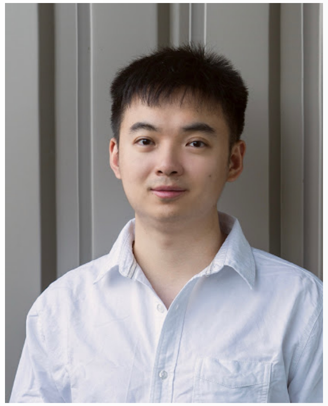
EMNLP 2021 Tutorial

Knowledge-Enriched Natural Language Generation





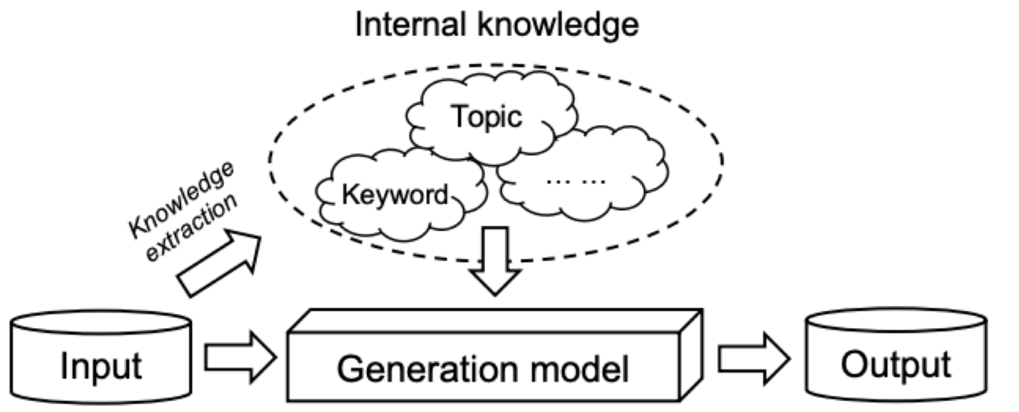

Wenhao Yu[1], Meng Jiang[1], Zhiting Hu[2], Qingyun Wang[3], Heng Ji[3,4], Nazneen Rajani[5]

1 University of Notre Name 2 University of California San Diego
3 University of Illinois at Urbana-Champaign 4 Amazon 5 Salesforce Research

Knowledge-enhanced NLG (Overall)

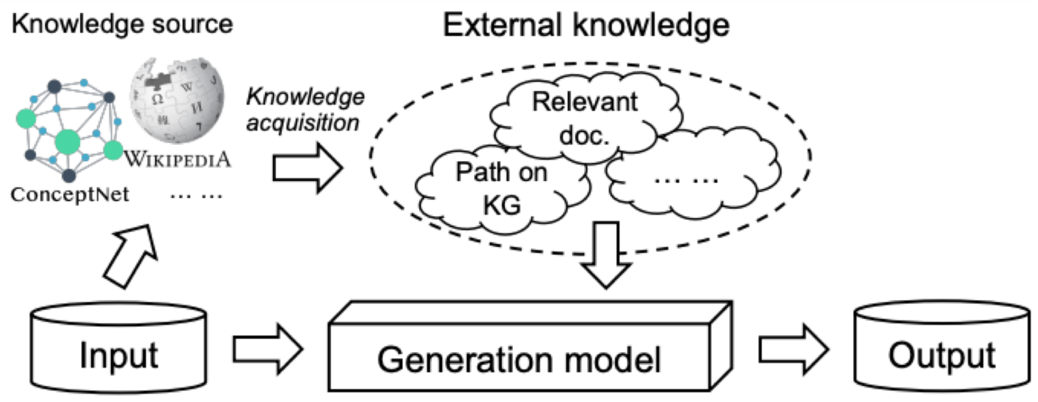
Knowledge-enhanced NLG Methods

Internal Knowledge-enhanced Methods



Internal knowledge creation takes place within the input text(s)

External Knowledge-enhanced Methods



External knowledge acquisition occurs when knowledge is provided from outside sources

Knowledge-enhanced NLG (Overall)

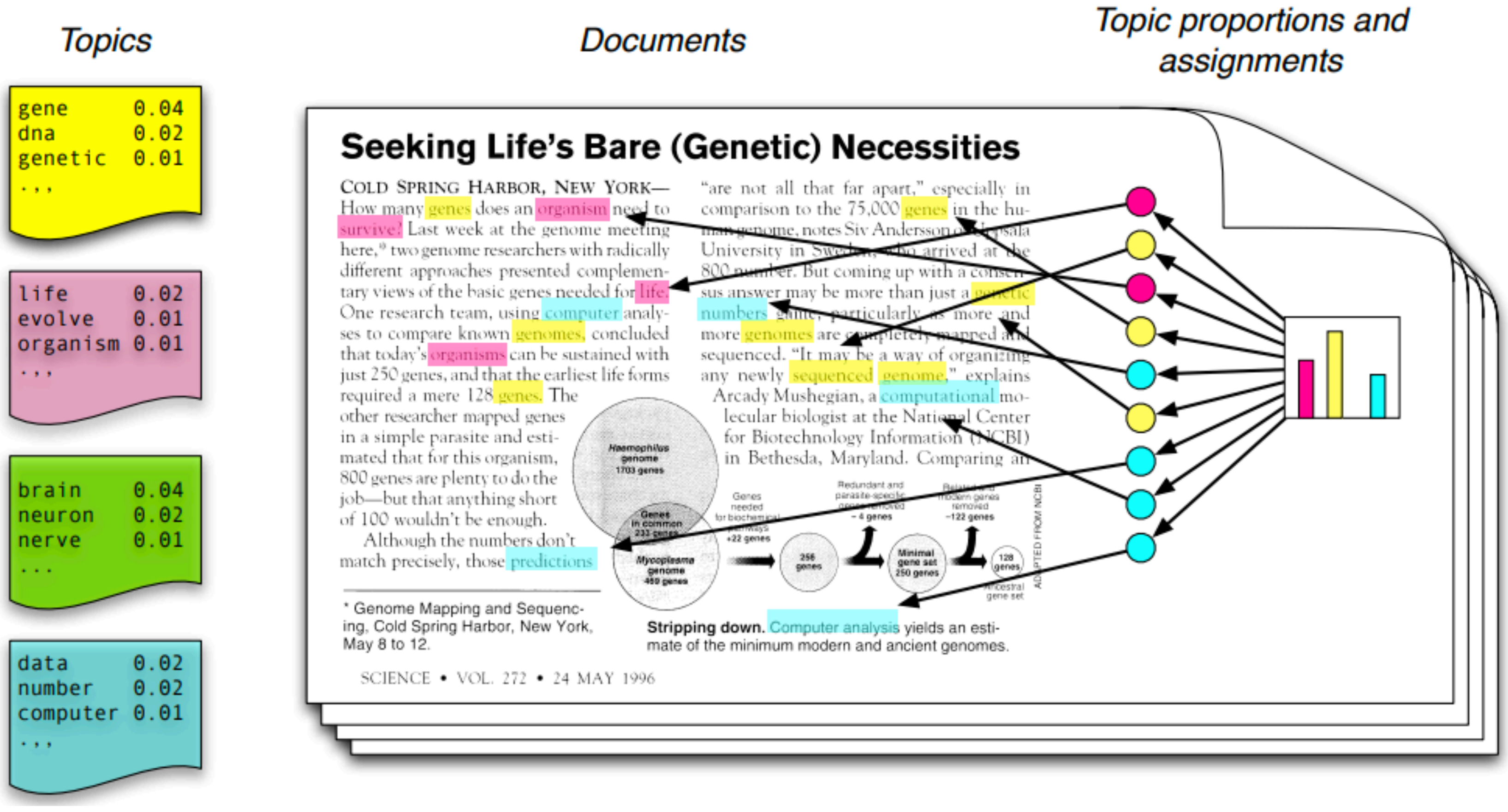
Knowledge-enhanced NLG Methods

- Internal Knowledge-enhanced Methods
- External Knowledge-enhanced Methods

	knowledge source	knowledge extraction and representation
Internal Knowledge-enhanced Methods	Topic	E.g., LDA, CNN, Neural topic
	Keyword	E.g., Keyword assignment, keyword extraction
	Linguistic features	E.g., POS tag, NER tag, Dependency parsing
External Knowledge-enhanced Methods	Knowledge base	E.g., KB memory network, KB related tasks
	Knowledge graph	E.g., KGE, Path finding and reasoning, Graph neural network
	Grounded text	E.g., Retrieval-augmented, Background based methods

Topic-enhanced NLG methods

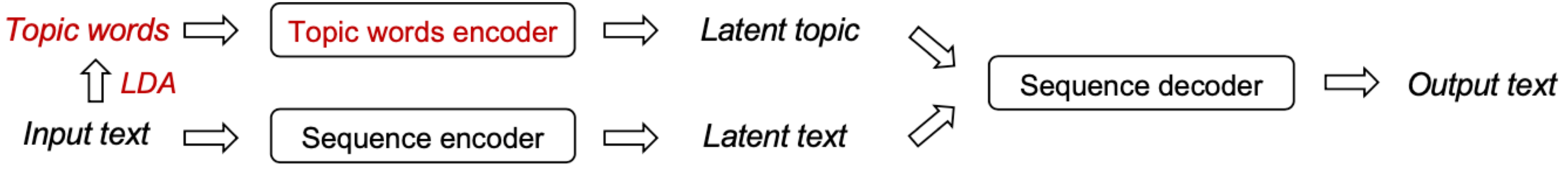
- Topic, which can be considered as a representative or compressed form of text, has been often used to maintain the semantic coherence and guide the NLG.



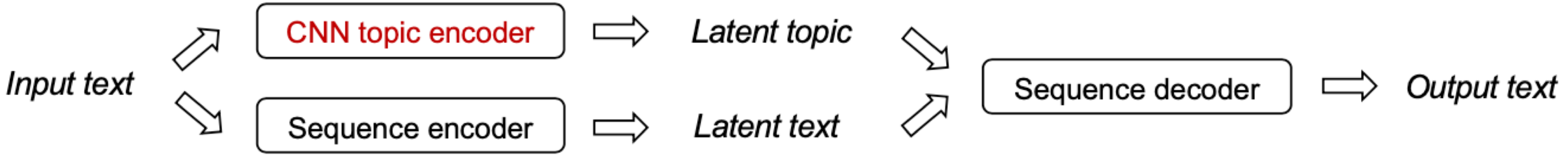
LDA topic modeling

- Each **topic** is a distribution over words.
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

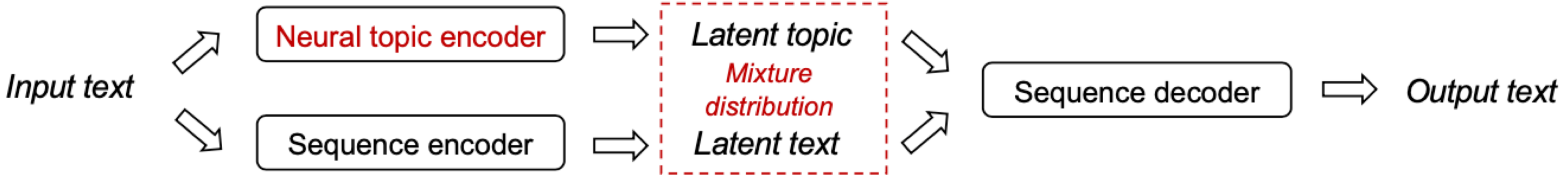
Topic-enhanced NLG methods



(M1) Leverage topic words from generative topic models



(M2) Jointly optimize generation model and CNN topic model



(M3) Enhance NLG by neural topic models with variational inference

Topic-enhanced NLG methods

Important applications

- **Dialogue system.** A vanilla Seq2Seq often generates trivial response, such as "I do not know", "I see". These responses are boring with very little information, quickly leading the conversation to an end.

- **Machine translation.** Though the input and output languages are different the contents are the same, and globally, under the same topic.

- **Paraphrase.** Naturally, paraphrases concern the same topic, which can serve as an auxiliary guidance to promote the preservation of source semantic.

Topic-enhanced NLG methods

- Topic Aware Neural Response Generation, In AAAI 2017
- Application: Dialogue system
- Motivation: natural and fluent ✓ informative and interesting ✗

DIALOGUE SYSTEM

You haven't been given an assignment in this case

I don't know what you are talking about

DIALOGUE SYSTEM

You programmed me to gather intelligence.
That's all I've ever done.

I see.

Figure: Two generated responses from a vanilla Seq2Seq model

Topic-enhanced NLG methods

- Topic Aware Neural Response Generation, In AAAI 2017
- Solution: extract topic from input -> incorporate topic into Seq2Seq

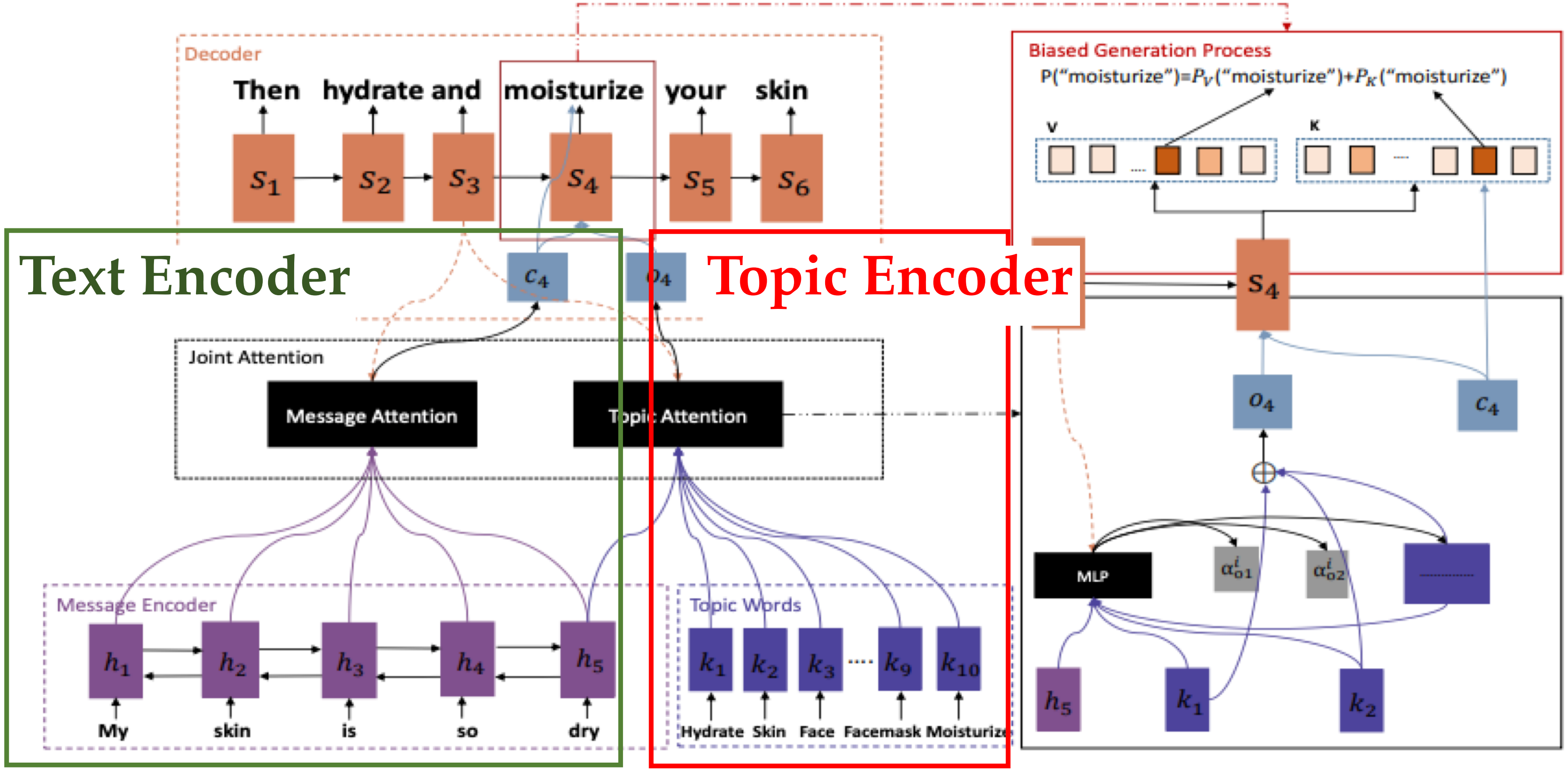


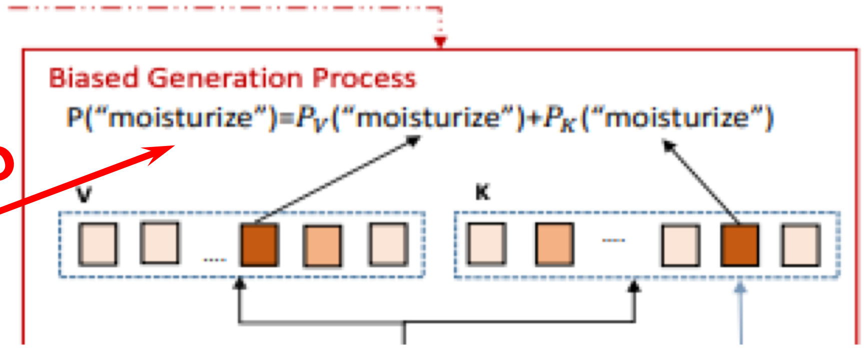
Figure: Proposed framework of topic-enhanced Seq2Seq model

Topic-enhanced NLG methods

- Topic Aware Neural Response Generation, In AAAI 2017
- Dataset: Baidu Tieba (not public);   Metric: Perplexity ↓; Distinct-k ↑

Models	PPL-D	PPL-T	distinct-1	distinct-2
S2SA	147.04	133.11	604/.091	1168/.207
S2SA-MMI	147.04	133.11	603/.151	1073/.378
S2SA-TopicConcat	150.45	132.12	898/.116	2197/.327
S2SA-TopicAttention	133.81	**119.55**	894/.106	2057/.277
TA-Seq2Seq	134.63	**122.82**	1355/.161	2970/.401

w/o



- Distinct-k measures the total number of unique k-grams in the entire corpus.
- The higher distinct-k indicates higher diversity (more unique k-grams are generated)

Topic-enhanced NLG methods

- Topic Aware Neural Response Generation, In AAAI 2017
- Dataset: Baidu Tieba (not public); Metric: Perplexity ↓; Distinct-k ↑

Models	PPL-D	PPL-T	distinct-1	distinct-2
S2SA	147.04	133.11	604/.091	1168/.207
S2SA-MMI	147.04	133.11	603/.151	1073/.378
S2SA-TopicConcat	150.45	132.12	898/.116	2197/.327
S2SA-TopicAttention	133.81	**119.55**	894/.106	2057/.277
TA-Seq2Seq	134.63	**122.82**	1355/.161	2970/.401

Message	TA-Seq2Seq	S2SA-MMI	S2SA
你也玩摄影 You also like photography	我不玩**摄影**，只是想拍个**照片**而已 I don't work on **photography**. I just want to take a few **pictures**.	额额额... Well well well....	额... Well...
我的左脸颊上有一些红血丝 There is some redness on my left cheek.	我也是**敏感肌** My **skin** is also very **sensitive**.	我也是啊 Me too.	啊啊啊啊 Oh my god oh my god
大学生的话可以去实习吗 Can a college student apply for an internship	可以的，如果你着急的话可以先填个**申请表** Yes, you can fill an **application form** first if you are in a hurry.	可以的可以的 Yes, you can.	可以的 Yes.
我的皮肤好干 My skin is so dry.	那就**补水保湿**吧 Then **hydrate** and **moisturize** your skin.	我也是啊 Me too.	啊啊啊啊 Oh my god oh my god.

Topic-enhanced NLG methods

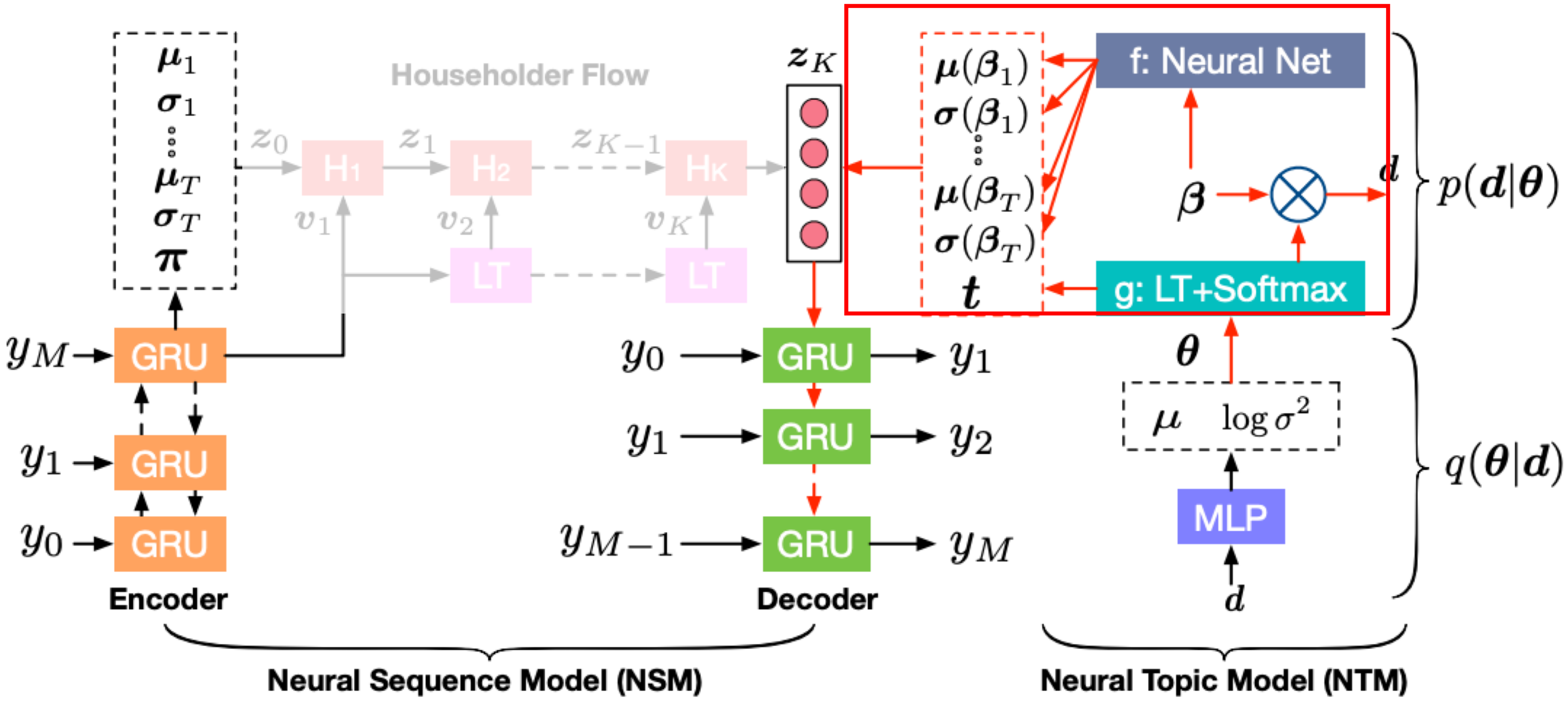
- Topic-Guided Variational Autoencoders for Text Generation, In NAACL 2019

Motivations:

- (1) LDA models may **fail to find proper topics** that the NLG task requires.
- (2) LDA models are **separated from the training process of generation**, so they cannot adapt to the diversity of dependencies between input and output sequences.

Topic-enhanced NLG methods

- Topic-Guided Variational Autoencoders for Text Generation, In NAACL 2019



Topic-enhanced NLG methods

- Topic-Guided Variational Autoencoders for Text Generation, In NAACL 2019

Metric	Methods	APNEWS				IMDB				BNC			
		B-2	B-3	B-4	B-5	B-2	B-3	B-4	B-5	B-2	B-3	B-4	B-5
test-BLEU	VAE	0.564	0.278	0.192	0.122	0.597	0.315	0.219	0.147	0.479	0.266	0.169	0.117
	VAE+HF (K=1)	0.566	0.280	0.193	0.124	0.593	0.317	0.218	0.148	0.475	0.268	0.165	0.112
	VAE+HF (K=10)	0.570	0.279	0.195	0.123	0.610	0.322	0.221	0.147	0.483	0.270	0.169	0.110
	TGVAE (K=0, T=10)	0.582	0.320	0.203	0.125	0.627	0.362	0.223	0.159	0.517	0.282	0.181	0.115
	TGVAE (K=1, T=10)	0.581	0.326	0.202	0.124	0.623	0.358	0.224	0.160	0.519	0.282	0.182	0.118
	TGVAE (K=10, T=10)	0.584	0.327	0.202	0.126	0.621	0.357	0.223	0.159	0.518	0.283	0.173	0.119
	TGVAE (K=10, T=30)	0.627	0.335	0.207	0.131	**0.655**	0.369	**0.243**	**0.165**	0.528	**0.291**	0.182	0.119
	TGVAE (K=10, T=50)	**0.629**	**0.340**	**0.210**	**0.132**	0.652	**0.372**	0.239	0.160	**0.535**	0.290	**0.188**	**0.120**
self-BLEU	VAE	0.866	0.531	0.233	-	0.891	0.632	0.275	-	0.851	0.51	0.163	-
	VAE+HF (K=1)	0.865	0.533	0.241	-	0.899	0.641	0.278	-	0.854	0.515	0.163	-
	VAE+HF (K=10)	0.873	0.552	0.219	-	0.902	0.648	0.262	-	0.854	0.520	0.168	-
	TGVAE (K=0, T=10)	0.847	0.499	0.161	-	0.878	0.572	0.234	-	0.832	0.488	0.160	-
	TGVAE (K=1, T=10)	0.847	0.495	0.160	-	0.871	0.571	0.233	-	0.828	0.483	0.150	-
	TGVAE (K=10, T=10)	0.839	0.512	0.172	-	0.889	0.577	0.242	-	0.829	0.488	0.151	-
	TGVAE (K=10, T=30)	0.811	0.478	0.157	-	0.850	0.560	0.231	-	0.806	0.473	0.150	-
	TGVAE (K=10, T=50)	**0.808**	**0.476**	**0.150**	-	**0.842**	**0.559**	**0.227**	-	**0.793**	**0.469**	**0.150**	-

VAE: RNN with variational autoencoder; HF: householder flow; TGVAE: topic guided variational autoencoder

Topic-enhanced NLG methods (discussion)

- Advantages and disadvantages of different topic-enhanced methods

- LDA topic

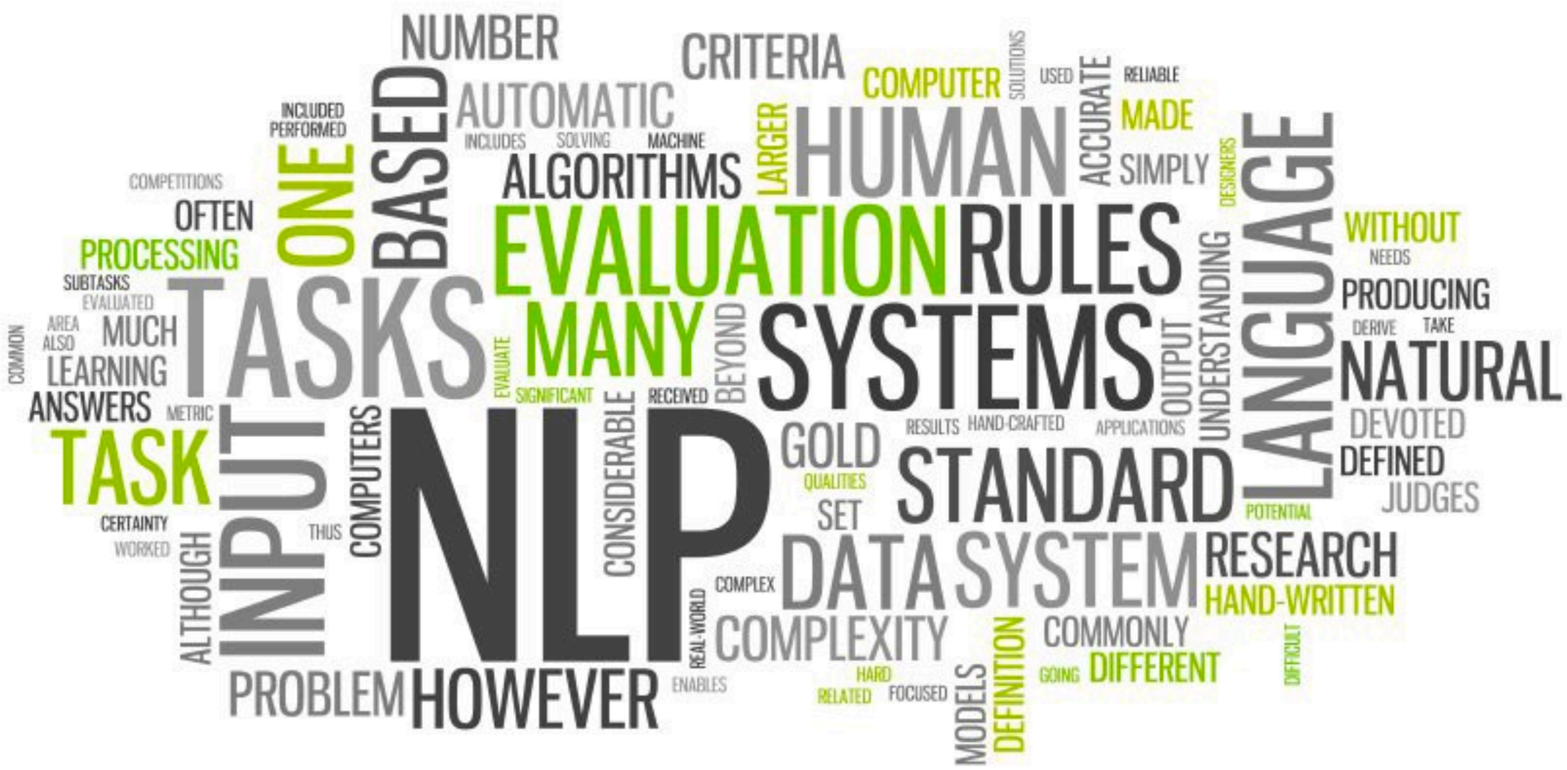
Pros: LDA has a strict probabilistic explanation with great interpretability
Cons: LDA models are separated from the generation training process

- Neural topic

Pros: They enable back propagation for joint optimization, contributing to more coherent topics, and can be scaled to large data sets.
Cons: topic distribution is assumed to be an isotropic Gaussian, which makes them incapable of modeling topic correlations.

Keyword-enhanced NLG methods

- Keyword (aka., key phrase, key term) is often referred as a sequence of one or more words, providing a compact representation of the content of a document.



LDA is based on a generative probabilistic model that associates a topic with a distribution over set of words, but those words would not normally be considered "keywords" in any way.

Keyword-enhanced NLG methods

- Keywords-Guided Abstractive Sentence Summarization, In AAAI 2020

- Applications:

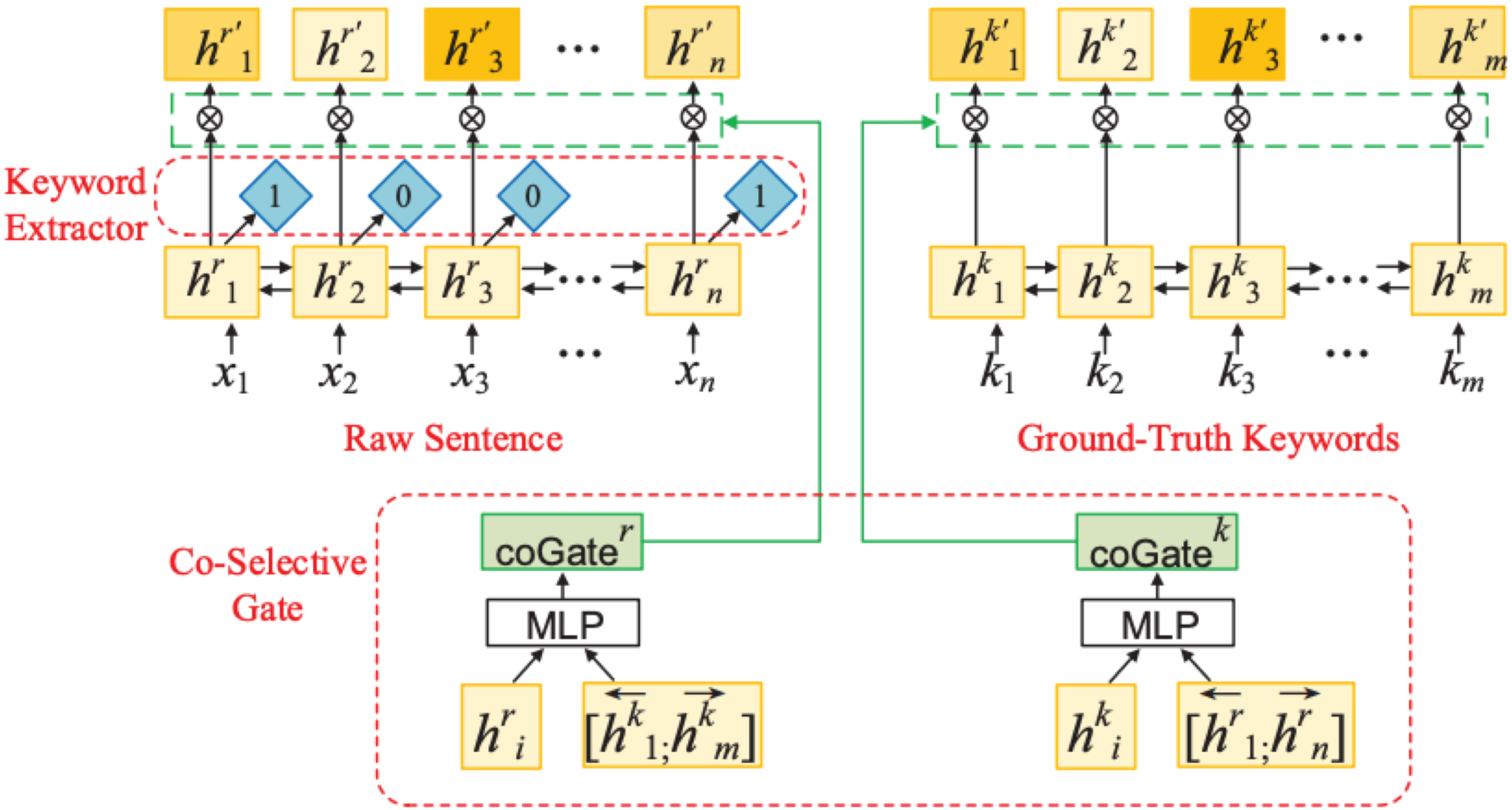
Vanilla Seq2Seq: hard to control and often misses salient information.

Summarization

Keyword: provide significant clues of the main points about the document.

Keyword-enhanced NLG methods

- Keywords-Guided Abstractive Sentence Summarization, In AAAI 2020



Keyword-enhanced NLG methods

- Keywords-Guided Abstractive Sentence Summarization, In AAAI 2020
- Dataset: Gigawords   Metric: ROUGE score

Method		R-1	R-2	R-L
ABS		37.41	15.87	34.70
SEASS		46.86	24.58	43.53
PG		46.97	24.63	43.66
KIGN		46.18	23.93	43.44
Bottom-up		45.80	23.61	42.54
Co-Selective	Concat+DualPG	47.05	24.39	43.77
	Gated+DualPG	47.13	24.87	44.34
	Hier+DualPG	**47.14**	**25.06**	**44.39**

Linguistic feature-enhanced NLG methods

- Why does linguistic features include?
  - Lemma; POS tag; NER tags; dependency parsing; semantic parsing
- How to include linguistic features into NLG?
  - Fused encoder (often used for POS tags, NER tags -> See below figure)
  - Separate encoder (often used for dependency graphs -> GNN)

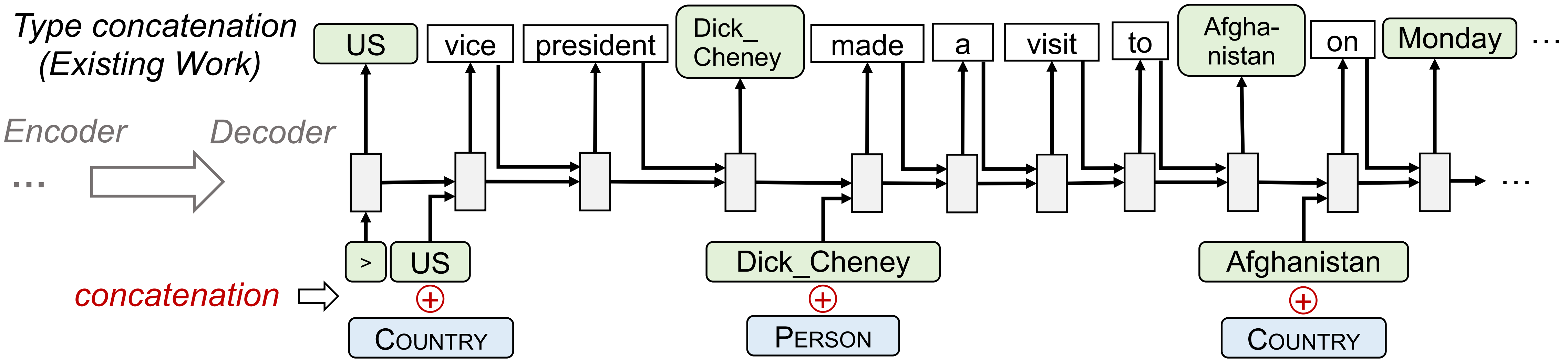
Entity Types leak more information than we think

- Accurate contexts depend on the type of word

Newark (Name)	Say hello to Newark for me!	✓
	I just arrived at Newark.	✗
Newark(Location)	Say hello to Newark for me!	✗
	I just arrived at Newark.	✓

Linguistic feature-enhanced NLG methods

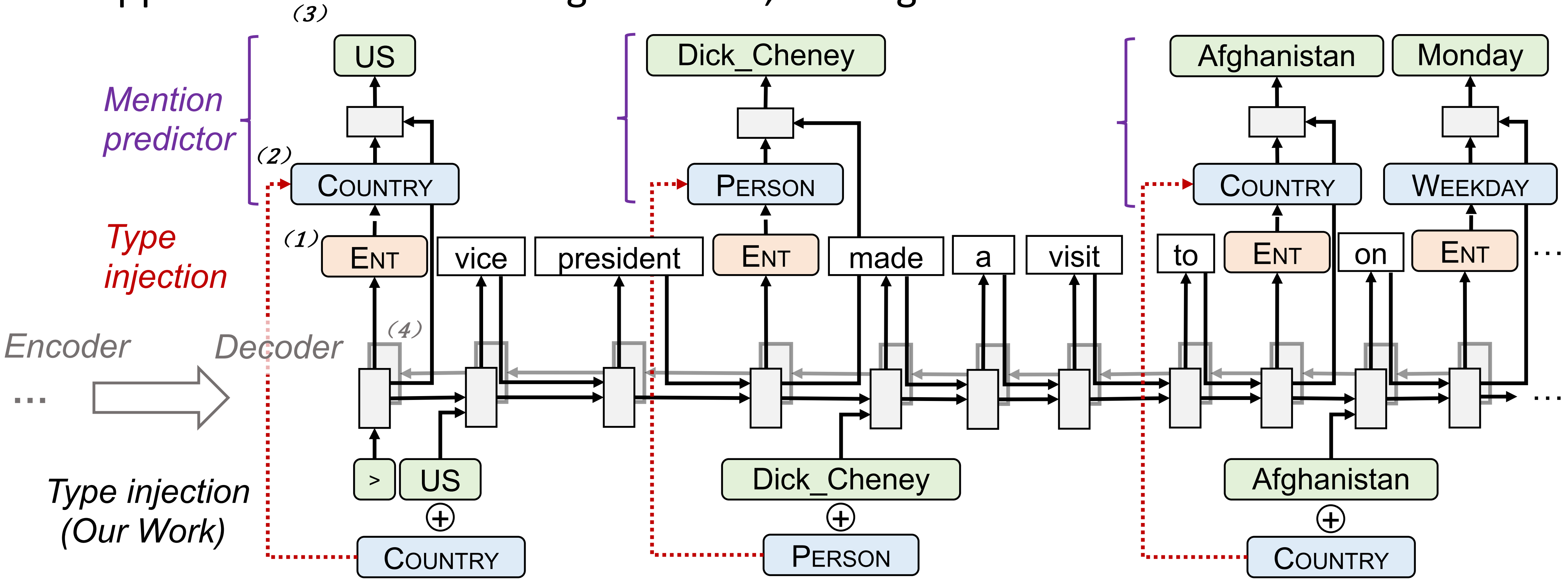
- Entity Types serve as a guide to generate more accurate context words.



Concatenating entity mention and type embeddings is a straightforward way to use type information.

Linguistic feature-enhanced NLG methods

- Injecting Entity Types into Entity-Guided Text Generation, In EMNLP 2021
- Application: Word-to-text generation, News generation



Steps: (1) predicting the <Ent> token (i.e., entity indicator) (2) injecting the entity types (3) predicting the entity mention using the type embedding and hidden state by a mention predictor (4) combine with an entity enhanced NLU module

Linguistic feature-enhanced NLG methods

- Injecting Entity Types into Entity-Guided Text Generation, In EMNLP 2021
- Dataset: Gigaword, New York times; Metric: ROUGE ↑; BLUE ↑

Table 3: Our *InjType* can outperform various baseline models enhanced by type embedding concatenation.

Methods	GIGAWORDS			NYT		
	ROUGE-2	ROUGE-L	BLEU-4	ROUGE-2	ROUGE-L	BLEU-4
Seq2Seq	8.83±0.15	31.43±0.13	12.21±0.30	8.83±0.15	31.43±0.13	12.21±0.30
SeqAttn	9.10±0.13	36.62±0.11	16.17±0.28	5.95±0.15	29.67±0.06	11.86±0.15
CopyNet	9.44±0.11	36.96±0.10	16.40±0.24	6.25±0.14	30.58±0.09	11.96±0.14
GPT-2	9.04±0.20	31.30±0.16	15.66±0.40	5.86±0.20	24.19±0.14	10.89±0.22
UniLM	11.77±0.18	36.54±0.15	17.66±0.35	7.47±0.15	30.66±0.13	12.90±0.20
InjType	**13.37±0.12**	**41.16±0.31**	**18.55±0.09**	**8.55±0.09**	**31.53±0.17**	**13.14±0.03**
⊢ w/o MP	9.39±0.16	38.34±0.10	16.36±0.25	6.52±0.09	30.10±0.08	12.19±0.10
⊢ w/o NLU	12.85±0.18	40.65±0.37	18.24±0.26	8.13±0.10	30.80±0.36	13.10±0.09

KG-enhanced text generation methods

- Knowledge graph (KG), as a type of structured human knowledge consisting of entities†, relations, and semantic descriptions. People can easily traverse links to discover how entities are interconnected to express certain knowledge.

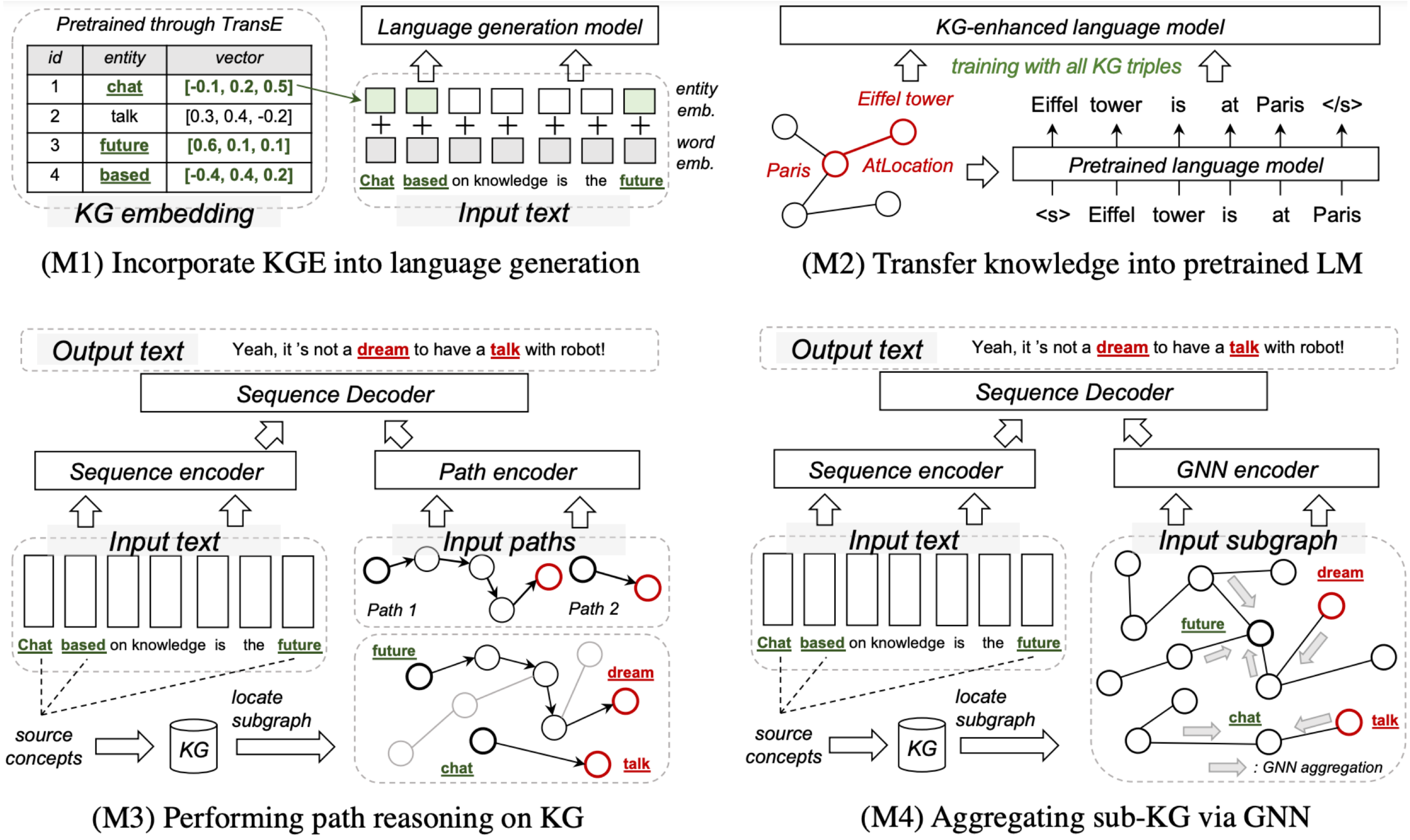
- KG definition: A KG is defined as $\mathcal{G} = (\mathcal{U}, \mathcal{E}, \mathcal{R})$, where $\mathcal{U}$ is the set of entity nodes and $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{R} \times \mathcal{U}$ is the set of typed edges between nodes in $\mathcal{U}$ with a certain relation in the relation schema $\mathcal{R}$.

Topic-enhanced NLG methods

Important applications

- **Commonsense reasoning.** It often needs to exploit both structural and semantic information of the commonsense KG and perform reasoning over multi-hop relational paths, in order to augment the limited information for commonsense reasoning.
- **Dialogue system.** A dialogue may shift focus from one entity to another, breaking one discourse into several segments, which can be represented as a linked path connecting the entities and their relations.
- **Creative text generation.** This task can be found in both scientific and story-telling domains. Scientific writing aims to explain natural processes and phenomena step by step, so each step can be reflected as a link on KG and the whole explanation is a path. In story generation, the implicit knowledge in KG can facilitate the understanding of storyline and better predict what will happen in the next plot.

KG-enhanced text generation methods



- M1: KGE into NLG

  [Zhou 2018 IJCAI]

- M2: KG into PLMs

  [Guan 2020 TACL]

- M3: Path Reasoning

  [Liu 2019 EMNLP]

  [Ji 2020 EMNLP]

- M4: GNN on sub-KG

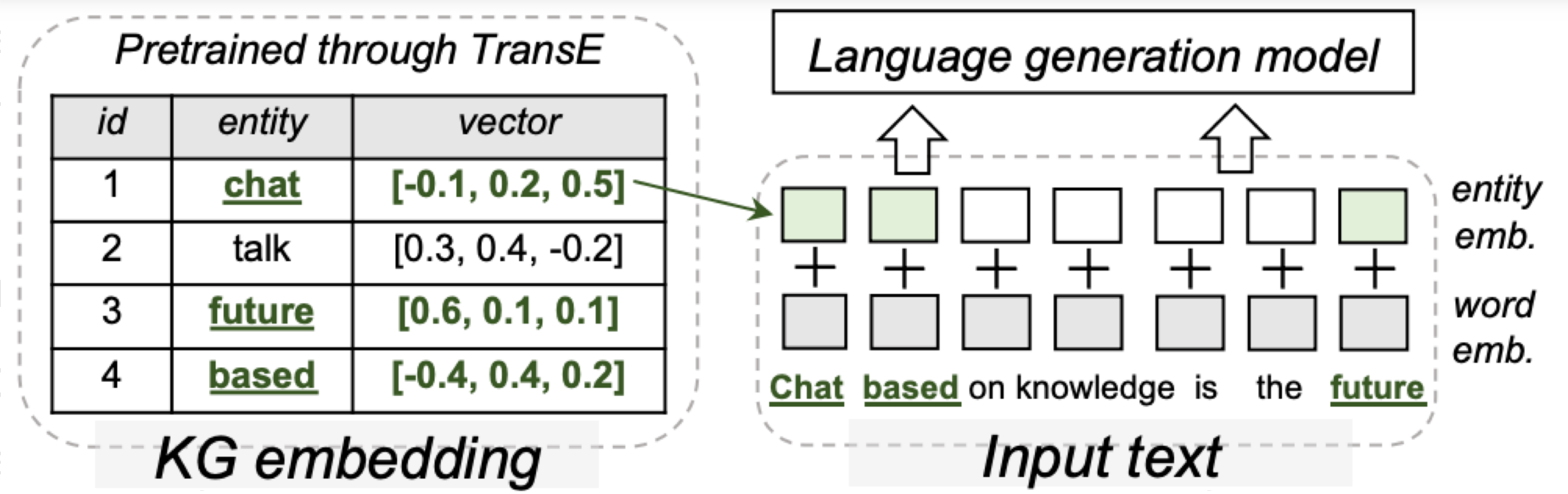
  [Zhou 2018 IJCAI]

  [Zhang 2020 ACL]

KG-enhanced text generation methods

- **M1: Incorporate Knowledge Graph Embeddings into NLG**

- What is knowledge graph embedding (KGE)?
  - Goal: KGE represe
    dimensionality wl

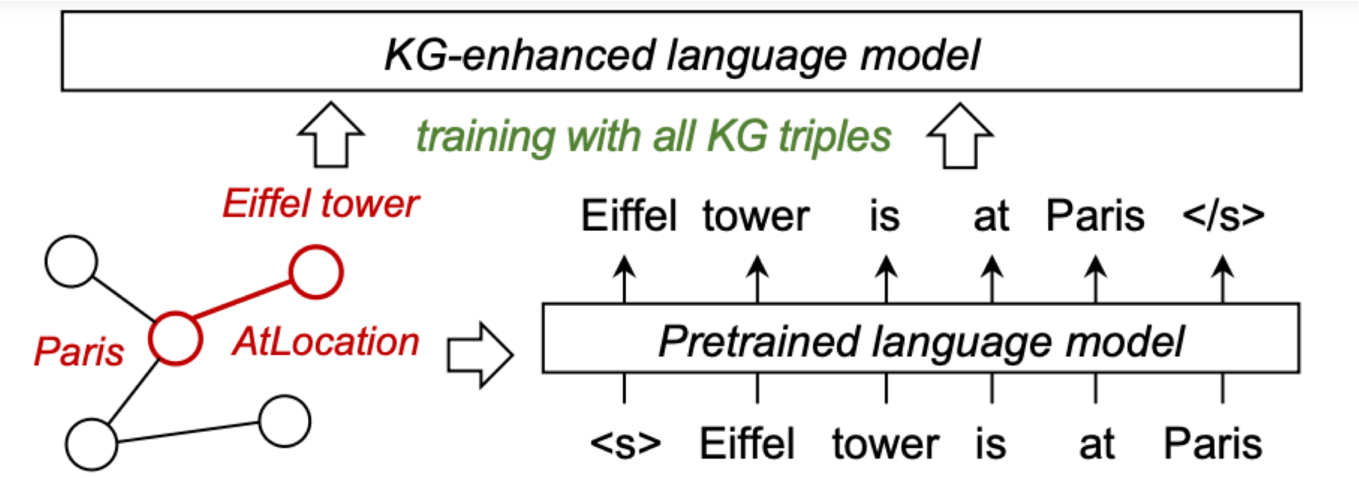


- What are the com
  - TransE: Given a K(
    embedded entitie
  - Example: Tokyo + IsCapitalOf ≈ Japan.

KG-enhanced text generation methods

- **M2: Transfer Knowledge into LMs with Knowledge Triplet Information**

Knowledge Bases	Original Triples	Examples of Transformed Sentences
ConceptNet	(eiffel tower, **AtLocation**, paris) (telephone, **UsedFor**, communication)	eiffel tower **is at** paris. telephone **is used for** communication.
ATOMIC	(PersonX dates for years, **oEffect**, continue dating) (PersonX cooks spaghetti, **xIntent**, to eat)	PersonX dates for years. **PersonY will** continue dating. PersonX cooks spaghetti. **PersonX wants** to eat.

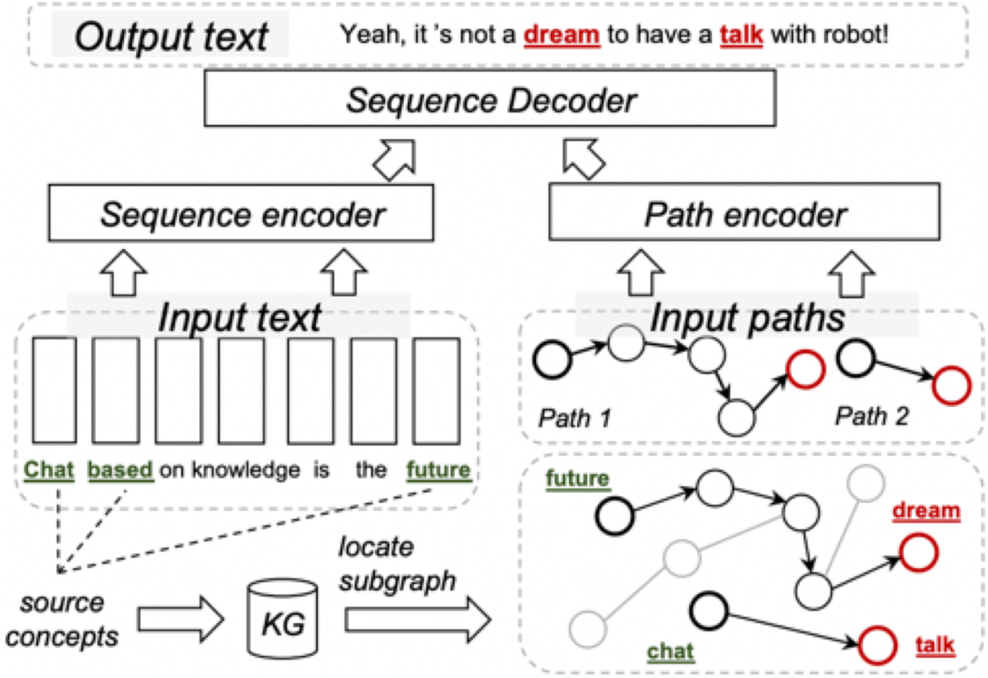


- A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation, TACL 2020

KG-enhanced text generation methods

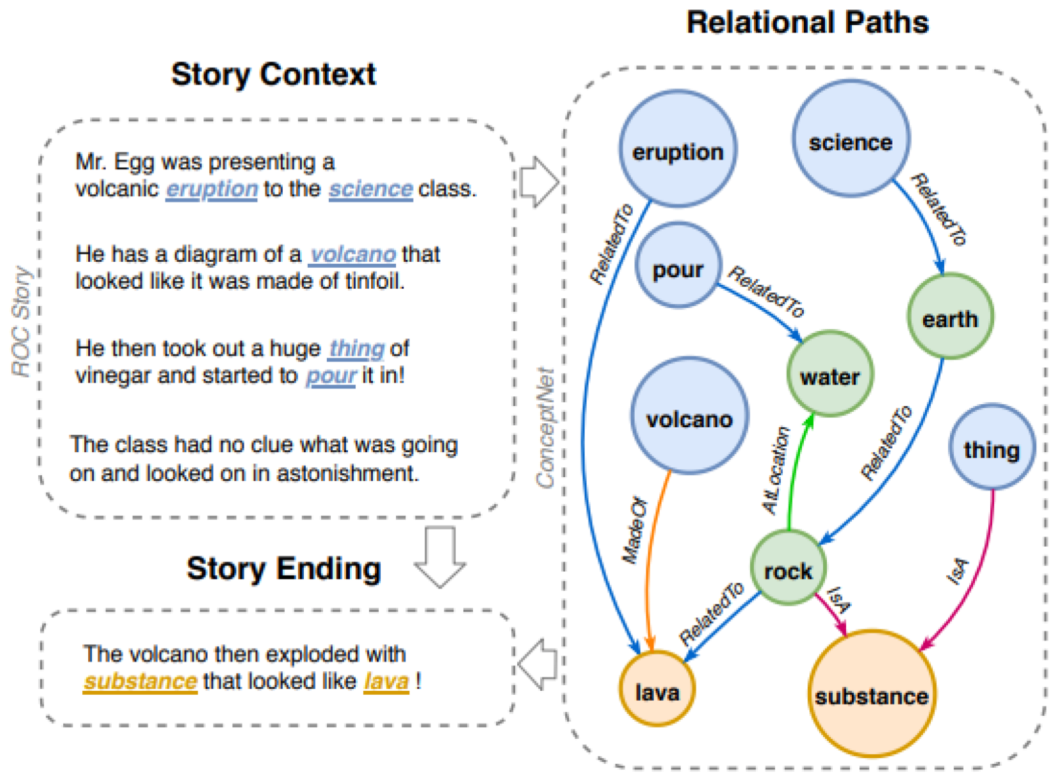
- **M3: Perform Reasoning over KG via Path Finding Strategies**

- Path routing and ranking (PRA algrithom)
  - PRA uses random walks to perform multiple bounded depth-first search processes to find relational paths on the KG, then integrate the path into Seq2Seq models
- Neural network based path scoring/finding



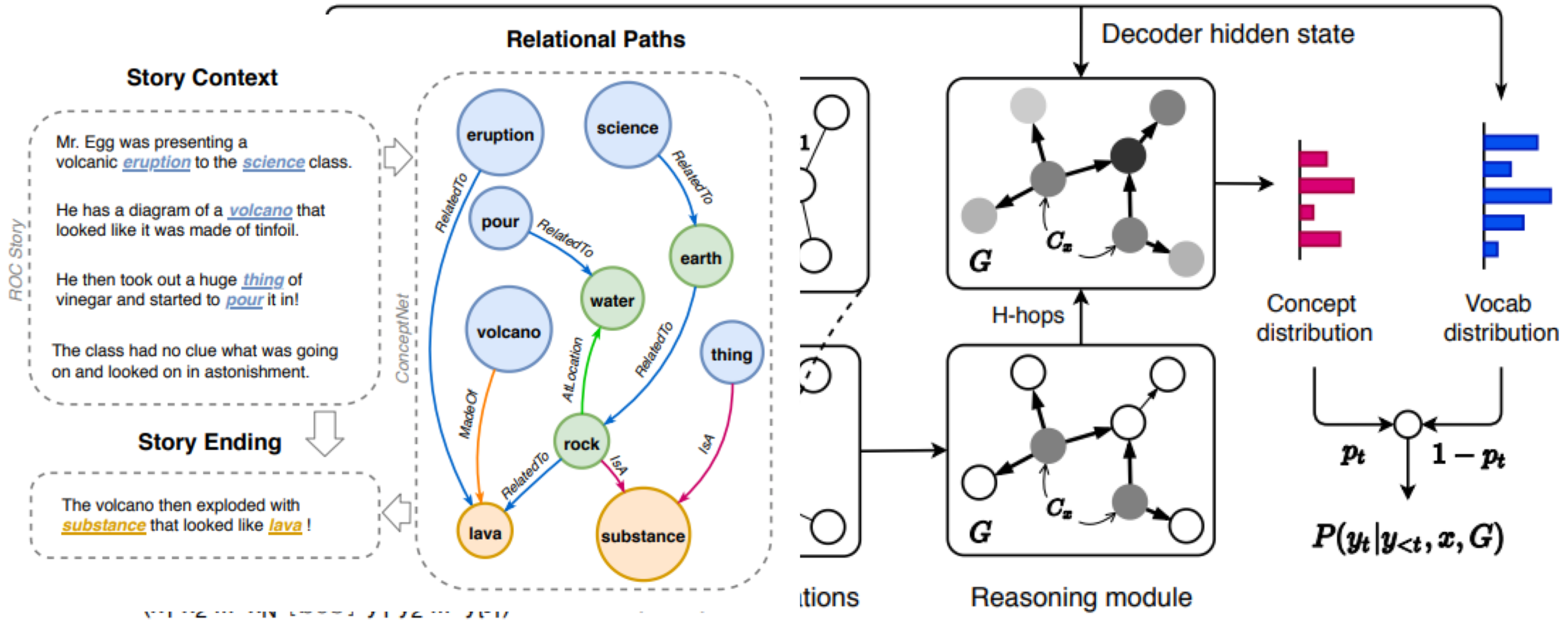
KG-enhanced text generation methods

- Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, In EMNLP 2020
- Application: Generative commonsense reasoning (e.g., story, alpha-NLG)
- Motivation: To reason over multi-hop relational paths where multiple conec-ted triples provide chains of evidence for grounded text generation.



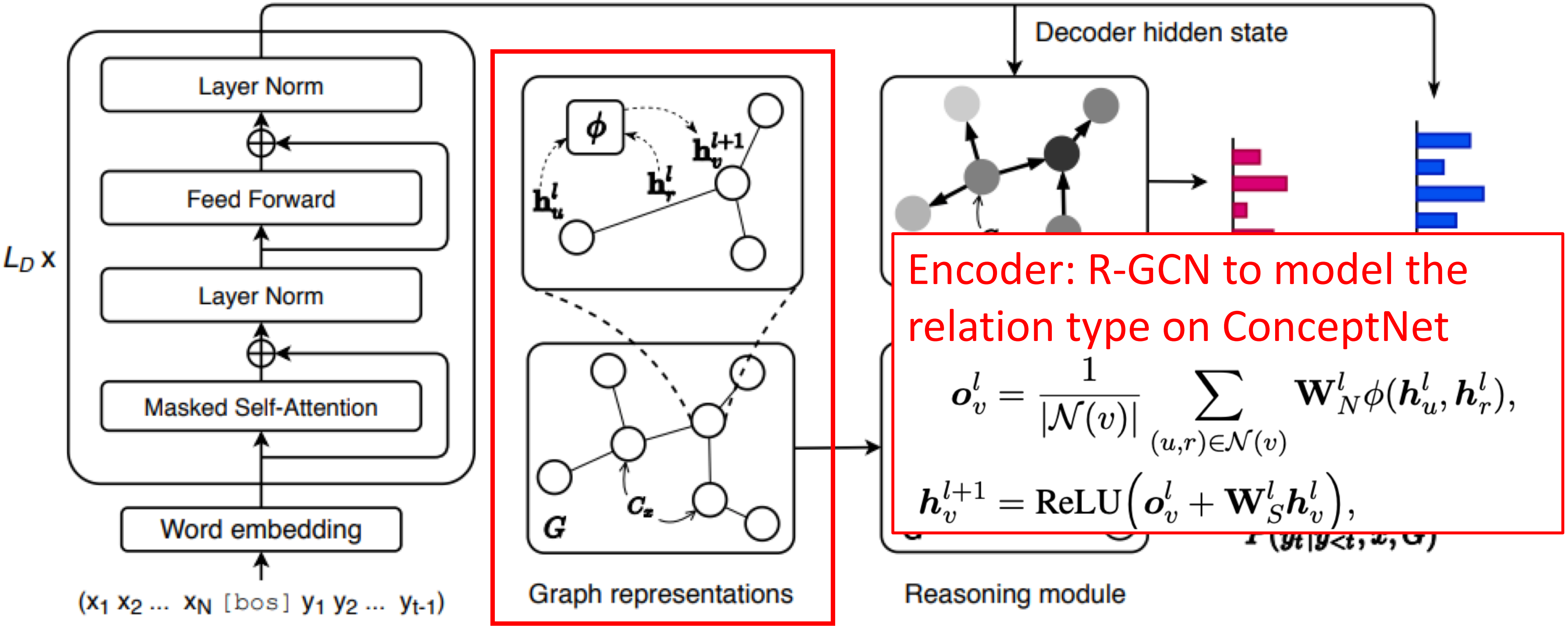
KG-enhanced text generation methods

- Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, In EMNLP 2020



KG-enhanced text generation methods

- Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, In EMNLP 2020



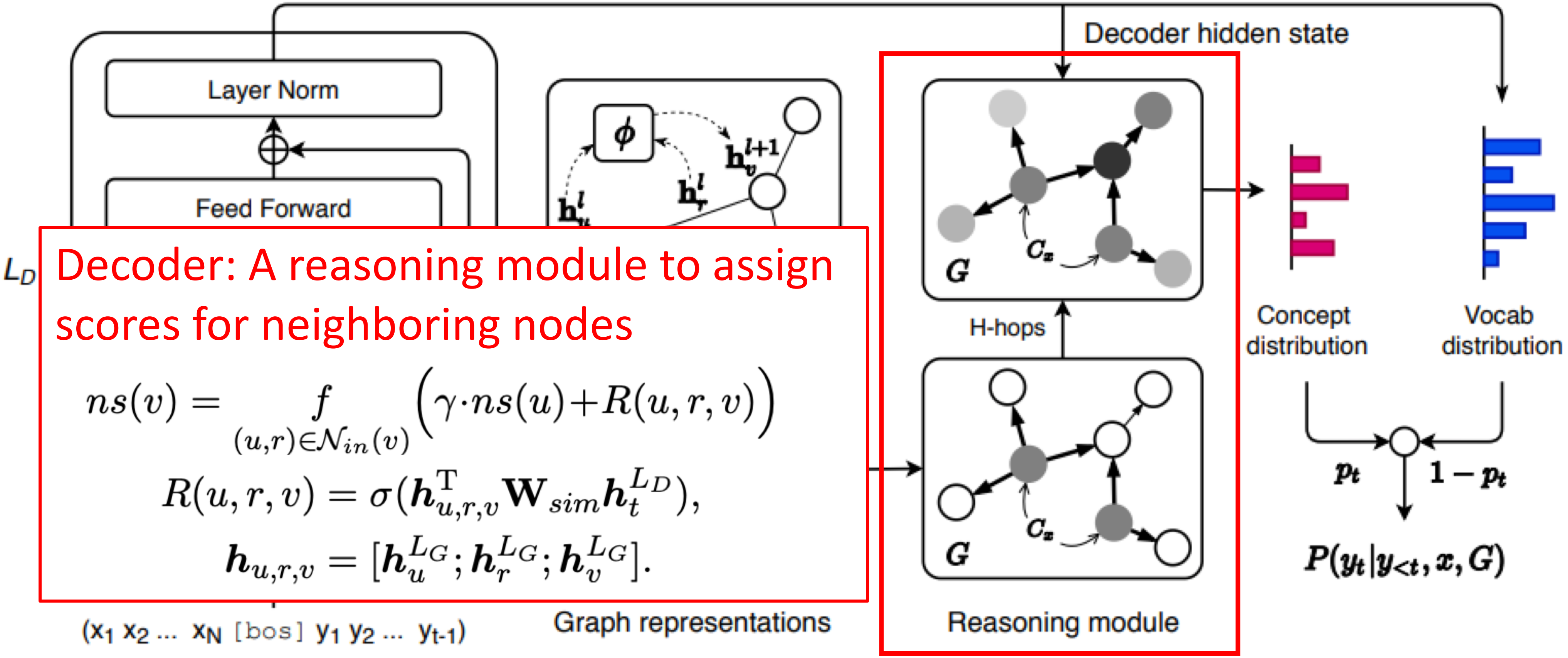
Encoder: R-GCN to model the relation type on ConceptNet

$$o_v^l = \frac{1}{|\mathcal{N}(v)|} \sum_{(u,r)\in\mathcal{N}(v)} \mathbf{W}_N^l \phi(h_u^l, h_r^l),$$

$$h_v^{l+1} = \mathrm{ReLU}\left(o_v^l + \mathbf{W}_S^l h_v^l\right),$$

KG-enhanced text generation methods

- Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, In EMNLP 2020



Decoder: A reasoning module to assign scores for neighboring nodes

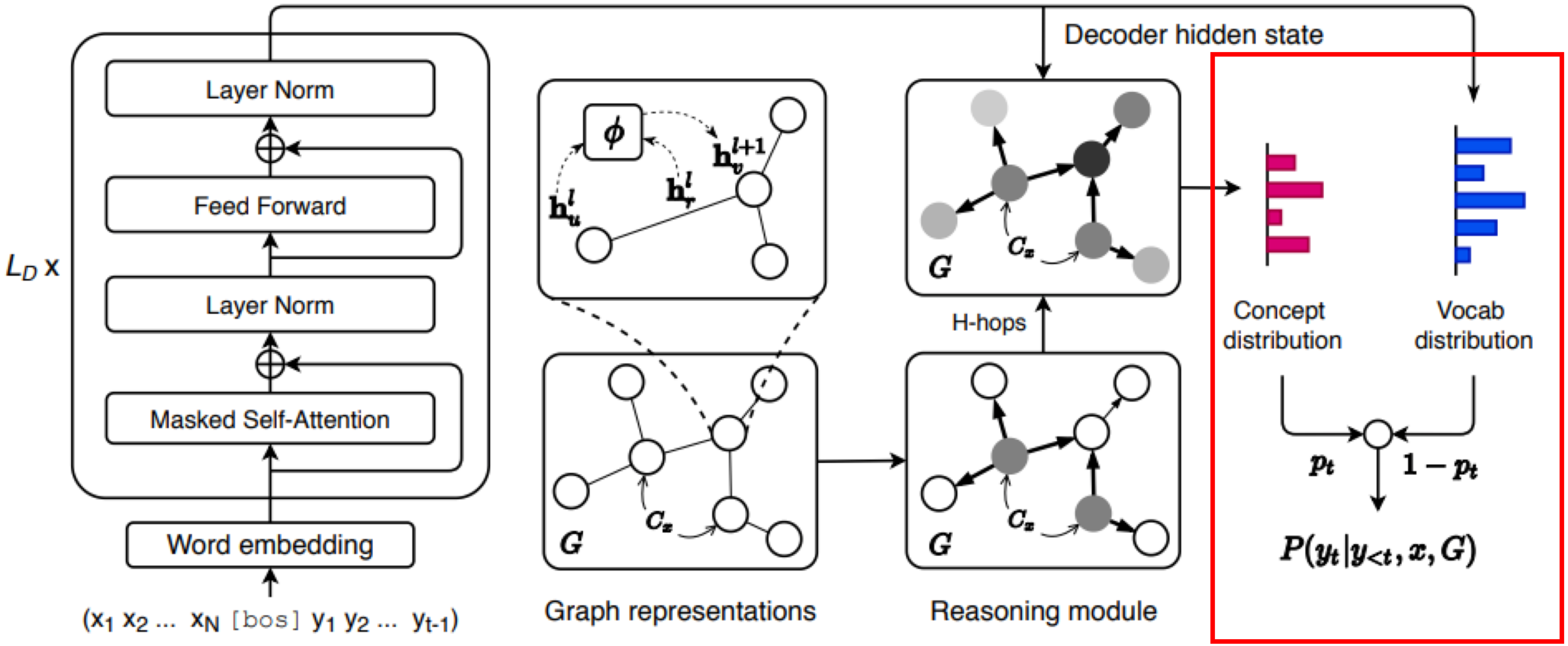
$$ns(v) = \underset{(u,r)\in\mathcal{N}_{in}(v)}{f}\Big(\gamma \cdot ns(u) + R(u,r,v)\Big)$$

$$R(u,r,v) = \sigma(\boldsymbol{h}_{u,r,v}^{\mathrm{T}} \mathbf{W}_{sim} \boldsymbol{h}_t^{L_D}),$$

$$\boldsymbol{h}_{u,r,v} = [\boldsymbol{h}_u^{L_G}; \boldsymbol{h}_r^{L_G}; \boldsymbol{h}_v^{L_G}].$$

KG-enhanced text generation methods

- Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, In EMNLP 2020



KG-enhanced text generation methods

- Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, In EMNLP 2020
- Dataset: ROCStories, alpha-NLG, EG.   Metric: BLEU, METEOR, ROUGE

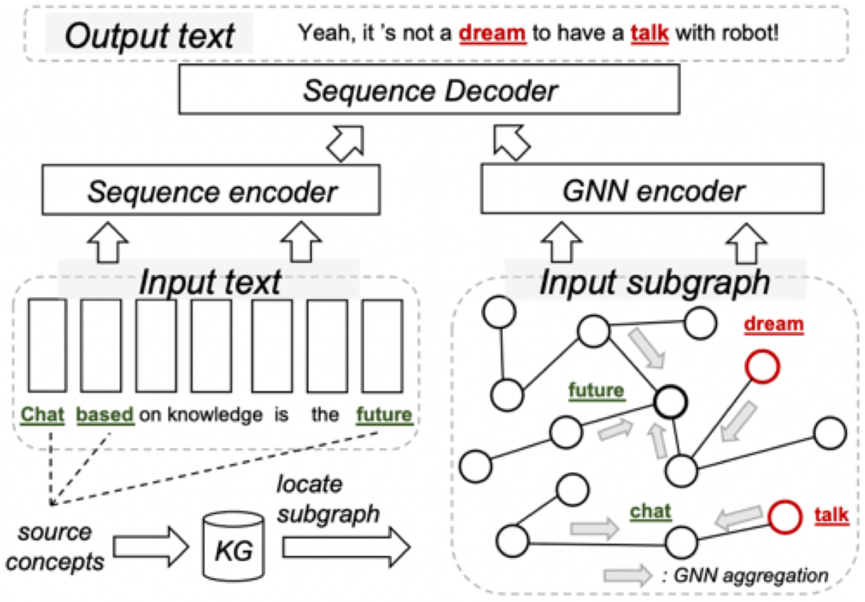
Models	EG				αNLG			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
Seq2Seq	6.09	24.94	26.37	32.37	2.37	14.76	22.03	29.09
COMeT-Txt-GPT2	N/A	N/A	N/A	N/A	2.73†	18.32†	24.39†	32.78†
COMeT-Emb-GPT2	N/A	N/A	N/A	N/A	3.66†	19.53†	24.92†	32.67†
GPT2-FT	15.63	38.76	37.32	77.09	9.80	25.82	32.90	57.52
GPT2-OMCS-FT	15.55	38.28	37.53	75.60	9.62	25.83	32.88	57.50
GRF	**17.19**	**39.15**	**38.10**	**81.71**	**11.62**	**27.76**	**34.62**	**63.76**

Table 3: Automatic evaluation results on the test set of EG and αNLG. Entries with N/A mean the baseline is not designated for this task. †: we use the generation results from Bhagavatula et al. (2020).

KG-enhanced text generation methods

- **M4: Improve the Graph Embeddings with Graph Neural Networks.**

- KG definition: A KG is defined as $\mathcal{G} = (\mathcal{U}, \mathcal{E}, \mathcal{R})$, where $\mathcal{U}$ is the set of entity nodes and $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{R} \times \mathcal{U}$ is the set of typed edges between nodes in $\mathcal{U}$ with a certain relation in the relation schema $\mathcal{R}$.



- Graph neural network (GNN):

$$\mathbf{u}^{(k)} = \textsc{Combine}_k\left(\mathbf{u}^{(k-1)}, \textsc{Aggregate}_k\left(\left\{\left(\mathbf{u}_i^{(k-1)}, \mathbf{e}_{ij}^{(k-1)}, \mathbf{u}_j^{(k-1)}\right) : \forall (u_i, e_{ij}, u_j) \in \mathcal{N}(u)\right\}\right)\right),$$

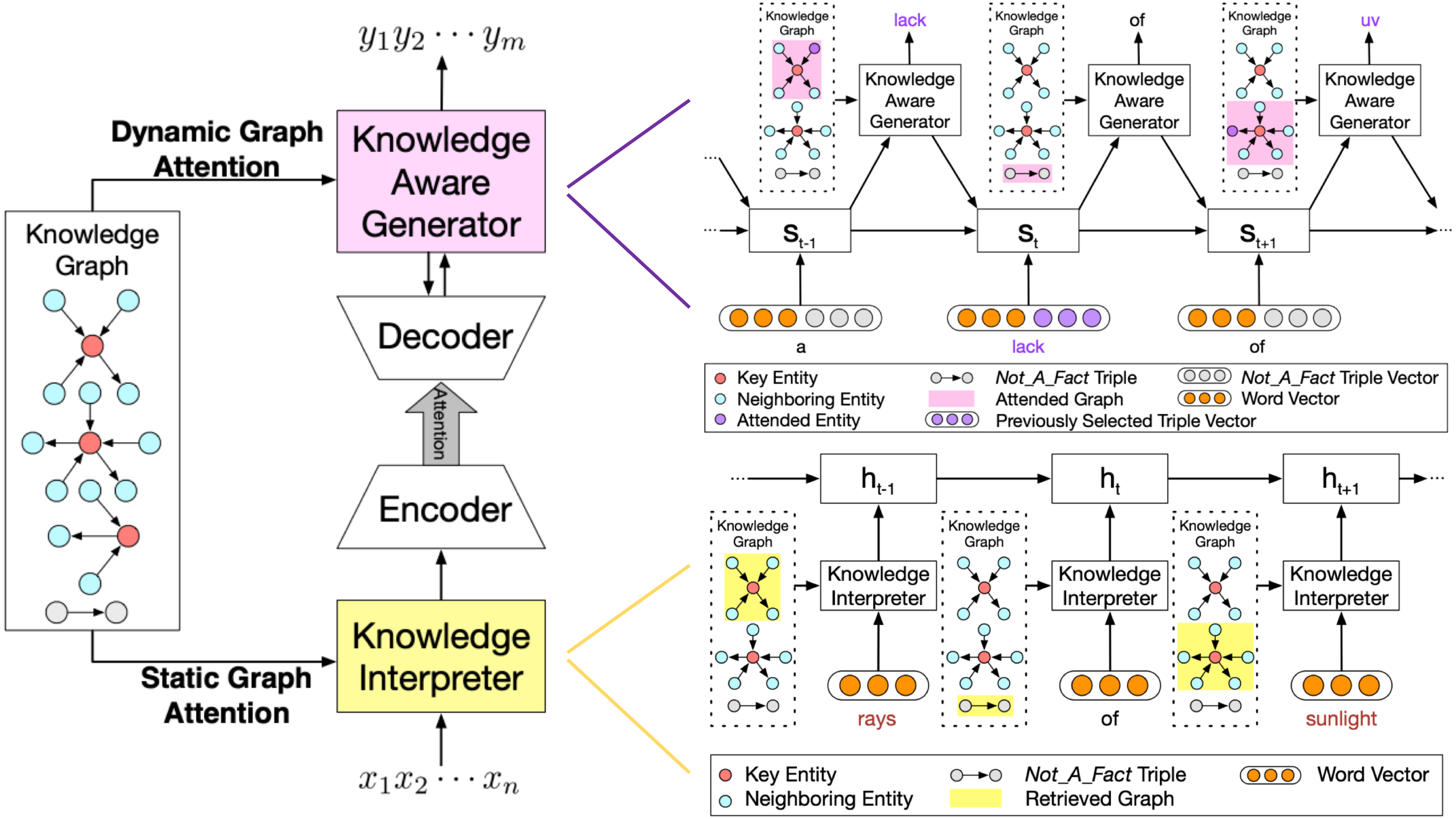
$$\mathbf{h}_G = \textsc{Readout}\left(\left\{\mathbf{u}^{(K)} : u \in \mathcal{U}\right\}\right).$$

KG-enhanced text generation methods

- Commonsense Knowledge Aware Conversation Generation with Graph Attention, In IJCAI 2018
- Application: Dialogue system

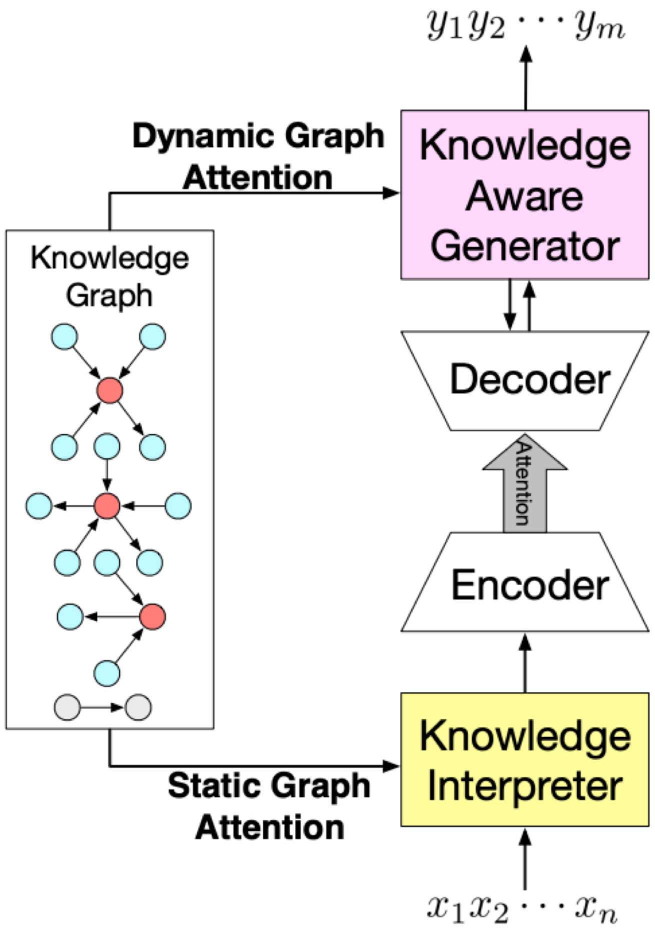
KG-enhanced text generation methods

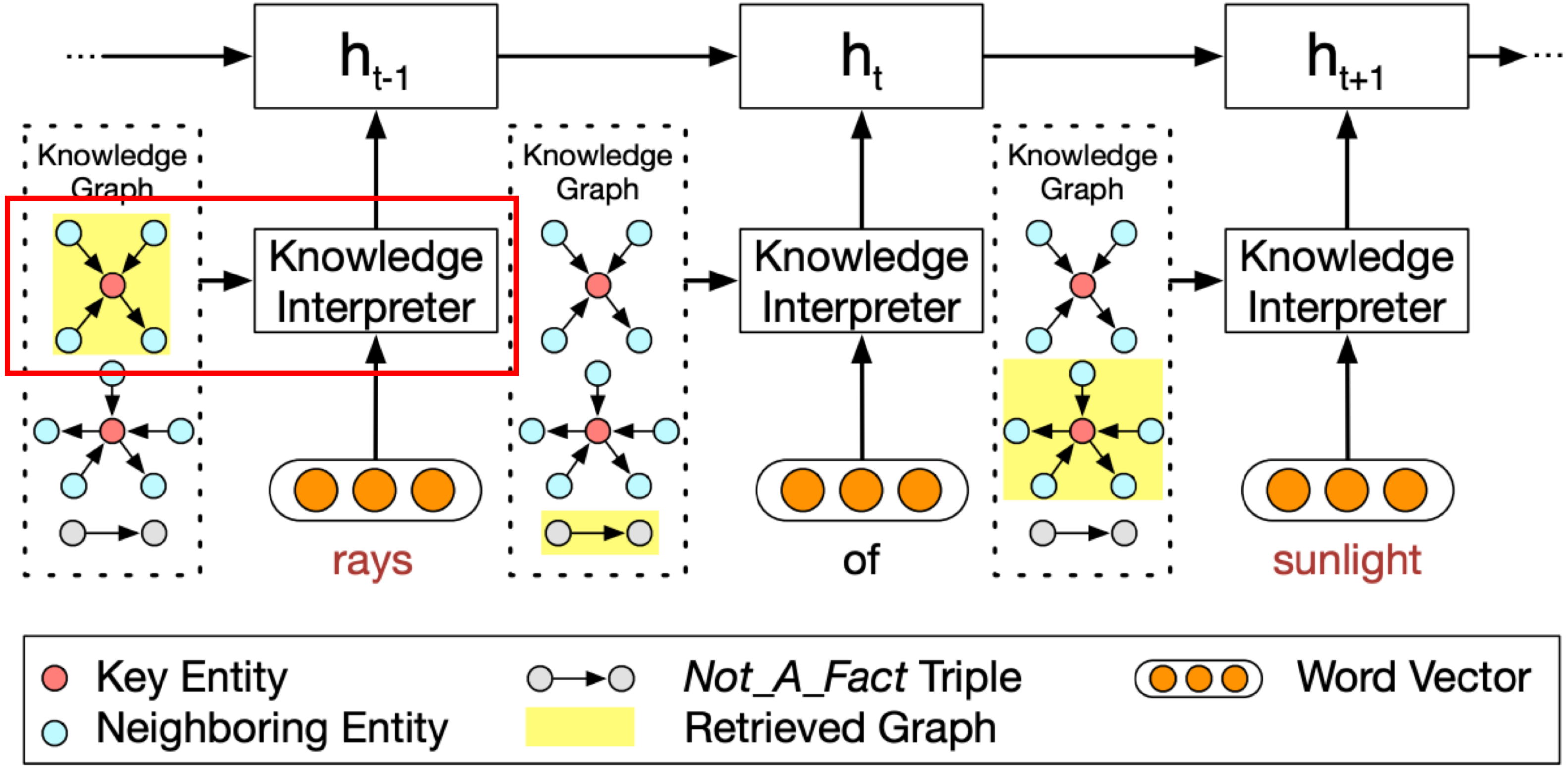
- Commonsense Knowledge Aware Conversation Generation with Graph Attention, In IJCAI 2018



KG-enhanced text generation methods

- Commonsense Knowledge Aware Conversation Generation with Graph Attention, In IJCAI 2018



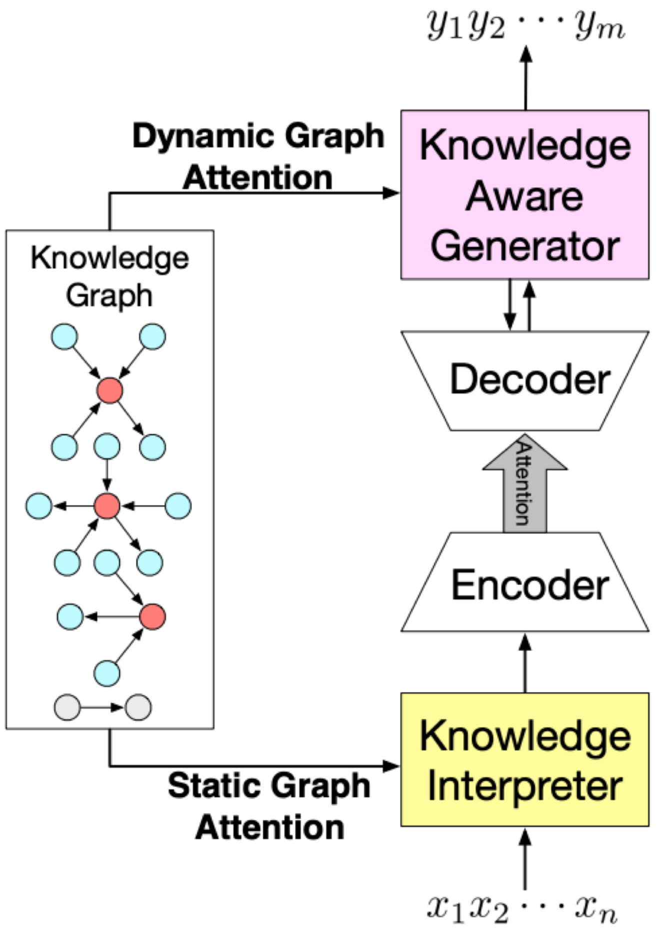


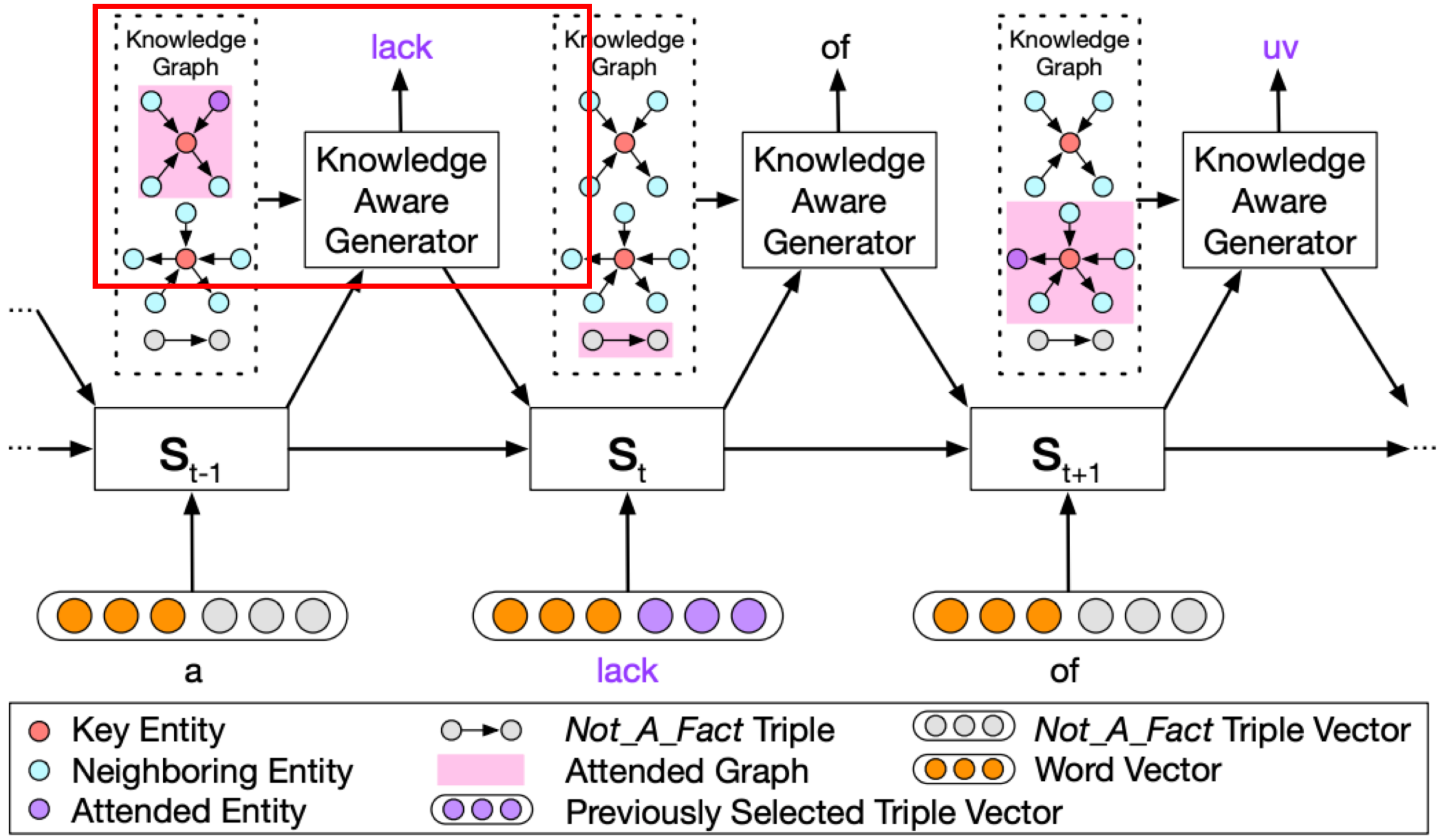
Encoder: word embedding from text + entity embedding from KG

* Only one-hop relation is used in this work

KG-enhanced text generation methods

- Commonsense Knowledge Aware Conversation Generation with Graph Attention, In IJCAI 2018





Decoder: encoder-decoder attention + knowledge graph attention

KG-enhanced text generation methods

- Commonsense Knowledge Aware Conversation Generation with Graph Attention, In IJCAI 2018
- Dataset: Reddit-1M + ConceptNet   Metric: Perplexity ↓; Entropy ↑

Model	Overall		High Freq.		Medium Freq.		Low Freq.		OOV	
	ppx.	ent.	ppx.	ent.	ppx.	ent.	ppx.	ent.	ppx.	ent.
Seq2Seq	47.02	0.717	42.41	0.713	47.25	0.740	48.61	0.721	49.96	0.669
MemNet	46.85	0.761	41.93	0.764	47.32	0.788	48.86	0.760	49.52	0.706
CopyNet	40.27	0.96	36.26	0.91	40.99	0.97	42.09	0.96	42.24	0.96
CCM	**39.18**	**1.180**	**35.36**	**1.156**	**39.64**	**1.191**	**40.67**	**1.196**	**40.87**	**1.162**

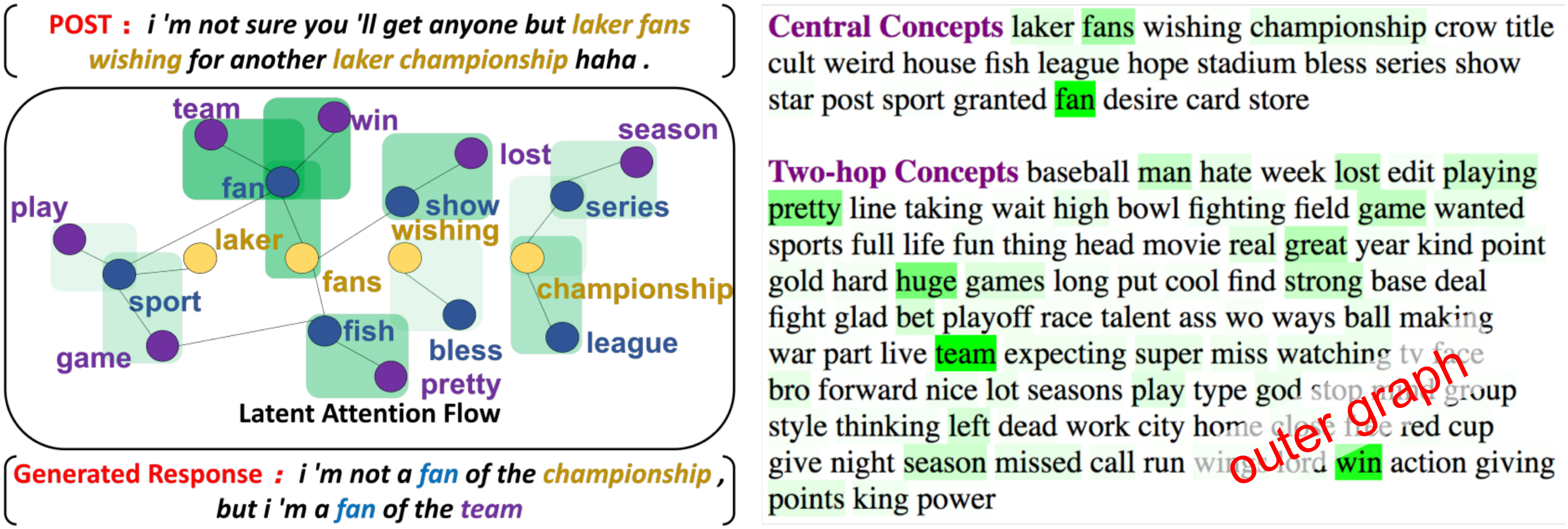
Table 2: Automatic evaluation with *perplexity* (ppx.), and *entity score* (ent.).

Model	Overall		High Freq.		Medium Freq.		Low Freq.		OOV	
	app.	inf.	app.	inf.	app.	inf.	app.	inf.	app.	inf.
CCM vs. Seq2Seq	0.616	0.662	0.605	0.656	0.549	0.624	0.636	0.650	0.673	0.716
CCM vs. MemNet	0.602	0.647	0.593	0.656	0.566	0.640	0.622	0.635	0.626	0.657
CCM vs. CopyNet	0.600	0.640	0.606	0.669	0.586	0.619	0.610	0.633	0.596	0.640

Table 3: Manual evaluation with *appropriateness* (app.), and *informativeness* (inf.). The score is the percentage that CCM wins its competitor after removing "Tie" pairs. CCM is significantly better (sign test, p-value < 0.005) than all the baselines on all the test sets.

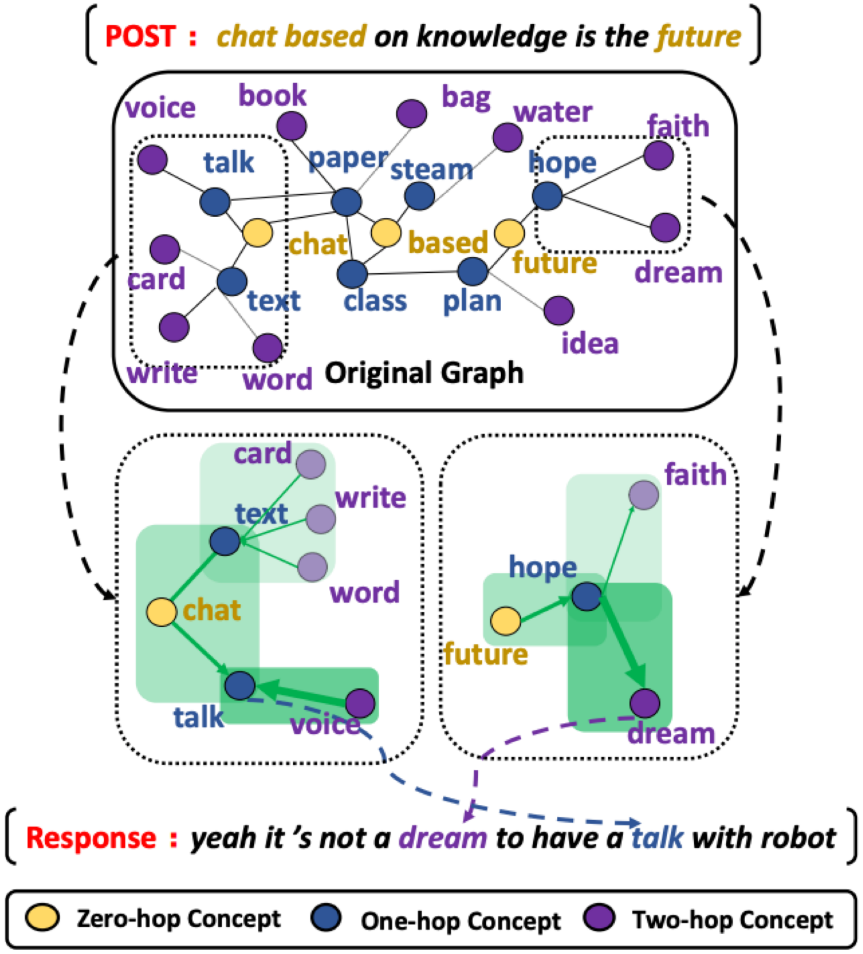
KG-enhanced text generation methods

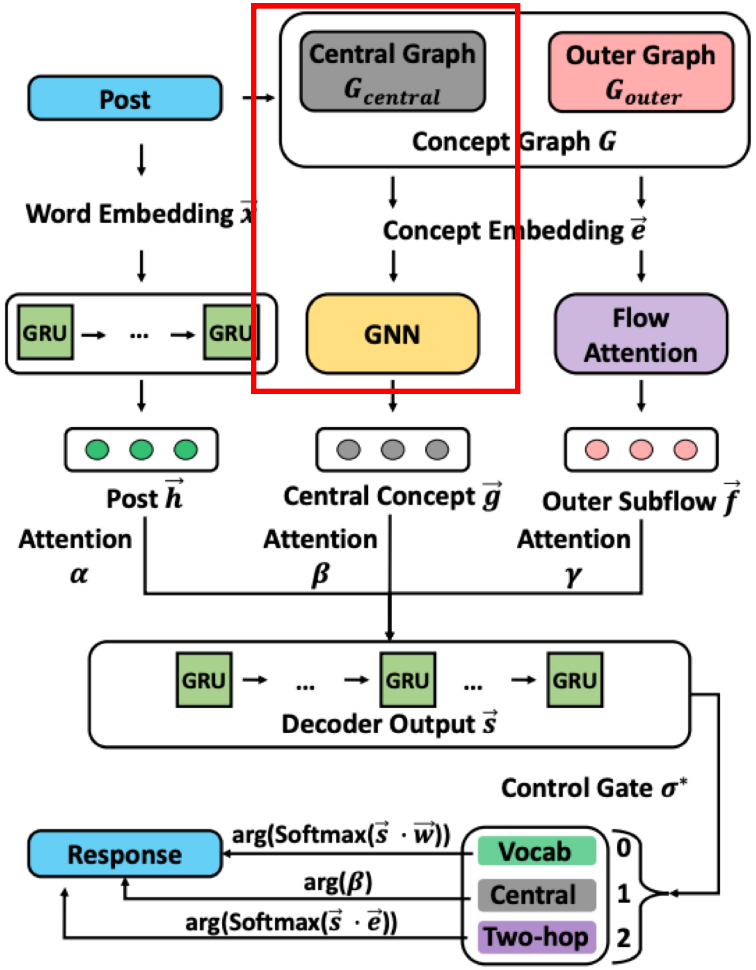
- Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs, In ACL 2020.
- Application: Dialogue system
- Motivation: Concept shift in human conversations has not been modeled.



KG-enhanced text generation methods

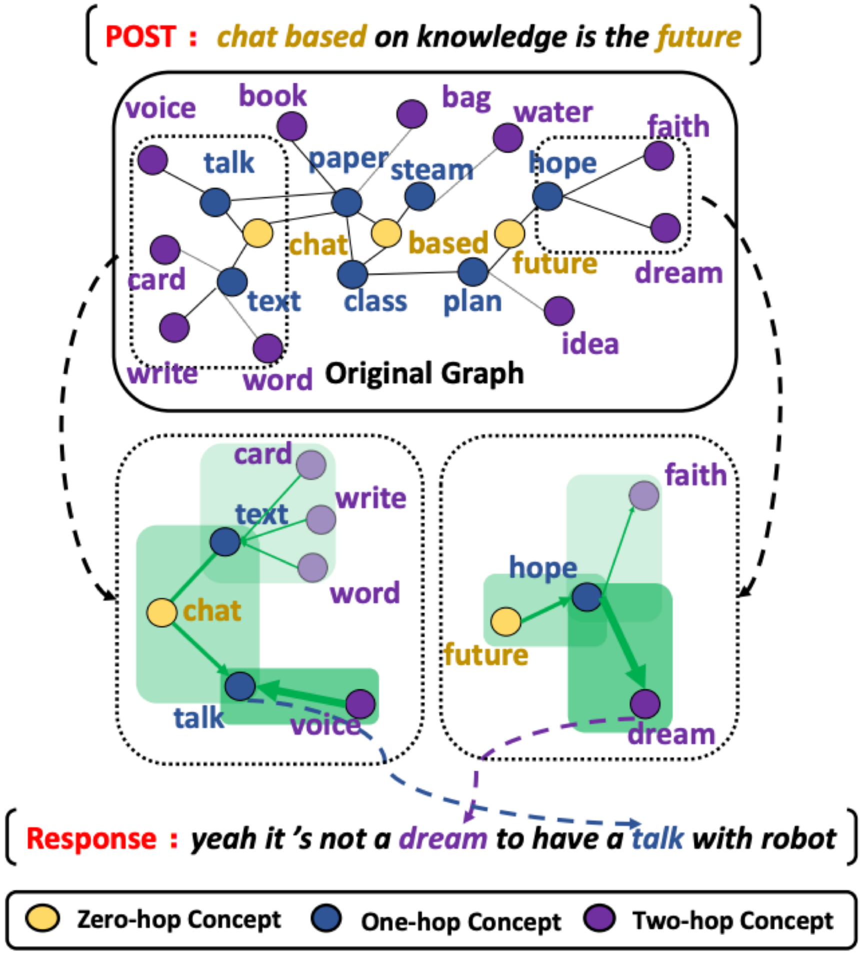
- Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs, In ACL 2020

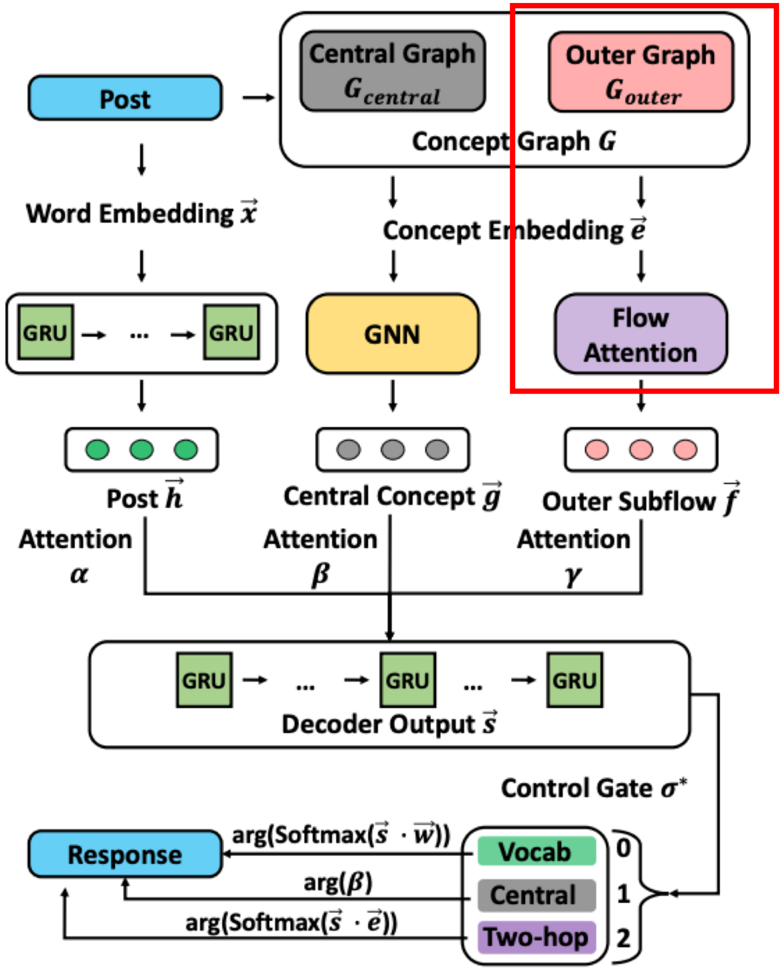




KG-enhanced text generation methods

- Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs, In ACL 2020





KG-enhanced text generation methods

- Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs, In ACL 2020
- Dataset: Reddit-1M + ConceptNet; Metric: BLEU; Nist; ROUGE

Model	Bleu-4	Nist-4	Rouge-1	Rouge-2	Rouge-L	Meteor	PPL
Seq2Seq	0.0098	1.1069	0.1441	0.0189	0.1146	0.0611	48.79
MemNet	0.0112	1.1977	0.1523	0.0215	0.1213	0.0632	47.38
CopyNet	0.0106	1.0788	0.1472	0.0211	0.1153	0.0610	43.28
CCM	0.0084	0.9095	0.1538	0.0211	0.1245	0.0630	42.91
GPT-2 (lang)	0.0162	1.0844	0.1321	0.0117	0.1046	0.0637	29.08*
GPT-2 (conv)	0.0124	1.1763	0.1514	0.0222	0.1212	0.0629	24.55*
ConceptFlow	**0.0246**	**1.8329**	**0.2280**	**0.0469**	**0.1888**	**0.0942**	**29.90**

Table: Relevance Between Generated and Golden Responses.

KG-enhanced text generation methods

Tasks	Methods	Cat.	Dataset Information		Effect of KG			KG source
			Name	#Instance	w/o KG	with KG	ΔBLEU	
Common-sense reasoning	KG-BART	M4	CommonGen	77,449	28.60	30.90	+2.30	ConceptNet
	CE-PR	M3	ComVE	30,000	15.70	17.10	+1.60	ConceptNet
	GRF	M4	αNLG-ART	60,709	9.62	11.62	+2.00	ConceptNet
	MGCN	M3	EntDesc	110,814	24.90	30.00	+4.30	Self-built KG
Story generation	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	KEPM	M2	ROCStories (split-2)	98,162	14.10	14.30	+0.20	ConceptNet & ATOMIC
	MRG	M3	VisualStory	50,000	3.18	3.23	+0.05	ConceptNet
Scientific writing	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Self-built KG
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Self-built KG
Dialogue system	ConceptFlow	M4	Reddit-10M	3,384K	1.62	2.46	+0.84	ConceptNet
	AKGCM	M3	EMNLP dialog	43,192	32.45	30.84	-1.61	Self-built KG
	AKGCM	M3	ICLR dialog	21,569	6.74	6.94	+0.20	Self-built KG
Question answering	MHPGM	M3	NarrativeQA	46,765	19.79	21.07	+1.28	Self-built KG

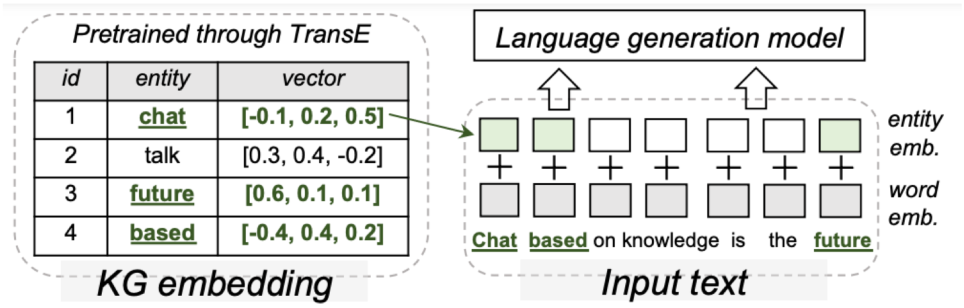
Observation 1: KG makes largest improvement on commonsense reasoning tasks

Observation 2: ConceptNet is the most popular used KG.

Table: Tasks, datasets and KG sources used in different KG-enhanced papers.

Dataset and code links: https://github.com/wyu97/KENLG-Reading

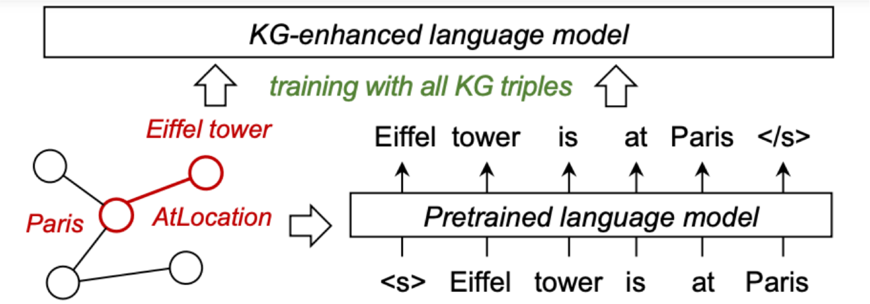
KG-enhanced text generation methods



(M1) Incorporate KGE into language generation

M1 (Incorporate Knowledge Graph Embeddings into Language Generation):

- Pros: (i) Easy to use (by simple vector concatenation)
- Cons: (i) Text representation and KG representation are from two vector space
  (ii) KGE can only capture one-hop relations

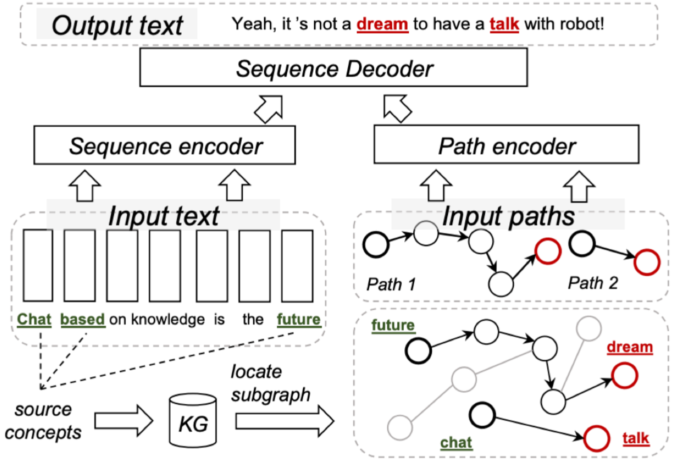


(M2) Transfer knowledge into pretrained LM

M2 (Transfer Knowledge into Language Model with Knowledge Triplet Information):

- Pros: (i) Easy to use (train with any pre-trained LM)
  (ii) KG knowledge is embedded into LMs
- Cons: (i) Only capture one-hop relations

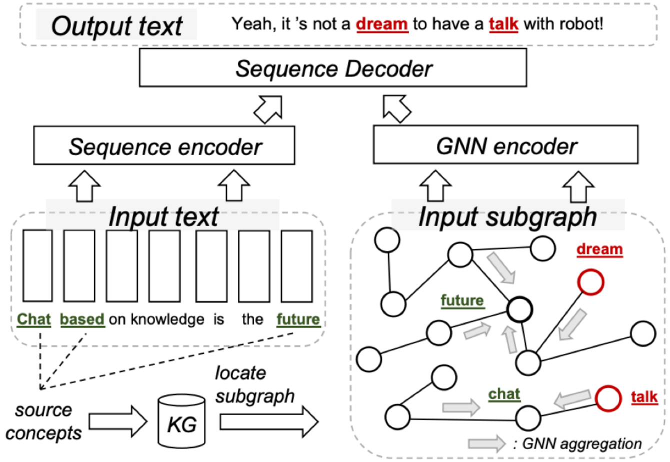
KG-enhanced text generation methods



(M3) Performing path reasoning on KG

M3 (Perform Reasoning over Knowledge Graph via Path Finding Strategies.):

- Pros: (i) Multi-hop reasoning
  (ii) Better interpretability
- Cons: (i) Only one path is considered
  (ii) Large complexity and hard to train

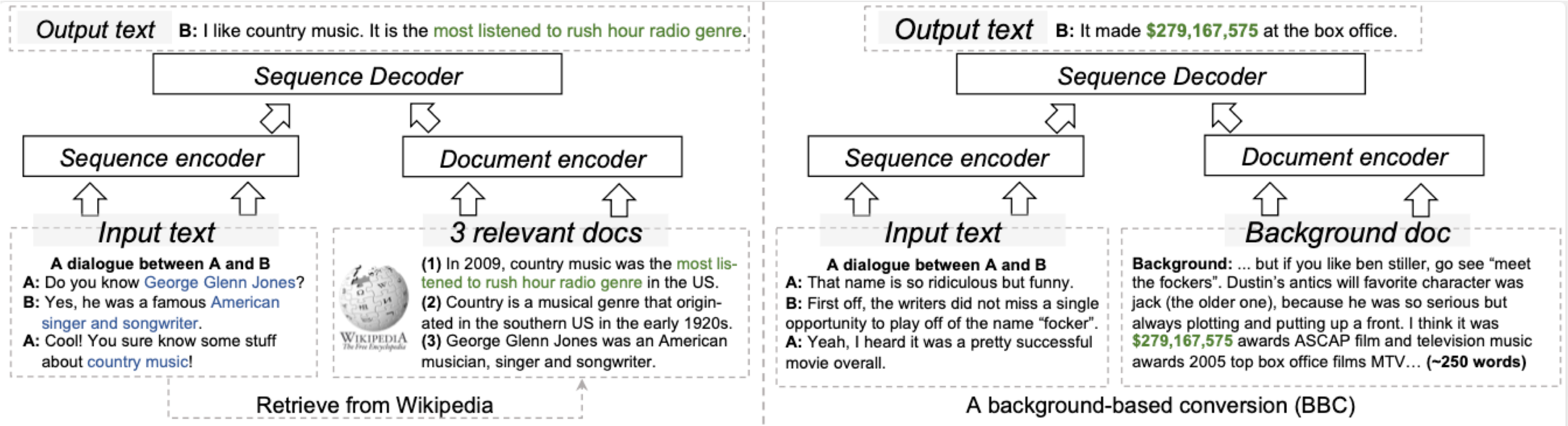


(M4) Aggregating sub-KG via GNN

M4 (Improve the Graph Embeddings with Graph Neural Networks):

- Pros: (i) Multi-hop relations
  (ii) Joint optimization of Seq2Seq and GNN
- Cons: (i) High computation cost

Grounded text-enhanced NLG methods



(a) M1: Retrieve relevant documents, use them for generation

(b) M2: Read background document and generate output

M1: Retrieval-augmented NLG

- [Lewis et al. 2020 Neurips]
- [Wang et al. 2021 ACL]

M2: Background-based NLG

- [Qin et al. 2019 ACL]
- [Meng et al. 2020 AAAI]

Grounded text-enhanced NLG methods

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, In Neruips 2020
- Motivation: Large pre-trained LMs cannot easily expand or revise their memory, can't straightforwardly provide insight into their predictions, and may produce "hallucinations".

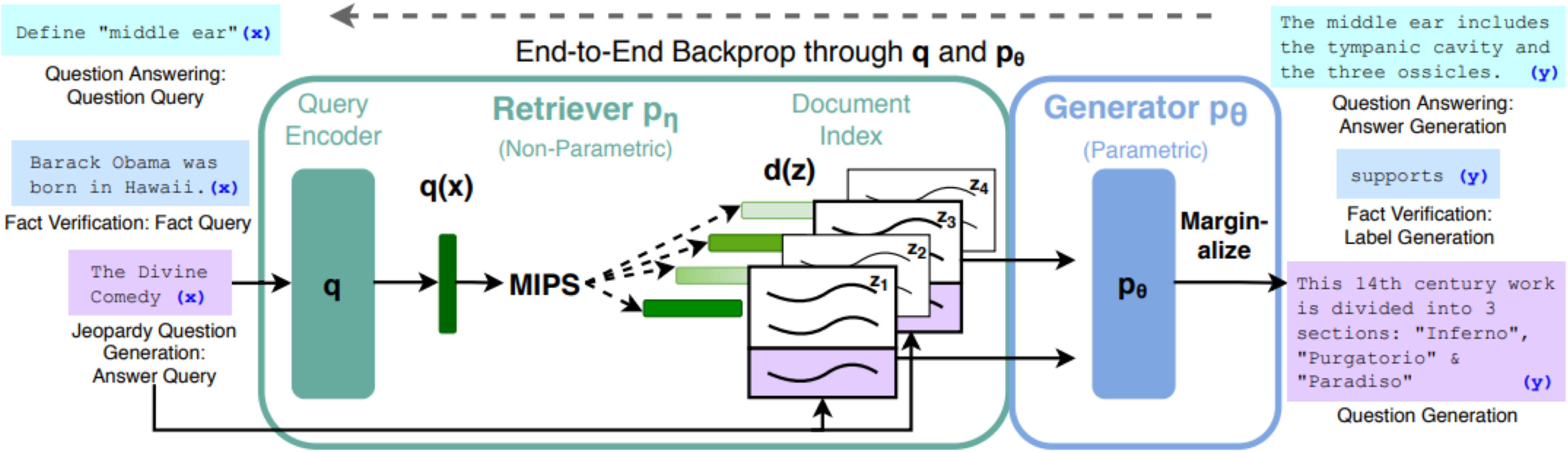


Figure: RAG combines a pre-trained retriever (DPR) with a pre-trained seq2seq model (BART) and fine-tune end-to-end.

Grounded text-enhanced NLG methods

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, In Neruips 2020
- Dataset: Trivial QA, MS-MARCO   Metric: Exact match (for ODQA); BLEU, ROUGE

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set. See Appendix D for further details.

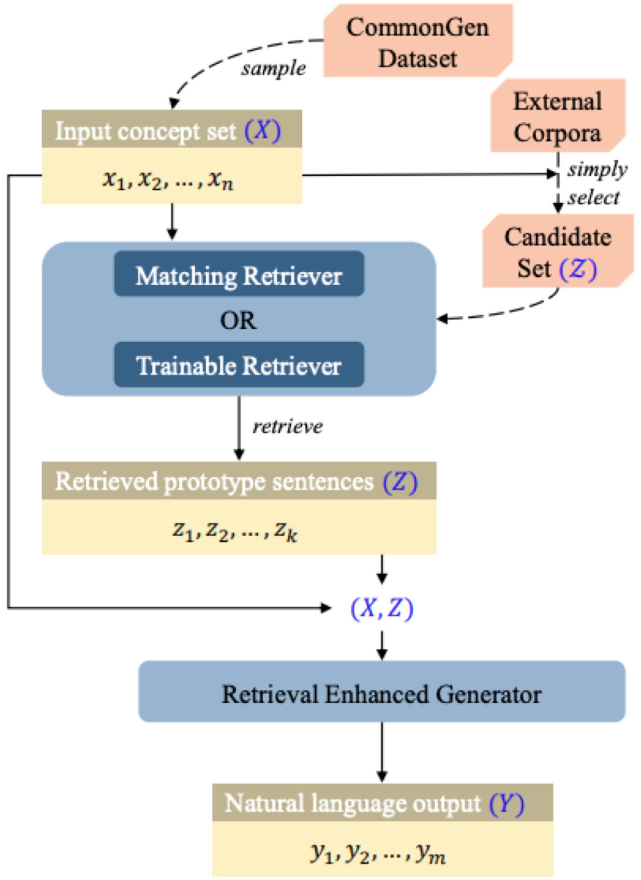
	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] *Uses gold context/evidence. Best model without gold access underlined.

Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label Acc.	FVR2 Label Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	40.1	41.5	72.5	89.5
RAG-Seq.	14.7	21.4	40.8	44.2		

Grounded text-enhanced NLG methods

- Retrieval Enhanced Model for Commonsense Generation, In ACL 2021
- Motivation: It is challenging to organize provided concepts into the most plausible scenario, avoid violation of commonsense.



Grounded text-enhanced NLG methods

- Retrieval Enhanced Model for Commonsense Generation, In ACL 2021
- Task: CommonGen   Metric: BLEU, CIDEr, SPICE

Model	BLEU-4	CIDEr	**SPICE**	SPICE(v1.0)
GPT-2 (Radford et al., 2019)	26.833	12.187	23.567	25.90
BERT-Gen (Bao et al., 2020)	23.468	12.606	24.822	27.30
UniLM (Dong et al., 2019)	30.616	14.889	27.429	30.20
BART (Lewis et al., 2020)	31.827	13.976	27.995	30.60
T5-base (Raffel et al., 2020)	18.546	9.399	19.871	22.00
T5-large (Raffel et al., 2020)	31.962	15.128	28.855	31.60
EKI-BART (Fan et al., 2020)	35.945	16.999	29.583	32.40
KG-BART (Liu et al., 2021)	33.867	16.927	29.634	32.70
CALM(T5-base) (Zhou et al., 2021)	-	-	-	33.00
RE-T5 (ours)	**40.863**	**17.663**	**31.079**	**34.30**

Table 2: Test results on CommonGen benchmark. All results except CALM are based on the latest human references(v1.1). v1.0 indicates evaluation with old evaluation protocol.[2]

Grounded text-enhanced NLG methods

- Retrieval Enhanced Model for Commonsense Generation, In ACL 2021
- Task: CommonGen   Metric: BLEU, CIDEr, SPICE

Concept Set:
trailer shirt side sit road

T5:
A man sits on the side of a trailer and a shirt.

Trainable Retriever:
(1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer.
(2)Teenagers in matching shirts stand at the side of the road holding trash bags.
(3)A man in a white shirt and black pants standing at the side or the road.
RE-T5(trainable retriever):
a man in a white shirt and black pants sits on the side of a trailer on the road.

Figure: An example of sentences generated based on the retrieved sentences.

Grounded text-enhanced NLG methods

- Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading, In ACL 2019
- Task: Dialogue system

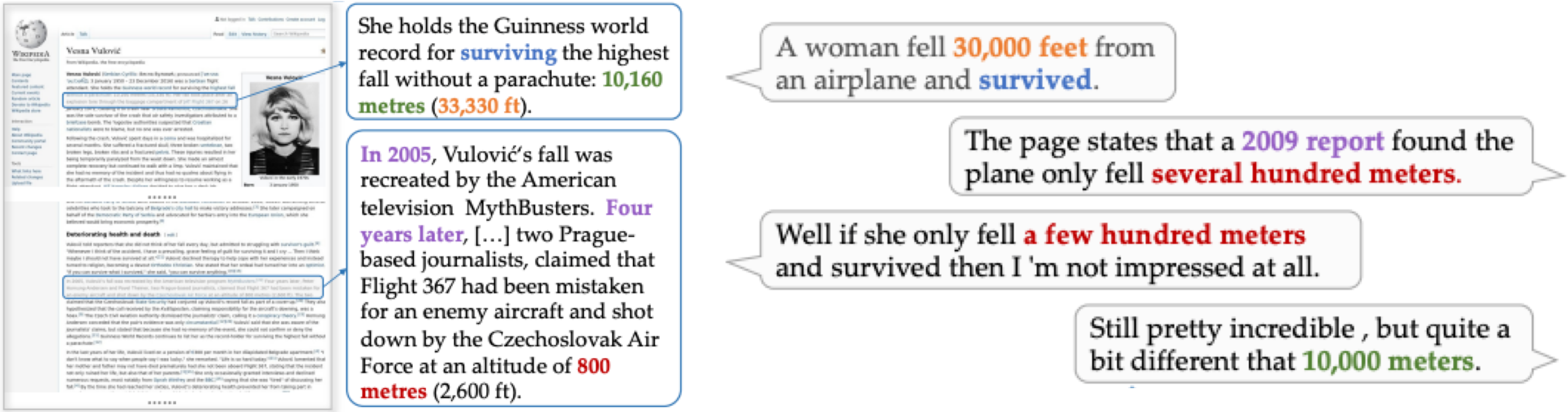


Figure: Users discussing a topic defined by a Wikipedia article. In this real-world example from our Reddit dataset, information needed to ground responses is distributed throughout the source document.

Grounded text-enhanced NLG methods

- Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading, In ACL 2019

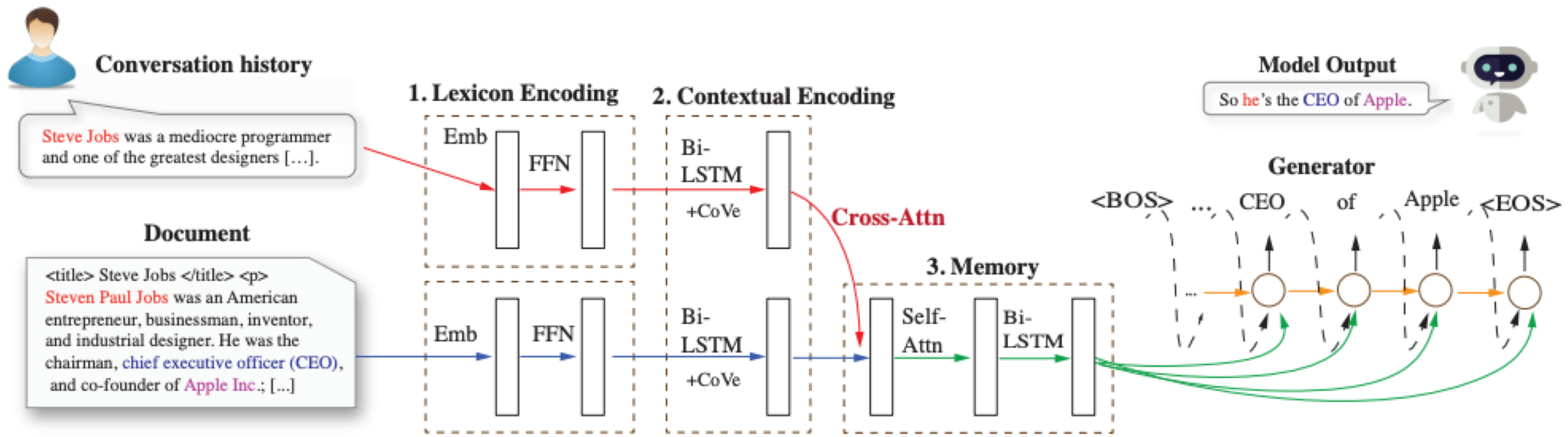


Figure: Model Architecture for Response Generation with on-demand Machine Reading

Grounded text-enhanced NLG methods

- Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading, In ACL 2019
- Dataset: Reddit Metric: NIST; BLEU; F1; Distinct-k …

	Appropriateness			Grounding			Diversity			
	NIST	BLEU	METEOR	Precision	Recall	F1	Entropy-4	Distinct-1	Distinct-2	Len
Human	2.650	3.13%	8.31%	2.89%	0.45%	0.78%	10.445	0.167	0.670	18.757
SEQ2SEQ	2.223	1.09%	7.34%	1.20%	0.05%	0.10%	9.745	0.023	0.174	15.942
MEMNET	2.185	1.10%	7.31%	1.25%	0.06%	0.12%	9.821	0.035	0.226	15.524
CMR-F	**2.260**	1.20%	7.37%	1.68%	0.08%	0.15%	9.778	0.035	0.219	15.471
CMR	2.213	**1.43%**	7.33%	2.44%	0.13%	0.25%	9.818	0.046	0.258	15.048
CMR+W	2.238	1.38%	**7.46%**	**3.39%**	**0.20%**	**0.38%**	**9.887**	**0.052**	**0.283**	15.249

Table: Automatic Evaluation results on Reddit dataset.

Grounded text-enhanced NLG methods

Evidence sources	Tasks	Methods	Dataset Information		Retrieval space (d/s)	# Retri-eved d/s
			Name	#Instance		
Wikipedia	Dialogue system	MemNet	Wizard of Wikipedia (WoW)	22,311	5.4M/93M	7
		SKT				7
	Question answering	RAG	MS-MARCO	267,287	21M/-	10
		BART+DPR	ELI5	274,741	3.2M/-	[illegible]
		RT+C-REALM			3.2M/-	7
	Argument generation	H&W	ChangeMyView	287,152	5M/-	10
		CANDELA			5M/-	10
Online platform (e.g., Amazon)	Dialogue (for business)	AT2T	Amazon books	937,032	-/131K	10
		KGNCM	Foursquare	1M	-/1.1M	10
Gigawords	Summari-zation	R^{3}Sum	Gigawords	3.8M	-/3.8M	30
		BiSET			-/3.8M	30

challenging

Table: Tasks, datasets and evidence sources used in retrieve-then-generate papers.