WebNLG
A Benchmark for Microplanning

Claire Gardent

CNRS/LORIA and Université de Lorraine, Nancy

Amazon, Cambridge
08 June 2017
Joint Work with

Anastasia Shimorina  Shashi Narayan  Laura Perez-Beltrachini

Funded by the French ANR Project WebNLG
http://talc1.loria.fr/webnlg/stories/about.html
Microplanning in NLG: How to say it?

Data ⇒ Fluent text

(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation Fighter_pilot)

John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot

- Generating Referring Expressions: Describing entities
- Lexicalisation: Choosing lexical items
- Surface Realisation: Choosing syntactic structures
- Aggregation: Avoiding repetition
- Sentence segmentation: Segmenting the content into sentence size chunks
Generating Referring Expressions: Describing entities

Data

(John_E_Blaha birthDate 1942-08-26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation Fighter_pilot)

*John E Blaha* was born in San Antonio on 1942-08-26. *He worked as a fighter pilot*
Lexicalisation: Choosing lexical items

Data

(John_E_Blahoma birthDate 1942_08_26)

*John E Blaha was born on* 1942-08-26

*John E Blaha’s birthdate is* 1942-08-26.
Surface Realisation: Choosing syntactic structures

Data

(John_E_Blah birthPlace San_Antonio)
(John_E_Blah birthDate 1942_08_26)
(John_E_Blah occupation Fighter_pilot)

John E Blaha, *(born in San Antonio)* \text{APPOS}, on 1942-08-26 worked as a fighter pilot

John E Blaha *(was born in San Antonio)* \text{VP} on 1942-08-26. He worked as a fighter pilot

John E Blaha *(who was born in San Antonio on 1942-08-26)* \text{RELX} worked as a fighter pilot
Aggregation: Avoiding repetition

Data

(John_E_Blahoma birthDate 1942_08_26)
(John_E_Blahoma birthPlace San_Antonio)
(John_E_Blahoma occupation Fighter_pilot)

*John E Blaha*, born in San Antonio on 1942-08-26, worked as a fighter pilot

?? *John E Blaha* was born in San Antonio. *John E Blaha* was born on 1942-08-26. *John E Blaha* worked as a fighter pilot
Sentence segmentation: Segmenting the content into sentence size chunks

Data

(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation Fighter_pilot)

[John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot]$_s$

[John E Blaha was born in San Antonio on 1942-08-26]$_s$. [He worked as a fighter pilot]$_s$
Outline

1 Existing Benchmarks

2 The WebNLG Framework
   - Creating Data
   - Associating Data with Text
   - Comparing Benchmarks

3 The WebNLG Challenge
Existing Benchmarks
Data-to-Text Corpora

Domain specific

Constructed from expert linguistic annotations.

Crowdsourced
Domain Specific Benchmark

- (Chen et al. 2008): Soccer Games
  1,539 data-text pairs, Vocabulary of 214 words.
- (Liang et al. 2009): Weather forecasts
  29,528 data-text pairs, Vocabulary of 345 words.
- (Ratnaparkhi et al. 2000): Air travel domain
  5,426 data-text pairs, Vocabulary of 927 words.

Strongly stereotyped text with restricted syntax and lexicon.
Benchmarks constructed from expert linguistic annotations

Unordered dependency trees / Newspaper text

Banarescu et al. 2012.
Abstract Meaning Representations / News and Discussion Forum

- Linguistic input
- Focus on surface realisation
  No sentence segmentation, restricted REG and lexicalisation
- Manual annotation of text with complex linguistic structure is expensive (time and expertise)
Crowdsourced

(Wen et al. 2016, Novikova and Rieser 2016): Dialog acts

\[
\text{recommend(name=caerus 33;type=television;}
\text{screensizerange=medium;family=t5;hasusbport=true)}
\]

The caerus 33 is a medium television in the T5 family that’s USB-enabled.

- ✓ Low cost (no expert linguist required)
- × Data synthetised from toy ontology
- × Limited Data Variety: input = tree of depth one
The WebNLG Framework
The WebNLG Approach

- RDF KB – Content Selection → Data
  - “Real” data: automatically extracted from RDF KB
  - “Varied” data: data of various shapes and sizes
- Text produced by crowdworkers

Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini
Creating Training Corpora for NLG Micro-Planning
ACL, 2017.
DBPedia

Data stored as RDF triples of the form (subject, property, object)

(Alan.Bean mission Apollo.12)
(Apollo.12 crewMember Peter.Conrad)
(Apollo.12 operator NASA)
(Alan.Bean birthDate 1932-03-15)
(Alan.Bean birthPlace Wheeler,Texas)
(Wheeler,Texas country USA)

6.2M entities, 739 classes, 2,695 properties
Content Selection
Data Shape and NL Syntax

**CHAIN**

<table>
<thead>
<tr>
<th>Discourse-Based</th>
<th>Coherence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A participated in mission B operated by C.</td>
<td>A participated in mission B which was operated by C.</td>
</tr>
</tbody>
</table>

**SIBLING**

<table>
<thead>
<tr>
<th>Topic-Based</th>
<th>Coherence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A was born in E. She worked as an engineer.</td>
<td>A was born in E and worked as an engineer.</td>
</tr>
</tbody>
</table>
Content Selection Procedure

Step 1: Learn bigram models of RDF-properties

Step 2: Use these models and Integer Linear Programming to extract data units

- that are subtrees of the DBPedia graph
- that maximise coherence
- that have various shapes and sizes

Laura Perez-Beltrachini, Rania Mohammed Sayed and Claire Gardent
Building RDF Content for Data-to-Text Generation

*COLING, 2016.*
Bi-grams of RDF Properties

S(IBLING) bi-grams
mission-birthDate
mission-birthPlace
birthDate-birthPlace
crewMember-operator

C(HAIN) bi-grams
mission-crewMember
mission-operator
birthPlace-country

Claire Gardent
“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)
Extracting Data Units

\[ x_t = x^p_{s,o} = \begin{cases} 
1 & \text{if the triple is preserved} \\
0 & \text{otherwise}
\end{cases} \]

\[ y_{t_1,t_2} = \begin{cases} 
1 & \text{if the pair of triples is preserved} \\
0 & \text{otherwise}
\end{cases} \]
Objective Function

s- and c-Model

\[ S(X) = \sum_{Y} y_{t_i,t_j} \cdot P(t_i, t_j) \]

m-Model

\[ S(X) = \gamma \sum_{Y} y_{t_i,t_j} \cdot P(t_i, t_j) + (1 - \gamma) \sum_{Z} z_{t_k,t_l} \cdot P(t_k, t_l) \]
Consistency Constraints.

Bigram → Triple

\[ \forall i, j \ (y_{i,j} \leq x_i \ \text{and} \ y_{i,j} \leq x_j) \]

Triple → Bigram

\[ y_{i,j} + (1 - x_i) + (1 - x_j) \geq 1 \]
Tree constraints

Each object has at most one subject

\[ \forall o \in Soln, \sum_{s,p} x_{s,o}^p \leq 1 \]

All triples are connected

\[ \forall o \in Soln, \sum_{s,p} x_{s,o}^p - \frac{1}{|X|} \sum_{u,p} x_{o,u}^p \geq 0 \]
Crowdsourcing Text
Associating Data with Text

1. Clarifying RDF properties
   (Allan_Bean crew1up Apollo_12)  
   ⇒ (Allan_Bean commander Apollo_12)

2. Getting verbalisations for single triples.
   (John_E_Blaha birthDate 1942_08_26)  
   ⇒ ??

3. Getting verbalisations for input containing more than one triple.
   Make a text out of $n$ clauses
   John E Blaha was born in San Antonio.  
   John E Blaha was born on 1942-08-26.  
   ⇒ ??

4. Verifying the quality of the collected texts.
Monitoring Crowdworkers

- *A priori* automatic checks. 12 custom javascript validators implemented in the CrowdFlower platform
  - Minimal text length
  - Minimal match triple/text
  - No exact match
  - No cut and paste
  - ...

- *A posteriori* manual checks to remove incorrect verbalisations
- Continuous monitoring of crowdworkers (bans, bonuses)
Verifying the quality of the collected texts

*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)?*

5 judgments / question
Reject text if it received three negative answers in at least one criterion.
Total corpus loss: 8.7%

**Rejected example**

(AEK_Athens_F.C. manager Gus_Poyet)
(Gus_Poyet club Chelsea_F.C.)

*AEK Athens F.C. are managed by Gus Poyet, who is in Chelsea_F.C.*
Evaluation
Evaluation

Content selection

*Are the created data units coherent and varied?*

Benchmark Comparison

*How does a WebNLG corpus compares with Wen’s Dataset?*
Evaluating the Results of Content Selection

Are the created data units coherent and varied?

Experiment

- 3 DBPedia categories: Monument, University, Astronaut
- 5 entity graphs per category
- 10 best solutions produced by each model
Diversity

Input shapes

- 75 distinct shapes
- Nb of instances per shape: Min = 1, Max = 24, Avg = 5.31

Average Overlap

\[
\frac{\sum_{i,j} O(s_i, s_j)}{N}
\]

\[
O(s_i, s_j) = \frac{\text{Nb. of common properties}}{\text{Total nb of triples}}
\]
### Overlap within Models

<table>
<thead>
<tr>
<th></th>
<th>Depth 1</th>
<th>Depth 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>s-Model</td>
<td>c-Model</td>
<td>m-Model</td>
</tr>
<tr>
<td>n3</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td>n4</td>
<td>0.29</td>
<td>0.21</td>
</tr>
<tr>
<td>n5</td>
<td>0.29</td>
<td>0.23</td>
</tr>
<tr>
<td>n6</td>
<td>0.27</td>
<td>0.23</td>
</tr>
<tr>
<td>n7</td>
<td>0.34</td>
<td>0.25</td>
</tr>
<tr>
<td>n8</td>
<td>0.36</td>
<td>0.26</td>
</tr>
<tr>
<td>n9</td>
<td>0.34</td>
<td>0.27</td>
</tr>
<tr>
<td>n10</td>
<td>0.39</td>
<td>0.30</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.31</td>
<td>0.24</td>
</tr>
</tbody>
</table>
### Overlap across Models

<table>
<thead>
<tr>
<th></th>
<th>Depth 2</th>
<th>Depth1 vs. Depth 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c-Model</td>
<td>s-Model</td>
</tr>
<tr>
<td>m-Model</td>
<td>0.21</td>
<td>0.10</td>
</tr>
<tr>
<td>n3</td>
<td>0.25</td>
<td>0.15</td>
</tr>
<tr>
<td>n4</td>
<td>0.25</td>
<td>0.16</td>
</tr>
<tr>
<td>n5</td>
<td>0.23</td>
<td>0.17</td>
</tr>
<tr>
<td>n6</td>
<td>0.25</td>
<td>0.19</td>
</tr>
<tr>
<td>n7</td>
<td>0.26</td>
<td>0.20</td>
</tr>
<tr>
<td>n8</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td>n9</td>
<td>0.25</td>
<td>0.27</td>
</tr>
<tr>
<td>n10</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>0.24</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Irrelevant Properties

E.g., leader for category Astronaut

Baseline: Random extraction of subtrees from entity graph

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th># Solns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>d1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>2</td>
<td>0.44</td>
<td>400</td>
</tr>
<tr>
<td>s-Model</td>
<td>0</td>
<td>1.75</td>
<td>0.31</td>
<td>271</td>
</tr>
<tr>
<td><strong>d2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>2</td>
<td>0.73</td>
<td>218</td>
</tr>
<tr>
<td>c-Model</td>
<td>0</td>
<td>1.94</td>
<td>0.59</td>
<td>382</td>
</tr>
<tr>
<td>m-Model</td>
<td>0</td>
<td>1.25</td>
<td>0.43</td>
<td>152</td>
</tr>
<tr>
<td>s-Model</td>
<td>0.07</td>
<td>1.29</td>
<td>0.54</td>
<td>123</td>
</tr>
</tbody>
</table>
Human Evaluation

<table>
<thead>
<tr>
<th></th>
<th>BL</th>
<th>s-Model</th>
<th>c-Model</th>
<th>m-Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coherent (3)</td>
<td>6</td>
<td>18</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Medium (2)</td>
<td>15</td>
<td>11</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>Low (1)</td>
<td>10</td>
<td>2</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td><strong>Avg</strong></td>
<td>1.87</td>
<td>2.52</td>
<td>2.27</td>
<td>2.43</td>
</tr>
</tbody>
</table>

23 pairs of data units
Size 3 to 10
Three categories
10 judgements for each pair
Comparing Benchmarks
Comparing Benchmarks

RNNNLG (Wen et al. 2016)

recommend(name=caerus 33;type=television;
screensizerange=medium;family=t5;hasusbport=true)

*The caerus 33 is a medium television in the T5 family that’s USB-enabled.*

WebNLG

(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation Fighter_pilot)

*John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot*
Properties

<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. Properties</td>
<td>172</td>
<td>108</td>
</tr>
</tbody>
</table>

A larger number of properties is more likely to induce texts with greater **lexical variety**.

- X title Y / X *served as* Y
- X nationality Y / X’s *nationality is* Y
- X country Y / X *is in* Y
- X nationality USA / X *is American*
Input Patterns

<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. Input Patterns</td>
<td>2108</td>
<td>2155</td>
</tr>
<tr>
<td>Nb Input Pattern / Nb. Input</td>
<td>0.41</td>
<td>0.09</td>
</tr>
</tbody>
</table>

A larger number of input patterns is more likely to induce texts with greater syntactic variety.

country-location-startDate  ⇒  passive, apposition, deverbal nominal
108 St. Georges Terrace is located in Perth, Australia. Its construction began in 1981.

almaMater-birthPlace-selection  ⇒  passive, VP coordination
William Anders was born in British Hong Kong, graduated from AFIT in 1962, and joined NASA in 1963.
Neural Generation

(Vinyals et al. 2015) Multi-layered sequence-to-sequence model with attention mechanism.

- 13K data-text pairs
- 3-layer LSTMs with 512 units each
- batch size of 64
- learning rate of 0.5.

<table>
<thead>
<tr>
<th></th>
<th>WEBNLG</th>
<th>RNNLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocab (Input/Output)</td>
<td>520 / 2430</td>
<td>140 / 1530</td>
</tr>
<tr>
<td>Perplexity</td>
<td>27.41</td>
<td>17.42</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.19</td>
<td>0.26</td>
</tr>
</tbody>
</table>
The WebNLG Challenge
The WebNLG Challenge

21,855 data/text pairs
8,372 distinct data input
9 DBpedia categories: Astronaut, University, Monument, Building, ComicsCharacter, Food, Airport, SportsTeam and WrittenWork
CC Attribution-Noncommercial-Share Alike 4.0 International licence

Baseline
OpenNMT sequence-to-sequence model with attention mechanism
BLEU = 54.03
Schedule

14 April 2017: Release of Training and Development Data
30 April 2017: Release of Baseline System
22 August 2017: Release of Test Data
25 August 2017: Entry submission deadline
5 September 2017: Results of automatic evaluation and system presentations (at INLG 2017)
30 September 2017: Results of human evaluation

37 downloads from 15 countries
Summary

- Generation
- Multilingual
- Discourse
Summary

- Generation
- Multilingual
- Discourse

THANKS!