**Barack’s Wife Hillary:**

Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV* Nelson F. Liu‡§ Matthew E. Peters§ Matt Gardner§ Sameer Singh*

*University of California, Irvine, CA, USA, †University of Washington, Seattle, WA, USA, §Allen Institute for Artificial Intelligence, Seattle, WA, USA

For any questions, email: rlogan@uci.edu

**Summary**

- Traditional language models have limited ability to generate factually correct text.
- We introduce the knowledge graph language model (KGLM), a neural language model with mechanisms for generating information from a knowledge graph.
- We collect the Linked WikiText-2 dataset, which aligns WikiText-2 to the Wikidata knowledge graph.
- Experiments show that the KGLM has better perplexity than AWD-LSTM-LM, and better fact-completion capabilities than GPT-2 small despite being trained on less data.

**Motivating Example**

**Linked WikiText-2 Dataset**

**Example Annotation**

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Super Mario Land is a new platform video game developed and published by Nintendo as a launch title for their Game Boy handheld game console.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mention Type</td>
<td>Entity Mentioned</td>
</tr>
<tr>
<td>new</td>
<td>new</td>
</tr>
<tr>
<td>2019</td>
<td>pub, date</td>
</tr>
</tbody>
</table>

**Dataset Statistics**

- Documents: 600
  - Dev: 60
  - Test: 60
- Tokens: 2M
  - 208K: 238K
- Vocabulary: 33K
  - Unknown: -
- Mention Tokens: 207K
  - 12K: 21K
- Unique Entities: 207K
  - 21K: 24K
- Unique Relations: 123K
  - 41K: 5.4K

**Fact Completion**

<table>
<thead>
<tr>
<th>Input Sentence</th>
<th>Gold</th>
<th>GPT-2</th>
<th>KGLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Birthplace</td>
<td>Paris Hilton was born in New York City in 1960</td>
<td>New York City</td>
<td>New York</td>
</tr>
<tr>
<td>Correct</td>
<td>Arnold Schwarzenegger was born on 1947-07-30</td>
<td>1947-07-30</td>
<td>1947-07-30</td>
</tr>
<tr>
<td>KGLM</td>
<td>Bob Dylan was born in 1941</td>
<td>Duluth</td>
<td>New Duluth</td>
</tr>
<tr>
<td>Correct</td>
<td>Ulysses is a book that was written by James Joyce</td>
<td>James Joyce</td>
<td>a James Joyce</td>
</tr>
<tr>
<td>GPTv2 Correct</td>
<td>St. Louis is a city in the state of Missouri</td>
<td>Missouri</td>
<td>Missouri</td>
</tr>
<tr>
<td>Both Wrong</td>
<td>Kanye West is married to Kim Kardashian</td>
<td>Kim Kardashian</td>
<td>Kim Kardashian</td>
</tr>
<tr>
<td>Wrong</td>
<td>The capital of India is New Delh</td>
<td>New Delhi</td>
<td>the a</td>
</tr>
<tr>
<td>Wrong</td>
<td>Madonna is married to Carlos Leon</td>
<td>Carlos Leon</td>
<td>a Alex</td>
</tr>
</tbody>
</table>

**Resources**

- Code: [github.com/rloganiv/kglm-model](https://github.com/rloganiv/kglm-model)
- Dataset: [rloganiv.github.io/linked-wikitext-2](https://rloganiv.github.io/linked-wikitext-2)

**Perplexity**

- PPL: 85.4
  - UPP: 189.2
- EntityNLM*: 76.1
- EntityCopyNet*: 144.0
- AWD-LSTM*: 74.8
  - UPP: 165.8
- KGLM*: 44.1

*Obtained using importance sampling

**Unknown Penalty**

\[ P_{\text{UPP}}(v_{\text{unk}}) = \frac{P(v_{\text{unk}})}{\text{Vocab} / \text{Vocab}} \]