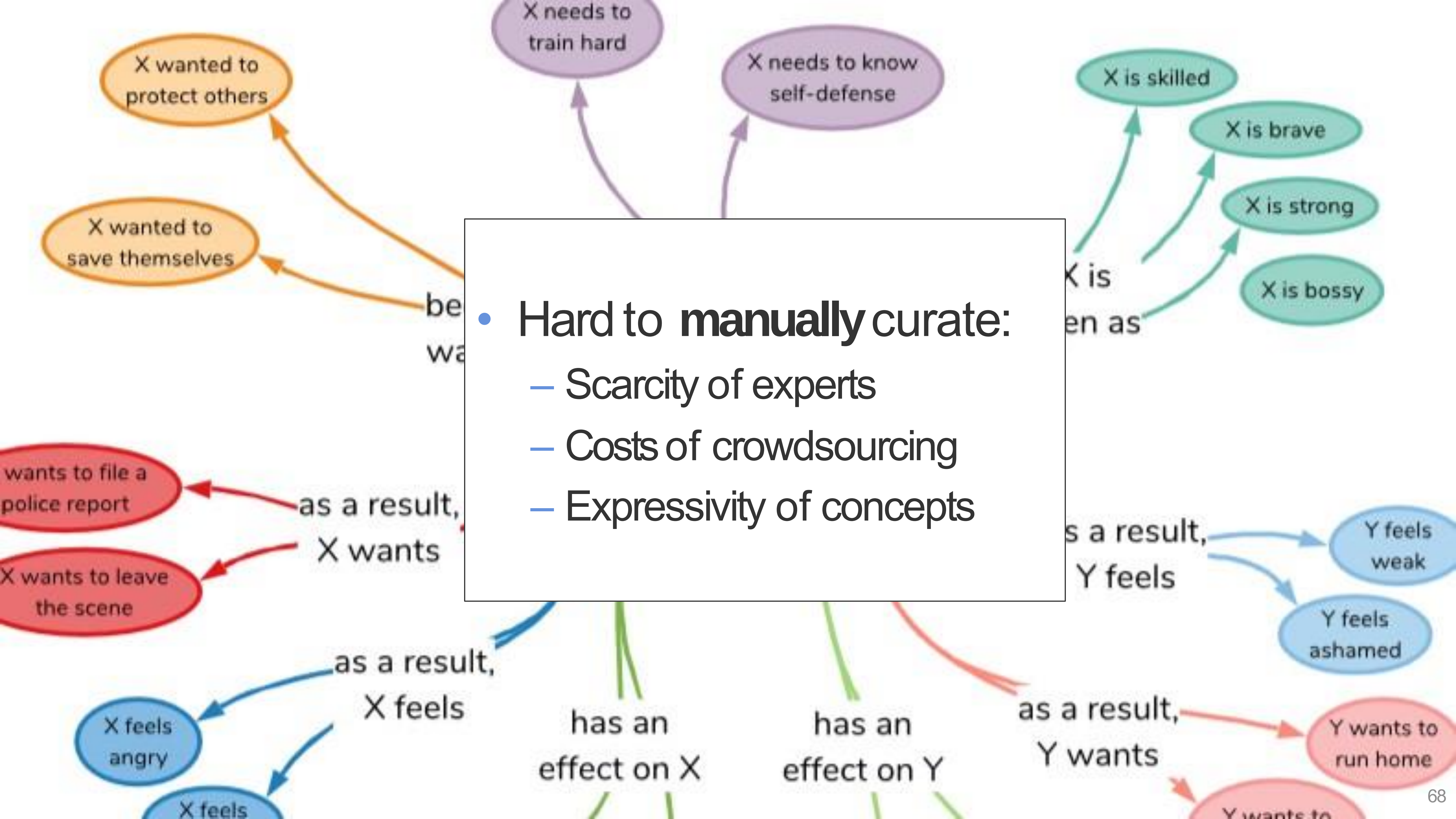
Knowledge Graphs



Slide: Bosselut, 2019

Familiarity Questions

1. What is a knowledge graph ?
2. Provide examples of some.



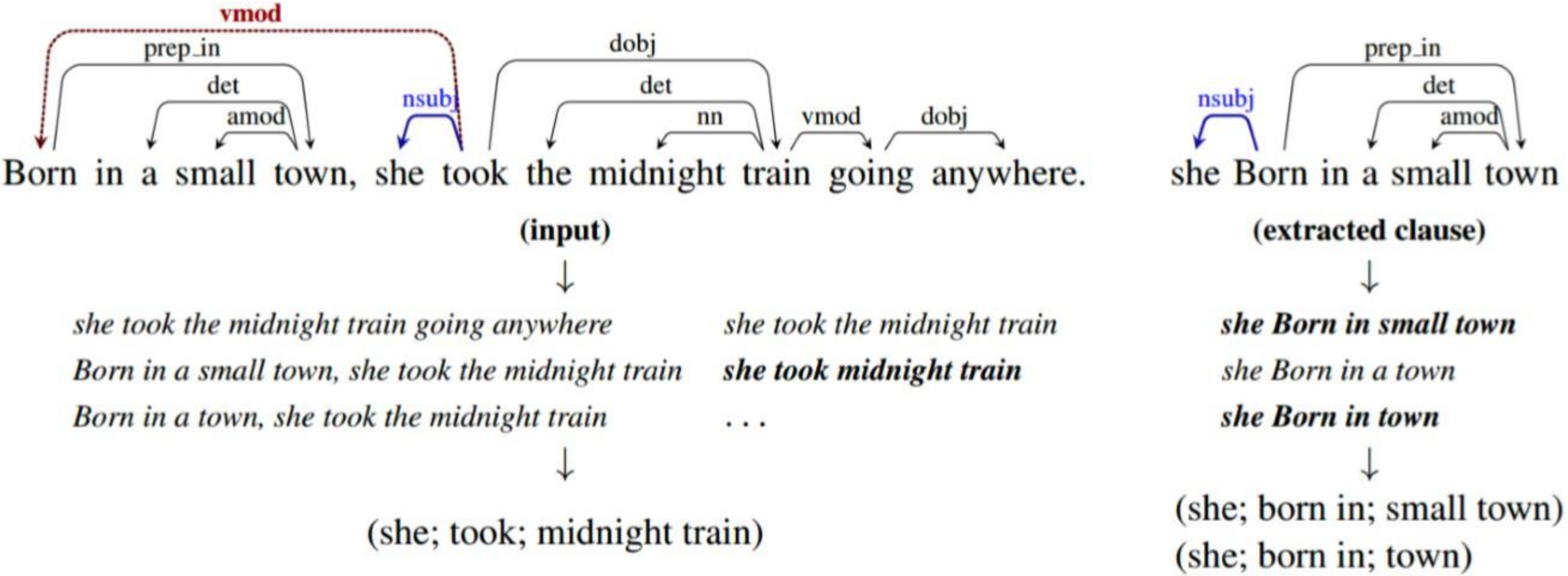
- Hard to **manually** curate:
  - Scarcity of experts
  - Costs of crowdsourcing
  - Expressivity of concepts

Commonsense Knowledge Graphs

- **Lots of entities**
  - A house, fish, plate, …..
  - Selling a house around Halloween
  - Selling a house because it is haunted
  - Selling a haunted house around Halloween
- **Lots of relations**
  - You can eat fish
  - A fish can be on a plate
  - You can sell your house
  - A fish probably won't buy it

Slide: Bosselut, 2019

Extractive Knowledge Graph Construction



https://nlp.stanford.edu/software/openie.html

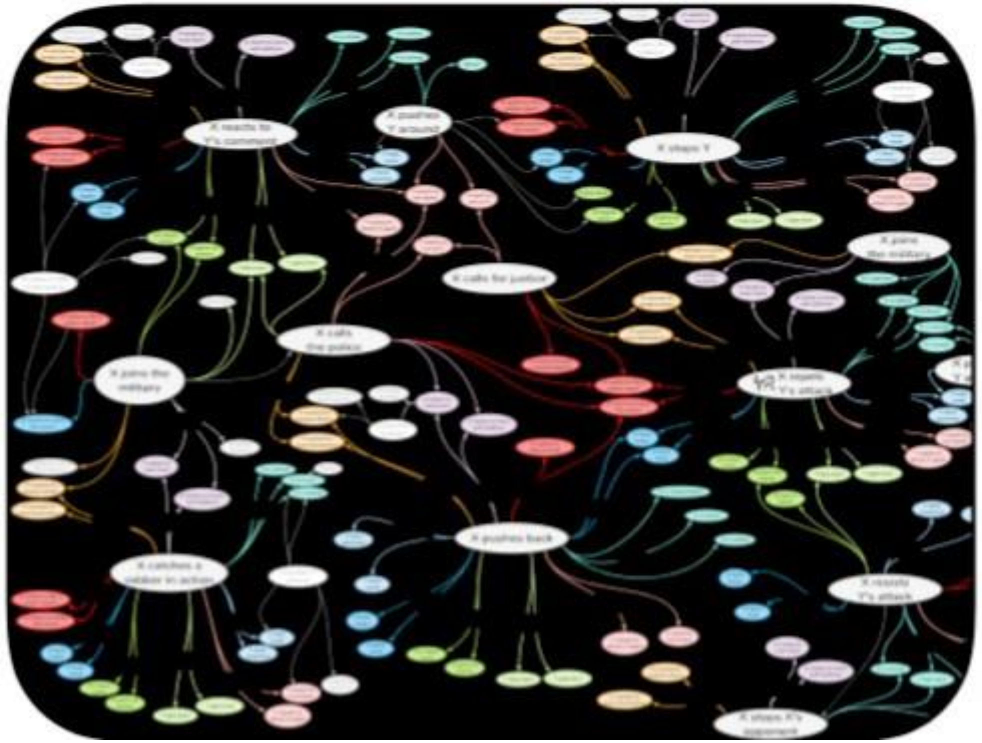
Issues: Extractive Knowledge Graph Construction

- Knowledge (particularly commonsense) is **immeasurably vast**, making it difficult to manually enumerate in all its forms
- Knowledge can be **assumed**, therefore **not written directly** in text
  - Extractive methods won't cover the cases we need
- Open text is the **most abundant, low-cost resource** we have

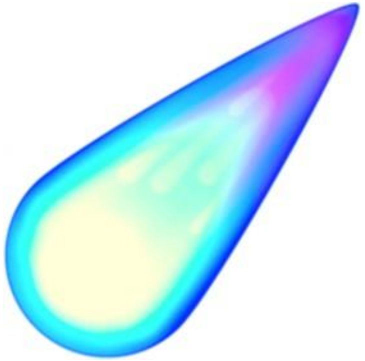
Slide: Bosselut, 2019

Pretrained Language Model

Seed Knowledge Graph

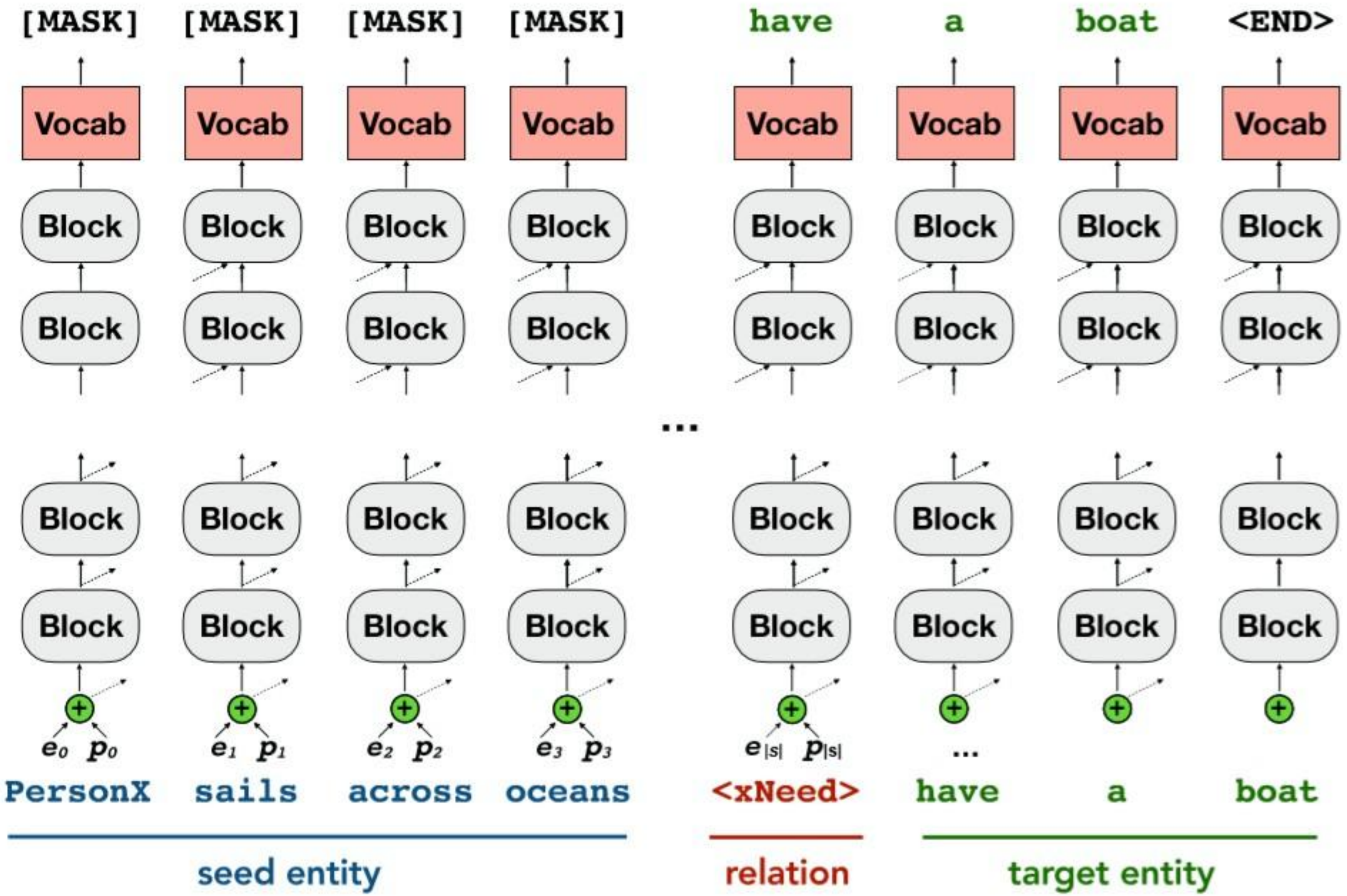


seed set of tuples



COMET

Given a **seed entity** and a **relation**, learn to generate the **target entity**



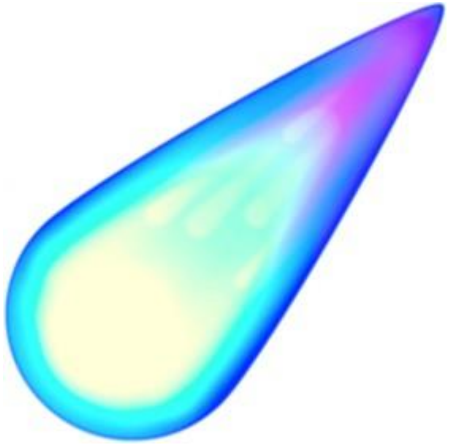
ATOMIC as seed data

~78% of tuple endings are rated correct by human workers

Human evaluation of gold annotations is ~86%

Model is able to generated high quality tuples

Percentage of generated tuples rated correct by human evaluators	Percentage of test set tuples rated correct by human evaluators
77.53	86.18



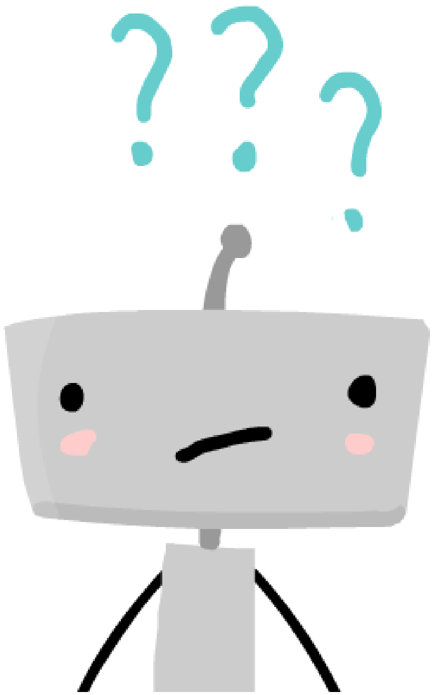
+

+ **Constraint Decoding**

Briefly…

- Transformer models for **generation** are **just attention based** autoregressive decoders
- **Perplexity** is not the only way to train LMs
- Long text generation requires learning a **better discourse**
- Introducing **implicit evaluators** for teaching discourse is beneficial
- Humans start writing stories with an **outline,** so **should agents**
- **Transformer based commonsense knowledge** can help create more human like stories/text

Some of the Challenges in Long Text Generation



- Training **corpora**
- Learning **discourse**, **reference**, etc.
- Loss-Function for high level semantics of the long text
- Repetitions and dull sentences (modal collapse)
- Maintaining **coherence between paragraphs**
- Sub-optimal **evaluation** metrics
- Word-by-word generation is **sub-optimal**, can't see the global context!
- Long text generation suffers from lack of implicit "**planning**" !
- Biased pre-trained language models
- Domain transfer is ridiculously hard
- Outdated generation methods: beam-search, or sampling
- Softmax bottleneck issues
- Left-to-write generation

Question set C [write up the answers]

1. What kind Information Extraction / Information Retrieval was leveraged for COMET (if any)

2. Can we use a system like COMET for generation of
   - Wikipedia articles
   - Generating long poems in Hindi
   - Translating English to Hindi

3. COMET is based on a CNN architecture. (True or False)

Discrete Metrics Don't Work For Text Generation

		BLEU	ROUGE	CIDEr	SPICE	METEOR	Word Movers
Original	a man wearing a red life jacket is sitting in a canoe on a lake	1.00	1.00	10.0	1.00	1.00	1.00
Candidate	a man wearing a life jacket is in a small boat on a lake	0.45	0.67	2.19	0.40	0.28	0.19
Synonyms	guy wearing a red life vest is in a small boat on a lake	**0.20**	0.57	**0.65**	**0.0**	**0.17**	**0.10**
Word Order	in a small boat on a lake a man is wearing a life jacket	**0.26**	**0.38**	**1.32**	0.40	0.26	0.19

Kilickaya, et.al., 2017

Further Reading

BERT-PLI: https://www.ijcai.org/Proceedings/2020/0484.pdf

Structured Document Retrieval: http://www.cwr.cl/documentRetrival.pdf

ElasticBERT: https://medium.com/analytics-vidhya/elasticbert-information-retrieval-using-bert-and-elasticsearch-51fef465b9ae

Easing Legal News Monitoring with Learning to Rank and BERT - https://www.youtube.com/watch?v=PUqyvKid9TY

Legal Area Classification: A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments - https://arxiv.org/pdf/1904.06470.pdf

Simple Applications of BERT for Ad Hoc Document Retrieval - https://arxiv.org/pdf/1903.10972.pdf