Creating Training Corpora for NLG Micro-Planning

Claire Gardent, Anastasia Shimorina, Shashi Narayan, Laura Perez-Beltrachini

Presented by: Omar Elabd
Final Product

<originaltripleset>
  <otriple>Buzz_Aldrin | mission | Apollo_11</otriple>
  <otriple>Buzz_Aldrin | timeInSpace | 52.0</otriple>
  <otriple>Apollo_11 | operator | NASA</otriple>
</originaltripleset>

<modifiedtripleset>
  <mtriple>Buzz_Aldrin | was a crew member of | Apollo_11</mtriple>
  <mtriple>Buzz_Aldrin | timeInSpace | "52.0"(minutes)</mtriple>
  <mtriple>Apollo_11 | operator | NASA</mtriple>
</modifiedtripleset>

<lex comment="good" lid="Id1">Buzz Aldrin, as part of the NASA operated Apollo 11 program, spent 52 minutes in space.</lex>

<lex comment="good" lid="Id2">On the NASA operated Apollo 11 program, crew member Buzz Aldrin spent 52.0 minutes in space.</lex>

Introduction

• Authors generated a dataset consisting of data and text pairs.
• The data is in the form of RDF triples from DBpedia (which is a knowledge based).
• The sentences were generated from the RDF triples using crowd workers on the CrowdFlower platform.
Motivation

• In general, these datasets are useful for Micro-Planners (i.e. data-to-text generation systems)
  • Generating Referring Expressions
  • Lexicalization
  • Aggregation
  • Surface Realization
  • Sentence Segmentation

• Current data-text corpora are domain specific and crafted by experts
  • Results in stereotyped texts by generators

• Wen et al. created a dataset from a knowledge base using crowdsourced methods (RNNLG)
RNNLG Example Dataset

```
inform(name=satellite eurus 65; type=laptop;
memory=4 gb; isforbusinesscomputing=false;
drive range=medium)
```

"the satellite eurus 65 is a laptop designed
for home use with 4 gb of memory and a medium
sized hard drive"

"satellite eurus 65 is a laptop which has a 4
gb memory, is not for business computing, and
is in the medium drive range"

<table>
<thead>
<tr>
<th></th>
<th>WEBNLG</th>
<th>RNNLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nb. Input</td>
<td>5068</td>
<td>22225</td>
</tr>
<tr>
<td>Nb. Data-Text Pairs</td>
<td>13339</td>
<td>30842</td>
</tr>
<tr>
<td>Nb. Domains</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Nb. Attributes</td>
<td>172</td>
<td>108</td>
</tr>
<tr>
<td>Nb. Input Patterns</td>
<td>2108</td>
<td>2155</td>
</tr>
<tr>
<td>Nb. Input / Nb Input Pattern</td>
<td>2.40</td>
<td>10.31</td>
</tr>
<tr>
<td>Nb. Input Shapes</td>
<td>58</td>
<td>6</td>
</tr>
</tbody>
</table>

the satellite eurus 65 is a laptop designed for home use with 4 gb of memory and a medium sized hard drive.

satellite eurus 65 is a laptop which has a 4 gb memory, is not for business computing, and is in the medium drive range.
• Buzz Aldrin, as part of the NASA operated Apollo 11 program, spent 52 minutes in space.
• On the NASA operated Apollo 11 program, crew member Buzz Aldrin spent 52.0 minutes in space.
Data Shape - Comparison

- A participated in mission B operated by C.
- A participated in mission B which was operated by C.
- A was born in E. She worked as an engineer
- A was born in E and worked as an engineer

Data Shape – Take Home

• In general, trees of deeper depth allows for more various syntactic constructs to be learned by generators.
1. Retrieve RDF triples from DBpedia
2. Clean up property names to be less ambiguous
3. Use CrowdFlower platform to generate sentences
4. Validate generated sentences using CrowdFlower

<originaltripleset>
  <otriple>Buzz_Aldrin | mission | Apollo_11</otriple>
  <otriple>Buzz_Aldrin | timeInSpace | 52.0</otriple>
  <otriple>Apollo_11 | operator | NASA</otriple>
</originaltripleset>

<modifiedtripleset>
  <mtriple>Buzz_Aldrin | was a crew member of | Apollo_11</mtriple>
  <mtriple>Buzz_Aldrin | timeInSpace | "52.0"(minutes)</mtriple>
  <mtriple>Apollo_11 | operator | NASA</mtriple>
</modifiedtripleset>

<lex comment="good" lid="Id1">Buzz Aldrin, as part of the NASA operated Apollo 11 program, spent 52 minutes in space.</lex>
<lex comment="good" lid="Id2">On the NASA operated Apollo 11 program, crew member Buzz Aldrin spent 52.0 minutes in space.</lex>
Process – #1 Data Selection/Retrieval

• Authors adopted a procedure by Perez-Beltrachini et al. (2016)

1. Start with a broad category (e.g. Astronomy)
2. Compute probabilities of RDF properties co-occurring together
   • They used the SRILM toolkit
3. Content selection can be formulated as an Integer Linear Programming (ILP) problem
   • Attempts to maximize coherence and variability of input shape
Process - #1 Data Selection/Retrieval

“mission” (1-gram)
“mission - birthPlace” (2-gram)
“mission - birthPlace - birthDate” (3-gram).

SRILM toolkit

-1.329421 mission (1-gram),
-0.8845956 mission - birthPlace (2-gram),
-0.5842706 mission - birthPlace - birthDate (3-gram)

ILP program

Process – #2 Cleanup

A new “modifiedtripleset” was created where RDF properties were clarified manually.

```
<originaltripleset>
  <otriple>Buzz_Aldrin | mission | Apollo_11</otriple>
  <otriple>Buzz_Aldrin | timeInSpace | 52.0</otriple>
  <otriple>Apollo_11 | operator | NASA</otriple>
</originaltripleset>
<modifiedtripleset>
  <mtriple>Buzz_Aldrin | was a crew member of | Apollo_11</mtriple>
  <mtriple>Buzz_Aldrin | timeInSpace | "52.0"(minutes)</mtriple>
  <mtriple>Apollo_11 | operator | NASA</mtriple>
</modifiedtripleset>
```
Process – #3 Sentence Generation

• For single triples
  • Crowd workers were asked to generate a sentence based on cleaned up triple.
    
    `<mtriple>Apollo_11 | operator | NASA</mtriple>`
    
    Apollo 11 was operated by NASA

• For sets of triples
  • Crowd workers were asked to merge sentences together into a natural sounding text.
    
    “Buzz Alderin was a crew member of Apollo 11”
    “Apollo 11 was operated by NASA”
    
    Apollo 11 was operated by NASA
Process - #4 Validation

• Authors used CrowdFlower again to validate the generated sentences for coherence.

• Crowd workers were asked three questions:
  • Does the text sound fluent and natural?
  • Does the text contain all and only the information from the data?
  • Is the text good English (no spelling or grammatical mistakes)?
How do you test which dataset is better?
Results – Part-of-Speech Tagger

• Ran Stanford Part-Of-Speech Tagger and Parser v3.5.2
  • WEBNLG has a higher corrected type-token ratio (CTTR) which indicates greater lexical variety
  • WEBNLG has a higher lexical sophistication

<table>
<thead>
<tr>
<th></th>
<th>WEBNLG</th>
<th>RNNLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nb. Text / Input</td>
<td>2.63</td>
<td>1.38</td>
</tr>
<tr>
<td>Text Length</td>
<td>24.36/23/4/80</td>
<td>18.37/19/1/76</td>
</tr>
<tr>
<td>(avg/median/min/max)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nb. Sentence / Text</td>
<td>1.45/1/1/6</td>
<td>1.25/1/1/6</td>
</tr>
<tr>
<td>(avg/median/min/max)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nb. Tokens</td>
<td>290479</td>
<td>531871</td>
</tr>
<tr>
<td>Nb. Types</td>
<td>2992</td>
<td>3524</td>
</tr>
<tr>
<td>Lexical Sophistication</td>
<td>0.69</td>
<td>0.54</td>
</tr>
<tr>
<td>CTTR</td>
<td>3.93</td>
<td>3.42</td>
</tr>
</tbody>
</table>

Results – Neural Generation

• Basic premise: Richer and more varied datasets are harder to learn.

• Ran an out of the box sequence-to-sequence model
  • 3-layer LSTM with 512 units, Batch size of 64, Learning rate of 0.5
  • Similar amount of data from RNNLG and WEBNLG used for training (13K data-text pairs)
  • 3:1:1 training, validation, test split
  • Two modes of delexicalization, **Fully** and **Name only**
    • **Fully**: Buzz Aldrin participated in Apollo 11 → Astronaut participated in Mission
    • **Name only**: Buzz Aldrin participated in Apollo 11 → Astronaut participated in Apollo 11

• Code used available at:
  [https://github.com/tensorflow/nmt/tree/master/nmt](https://github.com/tensorflow/nmt/tree/master/nmt)
## Results

<table>
<thead>
<tr>
<th>Vocab size</th>
<th>Delexicalisation Mode</th>
<th>WEBNLG</th>
<th>RNNNLG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fully</td>
<td>520, 2430</td>
<td>140, 1530</td>
</tr>
<tr>
<td></td>
<td>Name only</td>
<td>1130, 2940</td>
<td>570, 1680</td>
</tr>
<tr>
<td>Perplexity</td>
<td>Fully</td>
<td>27.41</td>
<td>17.42</td>
</tr>
<tr>
<td></td>
<td>Name only</td>
<td>25.39</td>
<td>23.93</td>
</tr>
<tr>
<td>BLEU</td>
<td>Fully</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>Name only</td>
<td>0.10</td>
<td>0.27</td>
</tr>
</tbody>
</table>

References


