EMNLP 2021 Tutorial

Knowledge-Enriched Natural Language Generation

Wenhao Yu¹, Meng Jiang¹, Zhiting Hu², Qingyun Wang³, Heng Ji³,⁴, Nazneen Rajani⁵

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

External Knowledge-enhanced Methods

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

External knowledge acquisition occurs when knowledge is provided from outside sources
Knowledge-enhanced NLG (Overall)

Knowledge-enhanced NLG Methods

Internal Knowledge-enhanced Methods

Knowledge base

Knowledge graph

Grounded text

External Knowledge-enhanced Methods

knowledge source

Topic

Keyword

Linguistic features

knowledge extraction and representation

E.g., LDA, CNN, Neural topic

E.g., Keyword assignment, keyword extraction

E.g., POS tag, NER tag, Dependency parsing

E.g., KB memory network, KB related tasks

E.g., KGE, Path finding and reasoning, Graph neural network

E.g., Retrieval-augmented, Background based methods
Topic-enhanced NLG methods

- 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.

LDA topic modeling
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  ❌

You haven’t been given an assignment in this case

I don’t know what you are talking about

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 ↑

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 ↑

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

### Examples

#### Message
- 你也玩摄影
- You also like photography
- 我的左脸颊上有一些红血丝
- There is some redness on my left cheek.
- 大学生的话可以去实习吗
- Can a college student apply for an internship
- My skin is so dry.

#### TA-Seq2Seq
- 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.
- My skin is also very sensitive.
- 我也是敏感肌
- Can a college student apply for an internship
- Then hydrate and moisturize your skin.

#### S2SA-MMI
- 额额额...
- Well well well...

#### S2SA
- 额...
- Well...

#### S2SA
- Me too.
- 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-Guided Variational Autoencoders for Text Generation, In NAACL 2019
### Topic-enhanced NLG methods

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Methods</th>
<th>APNEWS</th>
<th>IMDB</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B-2</td>
<td>B-3</td>
<td>B-4</td>
</tr>
<tr>
<td><strong>test-BLEU</strong></td>
<td>VAE</td>
<td>0.564</td>
<td>0.278</td>
<td>0.192</td>
</tr>
<tr>
<td></td>
<td>VAE+HF (K=1)</td>
<td>0.566</td>
<td>0.280</td>
<td>0.193</td>
</tr>
<tr>
<td></td>
<td>VAE+HF (K=10)</td>
<td>0.570</td>
<td>0.279</td>
<td>0.195</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=0, T=10)</td>
<td>0.582</td>
<td>0.320</td>
<td>0.203</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=1, T=10)</td>
<td>0.581</td>
<td>0.326</td>
<td>0.202</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=10, T=10)</td>
<td>0.584</td>
<td>0.327</td>
<td>0.202</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=10, T=30)</td>
<td>0.627</td>
<td>0.335</td>
<td>0.207</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=10, T=50)</td>
<td><strong>0.629</strong></td>
<td><strong>0.340</strong></td>
<td><strong>0.210</strong></td>
</tr>
<tr>
<td><strong>self-BLEU</strong></td>
<td>VAE</td>
<td>0.866</td>
<td>0.531</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>VAE+HF (K=1)</td>
<td>0.865</td>
<td>0.533</td>
<td>0.241</td>
</tr>
<tr>
<td></td>
<td>VAE+HF (K=10)</td>
<td>0.873</td>
<td>0.552</td>
<td>0.219</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=0, T=10)</td>
<td>0.847</td>
<td>0.499</td>
<td>0.161</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=1, T=10)</td>
<td>0.847</td>
<td>0.495</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=10, T=10)</td>
<td>0.839</td>
<td>0.512</td>
<td>0.172</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=10, T=30)</td>
<td>0.811</td>
<td>0.478</td>
<td>0.157</td>
</tr>
<tr>
<td></td>
<td>TGVAE (K=10, T=50)</td>
<td><strong>0.808</strong></td>
<td><strong>0.476</strong></td>
<td><strong>0.150</strong></td>
</tr>
</tbody>
</table>

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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS</td>
<td>37.41</td>
<td>15.87</td>
<td>34.70</td>
</tr>
<tr>
<td>SEASS</td>
<td>46.86</td>
<td>24.58</td>
<td>43.53</td>
</tr>
<tr>
<td>PG</td>
<td>46.97</td>
<td>24.63</td>
<td>43.66</td>
</tr>
<tr>
<td>KIGN</td>
<td>46.18</td>
<td>23.93</td>
<td>43.44</td>
</tr>
<tr>
<td>Bottom-up</td>
<td>45.80</td>
<td>23.61</td>
<td>42.54</td>
</tr>
<tr>
<td>Concat+DualPG</td>
<td>47.05</td>
<td>24.39</td>
<td>43.77</td>
</tr>
<tr>
<td>Gated+DualPG</td>
<td>47.13</td>
<td>24.87</td>
<td>44.34</td>
</tr>
<tr>
<td>Hier+DualPG</td>
<td>47.14</td>
<td>25.06</td>
<td>44.39</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>ROUGE-2</th>
<th>GIGAWORDS ROUGE-L</th>
<th>BLEU-4</th>
<th>ROUGE-2</th>
<th>NYT ROUGE-L</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2Seq</td>
<td>8.83±0.15</td>
<td>31.43±0.13</td>
<td>12.21±0.30</td>
<td>8.83±0.15</td>
<td>31.43±0.13</td>
<td>12.21±0.30</td>
</tr>
<tr>
<td>SeqAttn</td>
<td>9.10±0.13</td>
<td>36.62±0.11</td>
<td>16.17±0.28</td>
<td>5.95±0.15</td>
<td>29.67±0.06</td>
<td>11.86±0.15</td>
</tr>
<tr>
<td>CopyNet</td>
<td>9.44±0.11</td>
<td>36.96±0.10</td>
<td>16.40±0.24</td>
<td>6.25±0.14</td>
<td>30.58±0.09</td>
<td>11.96±0.14</td>
</tr>
<tr>
<td>GPT-2</td>
<td>9.04±0.20</td>
<td>31.30±0.16</td>
<td>15.66±0.40</td>
<td>5.86±0.20</td>
<td>24.19±0.14</td>
<td>10.89±0.22</td>
</tr>
<tr>
<td>UniLM</td>
<td>11.77±0.18</td>
<td>36.54±0.15</td>
<td>17.66±0.35</td>
<td>7.47±0.15</td>
<td>30.66±0.13</td>
<td>12.90±0.20</td>
</tr>
<tr>
<td>InjType</td>
<td>13.37±0.12</td>
<td>41.16±0.31</td>
<td>18.55±0.09</td>
<td>8.55±0.09</td>
<td>31.53±0.17</td>
<td>13.14±0.03</td>
</tr>
<tr>
<td>⊥ w/o MP</td>
<td>9.39±0.16</td>
<td>38.34±0.10</td>
<td>16.36±0.25</td>
<td>6.52±0.09</td>
<td>30.10±0.08</td>
<td>12.19±0.10</td>
</tr>
<tr>
<td>⊥ w/o NLU</td>
<td>12.85±0.18</td>
<td>40.65±0.37</td>
<td>18.24±0.26</td>
<td>8.13±0.10</td>
<td>30.80±0.36</td>
<td>13.10±0.09</td>
</tr>
</tbody>
</table>
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 $\mathcal{R}$.
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 represents entities and relations in a low-dimensional vector space to reduce data dimensionality while preserving the inherent structure of the KG.

• What are the commonly used KGE methods?
  - **TransE**: Given a KG edge \( e \), the relation is seen as a translation vector \( r \) so the embedded entities \( e_1 \) and \( e_2 \) can be connected with low error, namely \( e_1 + r \approx e_2 \).
  - Example: Tokyo + IsCapitalOf ≈ Japan.

<table>
<thead>
<tr>
<th>id</th>
<th>entity</th>
<th>vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chat</td>
<td>[-0.1, 0.2, 0.5]</td>
</tr>
<tr>
<td>2</td>
<td>talk</td>
<td>[0.3, 0.4, -0.2]</td>
</tr>
<tr>
<td>3</td>
<td>future</td>
<td>[0.6, 0.1, 0.1]</td>
</tr>
<tr>
<td>4</td>
<td>based</td>
<td>[-0.4, 0.4, 0.2]</td>
</tr>
</tbody>
</table>
**KG-enhanced text generation methods**

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

<table>
<thead>
<tr>
<th>Knowledge Bases</th>
<th>Original Triples</th>
<th>Examples of Transformed Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet</td>
<td>(eiffel tower, <strong>AtLocation</strong>, paris) (telephone, <strong>UsedFor</strong>, communication)</td>
<td>eiffel tower <em>is at</em> paris.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>telephone <em>is used for</em> communication.</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>(PersonX dates for years, <strong>oEffect</strong>, continue dating) (PersonX cooks spaghetti, <strong>xintent</strong>, to eat)</td>
<td>PersonX dates for years. PersonY will continue dating. PersonX cooks spaghetti. PersonX wants to eat.</td>
</tr>
</tbody>
</table>

**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 algorithm)
  • 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 connected 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

\[ o_v^l = \frac{1}{|\mathcal{N}(v)|} \sum_{(u,r) \in \mathcal{N}(v)} W^l_{KN}(h_u^l, h_r^l), \]

\[ h_v^{l+1} = \text{ReLU}(o_v^l + W^l_{ES} h_v^l), \]
• Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, In EMNLP 2020

Decoder: A reasoning module to assign scores for neighboring nodes

\[ n_s(v) = \frac{1}{f_{(u,r) \in N_{in}(v)}} \left( \gamma \cdot n_s(u) + R(u, r, v) \right) \]

\[ R(u, r, v) = \sigma(h_{u,r,v}^T W_{sim} h_{L_D}^r) \]

\[ h_{u,r,v} = [h_u^L_G, h_r^L_G, h_v^L_G] \]
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

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

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).
• **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):

$$
\mathbf{u}^{(k)} = \text{COMBINE}_k(\mathbf{u}^{(k-1)}, \text{AGGREGATE}_k(\{(\mathbf{u}^{(k-1)}_i, \mathbf{e}^{(k-1)}_{ij}, \mathbf{u}^{(k-1)}_j) : \forall (u_i, e_{ij}, u_j) \in \mathcal{N}(u)\})),$$

$$
\mathbf{h}_G = \text{READOUT}(\{\mathbf{u}^{(K)} : u \in \mathcal{U}\}).$$
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 ↑

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>High Freq.</th>
<th>Medium Freq.</th>
<th>Low Freq.</th>
<th>OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ppx.</td>
<td>ent.</td>
<td>ppx.</td>
<td>ent.</td>
<td>ppx.</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>47.02</td>
<td>0.717</td>
<td>42.41</td>
<td>0.713</td>
<td>47.25</td>
</tr>
<tr>
<td>MemNet</td>
<td>46.85</td>
<td>0.761</td>
<td>41.93</td>
<td>0.764</td>
<td>47.32</td>
</tr>
<tr>
<td>CopyNet</td>
<td>40.27</td>
<td>0.96</td>
<td>36.26</td>
<td>0.91</td>
<td>40.99</td>
</tr>
<tr>
<td>CCM</td>
<td>39.18</td>
<td>1.180</td>
<td>35.36</td>
<td>1.156</td>
<td>39.64</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>High Freq.</th>
<th>Medium Freq.</th>
<th>Low Freq.</th>
<th>OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCM vs. Seq2Seq</td>
<td>0.616</td>
<td>0.662</td>
<td>0.605</td>
<td>0.656</td>
<td>0.549</td>
</tr>
<tr>
<td>CCM vs. MemNet</td>
<td>0.602</td>
<td>0.647</td>
<td>0.593</td>
<td>0.656</td>
<td>0.566</td>
</tr>
<tr>
<td>CCM vs. CopyNet</td>
<td>0.600</td>
<td>0.640</td>
<td>0.606</td>
<td>0.669</td>
<td>0.586</td>
</tr>
</tbody>
</table>

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

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

Table: Relevance Between Generated and Golden Responses.
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 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 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

(M1) Incorporate KGE into language generation

(M2) Transfer knowledge into pretrained LM
KG-enhanced text generation methods

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 (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

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 Neurips 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 Neurips 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>NQ</th>
<th>TQA</th>
<th>WQ</th>
<th>CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed</td>
<td>34.5</td>
<td>-</td>
<td>37.4</td>
<td>-</td>
</tr>
<tr>
<td>Book</td>
<td>36.6</td>
<td>-</td>
<td>44.7</td>
<td>-</td>
</tr>
<tr>
<td>Open</td>
<td>40.4</td>
<td>-</td>
<td>40.7</td>
<td>46.8</td>
</tr>
<tr>
<td>Book</td>
<td>41.5</td>
<td>57.9</td>
<td>41.1</td>
<td>50.6</td>
</tr>
<tr>
<td>RAG-Token</td>
<td>44.1</td>
<td>55.2</td>
<td>45.5</td>
<td>50.0</td>
</tr>
<tr>
<td>RAG-Seq.</td>
<td>44.5</td>
<td>56.8</td>
<td>52.2</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Jeopardy B-1</th>
<th>MSMARCO R-L</th>
<th>FVR3 B-1</th>
<th>FVR2 Label Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SotA</td>
<td>-</td>
<td>49.8*</td>
<td>49.9*</td>
<td>76.8</td>
</tr>
<tr>
<td>BART</td>
<td>15.1</td>
<td>38.2</td>
<td>41.6</td>
<td>64.0</td>
</tr>
<tr>
<td>RAG-Tok.</td>
<td>17.3</td>
<td>40.1</td>
<td>41.5</td>
<td>72.5</td>
</tr>
<tr>
<td>RAG-Seq.</td>
<td>14.7</td>
<td>40.8</td>
<td>44.2</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>SPICE(v1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-Gen (Bao et al., 2020)</td>
<td>23.468</td>
<td>12.606</td>
<td>24.822</td>
<td>27.30</td>
</tr>
<tr>
<td>UniLM (Dong et al., 2019)</td>
<td>30.616</td>
<td>14.889</td>
<td>27.429</td>
<td>30.20</td>
</tr>
<tr>
<td>BART (Lewis et al., 2020)</td>
<td>31.827</td>
<td>13.976</td>
<td>27.995</td>
<td>30.60</td>
</tr>
<tr>
<td>T5-base (Raffel et al., 2020)</td>
<td>18.546</td>
<td>9.399</td>
<td>19.871</td>
<td>22.00</td>
</tr>
<tr>
<td>T5-large (Raffel et al., 2020)</td>
<td>31.962</td>
<td>15.128</td>
<td>28.855</td>
<td>31.60</td>
</tr>
<tr>
<td>EKI-BART (Fan et al., 2020)</td>
<td>35.945</td>
<td>16.999</td>
<td>29.583</td>
<td>32.40</td>
</tr>
<tr>
<td>KG-BART (Liu et al., 2021)</td>
<td>33.867</td>
<td>16.927</td>
<td>29.634</td>
<td>32.70</td>
</tr>
<tr>
<td>CALM(T5-base) (Zhou et al., 2021)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>33.00</td>
</tr>
<tr>
<td>RE-T5 (ours)</td>
<td><strong>40.863</strong></td>
<td><strong>17.663</strong></td>
<td><strong>31.079</strong></td>
<td><strong>34.30</strong></td>
</tr>
</tbody>
</table>

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 ...

<table>
<thead>
<tr>
<th></th>
<th>Appropriateness</th>
<th>Grounding</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIST</td>
<td>BLEU</td>
<td>METEOR</td>
</tr>
<tr>
<td>Human</td>
<td>2.650</td>
<td>3.13%</td>
<td>8.31%</td>
</tr>
<tr>
<td>SEQ2SEQ</td>
<td>2.223</td>
<td>1.09%</td>
<td>7.34%</td>
</tr>
<tr>
<td>MEMNET</td>
<td>2.185</td>
<td>1.10%</td>
<td>7.31%</td>
</tr>
<tr>
<td>CMR-F</td>
<td>2.260</td>
<td>1.20%</td>
<td>7.37%</td>
</tr>
<tr>
<td>CMR</td>
<td>2.213</td>
<td>1.43%</td>
<td>7.33%</td>
</tr>
<tr>
<td>CMR+W</td>
<td>2.238</td>
<td>1.38%</td>
<td>7.46%</td>
</tr>
</tbody>
</table>

Table: Automatic Evaluation results on Reddit dataset.
# Grounded text-enhanced NLG methods

<table>
<thead>
<tr>
<th>Evidence sources</th>
<th>Tasks</th>
<th>Methods</th>
<th>Dataset Information</th>
<th>Retrieval space (d/s)</th>
<th># Retrieved d/s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Name</td>
<td>#Instance</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Dialogue system</td>
<td>MemNet SKT</td>
<td>Wizard of Wikipedia (WoW)</td>
<td>22,311</td>
<td>5.4M/93M</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAG</td>
<td>MS-MARCO</td>
<td>267,287</td>
<td>21M/-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BART+DPR RT+C-REALM</td>
<td>ELI5</td>
<td>274,741</td>
<td>3.2M/-</td>
</tr>
<tr>
<td></td>
<td>Question answering</td>
<td>H&amp;W CANDELA</td>
<td>ChangeMyView</td>
<td>287,152</td>
<td>5M/-</td>
</tr>
<tr>
<td></td>
<td>Argument generation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online platform</td>
<td>Dialogue (for business)</td>
<td>AT2T KGNCM</td>
<td>Amazon books</td>
<td>937,032</td>
<td>-/131K</td>
</tr>
<tr>
<td>(e.g., Amazon)</td>
<td></td>
<td></td>
<td>Foursquare</td>
<td>1M</td>
<td>-/1.1M</td>
</tr>
<tr>
<td>Gigawords</td>
<td>Summarization</td>
<td>R³Sum BiSET</td>
<td>Gigawords</td>
<td>3.8M</td>
<td>-/3.8M</td>
</tr>
</tbody>
</table>

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