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>
```

Source Dataset: *Creating Training Corpora for Micro-Planners*. Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Proceedings of ACL 2017.

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 crowd sourced 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"

Source Dataset: Multi-domain Neural Network Language Generation for Spoken Dialogue Systems. Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Steve Young. Proceedings of the 2016 Conference on North American Chapter of the Association for Computational Linguistics (NAACL)

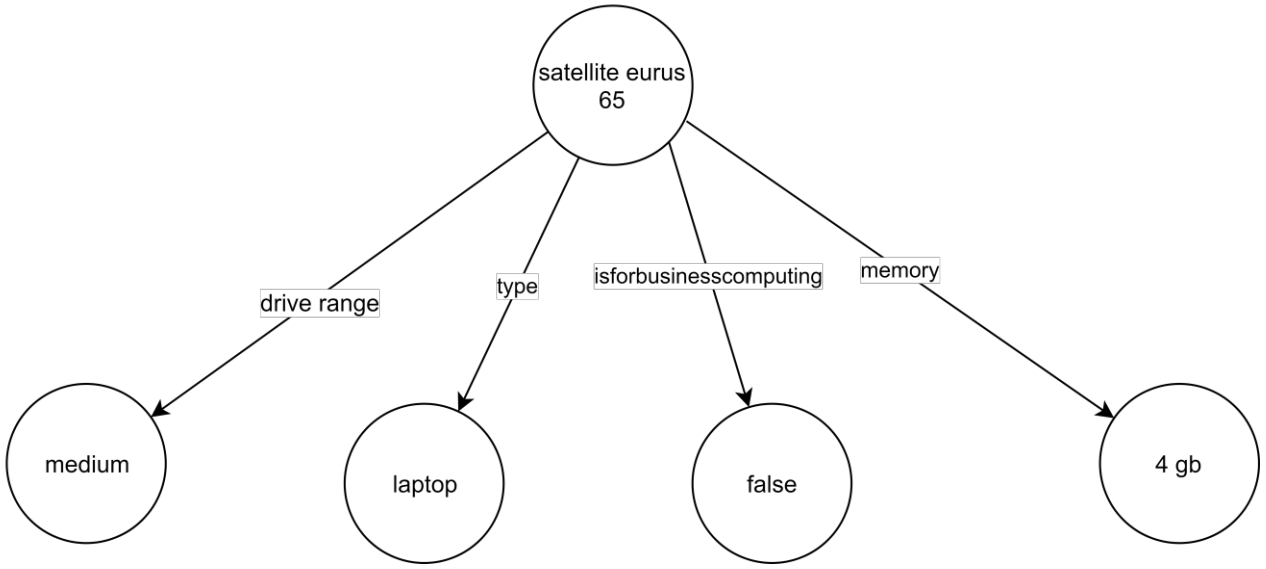
WEBNLG vs RNNLG

	WebNLG	RNNLG
Nb. Input	5068	22225
Nb. Data-Text Pairs	13339	30842
Nb. Domains	6	4
Nb. Attributes	172	108
Nb. Input Patterns	2108	2155
Nb. Input / Nb Input Pattern	2.40	10.31
Nb. Input Shapes	58	6

Source: *Creating Training Corpora for Micro-Planners*. Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Proceedings of ACL 2017.

Data Shape - RNNLG

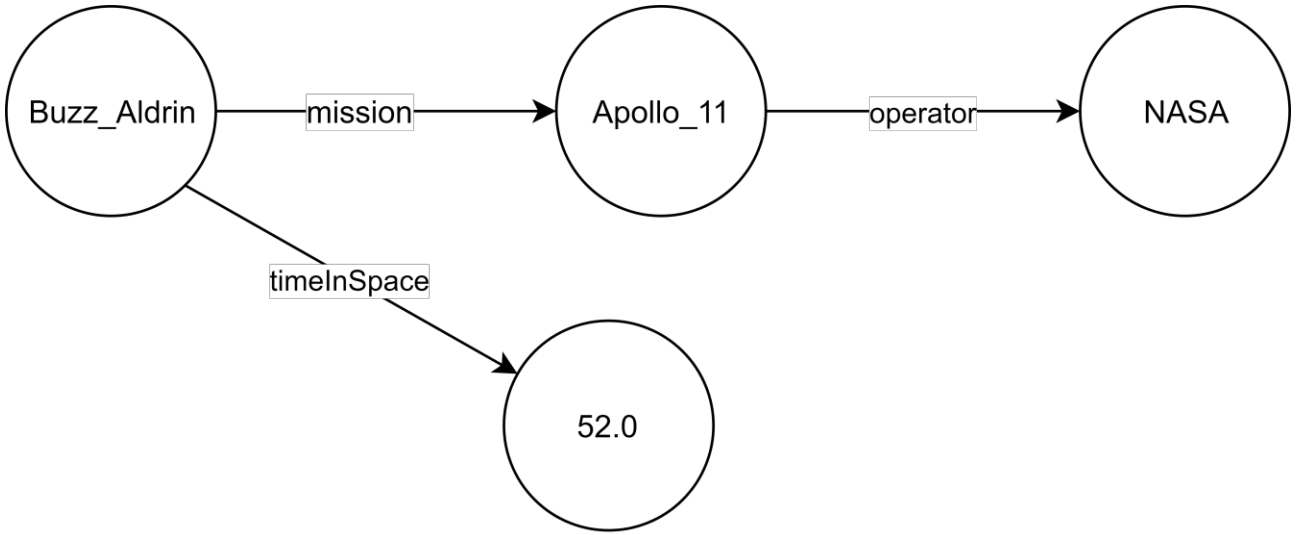
```
inform(name=satellite eurus 65; type=laptop; memory=4
gb; isforbusinesscomputing=false; drive range=medium)
```



- 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.

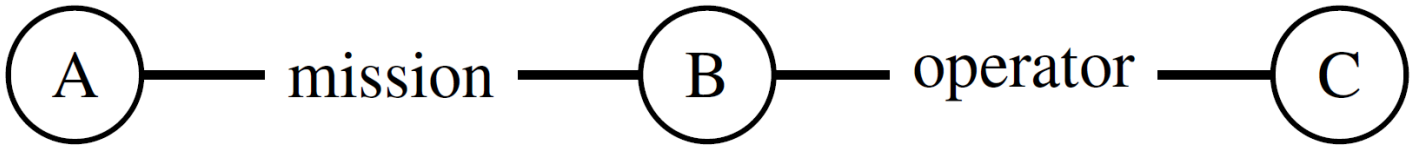
Data Shape - WebNLG

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



- 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

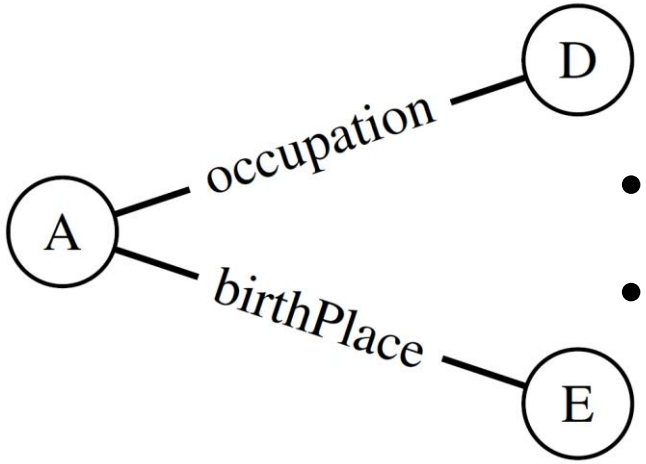


participial

- A participated in mission B operated by C.

passive subject relative clause

- A participated in mission B which was operated by C.



New clause with pronominal subject

- A was born in E. She worked as an engineer
- A was born in E and worked as an engineer

Coordinated verb phrase

Source: *Creating Training Corpora for Micro-Planners*. Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Proceedings of ACL 2017.

Data Shape – Take Home

- In general, trees of deeper depth allows for more various syntactic constructs to be learned by generators.

Process

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>
```

#1

#2

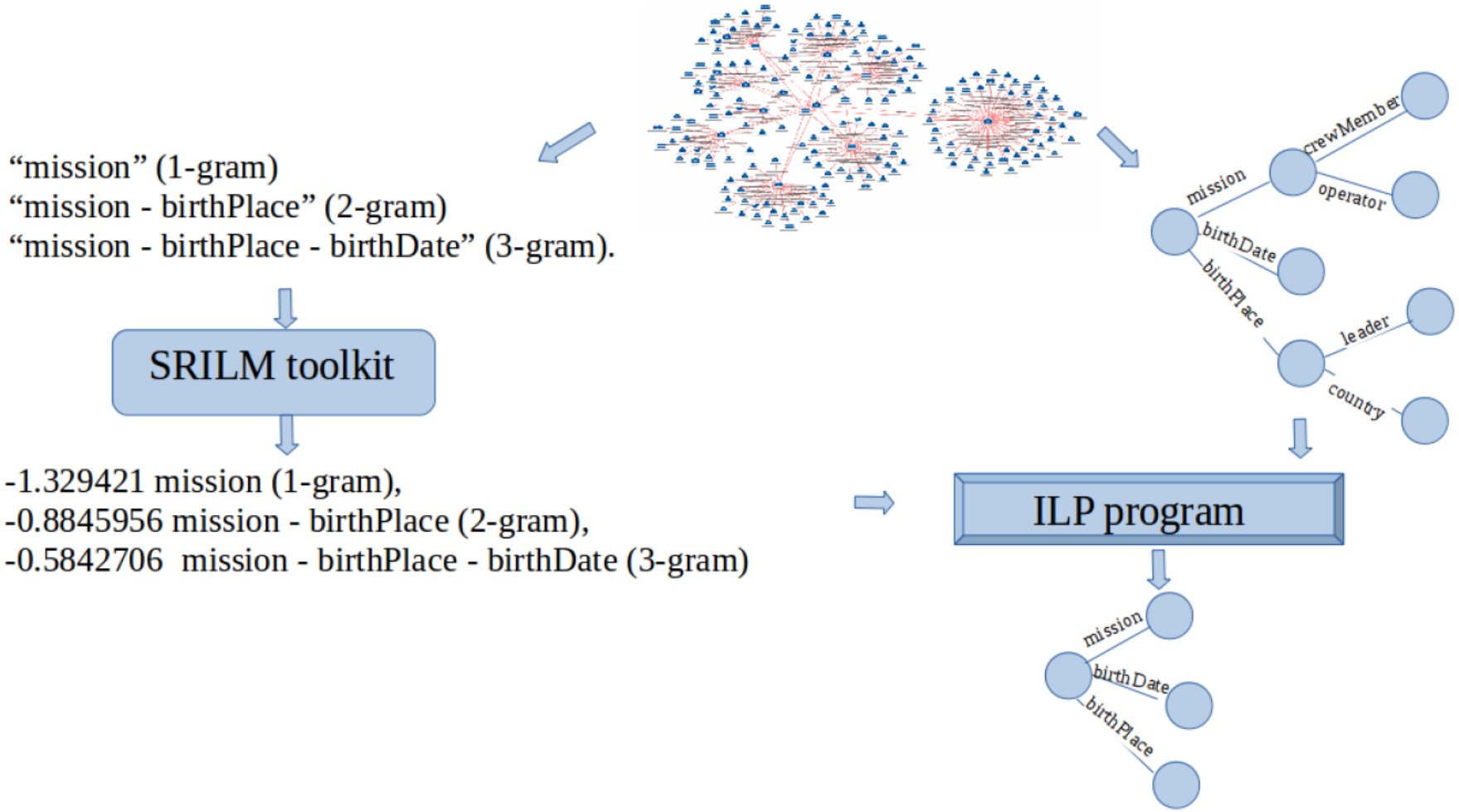
#3/4

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



Source: *Creating Training Corpora for Micro-Planners*. Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Proceedings of ACL 2017.

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.

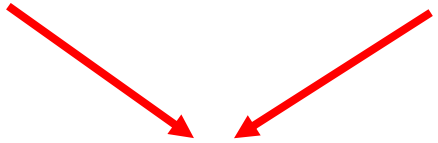
<mtriple>Apollo_11 | operator | NASA</mtriple>

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

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

Source: *Creating Training Corpora for Micro-Planners*. Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Proceedings of ACL 2017.

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

Results

	Delexicalisation Mode	WEBNLG	RNNLG
Vocab size	Fully	520, 2430	140, 1530
	Name only	1130, 2940	570, 1680
Perplexity	Fully	27.41	17.42
	Name only	25.39	23.93
BLEU	Fully	0.19	0.26
	Name only	0.10	0.27

Source: *Creating Training Corpora for Micro-Planners*. Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Proceedings of ACL 2017.

References

- *Creating Training Corpora for Micro-Planners*. Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Proceedings of ACL 2017.
- Gasic, M., Mrksic, N., Rojas-Barahona, L.M., Su, P., Vandyke, D., Wen, T., & Young, S.J. (2016). Multi-domain Neural Network Language Generation for Spoken Dialogue Systems. *HLT-NAACL*.
- Wen, Tsung-Hsien et al. “Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking.” *SIGDIAL Conference* (2015).
- Wen, Tsung-Hsien et al. “Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems.” *EMNLP* (2015).
- Wen, Tsung-Hsien et al. “Toward Multi-domain Language Generation using Recurrent Neural Networks.” (2015).