How to Talk to Stranger: Generating Medical Reports for First-Time Users

Dimitra Gkatzia, Verena Rieser and Oliver Lemon
d.gkatzia@napier.ac.uk

IEEE-FUZZ (@ WCCI 2016)

Edinburgh Napier INTERACTION LAB

July 27, 2016
Overview

• Introduction
Overview

- Introduction
- Data-to-text Generation
Overview

- Introduction
- Data-to-text Generation
- Motivation
Overview

- Introduction
- Data-to-text Generation
- Motivation
- Domain: first-aid provision
Overview

- Introduction
- Data-to-text Generation
- Motivation
- Domain: first-aid provision
- Methodology
Overview

- Introduction
- Data-to-text Generation
- Motivation
- Domain: first-aid provision
- Methodology
- Evaluation
Introduction
Data-to-text Generation
Motivation
Domain: first-aid provision
Methodology
Evaluation
Results
Overview

- Introduction
- Data-to-text Generation
- Motivation
- Domain: first-aid provision
- Methodology
- Evaluation
- Results
Overview

- Introduction
- Data-to-text Generation
- Motivation
- Domain: first-aid provision
- Methodology
- Evaluation
- Results
- Summary
Handling **first-time users** in the context of automatic report generation from time-series data in the health domain.
Handling **first-time users** in the context of automatic report generation from time-series data in the health domain.

We use **multi-objective optimisation** to account for the content preferences of multiple possible user types.
The breathing rate increased from 20 to 30 breaths per minute. The blood oxygen saturation dropped from 95% to 90%. The heart rate increased from 110 to 121 beats per minute.
Data-to-text Architecture (Reiter, 2007)

Event Input Data

Numeric Input Data

Signal Analysis

Patterns

Data Interpretation

Messages, Relations

Document Planning

Selected msgs, doc/rhet struct

Microplanning and Realisation

Text
Motivation

- NLG can enhance decision making when the underlying data is uncertain (Gkatzia et al., 2016).
Motivation

- NLG can enhance decision making when the underlying data is uncertain (Gkatzia et al., 2016).
- NLG can lead to more informed decisions in the health domain (van Meulen et al., 2010).
Four emergency scenarios provided by the University of Aberdeen, consisting of a description and physiological data.
Four emergency scenarios provided by the University of Aberdeen, consisting of a description and physiological data.

Physiological data: Breathing Rate, Blood Oxygen Saturation and Heart Rate.
Four emergency scenarios provided by the University of Aberdeen, consisting of a description and physiological data.

Physiological data: Breathing Rate, Blood Oxygen Saturation and Heart Rate.

Knowledge acquisition from experts in order to derive templates which correspond to time-series data.
Scenario:

A female aged 30 years has been rescued from a burning building by Fire Service personnel. She is conscious and breathing. She has no obvious burns but is suffering from smoke inhalation and is currently being treated with 100 % oxygen by fire crews. The following graphs show the measurements of her breathing rate, blood oxygen saturation and heart rate. The summary below describes the sensor data depicted on the graphs. Please rate the summary in terms of your preference.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Breathing Rate ($BR$)</th>
<th>Blood Oxygen Saturation ($SpO_2$)</th>
<th>Heart Rate ($HR$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) average</td>
<td>The breathing rate was $&lt;$average$&gt;$breaths per minute on average.</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>(2) trend_verbose</td>
<td>...</td>
<td>The Blood oxygen saturation $&lt;$trend$&gt;$from $&lt;$initialMeasSpo2$&gt;$% to $&lt;$finalMeasSpo2$&gt;$%.</td>
<td>...</td>
</tr>
<tr>
<td>(3) trend_succinct</td>
<td>...</td>
<td>...</td>
<td>Heart rate $&lt;$trend-succinct$&gt;$from $&lt;$initialMeasHR$&gt;$ to $&lt;$finalMeasHR$&gt;$.</td>
</tr>
<tr>
<td>(4) range_verbose</td>
<td>The breathing rate was between $&lt;$lowestBr$&gt;$ and $&lt;$highestBr$&gt;$breaths per minute.</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>(5) range_succinct</td>
<td>...</td>
<td>$SpO_2$ $&lt;$lowestSpo2$&gt;$% - $&lt;$highestSpo2$&gt;$%.</td>
<td>...</td>
</tr>
</tbody>
</table>
| (6) inference | ... | ... | Heart rate observation $<$inference$>$.
Data Collection (Gkatzia et al., 2014)

69 participants (34 males 35 females)  Each participant - 4 emergency scenarios

<table>
<thead>
<tr>
<th>Group</th>
<th>Level of training</th>
<th>Number of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>None</td>
<td>10</td>
</tr>
<tr>
<td>Group 2</td>
<td>First Aid at Work, Emergency</td>
<td>42</td>
</tr>
<tr>
<td>Group 3</td>
<td>Basic First Person on Scene, Intermediate First Person on Scene, Equivalent to BASP, Advanced First Aid, Combat Medical Technician 2</td>
<td>5</td>
</tr>
<tr>
<td>Group 4</td>
<td>Emergency Medical Technician, Ambulance Technician, Combat Medical Technician 1, Offshore Medic</td>
<td>0</td>
</tr>
<tr>
<td>Group 5</td>
<td>Paramedic, Nurse, Physician’s Assistant</td>
<td>4</td>
</tr>
<tr>
<td>Group 6</td>
<td>Medical Doctor</td>
<td>8</td>
</tr>
</tbody>
</table>
There weren’t any correlations between users’ background, expertise or demographics with template choices.
1. There weren’t any correlations between users’ background, expertise or demographics with template choices.
2. Users chose templates according to their preferences.
Clustering Users in terms of preferences

1. We used EM clustering algorithm (WEKA implementation) to cluster users in terms of preferences.
Clustering Users in terms of preferences

1. We used EM clustering algorithm (WEKA implementation) to cluster users in terms of preferences.

2. EM can automatically calculate the number of underlying clusters when it is unknown.
We used EM clustering algorithm (WEKA implementation) to cluster users in terms of preferences.

EM can automatically calculate the number of underlying clusters when it is unknown.

The clustering features were the template choices.
We used EM clustering algorithm (WEKA implementation) to cluster users in terms of preferences.

EM can automatically calculate the number of underlying clusters when it is unknown.

The clustering features were the template choices.

It resulted in two clusters: verbose and succinct.
## Cluster 1

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Template</th>
<th>Breathing Rate</th>
<th>( \text{SpO}_2 )</th>
<th>Heart Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 1:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoke</td>
<td>Average</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Trend verbose</td>
<td>33.3</td>
<td>25.9</td>
<td>3.7</td>
</tr>
<tr>
<td>Inhalation</td>
<td>Trend succinct</td>
<td><strong>51.8</strong></td>
<td><strong>66.6</strong></td>
<td><strong>92.5</strong></td>
</tr>
<tr>
<td>BR: incr</td>
<td>Range verbose</td>
<td>0</td>
<td>3.7</td>
<td>3.7</td>
</tr>
<tr>
<td>SpO\textsubscript{2}: decr</td>
<td>Range succinct</td>
<td>1.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HR: incr</td>
<td>Inference</td>
<td>0</td>
<td>3.7</td>
<td>0</td>
</tr>
<tr>
<td><strong>Scenario 2:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drowning</td>
<td>Average</td>
<td>24.0</td>
<td>20.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Trend verbose</td>
<td>0</td>
<td>4.0</td>
<td>16.0</td>
</tr>
<tr>
<td>BR: stable</td>
<td>Trend succinct</td>
<td>8.0</td>
<td>12.0</td>
<td><strong>72.0</strong></td>
</tr>
<tr>
<td>SpO\textsubscript{2}: stable</td>
<td>Range verbose</td>
<td>16.0</td>
<td>24.0</td>
<td>0</td>
</tr>
<tr>
<td>HR: incr</td>
<td>Range succinct</td>
<td><strong>40.0</strong></td>
<td><strong>36.0</strong></td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>Inference</td>
<td>12.0</td>
<td>4.0</td>
<td>8.0</td>
</tr>
<tr>
<td><strong>Scenario 3:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall down</td>
<td>Average</td>
<td>16.6</td>
<td>12.5</td>
<td>0</td>
</tr>
<tr>
<td>Stairs</td>
<td>Trend verbose</td>
<td>4.1</td>
<td>0</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>Trend succinct</td>
<td><strong>37.5</strong></td>
<td><strong>12.5</strong></td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>BR: stable</td>
<td>Range verbose</td>
<td>8.3</td>
<td>25.0</td>
<td>0</td>
</tr>
<tr>
<td>SpO\textsubscript{2}: stable</td>
<td>Range succinct</td>
<td>25.0</td>
<td><strong>45.8</strong></td>
<td>4.1</td>
</tr>
<tr>
<td>HR: decr</td>
<td>Inference</td>
<td>8.3</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td><strong>Scenario 4:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B i c y c l e accident</td>
<td>Average</td>
<td>4.1</td>
<td>16.6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Trend verbose</td>
<td>12.5</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>Trend succinct</td>
<td><strong>66.6</strong></td>
<td>25.0</td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>BR: incr</td>
<td>Range verbose</td>
<td>8.3</td>
<td>20.8</td>
<td>0</td>
</tr>
<tr>
<td>SpO\textsubscript{2}: stable</td>
<td>Range succinct</td>
<td>8.3</td>
<td><strong>37.5</strong></td>
<td>4.1</td>
</tr>
<tr>
<td>HR: incr</td>
<td>Inference</td>
<td>0</td>
<td>0</td>
<td>8.3</td>
</tr>
</tbody>
</table>
## Cluster 2

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Template</th>
<th>Breathing Rate</th>
<th>SpO₂</th>
<th>Heart Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 1:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoke</td>
<td>Average</td>
<td>2.3</td>
<td>0</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Trend verbose</td>
<td>93.0</td>
<td>93.0</td>
<td>79.0</td>
</tr>
<tr>
<td></td>
<td>Trend succinct</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BR</td>
<td>Range verbose</td>
<td>2.3</td>
<td>4.6</td>
<td>16.2</td>
</tr>
<tr>
<td>SpO₂</td>
<td>Range succinct</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HR</td>
<td>Inference</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td><strong>Scenario 2:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drowning</td>
<td>Average</td>
<td>34.0</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Trend verbose</td>
<td>4.8</td>
<td>0</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td>Trend succinct</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BR</td>
<td>Range verbose</td>
<td>41.4</td>
<td>58.5</td>
<td>0</td>
</tr>
<tr>
<td>SpO₂</td>
<td>Range succinct</td>
<td>2.4</td>
<td>2.4</td>
<td>0</td>
</tr>
<tr>
<td>HR</td>
<td>Inference</td>
<td>17.0</td>
<td>0</td>
<td>2.4</td>
</tr>
<tr>
<td><strong>Scenario 3:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall down</td>
<td>Average</td>
<td>23.0</td>
<td>38.4</td>
<td>0</td>
</tr>
<tr>
<td>Stairs</td>
<td>Trend verbose</td>
<td>35.8</td>
<td>0</td>
<td>87.1</td>
</tr>
<tr>
<td></td>
<td>Trend succinct</td>
<td>0</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>BR</td>
<td>Range verbose</td>
<td>20.5</td>
<td>58.9</td>
<td>7.6</td>
</tr>
<tr>
<td>SpO₂</td>
<td>Range succinct</td>
<td>2.5</td>
<td>2.5</td>
<td>0</td>
</tr>
<tr>
<td>HR</td>
<td>Inference</td>
<td>17.9</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td><strong>Scenario 4:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicycle accident</td>
<td>Average</td>
<td>10.5</td>
<td>39.4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Trend verbose</td>
<td>73.6</td>
<td>2.6</td>
<td>97.3</td>
</tr>
<tr>
<td></td>
<td>Trend succinct</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BR</td>
<td>Range verbose</td>
<td>15.7</td>
<td>55.2</td>
<td>0</td>
</tr>
<tr>
<td>SpO₂</td>
<td>Range succinct</td>
<td>0</td>
<td>2.6</td>
<td>0</td>
</tr>
<tr>
<td>HR</td>
<td>Inference</td>
<td>0</td>
<td>0</td>
<td>2.6</td>
</tr>
</tbody>
</table>
Hypothesis: A novice user would belong to one of the two clusters (but it’s unknown to us in which one), and therefore, optimising for both clusters simultaneously, the output summary would be more preferable than the output which is optimised for one of the two clusters.
**Framework**

1. Hypothesis: A novice user would belong to one of the two clusters (but it’s unknown to us in which one), and therefore, optimising for both clusters simultaneously, the output summary would be more preferable than the output which is optimised for one of the two clusters.

2. We use multi-objective optimisation with genetic algorithms.
Hypothesis: A novice user would belong to one of the two clusters (but it’s unknown to us in which one), and therefore, optimising for both clusters simultaneously, the output summary would be more preferable than the output which is optimised for one of the two clusters.

We use multi-objective optimisation with genetic algorithms.

Why not just aggregate all data and perform single-objective optimisation?
Hypothesis: A novice user would belong to one of the two clusters (but it’s unknown to us in which one), and therefore, optimising for both clusters simultaneously, the output summary would be more preferable than the output which is optimised for one of the two clusters.

We use multi-objective optimisation with genetic algorithms.

Why not just aggregate all data and perform single-objective optimisation?

We have shown before (Gkatzia et al., 2014) that this does not produce the preferable output because the preferences of the users are averaged and they cancel out each other (verbose vs. succinct).
Fitness Functions

- For each cluster: Preference elicitation through *logistic regression*
- Fitness functions:

\[
Fitness(cluster_n) = \arg \max P(t_{BR_i} \cap t_{SpO_2_i} \cap t_{HR_i})
\]  
(1)
• Every possible summary is encoded as a chromosome, which is essentially a feature vector.

The breathing rate increased from 20 to 30 breaths per minute. The blood oxygen saturation dropped from 95% to 90%. The heart rate increased from 110 to 121 beats per minute.

0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0
Maximum ranking method
• Maximum ranking method
• The initial population is sorted in two lists regarding the fitness functions
Ranking method

- Maximum ranking method
- The initial population is sorted in two lists regarding the fitness functions
- From each list, the eight fittest (i.e. with highest ranking) chromosomes are chosen + two of the twelve least fit chromosomes are chosen randomly in order to increase diversity of the population
Ten chromosomes as described previously.
Reproduction

- Ten chromosomes as described previously.
- Ten new chromosomes via one-point crossover.
Reproduction

- Ten chromosomes as described previously.
- Ten new chromosomes via one-point crossover.
- Five chromosomes are randomly selected from the new population and are mutated (i.e. one gene is changed randomly).
Reproduction

- Ten chromosomes as described previously.
- Ten new chromosomes via one-point crossover
- Five chromosomes are randomly selected from the new population and are mutated (i.e. one gene is changed randomly).
- The same process continues iteratively until the stopping criterion is met, i.e. when there is no improvement in terms of fitness of the top (most optimal) chromosome.
Ten chromosomes as described previously.

Ten new chromosomes via one-point crossover

Five chromosomes are randomly selected from the new population and are mutated (i.e. one gene is changed randomly).

The same process continues iteratively until the stopping criterion is met, i.e. when there is no improvement in terms of fitness of the top (most optimal) chromosome.

These parameters for reproduction are determined by an extensive trial and error process.
Choice of optimal solution (knee approach)

The optimal solutions from both populations are plotted in a graph in order to choose one single solution.
The output of the multi-objective optimisation system is compared against two meaningful baselines:

1) Cluster1-based
2) Cluster2-based

21 participants were recruited. All participants were unseen, i.e. they were not involved in the previous data collection. Each participant was shown an emergency scenario and a corresponding summary of the time-series data which was produced by one of the systems. The participants were asked to rate the summary in a 5-point Likert scale. Each participant repeated the process for 3 times and they were able to stop whenever they liked. 63 ratings were collected.
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  - 1) **Cluster1-based**
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  - 1) **Cluster1-based**
  - 2) **Cluster2-based**
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  1. **Cluster1-based**
  2. **Cluster2-based**
- 21 participants were recruited
- All participants were **unseen**, i.e. they were not involved in the previous data collection.
- Each participant was shown an emergency scenario and a corresponding summary of the time-series data which was produced by one of the systems.
- The participants were asked to rate the summary in a 5-point Likert scale.
- Each participant repeated the process for 3 times and they were able to stop whenever they liked.
- 63 ratings were collected.
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  1) **Cluster1-based**
  2) **Cluster2-based**
- 21 participants were recruited
- All participants were *unseen*, i.e. they were not involved in the previous data collection.
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  1) **Cluster1-based**
  2) **Cluster2-based**

- 21 participants were recruited
- All participants were *unseen*, i.e. they were not involved in the previous data collection.
- Each participant was shown an emergency scenario and a corresponding summary of the time-series data which was produced by one of the systems.
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  1) **Cluster1-based**
  2) **Cluster2-based**
- 21 participants were recruited
- All participants were *unseen*, i.e. they were not involved in the previous data collection.
- Each participant was shown an emergency scenario and a corresponding summary of the time-series data which was produced by one of the systems.
- The participants were asked to rate the summary in 5-point Likert scale.
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  1) **Cluster1-based**
  2) **Cluster2-based**
- 21 participants were recruited
- All participants were *unseen*, i.e. they were not involved in the previous data collection.
- Each participant was shown an emergency scenario and a corresponding summary of the time-series data which was produced by one of the systems.
- The participants were asked to rate the summary in 5-point Likert scale.
- Each participant repeated the process for 3 times and they were able to stop whenever they liked.
Evaluation with potential users

- The output of the multi-objective optimisation system is compared against two meaningful baselines:
  1) **Cluster1-based**
  2) **Cluster2-based**

- 21 participants were recruited
- All participants were **unseen**, i.e. they were not involved in the previous data collection.
- Each participant was shown an emergency scenario and a corresponding summary of the time-series data which was produced by one of the systems.
- The participants were asked to rate the summary in 5-point Likert scale.
- Each participant repeated the process for 3 times and they were able to stop whenever they liked.
- 63 ratings were collected.
**Results**

<table>
<thead>
<tr>
<th>System</th>
<th>Mean</th>
<th>Mode</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOO</td>
<td>3.75</td>
<td>4</td>
<td>0.89</td>
</tr>
<tr>
<td>Cluster1-based</td>
<td>2.9</td>
<td>2</td>
<td>1.17</td>
</tr>
<tr>
<td>Cluster2-based</td>
<td>3.22</td>
<td>4</td>
<td>1.34</td>
</tr>
</tbody>
</table>

**Table**: Mean, mode and standard deviation of user ratings. Significance with Mann Whitney U test in a pair-wise comparison is given in the following table.

<table>
<thead>
<tr>
<th>Systems</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOO vs. Cluster1-based</td>
<td>0.012*</td>
</tr>
<tr>
<td>MOO vs. Cluster2-based</td>
<td>0.24</td>
</tr>
<tr>
<td>Cluster1-based vs. Cluster2-based</td>
<td>0.356</td>
</tr>
</tbody>
</table>

**Table**: Significance (Mann Whitney U test) in a pair-wise comparison at $p<0.05$ is indicated as *.
Summary

- Data collection for a decision support system for medical emergencies
- Development of a multi-objective approach to data-to-text generation
- Contribution: Accounting for first-time users.

Future Work:
- Compare this approach with standard multi-objective optimisation algorithms such as NSGA.
- Transfer this approach to other domains.
Thank you!
References

Generating Verbal Descriptions from Medical Sensor Data: A Corpus Study on User Preferences.

*BCS HIS*

Dimitra Gkatzia, Helen Hastie and Oliver Lemon (2014)
Finding Middle Ground? Multi-objective Natural Language Generation from time-series data.

*EACL*

Dimitra Gkatzia, Oliver Lemon and Verena Rieser (2016)
Natural Language Generation enhances human decision-making with uncertain information

*ACL*

Ehud Reiter (2007)
An architecture for Data-to-text systems

*ENLG*
When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care

*Applied Cognitive Psychology, 24:77 89*