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_Blaha 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_Blaha birthPlace San_Antonio)

(John_E_Blaha birthDate 1942_08_26)

(John_E_Blaha occupation Fighter_pilot)

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

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

John E Blaha (who was born in San Antonio on 1942-08-26)$_{RELx}$ worked as a fighter pilot

Aggregation: Avoiding repetition

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

?? *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

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

Belz et al. 2011. Surface Realisation Shared Task.
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

```
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.

- ✓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

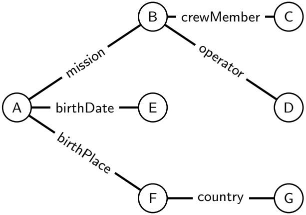
- RDF KB – $\boxed{\text{Content Selection}}$ $\rightarrow$ Data
  - "Real" data: automatically extracted from RDF KB
  - "Varied" data: data of various shapes and sizes
- Text produced by crowdworkers

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)
```

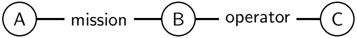


6.2M entities, 739 classes, 2,695 properties

Content Selection

Data Shape and NL Syntax

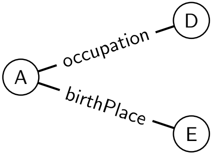
CHAIN



Discourse-Based Coherence

A participated in mission B operated by C.

A participated in mission B which was operated by C.



SIBLING

Topic-Based Coherence

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

A was born in E and worked as an engineer.

Content Selection Procedure

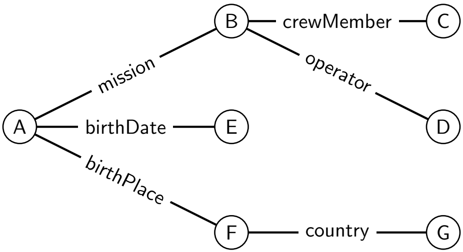
Step 1: Learn bigram models of RDF-properties

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



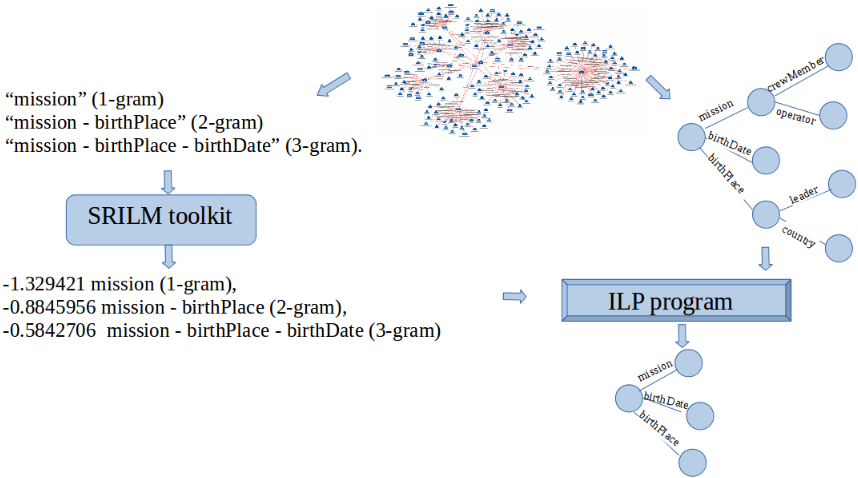
`S(IBLING)` bi-grams

`mission-birthDate`
`mission-birthPlace`
`birthDate-birthPlace`
`crewMember-operator`

`C(HAIN)` bi-grams

`mission-crewMember`
`mission-operator`
`birthPlace-country`

Creating Data



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

$$S(X) = \sum_Y y_{t_i,t_j} \; . \; P(t_i, t_j)$$

M-Model

$$S(X) = \gamma * \sum_Y y_{t_i,t_j} \; . \; P(t_i, t_j) + (1-\gamma) \sum_Z z_{t_k,t_l} \; . \; P(t_k, t_l)$$

Consistency Constraints.

Bigram $\rightarrow$ Triple

$$\forall i, j \; (y_{i,j} \leq x_i \textit{ and } y_{i,j} \leq x_j)$$

Triple $\rightarrow$ Bigram

$$y_{i,j} + (1 - x_i) + (1 - x_j) \geq 1$$

Tree constraints

Each object has at most one subject

$$\forall o \in Soln, \sum_{s,p} x^{p}_{s,o} \quad \leq \quad 1$$

All triples are connected

$$\forall o \in Soln, \sum_{s,p} x^{p}_{s,o} \quad - \quad \frac{1}{|X|} \quad \sum_{u,p} x^{p}_{o,u} \geq \quad 0$$

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

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

$$O(s_i, s_j) = \frac{\textit{Nb. of common properties}}{\textit{Total nb of triples}}$$

Overlap within Models

	Depth 1	Depth 2	
	S-Model	C-Model	M-Model
n3	0.18	0.16	0.24
n4	0.29	0.21	0.35
n5	0.29	0.23	0.27
n6	0.27	0.23	0.23
n7	0.34	0.25	0.27
n8	0.36	0.26	0.24
n9	0.34	0.27	0.25
n10	0.39	0.30	0.25
Avg.	0.31	0.24	0.26

Overlap across Models

	Depth 2	Depth1 vs. Depth 2	
	C-Model M-Model	S-Model C-Model	S-Model M-Model
n3	0.21	0.10	0.12
n4	0.25	0.15	0.19
n5	0.25	0.16	0.19
n6	0.23	0.17	0.21
n7	0.25	0.19	0.25
n8	0.26	0.20	0.23
n9	0.26	0.21	0.22
n10	0.25	0.27	0.20
Avg.	0.24	0.18	0.20

Irrelevant Properties

E.g., *leader* for category *Astronaut*

Baseline: Random extraction of subtrees from entity graph

		Min	Max	Avg	# Solns
d1	BL	0	2	0.44	400
	S-Model	0	1.75	**0.31**	271
d2	BL	0	2	0.73	218
	C-Model	0	1.94	0.59	382
	M-Model	0	1.25	**0.43**	152
	S-Model	0.07	1.29	0.54	123

Human Evaluation

	BL	S-Model	C-Model	M-Model
Coherent (3)	6	18	1	2
Medium (2)	15	11	20	13
Low (1)	10	2	9	15
Avg	1.87	2.52	2.27	2.43

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

	WEBNLG	RNNLG
Nb. Input	5068	22225
Nb. Properties	172	108

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

`X title Y` / *X served as Y*	Verb
`X nationality Y` / *X's nationality is Y*	Relational noun
`X country Y` / *X is in Y*	Preposition
`X nationality USA` / *X is American*	Adjective

Input Patterns

	WebNLG	RNNLG
Nb. Input	5068	22225
Nb. Input Patterns	2108	2155
Nb Input Pattern / Nb. Input	0.41	0.09

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.

	WebNLG	RNNLG
Vocab (Input/Output)	520 / 2430	140 / 1530
Perplexity	27.41	17.42
BLEU	0.19	0.26

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