Commonsense Knowledge Aware Conversation Generation with Graph Attention

IJCAI 2018

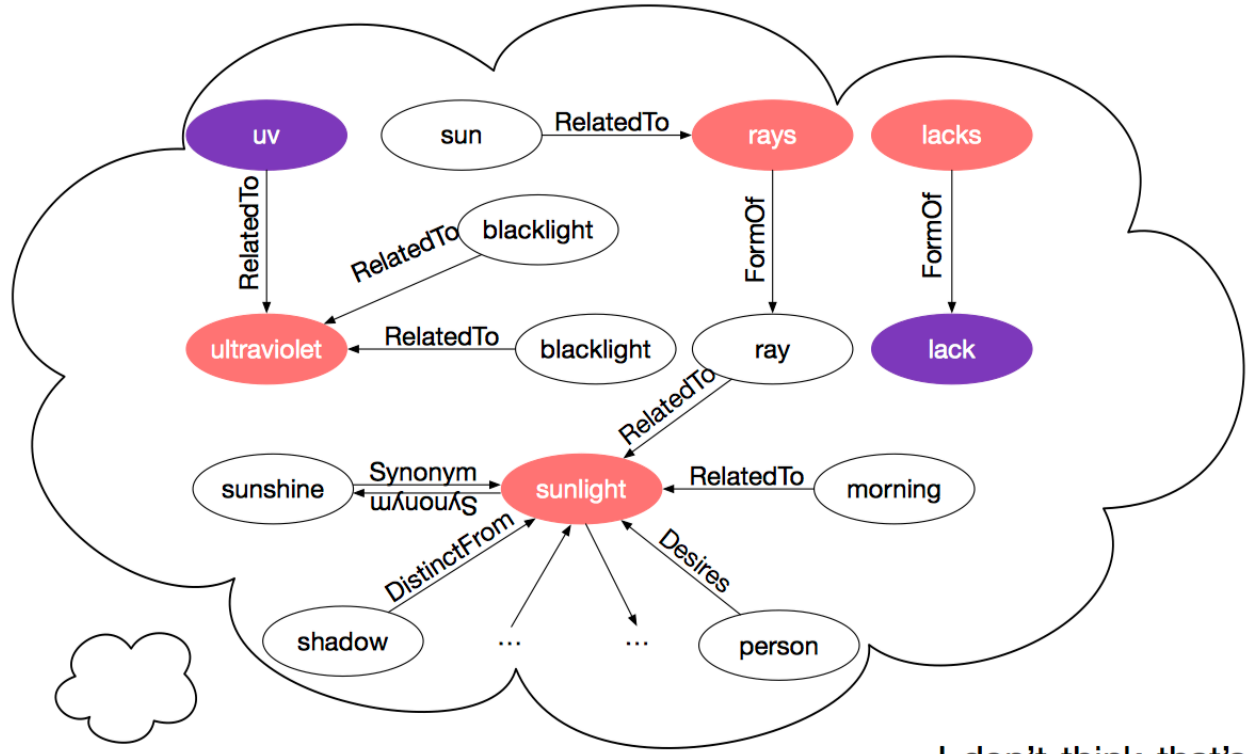
Hao Zhou, Tom Young, **Minlie Huang**, Haizhou Zhao, Jingfang Xu, Xiaoyan Zhu

Xiachong Feng

Outline

1. Overview
2. Problem
3. Motivation
4. Encoder-Decoder
5. Task Definition
6. Model-CCM
7. Experiment
8. Conclusion

Overview



Moonlight lacks the ultraviolet rays of sunlight. → I don't think that's a lack of uv.

Moonlight lacks the ultraviolet rays of sunlight. → I'm not sure what you're saying.

with/without considering commonsense knowledge

Problem

- **Conversational data**
  - Tend to generate **generic responses**, which are unable to respond **appropriately** and **informatively**
- **Unstructured texts or domain-specific knowledge triples**
  - highly dependent on the **quality of unstructured texts**
  - **limited** by the small-scale, domain-specific knowledge
  - make use of knowledge triples (entities) **separately** and **independently**, instead of treating knowledge triples as a whole in a graph

Motivation

- **Large-scale commonsense knowledge**
  - understand the **background information** of a given post
  - facilitate response generation with such knowledge
- **Attention mechanism**
  - **Static graph** attention mechanism ***(for encode)***
    - Encodes the **retrieved graphs** for a post to **augment** the semantic representation of the post
  - **Dynamic graph** attention mechanism ***(for decode)***
    - Reads the knowledge graphs and the triples in each graph for better **response generation**

Encoder-Decoder

- **Encoder**
  - Post: $X = x_1 x_2 \cdots x_n$
  - Hidden representations: $\boldsymbol{h}_t = \mathbf{GRU}(\boldsymbol{h}_{t-1}, \boldsymbol{e}(x_t)),$
- **Decoder**
  - Hidden representations: $\boldsymbol{s}_t = \mathbf{GRU}(\boldsymbol{s}_{t-1}, [\boldsymbol{c}_{t-1}; \boldsymbol{e}(y_{t-1})]),$
- **Probability distribution:**

$$\begin{aligned} y_t \sim \boldsymbol{o}_t &= P(y_t \mid y_{<t}, \boldsymbol{c}_t) \\ &= \mathrm{softmax}(\mathbf{W_o} \boldsymbol{s}_t). \end{aligned}$$

Task Definition

- **Given**
  - Post: $X = x_1 x_2 \cdots x_n$
  - Commonsense knowledge graphs: $G = \{g_1, g_2, \cdots, g_{N_G}\}$
    - *Each word corresponds to a graph in G*
    - *Each graph consists of a set of triples*

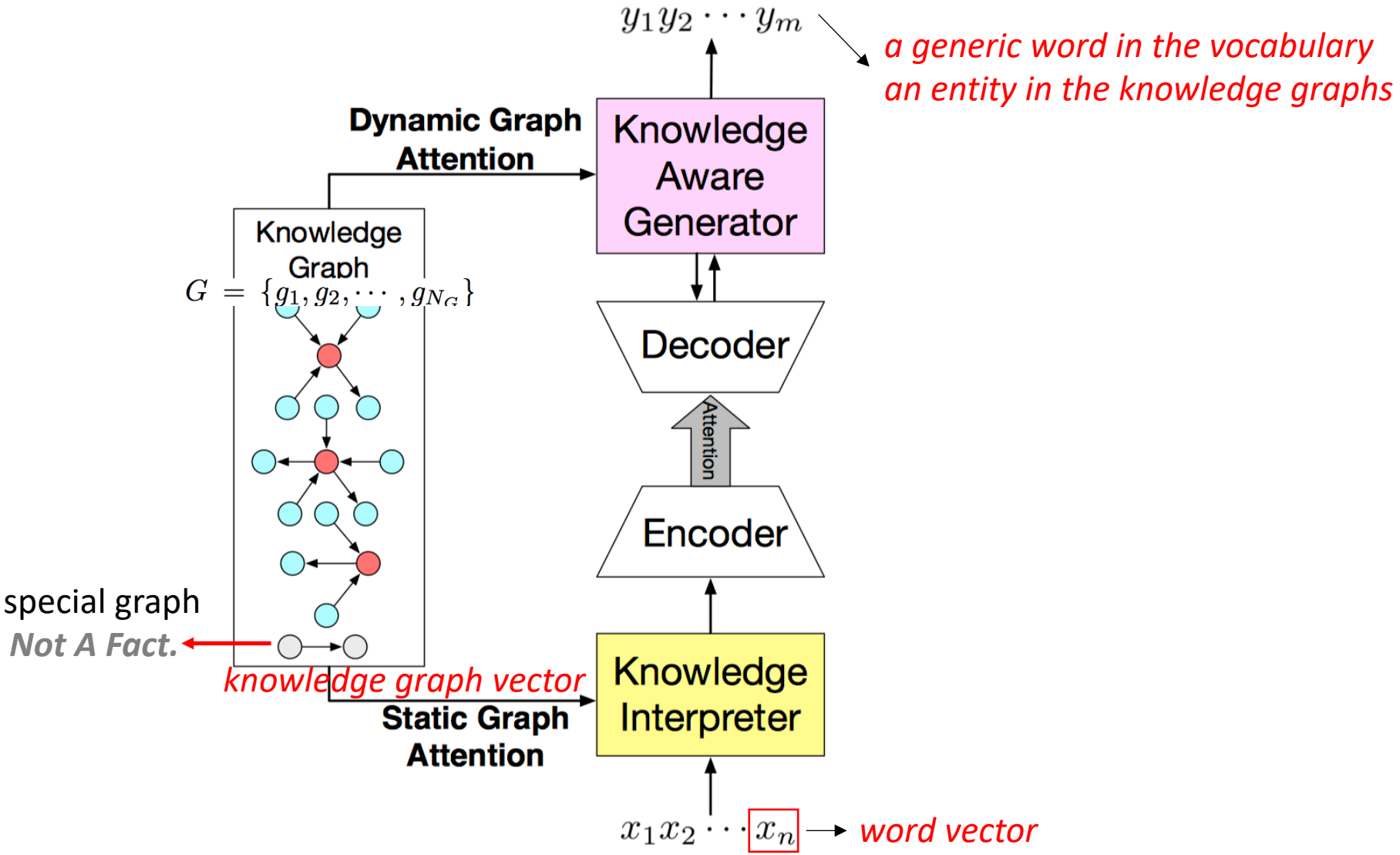
$$g_i = \{\tau_1, \tau_2, \cdots, \tau_{N_{g_i}}\}$$

    - *Each triple (head entity, relation, tail entity) is denoted as* $\tau = (h, r, t)$
    - *knowledge triple* τ *is represented by*

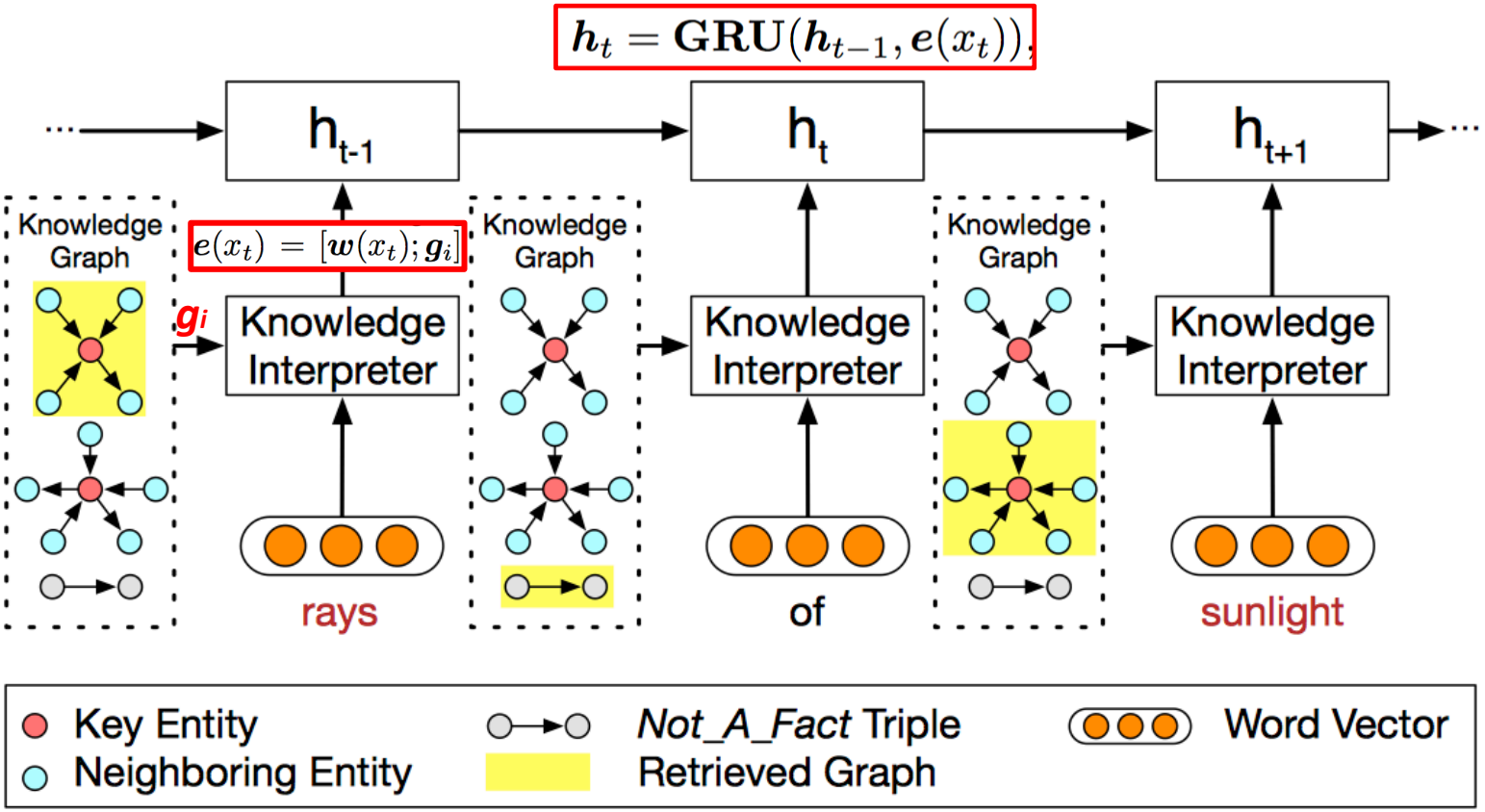
$$\boldsymbol{k} = (\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) = \mathbf{MLP}(\mathrm{TransE}(h, r, t))$$

- **Goal**
  - response: $Y = y_1 y_2 \cdots y_m$
- **Probability :** $P(Y|X,G) = \prod_{t=1}^{m} P(y_t|y_{<t}, X, G).$

Model-CCM


$y_1 y_2 \cdots y_m$
a generic word in the vocabulary
an entity in the knowledge graphs
Dynamic Graph Attention
Knowledge Aware Generator
Knowledge Graph
$G = \{g_1, g_2, \cdots, g_{N_G}\}$
Decoder
Attention
Encoder
special graph
Not A Fact.
knowledge graph vector
Knowledge Interpreter
Static Graph Attention
$x_1 x_2 \cdots x_n$
word vector

Knowledge Interpreter



Static Graph Attention

- **Generate a representation for a retrieved knowledge graph**
  - not only all **nodes** in a graph
  - but also **relations** between nodes
- **Input**: $\boldsymbol{K}(g_i) = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_{N_{g_i}}\}$

  $$\boldsymbol{k} = (\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) = \mathbf{MLP}(\mathrm{TransE}(h, r, t))$$
- **Output**: graph vector $\boldsymbol{g}_i$

$$\boldsymbol{g}_i = \sum_{n=1}^{N_{g_i}} \alpha_n^s [\boldsymbol{h}_n; \boldsymbol{t}_n],$$

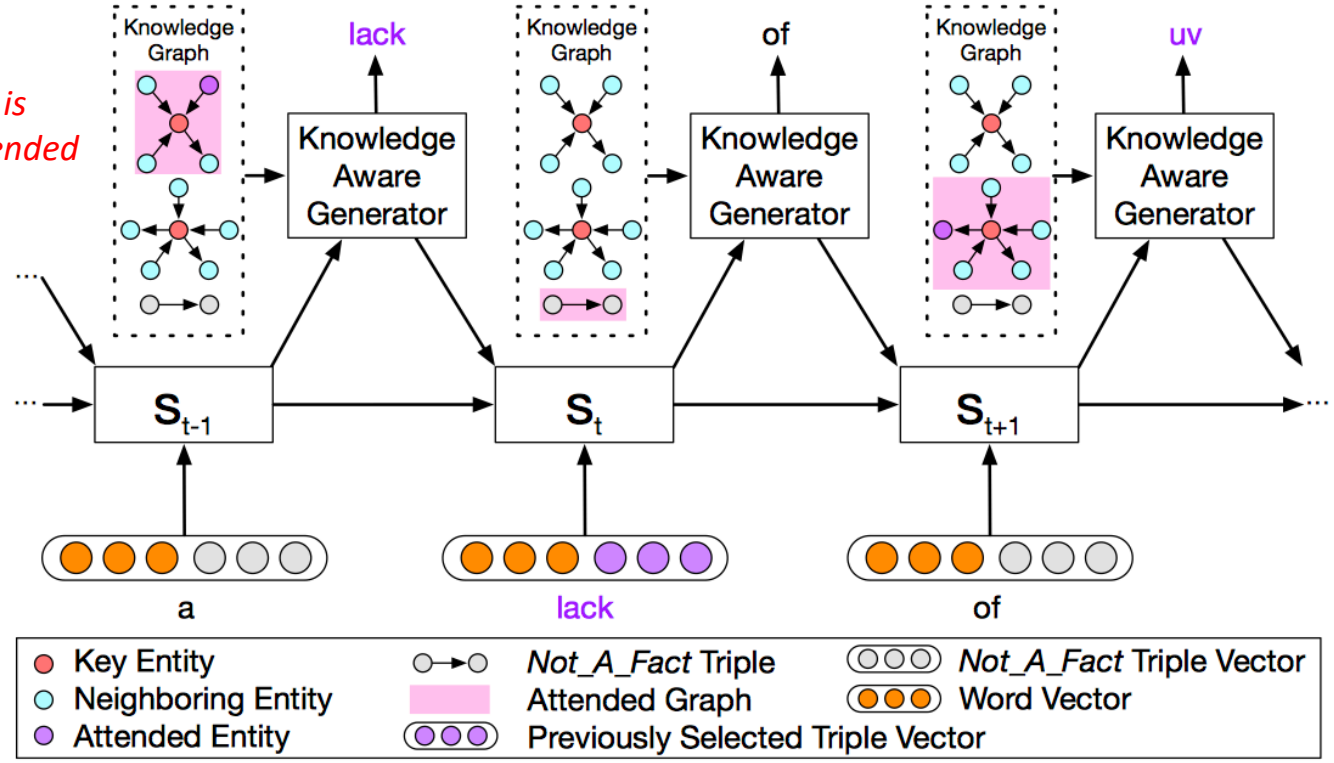
$$\alpha_n^s = \frac{\exp(\beta_n^s)}{\sum_{j=1}^{N_{g_i}} \exp(\beta_j^s)},$$

$$\beta_n^s = (\mathbf{W_r} \boldsymbol{r}_n)^\top \tanh(\mathbf{W_h} \boldsymbol{h}_n + \mathbf{W_t} \boldsymbol{t}_n),$$

Knowledge Aware Generator

triple's neighboring entity is used for word generation.

pink graph is mostly attended



$$s_{t+1} = \mathbf{GRU}(s_t, [c_t; c_t^g; c_t^k; e(y_t)])$$

- *vectors attended on encoder hidden*
- *vectors attended on knowledge graph vectors*
- *vectors attended on knowledge triple vectors*

$$e(y_t) = [w(y_t); k_j],$$

*concatenation of the **word vector** and the **previous knowledge triple vector***

Dynamic Graph Attention

- Given the **decoder state** s_t ,first attends on the **knowledge graph vectors** $\{g_1, g_2, \cdots, g_{N_G}\}$ to compute the probability of using of each graph g_i:

$$
\begin{aligned}
c_t^g &= \sum_{i=1}^{N_G} \alpha_{ti}^g g_i, \\
\alpha_{ti}^g &= \frac{\exp(\beta_{ti}^g)}{\sum_{j=1}^{N_G} \exp(\beta_{tj}^g)}, \\
\beta_{ti}^g &= V_b^\top \tanh(\mathbf{W_b} s_t + \mathbf{U_b} g_i),
\end{aligned}
$$

attentively reads all the knowledge graphs

*the **probability** of choosing knowledge graph gi at step t* →

- The model then attends on the knowledge triple vectors $K(g_i) = \{k_1, k_2, \cdots, k_{N_{g_i}}\}$ within each graph g_i

$$
\begin{aligned}
c_t^k &= \sum_{i=1}^{N_G} \sum_{j=1}^{N_{g_i}} \alpha_{ti}^g \alpha_{tj}^k k_j, \\
\alpha_{tj}^k &= \frac{\exp(\beta_{tj}^k)}{\sum_{n=1}^{N_{g_i}} \exp(\beta_{tn}^k)}, \\
\beta_{tj}^k &= k_j^\top \mathbf{W_c} s_t,
\end{aligned}
$$

attentively reads all the triples in each graph

Dynamic Graph Attention

$$
\begin{aligned}
\boldsymbol{a}_t &= [\boldsymbol{s}_t; \boldsymbol{c}_t; \boldsymbol{c}_t^g; \boldsymbol{c}_t^k], \\
\gamma_t &= \text{sigmoid}(\mathbf{V_o}^\top \boldsymbol{a}_t), \\
P_c(y_t = w_c) &= \text{softmax}(\mathbf{W_o} \boldsymbol{a}_t), \\
P_e(y_t = w_e) &= \alpha_{ti}^g \alpha_{tj}^k, \\
y_t \sim \boldsymbol{o}_t = P(y_t) &= \begin{bmatrix} (1-\gamma_t) P_g(y_t = w_c) \\ \gamma_t P_e(y_t = w_e) \end{bmatrix},
\end{aligned}
$$

balance the choice between an entity word and a generic word

Pc/Pe is the distribution over generic/entity words respectively

Loss Function

supervised signals on the knowledge aware generator layer to teacher-force the selection of an entity or a generic word.

$$L(\theta) = -\sum_{t=1}^{m} \boldsymbol{p}_t \log(\boldsymbol{o}_t) - \sum_{t=1}^{m} (q_t \log(\gamma_t) + (1-q_t)\log(1-\gamma_t)),$$

$q_t \in \{0, 1\}$ *is the true choice of an entity word or a generic word in Y*

Experiments

- **Commonsense Knowledge Base: ConceptNet**
  - **world facts** such as Paris is the capital of France
  - **informal relations** between common concepts such as “A dog is a pet”
- **Commonsense Conversation Dataset:**
  - reddit single-round dialogs
  - If a post-response pair can not be connected by any triple the pair will be removed.
  - **Four test sets**
    - **High-frequency pairs**: each post has all top 25% frequent words
    - **Medium-frequency pairs**: each post contains at least one word whose frequency is within the range of 25%-75%,
    - **Low-frequency pairs**: within the range of 75%-100%
    - **OOV pairs**: each post contains out-of-vocabulary words

Baselines

- **A seq2seq model (Seq2Seq),** which is widely used in open-domain conversational systems.
- **A knowledge-grounded model (MemNet),** where the memory units store TransE embeddings of knowledge triples.
- **A copy network (CopyNet) model,** which copies a word from knowledge triples or generates a word from the vocabulary.

Automatic Evaluation

- **Perplexity**
  - whether the content is **grammatical** and **relevant** in topic
- **Entity score**
  - the number of entities per response to measure the model's ability to select the concepts from the commonsense knowledge base in generation

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**

1. lowest perplexity on all the test sets
2. selects the most entities from the commonsense knowledge

rare concepts need more background knowledge to understand and respond

Manual Evaluation

- **Appropriateness**
  - whether the response is appropriate in grammar, topic, and logic
- **Informativeness**
  - whether the response provides new information and knowledge in addition to the post

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

Conclusion

- This work is the first attempt that uses **large-scale commonsense knowledge** in neural conversation generation.

- Instead of treating knowledge triples (or entities) separately and independently, we devise **static and dynamic graph attention mechanisms to treat the knowledge triples as a graph**

Thanks!