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

Dimitra Gkatzia
Institute of Informatics and Digital Innovation
School of Computing
Edinburgh Napier University
Edinburgh, UK, EH10 5DT
D.Gkatzia@napier.ac.uk

Verena Rieser and Oliver Lemon
Interaction Lab
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, UK, EH14 4AS
{v.t.rieser,o.lemon}@hw.ac.uk

***Abstract*—We propose a novel approach for handling *first-time* users in the context of automatic report generation from time-series data in the health domain. Handling first-time users is a common problem for Natural Language Generation (NLG) and interactive systems in general - the system cannot adapt to users without prior interaction or user knowledge. In this paper, we propose a novel framework for generating medical reports for first-time users, using multi-objective optimisation (MOO) to account for the preferences of multiple possible user types, where the content preferences of potential users are modelled as objective functions. Our proposed approach outperforms two meaningful baselines in an evaluation with prospective users, yielding large ($= .79$) and medium ($= .46$) effect sizes respectively.**

I. Introduction

First aid provision is often dependent on the use of sensors that measure physiological conditions. In this paper, we consider the use of such sensors in the context of a decision support system that employs data-to-text technology to assist in first aid provision. *Data-to-text generation* is the subfield of Natural Language Generation (NLG) that is concerned with the task of automatically generating text from non-linguistic data such as sensor data [1]. Previous studies have demonstrated that text descriptions can be more effective, understandable and helpful in decision making than graphical representations of data [2][3]. In addition, previous work has shown that it is important for these types of NLG systems to adapt their output to specific users or user groups, such as nurses and patients [4], or lecturers and students [5]. Handling first-time users becomes problematic when no prior information about a user exists. Furthermore, user preferences often vary significantly for the same utterance [6], which will naturally affect the performance of models based on population average. In addition, it is demonstrated that predictive models based on average population of users lead to worse rating predictions than models that are based on user groups or clusters of users [7].

In this paper, we propose a novel model for addressing *first-time users*, which is based on clusters of potential user types. Each clusters' preferences are modelled as an objective function. The derived functions are then optimised simultaneously using a multi-objective optimisation approach. Generally, learning algorithms can be divided into two categories: single-objective learning algorithms and multi-objective learning algorithms. Multi-objective optimisation can be applied to situations where optimal decisions are sought in the presence of trade-offs between conflicting objectives [8].

We consider the task of automatically generating short summaries from physiological time-series sensor data (Breathing Rate - BR, Blood Oxygen Saturation - SpO_2 and Heart Rate - HR) in the context of a first aid decision support system. In a medical emergency, a patient's survival often depends upon the prompt response and appropriate first aid provided by the first person on scene, also known as "bystander", who typically is a first-time user of such a decision support system. Therefore, in this paper, we explore handling first-time users, i.e. users with unknown preferences and background.

Our contributions to the field are as follows: we present a novel and efficient method for tackling the challenge of content selection using a multi-objective optimisation approach; we effectively account for first-time users; we apply our approach to the health domain; we present a comparison with a single-optimisation technique, and we discuss the similarities and differences with previous work.

In the next section, we refer to the related work on content selection from time-series data and user adaptation approaches. In Section III, we provide a description of the corpus used. In Section IV, we describe the user clustering analysis and the multi-objective optimisation approach. In Section V, we present the evaluation setup and in Section VI, we discuss the results obtained from prospective first-time users. Finally, in Section VII, we conclude the reported work and in Section VIII, we offer directions for future work.

II. Related Work

The work presented in this paper relates to several areas of NLG, evolutionary algorithms for NLG and user modelling. We review related work in these areas below.

a) Natural Language Generation:: Natural Language Generation from time-series data has been investigated for various tasks, including, but not limited to, weather forecast generation [9], [10], [11], sportcasting [12], [13], [14], narrative generation to assist children with communication needs

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.

Fig. 1: An example scenario from the dataset.

[15], student feedback generation [16]and report generation from students data[17] and report generation from medical time-series data [18][19][20][4].

The important tasks of report generation systems from time-series data are *content selection* (what to say), *surface realisation* (how to say it) and *information presentation* (Document Planning, Ordering etc.). In this work, we concentrate on content selection. There are several approaches to content selection which have previously studied including supervised learning [14], [16], unsupervised learning [12], Reinforcement Learning [21], multi-objective optimisation [5], Gricean Maxims [22], Integer Linear Programming [23], interest scores assigned to content [24], statistical approaches [25], a combination of statistical and template-based approaches [26] and many more. None of the aforementioned approaches uses multi-objective optimisation, with the exception of Gkatzia et al. [5].

The multi-objective optimisation approach by Gkatzia et al. [5] is similar to our proposed method in that we frame content selection as an optimisation task with two objective functions. The two methods differ in two ways. Firstly, the Gkatzia et al.'s approach takes into account *known* users, whereas our approach chooses content for *unknown first-time* users. Secondly, their model is based on function aggregation, which leads to sub-optimal solutions, as conflicting user preferences are smoothing out each other. Our model is therefore improved in that it overcomes this issue, by generating a set of optimal solutions (the Pareto set), which are then scored using both objective functions, i.e. the solution that scores best with both functions is chosen.

b) Evolutionary Algorithms and Multi-objective Optimisation for NLG: As discussed above, Gkatzia et al. [5] suggest a multi-objective approach to Natural Language Generation using Reinforcement Learning, where the reward function was the aggregated sum of two objective functions. However, the standard way of solving a multi-objective task is through generating a set of optimal solutions using genetic algorithms [8]. Although Multi-objective optimisation is novel for content selection, the use of genetic algorithms is not new to content selection. For instance, Duboue [27] present a content planner that uses genetic algorithms with a fitness function derived from available corpora. Genetic algorithms have been also applied for poem generation, e.g. [28].

c) Fuzzy Sets and Systems for Natural Language Generation: Fuzzy sets and systems aim to bridge the gap between uncertainties in data and their corresponding linguistic interpretation. There are several lines of research aiming to address this problem. For example, previous work has focused on addressing uncertainty in time-series data [17] and temporal uncertainty [29]. Our study addresses uncertainty derived from unknown first-time users.

d) User Adaptation: Research in NLG systems has shown the importance of adapting the system output to different user types. For instance, NLG systems employ different versions of a system for each user group [4], [20], [30] or employ User Models (UMs) to adapt their linguistic output to individual users [31], [32], [33].

A different direction has been followed by [34], [35] and [6], where an over-generate and rank approach to sentence generation has been suggested. In this approach, the over-generation phase can follow user- and domain-independent rules to generate a set of possible sentences and the ranking phase is responsible of ranking theal. sentences in accordance to a specific user's ratings and chooses the one that yields highest rating. This approach assumes that the knowledge about the user is already derived through user feedback (ratings) and many user ratings are required. For our domain, acquiring many ratings is not applicable, as we assume that users are first-time users. In addition, averaging existing user ratings is not an option, as Walker et al. also notice that user ratings for the same utterance differ significantly. This variability introduces noise to models that are developed using all available user ratings. To solve this issue, Dethlefs et al. [7] suggest clustering users in terms of similarity ratings and then develop cluster-specific models for predicting user ratings in order to improve accuracy. However, Dethlefs et al.'s approach is able to adapt to users' stylistic preferences only after nine ratings are provided, which means that for a first-time user the output will not be optimal. In our domain, we deal with first-time users with no prior ratings. In contrast to Dethlefs et als' approach, our approach suggests that optimising for all available clusters simultaneously using a multi-objective optimisation approach will meet users' expectations, because a first-time user will belong to one of the available clusters.

Raw Data - measurements per minute

factors	1'	2'	...	n'
Breathing rate	20	21	...	30
Blood Oxygen Saturation	95	95	...	90
Heart Rate	110	115	...	121

Trends from Data

factors	trend
Breathing rate	trend_increasing
Blood Oxygen Saturation	trend_decreasing
Heart Rate	trend_increasing

Example Summary 1

The breathing rate was 24 breaths per minute **on average**. Blood oxygen saturation **decreased** from 95% to 90% per minute. The heart rate **increased** from 110 to 121 beats per minute.

Example Summary 2

Resps ↑ from 20 to 30. SATS ↓ from 95% to 90%. Heart rate ↑ from 110 to 121.

TABLE I: The table on the top left shows an example of the time-series raw data and their measurements per minute. The table on the bottom left shows an example of described trends. The box on the right present potential summaries observed in the data. We see that some users prefer lengthy descriptions, whereas others prefer succinct.

Initially, we cluster potential users based on existing data (Section III). Second, we model group preferences using logistic regression in order to find a "middle ground" (i.e. a solution acceptable to all user groups) using multi-objective optimisation (MOO) (Section IV). Third, we evaluate this model with previously unknown users, i.e. different from the population in the initial data set (Section V). The results in Section VI show that joint optimisation significantly improves over models optimised for one user group only, and hence show the effectiveness of our approach.

III. Data

For our study, we used a previously obtained dataset which is described in [36]. The dataset consists of 280 instances of aligned sensor data from first aid scenarios (as the one shown in Figure 1) with the corresponding textual descriptions (each user provided preferences for four different first aid scenarios). The sensor data is numerical time-series data which measure three physiological conditions: breathing rate; blood oxygen saturation; and heart rate. On the top left of Table I an example of the time-series data can be seen.

The textual summaries were collected online from 70 participants with various levels of expertise, ranging from medical doctors to people with no prior experience or training in first aid provision. Therefore, there is variability in obtained summaries, which was found to correspond to different preferences rather than background, knowledge or occupation [36]. Due to this variation of preferences, the same time-series data can result in two very different summaries as it can be seen from the two examples summaries on the right of the Table I. The textual descriptions are based on templates (similar to [16]) which describe the time-series data in the following six ways:

1) <average>: referring to the average (e.g. "The breathing rate was 24 breaths per minute on average"),
2) <trend-verbose>: referring to the trend in a verbose way (e.g. "Blood oxygen saturation decreased from 95% to 90% per minute"),
3) <trend-succinct>: referring to the trend in a succinct way (e.g. "SATS ↓ from 95% to 90%"),
4) <range-verbose>: referring to the range of values observed in a verbose way (e.g. "The breathing rate was between 20 and 30 breaths per minute"),
5) <range-succinct>: referring to the range of values observed in a succinct way (e.g. "Resps 20-30"), and
6) <inference>: inference from the data (e.g. "The breathing rate is normal").

Table II show the bespoke templates in detail. Participants selected templates according to their preferences, thus the template choices correspond to the preferred content of each user. The next section will describe the overall methodology in detail.

IV. Methodology

The methodology is based on two assumptions, which are common for NLG and interactive systems: Firstly, we assume the presence of two or more user groups, where optimal generation decisions need to be made in the presence of trade-offs between two or more conflicting objectives (in our study, the objectives are the user preferences), such as report length and reading time. For example, some users prefer lengthy reports, whereas other users prefer succinct reports. In the former case, users will require more time to read the report than in the latter case. Secondly, we assume the need to generate for *unknown* first-time users, which could fall in any of the specified clusters. As such and given the data described in the previous section, we follow a four-step methodology as depicted in Figure 2:

1) We partition users based on their utterance choices, such that users with similar preferences belong to the same cluster, similar to Dethlefs et al. [7] (Section IV-A).
2) For each cluster, we derive an objective function based on the participants' preferences in this cluster using logistic regression (Section IV-B).
3) The derived objective functions are then used in a multi-objective optimisation (MOO) framework in order to derive a solution (textual description) that is preferable by all clusters of users (Section IV-C).
4) This framework outputs a set of optimal solutions, known as a *Pareto set*, rather than a single solution. All solutions

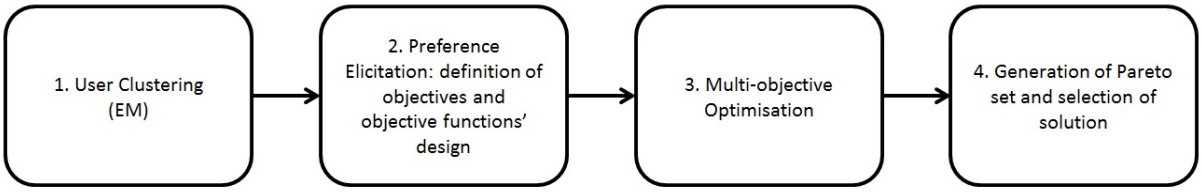

Fig. 2: Methodology

Reference	Breathing Rate (BR)	Blood Oxygen Saturation (SpO_2)	Heart Rate (HR)
(1) average	The breathing rate was <average>breaths per minute on average.	Blood oxygen saturation remained <average >% per minute.	The heart rate was <average>beats per minute on average.
(2) trend_verbose	The breathing rate <trend>from <initialMeasBr>to <finalMeasBr>breaths per minute.	The Blood oxygen saturation <trend>from <initialMeasSpo2>% to <finalMeasSpo2>%.	The heart rate <trend>from <initialMeasHR>to <finalMeasHR>beats per minute.
(3) trend_succinct	Resps <trendsuccinct>from <initialMeasBr>to <finalMeasBr>.	SATS <trendsuccinct>from <initialMeasSpo2>% to <finalMeasSpo2>%.	Heart rate <trendsuccinct>from <initialMeasHR>to <finalMeasHR>.
(4) range_verbose	The breathing rate was between <lowestBr>and <highestBr>breaths per minute.	The Blood oxygen saturation was between <lowestSpo2>% and <highestSpo2>%.	The heart rate was between <lowestHR>and <highestHR>beats per minute.
(5) range_succinct	Resps <lowestBr>-<highestBr>.	SpO2 <lowestSpo2>% -<highestSpo2>%.	Heart rate <lowestHR>-<highestHR>.
(6) inference	Breathing rate <inference>.	Blood oxygen saturation <inference>.	Heart rate observation <inference>.

TABLE II: The templates for the scenario in Figure 1.

are ranked by the available objective functions which facilitates the selection of one joint solution, as we describe in Section IV-D.

A. User Clustering

Previous results from a corpus study [36] showed that individual user characteristics, such as medical training level, gender, or experience with medical sensor data, do not have a significant effect on the template choice. Therefore, we conclude that categorizing users depending on these personal background factors will not necessarily yield distinctive user groups for NLG. For instance, users that have received training at work can have similar preferences in terms of content choice to medical doctors. We therefore consider automatic clustering to define user groups in terms of phrase choice, regardless of their training background, gender, or experience with sensors.

Cluster analysis groups a set of objects in such a way that objects in the same cluster are more similar (here in terms of their phrase choices) to each other than to those in other clusters [37]. For instance, people that prefer referring to the average value of time-series are more similar and thus they belong to the same cluster, whereas people that prefer to refer to the trend in a verbose way belong to a different cluster. In this way, users are grouped according to their preferences and regardless of their profession, gender, or level of training.

As discussed in Section II, Dethlefs et al. [7] show that predictive models based on groups of users result in more accurate rating predictions than models based on individual users. However, Dethlefs et al.'s approach addresses *known* users, in the sense that the user preferences are defined via previous ratings on generated text. In contrast, here, we deal with *unknown* first-time users (bystanders) and therefore, assigning a user into a group is not possible. In addition, Dethlefs et al.'s model solely addresses surface realisation, whereas we extend this to content selection.

We apply Expectation-Maximization (EM) clustering using the implementation provided by the WEKA toolkit [38]. EM is useful when the number of the clusters is *unknown* (or generally not obvious), as in our dataset. Consequently, we need a clustering algorithm that is able to *determine the number of clusters automatically*. EM[1] initially assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. It uses cross validation to determine the number of clusters following five steps: (1) set the number of clusters to 1; (2) split the training set into 10 folds randomly; (3) EM algorithm is applied 10 times as normally in cross validation; (4) average the log likelihood over all 10 results; and (5) if log likelihood has increased, the number of clusters is also increased by 1 and it repeats until convergence is achieved.

The clustering task is formulated as follows: given the

[1]A detailed description is given here: http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html.

feature vector of the time-series data and the template choices of a participant, assign him/her into a cluster. Each feature vector corresponds to a user and it is of the following format:

$$\{user_n, scenario_1, ...scenario_4\}$$

where,
$scenario_n = \{data_{BR}, data_{SpO_2}, data_{HR}, templates_{BR}, templates_{SpO_2}, templates_{HR}\}$

For each first-aid scenario, $data_{BR}$, $data_{SpO_2}$ and $data_{HR}$ correspond to the time-series data of the physiological factors and $templates_{BR}$, $templates_{SpO_2}$ and $templates_{HR}$ correspond to the templates chosen by the participant n. We remind the reader that there were four first-aid scenarios in our dataset.

The EM clustering results in two consistent user groups, where the first cluster consists of 27 participants and the second consists of 43 participants. We use a χ^2 test to validate the consistency of the clusters in terms of the scenarios and the template choices ($p < 0.05$). We observe that the users are grouped in terms of their preferences on the length of the descriptions. In particular, users in Cluster 1 prefer the succinct ways of referring to data whereas users in Cluster 2 prefer the verbose ways.

B. Preference Elicitation

Having defined the clusters, the next step is to acquire a preference function for each cluster. We use *iterated logistic regression* to estimate the probability of each template to be selected given all the previous choices. Previous work used *linear* regression in order to derive a model that can predict user ratings [39], [40]. Linear regression, however, assumes that there is linear relationship between the dependent and the independent variables, which is not true for our domain, where the aim is to predict a binary outcome (i.e. the probability of a template being selected or not). Therefore, we use logistic regression with maximum likelihood, which allows us to calculate the probability of an event/decision occurring. That is, each logistic regression model estimates the probability distribution of a specific template to be chosen for generation given the time-series data and the previously selected content. In this way, the combination of content is taken into account.

C. Content Selection as a Multi-objective Optimization Task

As discussed in Section II, *genetic algorithms* are used to solve this Multi-Objective Optimisation problem. A genetic algorithm designed for MOO consists of (a) *a fitness function*, which is essentially the objective to be optimised (in the case of Multi-objective Optimization there are two or more fitness functions); (b) *a population of chromosomes*, which is a set of solutions, (c) *a ranking method*, which determines which chromosomes are selected for reproduction, and (d) *genetic operators*, which determine how the population evolves through mutation and/or crossover.

(a) Objective or Fitness Functions: For each of the user clusters, we use logistic regression to model the user preferences described in Section IV-B. That is, the multiple logistic regressions are used to calculate the probability of each template to be selected within one cluster, given any previous decisions/preferences expressed (when applicable). For instance, if the first decision was to select a template that describes BR, then this decision will influence the template chosen for the next measurement, which is SpO_2. Therefore, our models consider the previously made decisions when estimating the content selection probabilities. Similarly, the regression model for the HR decision, includes both the decisions made for the BR and SpO_2 template. This approach is motivated by a previous corpus analysis, which found that current decisions about the content are influenced by the previous decisions [36].

Note that the order of describing measurements is fixed, i.e. first BR, then SpO_2, and finally HR. This ordering is standard in first aid, as the first aider would initially check the casualty's breathing rate, then the oxygen levels and finally the heart rate. Moreover, values for BR and SpO_2 are strongly related, and therefore should be mentioned in conjunction. However, our model *can also account for content selection and information ordering simultaneously*, as the fitness function is based on probabilities of content to be chosen, given previously selected content. In this regard, our approach is an improvement of a previously used approach [5], which only considers content selection.

The fitness function can thus be formulated as the joint probability of three templates occurring together, where each logistic function predicts the likelihood of one template being chosen for one of the three measurements, see Equation 1.

$$Fitness(cluster_n) = \arg\max P(t_{BR_i} \cap t_{SpO_{2_j}} \cap t_{HR_k}) \quad (1)$$

where n stands for the cluster ID (and can take values from $N = \{1, 2\}$) and i, j and k are the template ids (and can take values from $T = \{1, 2, ..., 6\}$).

(b) Encoding Population: To form a population, every possible summary is encoded as a *chromosome*, which is essentially a feature vector. Chromosomes are encoded as a vector of 18 features: each chromosome consists of 3 genes corresponding to BR, SpO_2 and HR, and each gene consists of 6 features, each one describing a template type. For example, the following summary can be represented as the chromosome below:

Summary:

> 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.

Chromosome:

$$\{0,1,0,0,0,0\},\{0,1,0,0,0,0\},\{0,1,0,0,0,0\}$$

The initial population is randomly generated and at a size of 20 chromosomes (which was determined by extensive trial and error).

(c) Ranking method: We use a *maximum ranking* method to rank the chromosomes in the population [41]. The initial population was sorted in two lists regarding the fitness functions, in order to keep a population with constant size. From each list, the eight fittest (i.e with highest ranking) chromosomes are chosen. In addition, two of the twelve least fit chromosomes were also chosen at random in order to increase diversity of the population.

(d) Reproduction: We then choose ten chromosomes (parents) for reproduction, i.e. the eight highest ranking chromosomes and two random chromosomes. Ten new chromosomes are reproduced via *one-point crossover*. One-point crossover is the process where a single crossover point is chosen and all the data before this point adopt the genes from the first parent and beyond this point from the second parent and vice versa (so as two chromosomes are reproduced by a pair of parents). Then, 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.

D. Choice of Optimal Solution

For a non-trivial multi-objective optimisation problem, there is no single solution that simultaneously optimises each objective. In this case, the objective functions are said to be conflicting, and there exists a (possibly infinite) number of *Pareto optimal* solutions. The choice of the unique solution from the Pareto set is based on the *knee*[2] approach [42], [43]. The idea is that the solution located in the knee (when plotting the solutions on a graph) scores well for all objectives (Figure 3).

Our approach finds a solution which *simultaneously* satisfies preferences for both user groups, rather than cancelling each other out, which was the case in the previously proposed multi-objective technique by [5].

[2]This is also known as the elbow approach in other fields.

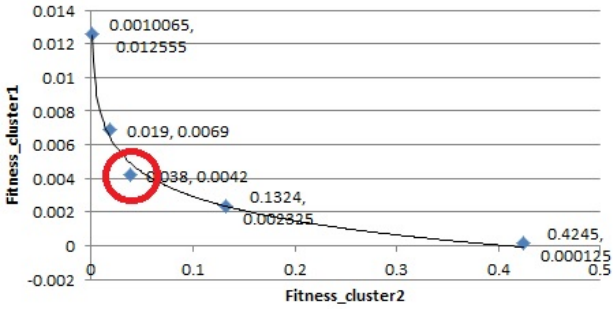

Fig. 3: Chromosomes plotted on a graph. The red circle indicates the knee.

V. Evaluation

In order to evaluate our methodology, the output of the multi-objective optimisation system is compared in a human evaluation with two meaningful baselines:

1) **Cluster1-based** optimises the content for the first cluster (which consists of 27 participants), and
2) **Cluster2-based** is optimised for the second cluster (which consist of 43 participants). This baseline roughly corresponds to the *majority baseline* since most users from the initial data collection were assigned to Cluster 2.

We recruited 21 new participants to perform the evaluations, advertising our study to students, colleagues and professional first responders. These participants were all previously *unseen*, i.e. from a different user population to the one in the initial data collection, and as such, preferences of these participants are *unknown*. Similarly to the setup for the initial data collection, each participant was presented with an emergency scenario and a summary of the time-series data as generated by one of our systems. Each participant was asked to rate the summary on a 5-point Likert scale (*Dislike, Slightly dislike, Neither like or dislike, Like overall/it's ok, Like very much*). The participants repeated this process three times for three different scenarios, collecting a total of 63 ratings. For each scenario, they were presented with a summary generated by a different system. The data will be made publicly available.

VI. Results

Table III shows the mean, mode and standard deviation of the humans' ratings. Results from a pair-wise Mann-Whitney U test (with Bonferroni adjustment) are shown in Table IV along with the effect size (Cohen's d).

System	Mean	Mode	standard deviation
MOO	**3.75**	**4**	**0.89**
Cluster1-based	2.9	2	1.17
Cluster2-based	3.22	**4**	1.34

TABLE III: Mean, mode and standard deviation of user ratings.

The participants rated the output from the MOO system higher than the other two systems. In particular, participants

Systems	p-value	Effect size
MOO vs. Cluster1-based	**0.012***	0.796
MOO vs. Cluster2-based	0.24	0.461
Cluster1 vs. Cluster2	0.356	0.246

TABLE IV: Significance (at $p<0.05$) is indicated as * as determined by a Mann Whitney U test and effect size (Cohen's d) for pair-wise comparison.

significantly preferred the MOO system to Cluster1-based system ($p < 0.05$, $power = 99.8\%$). They also rated the MOO system higher than the Cluster2-based system ($p = 0.24$, $power = 83.4\%$). We also report effect size in order to understand the magnitude of the differences found. The effect size of the differences between the MOO and Cluster1-based is large (≈ 0.8), which indicates reliable results and a strong preference towards the MOO system. There is also a medium effect between the MOO and Cluster2-based system ($= 0.461$). The effect size for Cluster2-based vs. Cluster1-based systems is small ($= 0.246$, $power = 41.4\%$ - note that if we increase α to .2, $power$ increases to 71.1%), which indicates that there is a tendency to rank the Cluster2-based system higher than the Cluster1-based system, however more ratings are needed for safe conclusions.

When observing the standard deviation of the ratings, it is evident that the ratings for the MOO system are more consistent than those for the other two systems. This shows that most users rated the MOO system consistently high. The Cluster2-based system has the highest standard deviation, which indicates that the ratings of this system are more variable, i.e. users in Cluster 1 would rate the Cluster1-based system much lower than the MOO system. Finally, we expect that the MOO method will perform much better in domains with more than two clusters, as users tend to rate the other cluster-based systems significantly lower than the MOO system.

In conclusion, our MOO-based approach is able to satisfy user preferences for *first-time* users, i.e. users a) whose preferences are not available and b) who potentially belong to any of the clusters (but have a higher likelihood of belonging to Cluster 2). In addition, we show that our approach makes three important contributions. Firstly, the proposed approach is able to optimally select content that is highly rated by multiple groups with conflicting preferences. Secondly, it can consider information ordering decisions as well as content selection. Thirdly, our approach can effectively address unknown first-time users, i.e. users that have not provided any ratings or have previously interacted with the system.

VII. Summary

We have shown that multi-objective optimisation is capable of addressing *unknown* first-time users of a data-to-text generation system. In particular, we used a combination of user clustering in terms of preferences and multi-objective optimisation to generate textual descriptions which can satisfy all potential user groups. We applied this framework within a medical decision support system for first aid provision, which is commonly used by first-time users ("bystanders"). It was shown that this approach outperforms single-objective optimisation approaches such as suggested by Gkatzia et al. [5], which adapt to a specific user group. To the best of our knowledge, this is the first approach which can handle *unknown* first-time users, which is an open common problem in many application areas within interactive setups.

VIII. Future Work

In future, we plan to conduct a task-based evaluation to explore whether the generated summaries can indeed support decision making whilst adapting to user preferences. We also plan to evaluate our multi-objective approach on other domains, such as student feedback generation, e.g. [16] and weather forecasts [12], [10], [9]. Furthermore, the above framework can be extended to learn online, i.e. every incoming data point/user rating can be used to re-cluster user groups (until convergence) and thus get a more accurate estimate of the underlying distribution of user preferences in this domain.

Finally, we are also interested in looking into generating descriptions of underlying data uncertainties (e.g. from sensor failures) and (medical) risks or data associated with probabilities, such as weather data. While there is considerable uncertainty arising from the underlying data, data-to-text systems fall short of communicating this uncertainty to the decision maker. Related research in decision support systems stresses the importance of conveying uncertainty using natural language in the decision making process, e.g. [44], [45].

Acknowledgment

This research received funding from the EPSRC GUI project Generation for Uncertain Information (EP/L026775/1).

References

[1] E. Reiter, "An architecture for data-to-text systems," in *11th European Workshop on Natural Language Generation (ENLG)*, 2007.

[2] A. S. Law, Y. Freer, J. Hunter, R. H. Logie, N. McIntosh, and J. Quinn, "A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit," *Journal of Clinical Monitoring and Computing*, pp. 19: 183–194, 2005.

[3] M. van den Meulen, R. Logie, Y. Freer, C. Sykes, N. McIntosh, and J. Hunter, "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," in *Applied Cognitive Psychology, 24: 77-89*, 2010.

[4] A. Gatt, F. Portet, E. Reiter, J. Hunter, S. Mahamood, W. Moncur, and S. Sripada, "From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management," *AI Communications*, vol. 22: 153-186, 2009.

[5] D. Gkatzia, H. Hastie, and O. Lemon, "Finding middle ground? Multi-objective natural language generation from time-series data." in *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014.

[6] M. Walker, A. Stent, F. Mairesse, and R. Prasad, "Individual and domain adaptation in sentence planning for dialogue," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 413–456, 2007.

[7] N. Dethlefs, H. Cuayahuitl, H. Hastie, V. Rieser, and O. Lemon, "Cluster-based prediction of user ratings for stylistic surface realisation," in *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014.

[8] K. Deb, *Multi-objective Optimization using Evolutionary Algorithms*. Wiley, 2001.

[9] A. Belz and E. Kow, "Extracting parallel fragments from comparable corpora for data-to-text generation," in *6th International Natural Language Generation Conference (INLG)*, 2010.

[10] G. Angeli, P. Liang, and D. Klein, "A simple domain-independent probabilistic approach to generation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

[11] S. Sripada, E. Reiter, I. Davy, and K. Nilssen, "Lessons from deploying NLG technology for marine weather forecast text generation," in *PAIS session of ECAI-2004:760-764*, 2004.

[12] I. Konstas and M. Lapata, "Unsupervised concept-to-text generation with hypergraphs," in *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2012.

[13] D. Chen, J. Kim, and R. Mooney, "Training a Multilingual Sportscaster: Using Perceptual Context to Learn Language," *Artificial Intelligence Research (JAIR)*, vol. 37, pp. 397 – 435, 2010.

[14] R. Barzilay and M. Lapata, "Collective content selection for concept-to-text generation," in *Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT - EMNLP)*, 2005.

[15] R. Black, J. Reddington, E. Reiter, N. Tintarev, and A. Waller, "Using NLG and sensors to support personal narrative for children with complex communication needs," in *NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, 2010.

[16] D. Gkatzia, H. Hastie, and O. Lemon, "Comparing Multi-label classification with Reinforcement Learning for Summarisation of Time-series data," in *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014.

[17] A. Ramos-Soto, A. Bugarin, and S. Barro, "On the role of linguistic descriptions of data in the building of natural language generation systems," *Fuzzy Sets and Systems*, vol. 285, pp. 31–51, 2016.

[18] A. Schneider, A. Mort, C. Mellish, E. Reiter, P. Wilson, and P.-L. Vaudry, *Proceedings of the 14th European Workshop on Natural Language Generation*. Association for Computational Linguistics, 2013, ch. MIME- NLG Support for Complex and Unstable Prehospital Emergencies, pp. 198–199. [Online]. Available: http://aclweb.org/anthology/W13-2128

[19] A. Wilbik, J. Keller, and J. Bezdek, "Linguistic prototypes for data from eldercare residents," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 1, pp. 110–123, 2014.

[20] J. Hunter, Y. Freer, A. Gatt, Y. Sripada, C. Sykes, and D. Westwater, "Btnurse: Computer generation of natural language shift summaries from complex heterogeneous medical data," *American Medical Informatics Association*, vol. 18(5):621-624, 2011.

[21] V. Rieser, O. Lemon, and X. Liu, "Optimising information presentation for spoken dialogue systems," in *48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

[22] S. Sripada, E. Reiter, J. Hunter, and J. Yu, "Generating english summaries of time series data using the gricean maxims," in *9th ACM international conference on Knowledge discovery and data mining (SIGKDD)*, 2003.

[23] G. Lampouras and I. Androutsopoulos, "Using integer linear programming in concept-to-text generation to produce more compact texts," in *51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.

[24] I. Androutsopoulos, G. Lampouras, and D. Galanis, "Generating Natural Language Descriptions from OWL Ontologies: the Natural OWL System," *Atrificial Intelligence Research*, vol. 48, pp. 671–715, 2013.

[25] P. Duboue and K. McKeown, "Statistical acquisition of content selection rules for natural language generation," in *Conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP)*, 2003.

[26] R. Kondadadi, B. Howald, and F. Schilder, "A statistical nlg framework for aggregated planning and realization," in *51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.

[27] P. Duboue and K. R. McKeown, "Content planner construction via evolutionary algorithms and a coprus-based fitness function," in *2nd International Natural Language Generation Conference (INLG)*, 2002.

[28] R. Manurung, G. Ritchie, and H. Thompson, "Using Genetic Algorithms to Create Meaningful Poetic Text," *Experimental Theoretical Artificial Intelligence*, vol. 24, no. 1, pp. 43–64, 2012.

[29] A. Gatt and F. Portet, "Multilingual generation of uncertain temporal expressions from data: A study of a possibilistic formalism and its consistency with human subjective evaluations," *Fuzzy Sets and Systems - Special Issue on Linguistic Description of Time Series*, vol. 285, pp. 73 – 93, 2016.

[30] S. Mahamood and E. Reiter, "Generating affective natural language for parents of neonatal infants," in *13th European Workshop on Natural Language Generation (ENLG)*, 2011.

[31] S. Janarthanam and O. Lemon, "Adaptive Generation in Dialogue Systems Using Dynamic User Modeling," *Computational Linguistics*, vol. 40, no. 4, pp. 883 – 920, 2014.

[32] C. A. Thompson, M. H. Goker, and P. Langley, "A personalised system for conversational recommendations," *Journal of Artificial Intelligence Research*, vol. 21, no. 1, 2004.

[33] I. Zukerman and D. Litman, "Natural language processing and user modeling: Synergies and limitations," *User Modeling and User-Adapted Interaction*, vol. 11, no. 2, 2001.

[34] O. Rambow, M. Rogati, and M. Walker, "Evaluating a Trainable Sentence Planner for a Spoken Dialogue System," in *39th Meeting of the Association for Computational Linguistics (ACL)*, 2001.

[35] M. Walker, O. Rambow, and M. Rogati, "SPoT: A trainable sentence planner," in *2nd meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL-HLT)*, 2001.

[36] D. Gkatzia, V. Rieser, A. McSporran, A. McGowan, A. Mort, and M. Dewar, "Generating verbal descriptions from medical sensor data: A corpus study on user preferences." in *BCS Health Informatics Scotland*, 2014.

[37] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264 – 323, 1999.

[38] I. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques." Morgan Kaufmann Publishers, 2005.

[39] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with paradise," *Natural Language Engineering*, vol. 6, no. 3, pp. 363 – 377, 2000.

[40] V. Rieser and O. Lemon, *Reinforcement Learning for Adaptive Dialogue Systems: A Data-driven Methodology for Dialogue Management and Natural Language Generation.* Theory and Applications of Natural Language Processing, Springer, 2011.

[41] D. J. Schaffer, "Multiple objective optimizatio using nondominated sorting in genetic algorithms," *Ecolutionary Computation*, vol. 2, no. 3, pp. 221–248, 1985.

[42] J. Branke, K. Deb, H. Dierolf, and M. Osswald, "Finding knees in multi-objective optimization," *Parallel Problem Solving from Nature - Lecture Notes in Computer Science*, vol. 3242, no. 2004, pp. 7722 – 731, 2004.

[43] J. Handl and J. Knowles, "Modes of problem solving with multiobjective optimization: Implications for interpreting the pareto set and for decision making," *Multiobjective Problem Solving from Nature. Springer Natural Computing Series, Springer-Verlang*, 2008.

[44] D. Spiegehalter and H. Riesch, "Don't know, can't know: embracing deeper uncertainties when analysing risk," *Phil. Trans. R. Soc. A*, vol. 369, pp. 4730–4750, 2011.

[45] National Research Council (US), *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. The National Academies Press, 2006. [Online]. Available: http://www.nap.edu/openbook.php?record_id=11699