MASTER

Summary generation for data from personal health services

Sylenko, A.Y.

Award date:
2017

Department of Mathematics and Computer Science
Architecture of Information Systems Research Group

Summary Generation for Data from Personal Health Services

Master Thesis

Aryna Sylenko

Supervisors:
Dr. Natalia Sidorova
Niek Tax, MSc.
Dr. Aki Härmä

UNITN Supervisor:
Prof. Fabio Casati

Final version

Eindhoven, July 2017

Abstract

Health is one of the most important aspects of the everyday life, and contemporary technologies allow to create effective ways of treatment and health monitoring. One of such technologies is the Personal Emergency Response Systems (PERS) that support the independent life of elderly people. On the one hand such technologies allow to live healthier and longer life, but on the contrary, this leads to such consequences as an aging population and information overload. One of the ways to tackle these problems is to use the data-to-text systems.

Data-to-text systems are summarization systems that generate a natural language summary out of raw data. They become more and more popular, since they are able to process a vast amount of data, analyze it, and generate a meaningful natural language summary. Application of data-to-text systems on data of PERS systems would allow to get an overview of the health status of a person based on his/her calls to the PERS call center.

This thesis project focuses on exploring how to generate a useful natural language summary based on the data obtained from a Personal Emergency Response System. To explore this, we describe the available NLG and data analysis techniques. Another question to be addressed is what should be the architecture of such a data-to-text system, so it would be able to scale and evolve over the time while supporting changes in a PERS system. To answer this question, the existing architectures are described. Furthermore, the examples of state of the art data-to-text systems are presented. These examples allow to learn more about the application of the described techniques in already existing systems. Next, this work addresses one of the main questions of the data-to-text systems which is what information should be shown in summary. To determine this, we have used user-defined rules, clustering technique and a scoring module.

As a result, we propose a data-to-text system that consists of an analytics stage, a message generation stage, a scoring stage and a summary generation stage. A set of candidate phrases for the summary is generated during the message generation stage based on the information obtained in the analytics stage. Candidate phrases are generated based on user-defined sentence templates. Sentences to be put in the summary are then scored in the scoring module, where each candidate phrase is scored with an importance scoring function that is configurable with domain knowledge. Further, selected phrases from the set of the candidate are used to generate a final summary in the summary generation module.

Finally, the possible evaluation methods of the automatically generated texts are discussed. In this work, the human experts and score analysis were used to evaluate the automatically generated texts. The evaluation has shown that application of data-to-text system on data provided by PERS systems is a promising technology. The majority of human experts found the system helpful to get insights of the health status, and 47% would subscribe to such a service for periodical updates. The main limitation that was found is the readability of the generated summaries.

Keywords: data-to-text, linguistic summarization, event logs summarization, healthcare event logs, Natural Language Generation.

Acknowledgements

First of all, I would like to thank the EIT Digital Master School for the opportunity to study as Service Design and Engineering student in two exceptional universities, University of Trento and Eindhoven University of Technology.

Secondly, I would like to thank my supervisors Natalia Sidorova, Niek Tax and Aki Härmä for all the feedback, help and dedication. This thesis would not have been possible without your support and guidance.

Furthermore, I would like to thank my supervisor from University of Trento Fabio Casati, all other employees from Philips who helped me either with word or deed, and all my friends for their continued support.

A very special thank goes to my family: my parents, my brother, my aunt, my parents in law and other family members. You supported me through the whole period of my life road and you gave me all the opportunities that I have, I am truly grateful that I have such a family!

And the most important thank goes to my partner Andrii Buriachevskyi. You are my constant source of support and encouragement. Every single day you help me to be a better version of myself. I am endlessly thankful for having you in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This master thesis is a concluding part of the Double degree Masters programme - EIT Digital Master School in Service Design and Engineering [53]. It was performed during the internship at Philips Research alongside with the Architecture of Information Systems group at TU/e and the University of Trento.

The introduction consists of four sections. The first and the second sections describe the background and motivate the research. The third section formulates the identified research questions. The following section discusses the research approaches that would be used to answer the research question. The final section introduces the structure of the thesis.

1.1 Problem

"Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" [41]. This is the official definition of health that was adopted by the World Health Organization (WHO) in 1948 and it has not been changed since that time [40]. This state of well-being is extremely important in everyday life, and health is one of the basic needs that allows to live a full and normal life. That is why governments, organizations and individuals spend a huge amount of time, effort, money and other resources aiming to directly or indirectly improve the healthcare situation. This allowed to create new and better ways of treatment and prevention means, in which technology has played an essential role.

Multiple services, wearable devices, and software alongside with hardware have provided opportunities for the healthcare improvement which leads to more sustainable life in general. This progress in medicine allowed to live a longer life and tackle many health related issues. This has two consequences: on the one hand it allows people with health issues to live a longer and healthier life, but on the other hand, it has increased the average life expectancy and this lead to the population aging. It is expected that from 2015 till 2030 the number of people of age 60 years and more will increase by 56 percent and will continue to grow [56], which is represented in Figure 1.1. This also means that the amount of collected data and need for additional services that assists medical personnel will grow proportionally as well.

The aging population is one of the factors that have led to the rising costs of healthcare. Healthcare costs growing faster than an economy and only in U.S. the portion of spendings devoted to healthcare has increased from around 5% in 1960 to 17.9% in 2010 and it is expected that it will grow to around 20% in 2020 [29]. One approach that would allow to cut healthcare costs and to handle the increasing amount of data is to provide technical assistance for various medical purposes where one of the solutions is to use summarization systems which allow to summarize the data generated by various healthcare related systems. Summarized data would support decision-making process and provide a better overview of the collected data. Data summarization could be done using natural language and graphical representation of the information, but some studies showed that in the medical domain data visualization process does not always improve the decision

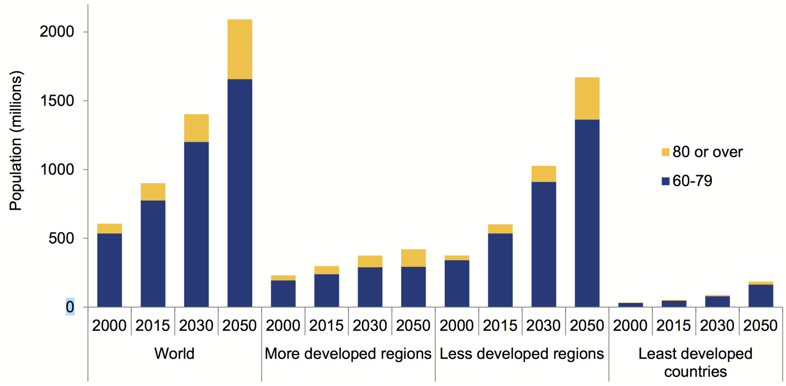

Figure 1.1: Dynamics of population aging [56].

making process [25], but the good quality summary in a natural language could be more effective than the data visualization [33]. Data-to-text systems are summarization systems that generate a natural language summary of the data. They allow to automate manual work of data analysis and writing summaries of the health situation.

Data-to-text systems which use Natural Language Generation techniques and data analysis become more and more popular because they allow to automatically generate summaries from various types of data [46]. They allow to easily deal with the information overload, process data more efficiently and provide the most vital information [19]. With the help of data-to-text systems, it became possible to generate a summary of health-related events that happened with the user [18], so that the caregiver or the person himself would be able to get an overview of the subscriber's current health situation. This summary could support the decisions making process and could help to provide continuous and appropriate care. Significant benefits of such systems are that they allow to process a vast amount of data, analyze it and generate meaningful natural language summary. Moreover, they allow to avoid mistakes that could be done by people. Such systems allow to see a bigger picture while taking less time to generate a summary than manual approach will [25]. Faster, easier and errorless analysis of vast amount of data will support the decision-making process. This will allow to cut costs and time spent on healthcare related procedures.

1.2 Background

This master thesis was written during my internship at Philips Research, where I was developing a system that would use data obtained from the of the Philips Lifeline Personal Emergency Response System (PERS).

Personal Emergency Response Systems are alert systems that allow elderly people to be connected to the call center in case if they face some health related issues or require assistance. Usually, such a system has additional features like fall detection and includes wearable using which elderly can connect to the call center. PERS system connects a user to the call center which stores particular data about each call. Examples of such systems are Philips Lifeline [36], GreatCall solutions that provide safety for seniors [21], medical alert systems and devices for seniors Medical Guardian [22], LifeFone medical alert system with fall detection and GPS [35], Alert1 [8] and many other similar systems. Such systems improve caring practices and assist safe and independent life of seniors. PERS systems were introduced in 1970 [39] and currently are being widely spread and used, serving to approximately 1.4 million users only in the United Kingdom [55]. The US is a particularly big market were only Philips Lifeline works with more than 7 million customers [36]. It is expected that PERS global market will reach 7.4 billion dollars by 2024 [6]. Kathryn Hyer and Linda Rudick found that use of Personal Emergency Response Systems by 48 people allowed to save 1,5 million dollars [27]. Moreover, PERS lead to a significant decrease in hospital admissions

[51] which also allows to save costs.

As long as PERS systems continue to be widely accepted and used, they generate a big amount of data and at the same time allow to save costs by reducing the need for the personal care and hospitalizations. Therefore they are an interesting data source for the data-to-text systems. PERS systems generate a vast amount of data, keeping track on health condition of each user. Collected data could be used to improve the business value of the service and provide more insights to the health situation of the system users while supporting the decision-making process for the caregivers and medical personnel. Data-to-text systems can generate a natural language summary of the subscriber's health situation based on the raw data from calls. An example of such a natural language summary generated manually based on the Lifeline data could be found in Figure 1.2.

CASETYPE	CASEOUTCOME	VL_LONG_TXT	dayssince
Incident	EmerAssistTrans	Dizziness	307
CourtesyCall	Completed	Sub in Hospital	305
Incident	EmerAssistTrans	Pain - Non-Chest	226
Incident	EmerAssistTrans	Bleeding/Injury	137
Incident	EmerAssistNoStatus	Illness	81
Incident	EmerAssistNoTrans	Fell	61
Incident	EmerAssistTrans	See Case Notes	48
Incident	EmerAssistNoTrans	Fell	24
Incident	EmerAssistNoTrans	Fell	6

Ms. X has been in our service for almost 10 months. After the emergency transportation at the beginning of the service subscription, there have been few incidents. However, there have been three fall-related incidents where the emergency assistance was needed. Although no transportation to the hospital was reported in those incidents, based on her age and recent history after the reported illness at the end of June, her estimated risk of hospitalization suggests that the risk is significantly higher than three months ago.

Figure 1.2: Example of a manually generated summary of a health situation of an elderly person based on the Lifeline data.

1.3 Research questions

The general goal of the thesis is to find out if it is possible to get a meaningful and more importantly, useful natural language summary out of the available data from the call centers. This summary should provide an overview of a health situation of the PERS system subscriber, so that caregiver, general practitioner or possibly another interested party could use it.

These could be formulated in the following research question:

- *How to generate a useful natural language summary out of the event logs provided by the call-center of Personal Emergency Response Systems?*

To guide through the process of finding the answer to the main research question the following research subquestions are identified:

- *How to generate a natural language representation of the data generated by Personal Emergency Response Systems?*
- *What should be the architecture of the system that will generate natural language summary?*
- *How to decide which information should be present in the generated summary?*
- *How to evaluate the summary that was generated from data provided by the Personal Emergency Response System?*

1.4 Reseach approach

The main goal of the thesis is to discover how does the meaningful natural language summary could be generated. To find out it and to answer the research questions the following steps were done:

- Literature review

  Continuous literature review throughout the whole process of the thesis development is a vital step that allows to learn more about state of the art data-to-text systems, Natural Language Generation methods, data analysis techniques, etc. Besides that literature review helped to get new ideas, find the ways to implement them and handle all the arising issues.

- Exploratory analysis

  At the very beginning, Philips Research has provided me with data from the Lifeline PERS system. This data contained logs of calls to the Lifeline call center and some data about subscribers. Therefore there was no need for the data collection step, but it was needed to get familiar with the provided data, to learn what data is available and would be potentially used to build a system. The available data was explored to find out what kind of information could be provided in the final summary. One of the means for the exploratory analysis was the creation of first simple prototype within the first couple of weeks. This allowed to explore the available data more precisely.

- Define what information should be present in the summary

  During this step, it was needed to understand what kind of information is necessary for each type of potential readers as well as to find out what information could be obtained from the available data and later on to find an overlap between these two findings. Afterwards, it was needed to decide what analysis techniques should be used to obtain the required information, as long as data provided by Philips contained raw call-center logs, which had to be processed.

- Summary generation

  After the data had been processed and analyzed, it was needed to generate a final summary based on the retrieved insights and parameters of the final text, such as length of the summary, where length is a configurable parameter that defines the final size of the generated summary. Insights were built based on the analyzed data. Text generation included internally developed scoring system. The automatically generated summary will be shown in the developed Graphical User Interface (GUI), so it would be easier to read the text and to control the process of summary generation.

- Get the feedback on the various aspects of the developed system from the domain experts

  Feedback was collected during three types of meetings: weekly meetings with the Philips Research group, meetings with the supervisors from the Eindhoven University of Technology and meetings with other domain experts from Philips that are familiar either with the data of PERS system or with the techniques that are used in the data-to-text systems. Feedback was obtained on various aspects of the thesis work such as developed prototypes, generated text, analysis techniques, etc. This feedback was used to iteratively improve the generated text and the prototype of the system.

- Evaluation of the generated summary

  This is the final step of the thesis which allows to answer to the last research question. It aims at evaluating the generated summary, based on multiple parameters. Moreover, the evaluation will allow to find out the downsides and advantages of the developed system and to learn what could be improved.

1.5 The Structure of the Thesis

The thesis is structured as follows. The second chapter aims to describe the Lifeline project and the task in more details, to describe the available data, types of users and insights, to describe in more details the requirements. The third chapter discusses the state of the art of data-to-text systems, available methods, and approaches which help to build data-to-text systems. Then in the fourth chapter, the chosen approach used to build the developed system will be presented. The implementation part, described in the fifth chapter, shows the process of the implementation and the architecture that was used to build the system alongside with the detailed description of each module. Evaluation chapter presents the results of the system evaluation and the conclusions chapter provides a summary of the performed work specifying whether the research question was answered and goals achieved, also mentioning possible developments of the system.

Chapter 2

Background

This chapter describes the Personal Emergency Response Systems like Lifeline in more details, the data provided by the PERS, and the desired output. It also describes the user groups and their specific requirements.

2.1 Personal Emergency Responce Systems

Personal Emergency Response Systems such as Lifeline, allow seniors and people with special needs to live independently and provide a piece of mind for their caregivers. The PERS described in this work is an alert system with fall detection that allows to connect its users with the call center when assistance required. In case if something happened the person that is subscribed to this system can reach the call center by himself/herself, or the system can make the call automatically. There are three types of calls:

- test calls
- health related calls
- neutral calls

Test calls are aimed to ensure the proper work of the systems, therefore subscribers are asked to make test calls from time to time. Health-related calls take place if the subscriber has any issues with the health. Neutral calls are all other calls which are not related to the health issues. An example of the neutral call is when subscriber requires some non-health related assistance.

2.2 Input data

Lifeline collects particular data about each conversation with the subscribers, so they will be able to provide more personalized service adapted to the needs of each user. Therefore the data required for the thesis was already collected by the Lifeline service.

2.2.1 Data description

Data were provided as separate files in a CSV format, where each file contained information about one subscriber of the Lifeline service. The amount of information about each subscriber varied and could include only a few lines of data as well as more than 100 records.

During the development of the first prototype, it was realized that initial data contains redundant fields, which will not be used during the system development. These have lead to the second iteration during which the new data was obtained.

The second iteration allowed to leave only the required fields and also add a new additional data, so it can be used during the data analysis phase with the aim to obtain new insights for the users. New data contained two type of files, first is the file that contained information about calls such as call type, the date when the case has happened, case name, annotation and amount of days since the data was accessed. The second type of files contained information about the subscriber, such as postal code, gender and the enrollment age of the person.

After some time of using the data, it was decided to update files with the additional information, fix minor issues and increase the number of files. Data was divided into two separate files where the first group of files contained information about the calls and the second contained information about the subscriber himself/herself. Subsequently, after the last iteration the following information was available for the further data analysis and manipulations:

- subscriber id

  ID of the person who is subscribed to the service.

- gender

  Specifies the gender of the subscriber.

- date of birth

  The field that specifies the birth date of the subscriber.

- postal code

  This field contains the postal code of the place where the subscriber lives.

- case id

  ID of the case that has happened to the subscriber.

- case description

  A textual description of the call reason. Examples of the possible values are: subscriber fell, had breathing problems, felt anxious and other. All the possible values will not be mentioned due to their big number and due to the fact that they could be changed or updated.

- case type

  Type of the case describes the nature of the call. The call could be test, health related or neutral. Test calls are used to verify the proper work of the service and neutral calls are the calls that are neither test or health related calls.

- case outcome

  This field represents the outcome of the call. For instance, it could specify if the emergency transportation was required or if the subscriber is OK.

- date created

  Represents the date and time when the call was performed.

A class diagram which represents the final structure of the data that is used in the thesis is presented in Figure 2.1:

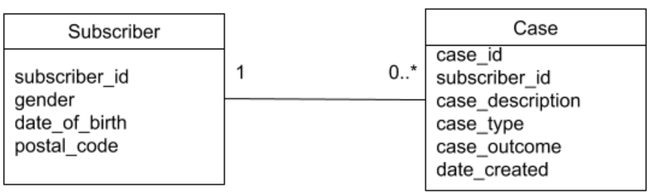

Figure 2.1: Class diagram of the PERS call center data.

This class diagram depicts the available data for subscribers and the cases, where case represents the call to the PERS call center. Each subscriber could have zero or more cases. Each case is unique and is related to only one subscriber.

2.3 Groups of potential summary readers

In most cases, two type of users are interested in getting the overview of the person's health status, they are family members and medical personnel. These user groups were also addressed in other data-to-text systems which are the part of the BabyTalk project [19] which will be described in more details in section 3.3. In the BabyTalk project the text is generated for parents and nurses about the health situation of the newborn babies. In the case of elderly people most of the time they have someone who takes care of them and it could be someone not only from their family but also an external caregiver. Therefore the user groups can be divided on caregivers, which include family caregiver or some external person and medical personnel, such as general practitioners. As long as elderly people want to stay independent, a third user group should be introduced, which is the user of the personal emergency system himself or herself. Based on the above considerations, three main groups of the potential readers of the automatically generated text were distinguished, they are:

- subscriber

  This is a person who is using the Personal Emergency Response System him or herself.

- caregiver

  A person who is taking care of the actual user of the system. It could be either professional caregiver, family member or any another person that assists a subscriber of the system.

- general practitioner

  The summary could also be potentially used by the general practitioner to get more acquainted with the patient's condition.

Each group of users has its specific requirements and needs. Summaries should be adjusted according to them, so users would get the most appropriate text that was generated taking into the account their needs. The requirements of each user group were defined alongside with the domain experts from Philips and could be found in Table 2.1.

User group	Requirements
subscriber	More gentle communication, more general story, long term trends, comparison with other people, finding patterns, messaging if nothing happens
caregiver	Better overview than for the subscriber required, with more precise timing, description of what has happened recently as well as a pretty long time ago, risk prediction
general practitioner	Comparison of the last period with the previous periods, so it is possible to see dynamics, more condensed messages

Table 2.1: Requirements of the user group.

2.4 Insights

Insights are the intermediate step between raw data provided by the call center and the actual summary which will be presented to the user. The insight could be defined the following way: it is a clause which provides information regarding one or few specific findings obtained during the data analysis process. For example, the insights could contain data about the most frequent reasons why the subscriber called to the call center or the number of health-related calls for the past two months. Examples of such insights are the following: "Mr Smith has been using our service for almost 3.5 years", "In the last two months Ms X has been in contact with our call center eight times", "Mr Smith fell less than other people in his age group".

Insights help to gain an understanding of the interaction between the subscriber and the service as well as to understand the health situation of the subscriber based on the available data. Generated insights could contain various information and their complexity could vary as well.

2.5 Automatically generated natural language summary

A summary is a final text that will be shown to the subscriber, caregiver or the general practitioner. This text is the result of the data analysis, insights extraction and aggregation process.

Final summary consists of the insights that were obtained during the previous step. Not all the insights will be present in the final summary, there is an additional processing step which makes a decision on which insights should be included. The length of the summary is adjustable and depends on the service provider that can define the desired size of the outcome summary.

The generated summary will be shown to one of the three main target users, which will be either the subscriber, his/her caregiver or general practitioner. This means that the generated text will be read by a human, therefore it needs to be adjusted, so it would be pleasant for the elderly person or the caregiver to read it. At the same time, it should be easy to read and comprehend the summary. To achieve these goals, the insights were already adjusted so the final text would meet the requirements mentioned above. To make a user experience better and the summary more human readable it was decided to include in the summary not only the insights but also the additional introduction and final clauses.

Chapter 3

Literature review

Currently, systems in healthcare domain generate a vast amount of data. This leads to information overload, which is a big problem, especially in systems which involve human decision-making process. These systems produce an extensive dataset based on which the decision should be made. Multiple technologies are used to support the process of decision-making aiming to make it easier. One of such technologies are the data-to-text systems which extract the most important findings about the data, which allow to get an overview of the overall data, so it would be easier for a human to interpret it. For the past years, the interest to the data-to-text systems has increased [19].

Further, different aspects of data-to-text systems will be discussed. In the coming sections, we discuss several data-to-text systems, typical data-to-text architectures, NLG techniques and some of the data extracting methods.

3.1 Architecture of the data-to-text systems

In the NLG domain, many different architectures have been suggested [14] and most of them had a similar structure. Therefore there was an intention to create one standard architecture that will make it easier for this kind of systems to evolve and to be reused. In 1994 Reiter noticed that many NLG systems follow similar architecture [45] and later in 2000 together with Dale in the book "Building Natural Language Generation Systems" [48] they described their vision of an architecture that could fit and standardize most of the NLG systems. The suggested architecture was designed as a pipeline that consists of three stages: document planning, microplanning, and surface realiser. The architecture could be found in Figure 3.1.

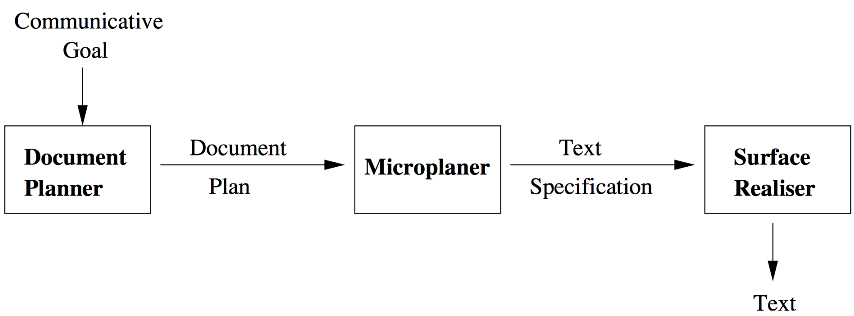

Figure 3.1: Architecture suggested by Reiter and Dale [48].

NLG systems have involved into the data-to-text systems. The difference between them is that the input data for the first architectures of the NLG systems was a knowledge base and the further evolved data-to-text systems brought a new challenge in terms of the data part, due to the fact that the input data for the data-to-text systems is a raw data that needs to be processed. Therefore, Reiter suggested to extend and adapt the suggested earlier architecture to the new needs. Two steps, signal analysis, and data interpretation were added to the previous architecture prior the document planning step [46]. The structure of the updated architecture is shown in Figure 3.2 and will be described in more details below. It is worthwhile to mention that suggested architecture could be adjusted according to the needs of the system, meaning that some of the stages could be skipped, subdivided or added in case if they are redundant, were already performed or some additional processing is needed for some specific system.

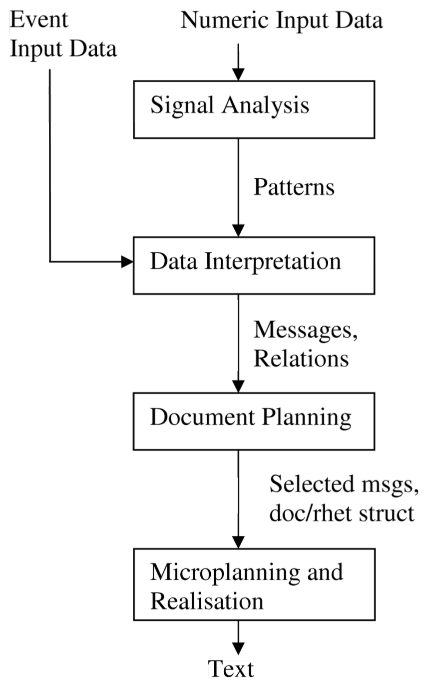

Figure 3.2: A new suggested architecture for the data-to-text systems [46].

The architecture consists of the following steps [46]:

- Signal analysis

  The task of this stage is to find patterns in the raw input data which will allow to shift from the numerical data to the discrete patterns, which could be presented in the form of ontology, taxonomy or any other suitable form. There is no one general approach or method for the signal/data analysis, they are defined individually for each system based on the domain, type of input data, and requirements.

- Data interpretation

  After the patterns were identified at the previous step, data interpretation will create messages using that patterns. At this stage the interrelations between patterns and trends which would allow to give a better overview of the data could also be found. In some domains such as medical, the knowledge of the domain experts could be a source of rules for the data acquisition part, which defines what are the patterns and trends that should be looked for, what is the normal and abnormal data, etc., which is the case for the BabyTalk project [46]. BabyTalk project consists of multiple data-to-text systems that generate a natural language summary based on the electronic record of the newborn babies. These data-to-text systems will be described in more details in section 3.3.

- Document planning

  The document planning stage is responsible for the decisions related to the structure of the final text, as well as for the decisions regarding which events should be mentioned in the final summary. This step is needed because the two previous steps could generate a huge number of messages, while the final text is of limited length, therefore some of the events have to be dropped out. One example of the task of this stage is to find interesting patterns based on the rules from the domain experts and the novelty of the found event [62].

- Microplanning and realization

  At this step, the actual text is generated based on the output of the previous step. It defines how to express the messages that were defined before, using natural language. One way to do it is to use predefined templates.

It was decided to use this architecture for the implementation of the data-to-text system due to the fact that this architecture splits up summary generation tasks into logical subunits which allows to easily change the logic of different summarization step without a need to change the other parts. This architecture also presents the summarization structure which will remain the same even if the input data are changed. Therefore particularly this architecture allows to easily change and adjust the system to new requirements.

This architecture has been successfully used to develop systems from BabyTalk project and commercial data-to-text engine which will be described further in the chapter.

3.1.1 AORTIS conceptual model

To be able to effectively collect, process and interpret clinical data a conceptual model AORTIS was developed [17]. AORTIS describes a model for the complex task of the sumaary creation of the patients' data using electronic health records (EHRs).

AORTIS conceptual model consists of five steps: Aggregation, Organization, Reduction and Transformation, Interpretation, Synthesis. In order to produce the most concise summary each step could be performed either automatically or manually by the doctor. The design of the model assumes the sequential flow of the data, where the content at each step could vary depending on the nature of the data. Every summarization process for different data is unique, therefore it may be the case that not all the steps will be needed to perform summarization. Therefore some steps could be omitted, could require less work, or the overall summarization process could be terminated early. The overall model is presented in Figure 3.3.

Below all five steps are described in more details:

- *Aggregation.* At this step, data is collected from all the available sources. Data could be stored in various formats such as numerical or textual and it could be stored in different places such as database or electronic health records (EHRs).

- *Organisation.* During this step, the data structuring and organization are performed, without changing the data. An example of such operations could be sorting and grouping of the data.

- *Reduction and Transformation.* This step allows to avoid information overload. The process of the summarization can proceed with one of the mentioned above steps: reduction or transformation. Reduction step leaves only part of the data which satisfies some criteria. For instance, if the data is numeric only, values above some threshold could be left and if the data is textual, only the most recent records could be left. By contrast, the transformation is intended to change the data in the way so it would be easier to comprehend it. Examples of the transformation could be finding the trends or representing the information in a graphical way.

- *Interpretation.* At this point, the data is being analyzed and interpreted using the medical knowledge. An example of such analysis could be a selection of the abnormal values of the particular attributes, for instance too many falls per months.

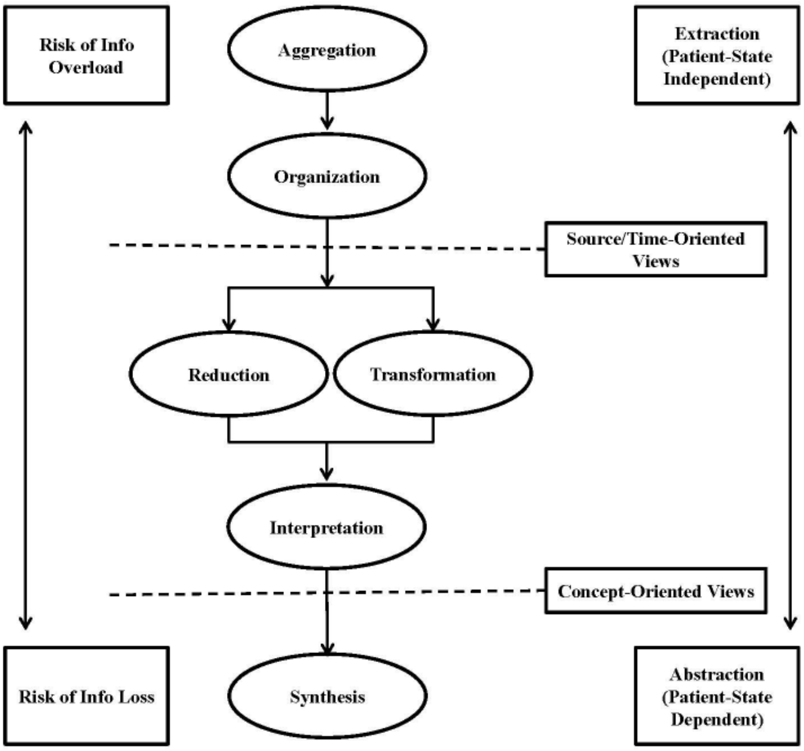

Figure 3.3: AORTIS conceptual model [17].

- *Synthesis.* This is a final phase of the AORTIS conceptual model, at which obtained data in combination with the domain knowledge could provide the description of the patient's situation and provide some suggestions if possible. Previous steps influence heavily on the final results and at this step, the actual summary is provided.

Aggregation, Organisation, Reduction and Transformation steps are similar to the Signal Analysis step in the Reuter's architecture, Interpretation step could belong either to Signal Analysis or Data Interpretation steps, Synthesis represents the Document planning, Microplanning and Realisation steps. This process flow makes it easier for the doctors to process the patient's data due to the fact that the data was preprocessed, key points were extracted and combined.

This model will not be used in the system development because it is focused more on the data interpretation steps and pays very few attention to the actual natural language generation. Moreover, it is not described how the text should be generated. Furthermore, this model was developed to deal with the clinical data, and the idea behind the developed system to be more general and not bound to one particular domain so that it could be potentially used for different kind of data.

3.2 Linguistic summarisation of data using fuzzy logic

Another approach to the linguistic data summaries was introduced by Yagger in 1982 [58] which was further developed and described in the article written in 1991 by Yager, Ford and Cañas [60]. The article describes a more implementable form of linguistic data summaries generation approach. This approach uses the concept of protoforms, where protoform (prototypical form) is a prototype for a sentence based on which the actual sentence can be built. An example of the protoform and the sentence build based on this protoform is the following: protoform "QR y's are S" and the sentence "Most Swedes are tall" [65]. The idea behind this approach is that the amount of numerical data in a database could be too large for the human to comprehend it, therefore it can be summarized in natural language sentences using the theory of fuzzy sets [63][66][59]. An example of such sentences are "Most Swedes are tall" or "Most students have high grades".

Yagger's approach consists of the following elements [60]:

- $Y = \{y_1, ..., y_n\}$ is a set of entities in available data, for which we want to perform a linguistic summarization. In current thesis, an example of such an entity is a PERS subscriber.
- V is an attribute that can take values from a set $X = \{x_1, ..., x_n\}$. V could be numeric or non-numeric and it characterizes a set of objects Y. V could be a case type, age, case outcome, etc. $V(y_i)$ represents the value of the attribute V for the object

The data that will be summarised is a collection of attributes V for the object Y which could be represented as $D = \{V(y_1), V(y_2), ..., V(y_n)\}$.

The actual summary consists of the following elements:

- a summarizer S, which could take linguistic values that expressed as a user-defined fuzzy subset of the base set X. Examples of such a summarizer could be "about 60", "over 80".
- is a quantity in agreement Q which specifies to which extent is the available data satisfies the summarizer S. It could take either relative or absolute user-defined values and some examples are: "more than 20", "at least 30%", "most".
- a measure of the summary validity T that indicates to which extent the quantity agreement Q satisfies the summarizer statement S, for instance, it will specify the truth of the claim "Most of the subscribers are over 70 years". The procedure for T calculation is the following:
  1. For each $d_i \in D$ calculate $S(d_i)$ which is the degree to which d_i satistfies the summarizer S
  2. Calculate r, where $r = \frac{1}{n} * \sum_{i=1}^{n} S(d_i)$. It represents the proportion of D that satisfies S.
  3. $T = Q(r)$, where T is the grade of membership of r in Q.

  Example of T calculation can be found in article [60].
- optional additional qualifier R which will be a supplement to the Q

Summaries could be written in the following forms:

$$\text{Q y's are S} \tag{3.1}$$

$$\text{QR y's are S} \tag{3.2}$$

which allow to define the statements as "Most of the time the subscriber had a small number of calls" and "Most of the time during the weekdays the subscriber had a small number of calls" respectively, where each statement has some truth value T.

This approach allows to define protoforms of the linguistic summaries that allow to describe the health status of the subscriber. An example of such a protoform could be "On Q of ys the subscriber had S" and others. The T value could be calculated according to the Zadeh's calculus of linguistically quantified propositions [64]. The approach that was described above was used to summarize the sensor data for eldercare [57].

The described approach is the static case where a database is a source of data. In later work [28], the method was extended to generate linguistic summaries from time series data. Because we do not deal with time series data in PERS systems, we leave the details of this approach out-of-scope for this thesis.

The fuzzy logic approach was not used in this thesis due to the fact that it literally provides the summary of the data, which is represented using generalisations like ”most”, ”less than 30%” and others. This allow to built sentences like ”most of the calls are falls related” or ”test calls are less than 30%”, meaning that we are able to build only sentences that generalise the data while the intention of the thesis is to build more broad types of summaries, which also can include more precise data. An example of the summary that can not be obtained using fuzzy logic is the following: ”Regarding the number of health related contacts, for the past 2 months Mr.Smith has faced 4 health-related issues while averaging 1 health related issues every 2 months”.

3.3 Existing systems

There are multiple already existing data-to-text systems. Some of them are focused on the medical domain and some can summarize the data from the different domains. The most suitable systems to the current master thesis in terms of technologies and domain-specific requirements would be presented below.

3.3.1 BabyTalk project

This is a project which includes multiple data-to-text systems, which generate natural language summaries based on the electronic record of the patient from the Neonatal Intensive Care Unit (NICU). Babytalk project was performed in collaboration with the NICU at the Edinburgh Royal Infirmary and the universities of Aberdeen and Edinburg, and Clevermed company [43]. The purpose of this project is to understand the process and the challenges of the natural language summary generation for the different user groups, time constraints, and contexts. Therefore during this project, multiple systems were designed and developed. All the systems have similar architectures, but their purpose is somewhat different. The systems that were developed are the following [25]:

- BT-45 is a system that generates a summary of the clinical data over the period of approximately 45 minutes, which intended to assist doctors and nurses in the decision making process [25] [43].
- BT-Family is a system that generates a summary of the clinical data over the 24 hours with a purpose to inform the parents regarding the health status of the baby [38].
- BT-Nurse generates a summary after the 12-hour shift of the nurse. The purpose of this summary is so that the nurse from the next shift could get familiar with the overall health status of the baby [25].

BT-45

This system was developed to demonstrate that it is feasible to build such a complex data-to-text system which combines the signal analysis techniques and the natural language generation. An input data consisted of time-series and information about the events presented in a structured way, for instance, an observation made by the medical staff [43]. The architecture of the BT-45 followed the standard architecture of the data-to-text systems [46]: the first step was the signal analysis at which the main trends and patterns were extracted, then data interpretation step was performed and afterwards the document planning and microplanning which chose the most important events which were combined in a coherent text. Ontology supported the communication between modules. The full architecture can be found in Figure 3.4.

BT-Family

The summary generated by this system will be shown to the parents of the children who have to stay in NICU. That brings new challenges to the summarization process. First of all the

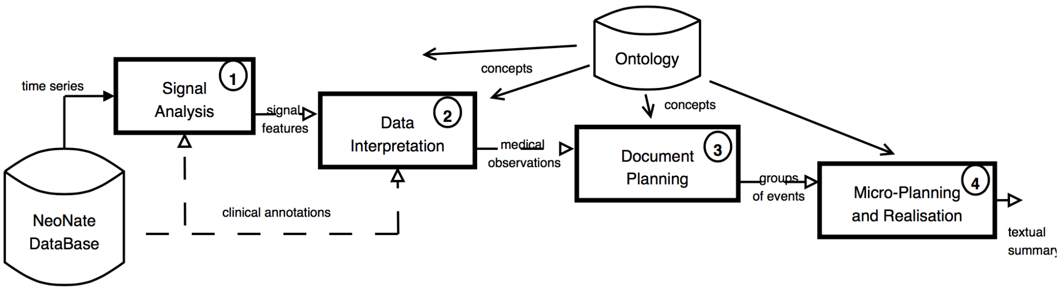

Figure 3.4: Architecture of the BT-45 system [43].

information should be presented to the average reader without a medical background, meaning that it should be relatively easy to understand it. Secondly, the connection between child and parents adds an emotional component and new challenges. Moreover, it is more complicated to communicate to such group of readers via the text which lacks the empathy of the person. Parents of babies in NICU continuously feel stress, and BT-Family aims at taking it into account. To do this, the BT-Family was developed using a stress prediction model [38].

There are some cases in which summary could be shown to the emotionally vulnerable readers, therefore it would be an improvement to the summary if it could be adjusted on the emotional level. The NLG community took into account the emotional part and explored the adjusting of the text according to the emotional aspect of the reader, which has led to the definition of the Affective Natural Language messages (ANLG) [13]. The architecture of the BT-Family is built upon the architecture of the BT-Nurse adding on top of it the effective extensions, which allow to produce a summary that takes into account the emotional condition of the reader [38]. The ANLG architecture of the system is presented in Figure 3.5.

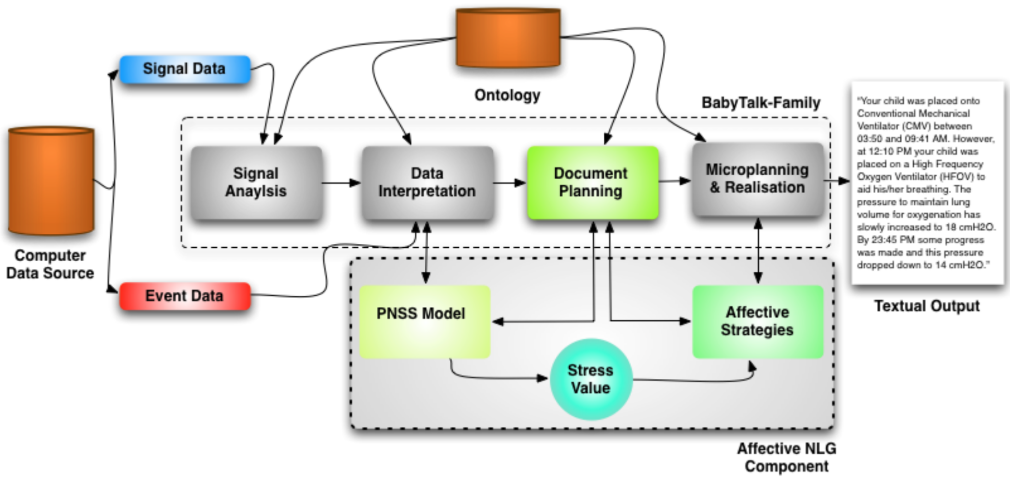

Figure 3.5: Architecture of the BT-Family system [38].

The system which was developed during this thesis work generates summaries for the different type of readers, therefore during the system development, it was also necessary to take into account the emotional part of different users. This is important, due to the fact that, for instance, the general practitioner will not take the information personally, while the elderly person could start to worry about some specific parts of a summary if they communicate bad news like the dramatic increase of health events compared to the previous period.

BT-Nurse

BT-Nurse [26] is a system which automatically generates a summary based on the patient's data recorded after a 12-hour nursing shift in a neonatal intensive care unit (NICU). The intention of the system is to provide the nurse from the next shift a summary of the shift of the previous nurse. No other input data from the nurses or some other sources is needed, only the data in the electronic records are used to create a summary [25]. Overall the system looks for the patterns and trends in the input data and afterwards integrates these findings with the data from the patient's record. BT-Nurse automatically decides which information is the most important and should be present in the final summary.

Example part of the summary created by the BT-Nurse is the following: *"Currently, the baby is on CMV in 27 % O2. Vent RR is 55 breaths per minute. Pressures are 20/4 cms H2O. Tidal volume is 1.5. SaO2 is variable within the acceptable range and there have been some desaturations. ... Between 00:30 and 03:15, SaO2 increased from 88 % to 97 %. Another ABG was taken at around 00:45. pH was 7.18. CO2 dropped to 7.95 kPa. Another blood gas was taken at about 06:15. Potential Problems: Purulent secretions during shift suggest risk of infection."* [25].

BT-Nurse is a data-to-text system which uses signal analysis, NLG techniques and artificial intelligence. It was developed using the standard data-to-text architecture described by Reiter and Robert [48] which is presented in Figure 3.6.

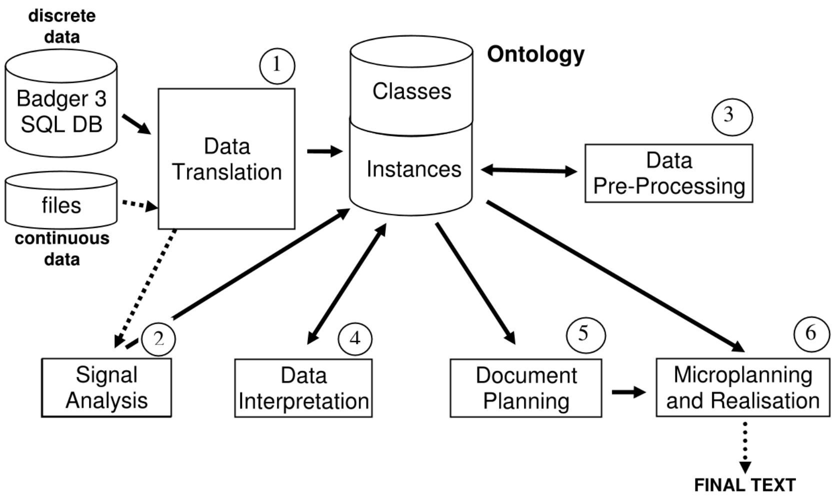

Figure 3.6: Architecture of the BT-Nurse system [25].

A domain ontology was defined that allow modules to communicate. It consists of entities, events, and relations. Entities represent the body systems, people, etc., events represent actions, observations, events, etc., a relation represents the order in which events have occurred, for example. The signal analysis allowed to reduce the amount of data, extracting the most interesting events. Alongside with the domain expert the normal range was defined to support the signal analysis process. The data were analyzed in multiple ways, for instance, the highest and lowest values for the sensors were defined as 95% and 5% of the values distribution respectively [25]. Also, the system determined the importance of the events, which has influenced on which events would be mentioned in the final summary. The importance was defined based on the range to which value belongs. Values treated as very high or very low had higher importance. As the next step data is being preprocessed and interpreted. Afterwards, the document planning step is performed at which all the events should be reduced to a small number of events that should be shown in the final summary. The key event algorithm was implemented to choose key events in the BT-Nurse system. It selects a limited number of the most vital events and then a paragraph around each sentence was built. Further, microplanning step performs the content selection, temporal planning, and aggregation which resulted in the final summary.

BT-Nurse is a well described data-to-text system which is very close to the system that was developed during the thesis, therefore it was a good example of the already implemented system based on which various details of the development approach can be learned. BT-Nurse example allowed to find out more about the data analysis methods, insights importance, information selection and evaluation step.

3.3.2 Arria

Arria is a commercial natural language generation platform which can generate insights based on various data sources adjusting the language to the target audience [3]. Arria can generate reports for different areas, one of which is the healthcare. The use case for the healthcare is a generation of the report for the neonatal care [4]. An interesting point is that this system uses the rules which were defined by the domain experts in order to be able to analyze and interpret data.

Architecture of the Arria NLG Engine

Arria takes as an input a non-linguistic data and presents it using natural language. To perform this transformation the standard architecture of data-to-text systems introduced by Reuter [46], is used [9]. A brief description and visualization of the Arria NLG Engine architecture could be found in Figure 3.7.

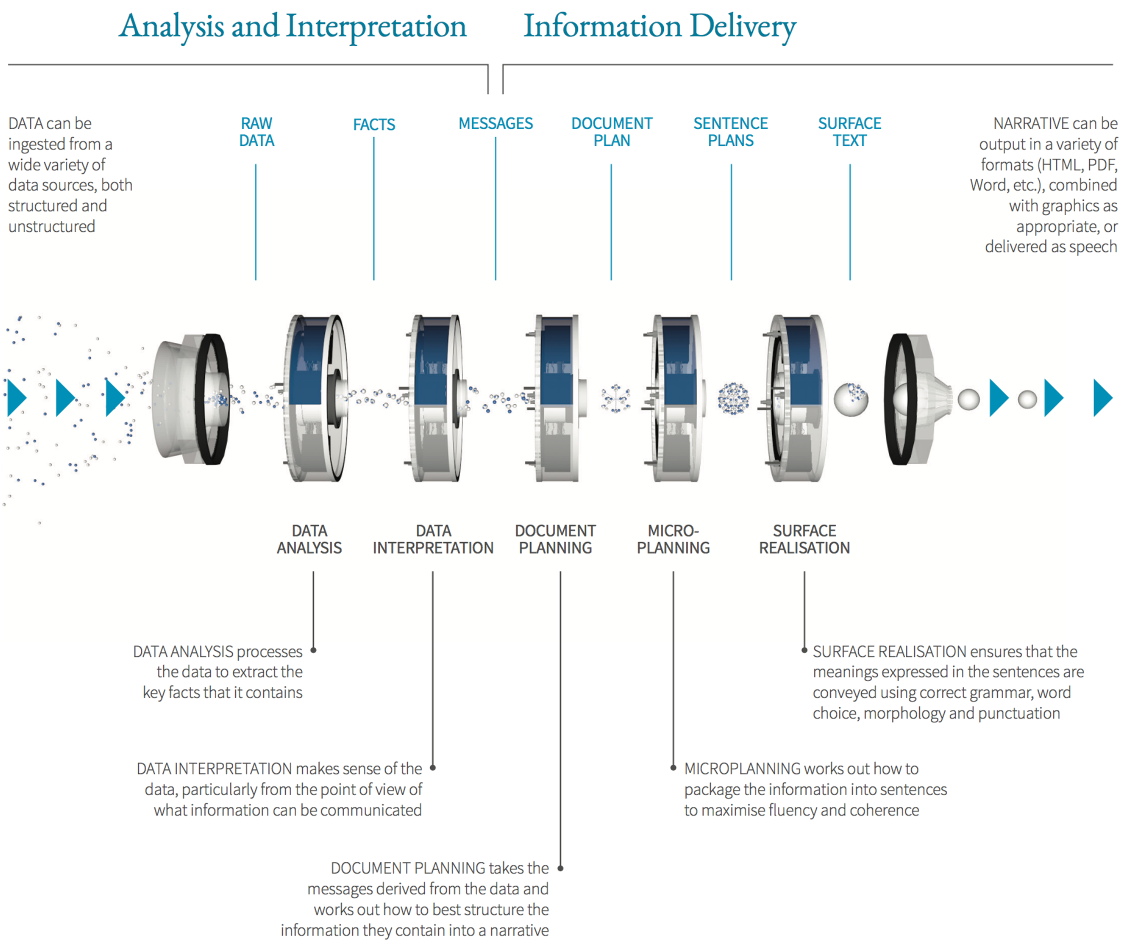

Figure 3.7: Arria NLG Engine architecture [9].

3.3.3 Other systems

One of the most successful applications of the NLG and data-to-text systems was in the weather forecasting. FoG was one of the first data-to-text systems in the weather forecasting domain [20]. Other examples from the same domain are the SumTime [54] and MULTIMETEO which also generates a weather forecast based on the structured data input, moreover, it provides a user with a possibility to modify the generated text but at the pretty basic level [12]. Examples from other domains are the following: ANA system which generates stock reports in a natural language based on stock data [31]; a system, that generates a summary of the patient's clinical history [23]; the DESCRIBER system, which describes the computer-generated visual scenes using natural language [52].

3.4 Data analysis

One of the core tasks that will highly influence the quality of the output summary is the analysis of the available data. Moreover, it will determine if the reader will get the most valuable information. This leads to the question how to analyze the data? The purpose of the data analysis part is to find the most significant events which will be interesting and useful to the reader. Statistical analysis is usually the first step in data investigation which allows to retrieve some useful information.

3.4.1 Frequent itemsets

Pattern mining is a fundamental data mining task which allows to find items in a data set which tend to occur together. This allows to discover interesting relationships between items in a data set. For this purpose can be used frequent itemsets mining and association rules mining. Frequent itemsets are the sets of items that occur more times than defined in threshold which is called support threshold [34]. This allows to find items that often occur together. Frequent items could be further used to find the association rules. Association rules are if-then rules, which allow to find associations between items. A good example of a successful application of the association rules mining is finding items which are frequently bought together in the supermarket, for instance, the rule $\{cheese, ham\} \Rightarrow \{bread\}$ could be discovered, meaning that if a customer buys cheese and ham, he would probably also buy bread. Regarding supermarkets, frequent items mining could help to define the promotions, products placement, etc. But supermarkets are not the only one who uses this technique, it is also used in many other areas. Particularly in medicine and the present case, pattern mining could allow to find some patterns in health events. This will allow to predict and prevent possible illnesses, for instance if a pattern $\{health\_event\_1, health\_event\_2\} \Rightarrow \{health\_event\_3\}$ will be found and afterwards the system will find that some subscriber had $health\_event\_1$ and $health\_event\_2$, so it would be possible to warn him out that most probably the $health\_event\_3$ will occur after some time. This allows to predict potential issues with health and prevent them.

Apriori algorithm

The most used and classical algorithm for finding frequent items is Apriori algorithm, which was introduced by Agrawal and Srikant [7]. Apriori uses bottom-up, approach meaning that at each step after the frequent subsets were found, a new item will be added to the subset. Each item in a set is compared to the minimum support, which is a user-defined threshold for the item to be treated as a frequent item. The algorithm will look for the frequent subsets that consist of 1 element, then 2, etc. until it won't be possible to create new item groups [61]. An example of the process that identifies the frequent itemsets using the Apriori algorithm can be found in Figure 3.4.

Many other algorithms used Apriori as a basic algorithm, but Apriori has some downsides such as a big number of subsets that it generates. Frequent-Pattern Tree Approach [24] and Max-Miner

Raw Transaction Data

Transaction No.	SKUs (Item No.)
1001234	1, 2, 3, 4
1001235	2, 3, 4
1001236	2, 3
1001237	1, 2, 4
1001238	1, 2, 3, 4
1001239	2, 4

One-Item Item Sets

Item Set (SKUs)	Support
1	3
2	6
3	4
4	5

Two-Item Item Sets

Item Set (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

Three-Item Item Sets

Item Set (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

Figure 3.8: Example of the Apriori algorithm application [15].

algorithm [10] try to overcome this issue using frequent-pattern tree data structure and to try to identify long, frequent itemsets early abandoning purely bottom-up approach respectively.

The application of the frequent itemset mining could provide useful information for the further data processing, analysis or summarization. Due to the time constraints and the fact that data analysis techniques are not the primary focus of the thesis it was decided not to use this analysis technique but to consider it for the further research.

3.4.2 Clustering

Clustering is another technique of data mining and statistical analysis that allows to simplify and process data in the way so it would be possible to find some patterns in the data. The purpose of the clustering is to group objects that are more similar to each other than with other objects. This allows to group together subscribers that are similar based on some parameter.

Multiple different clustering methods exist due to the fact that data is heterogeneous and therefore needs various grouping techniques. Moreover, similarity could be defined in a variety of ways where each could require different approaches.

Clustering methods could be divided into the following groups [50]:

- *Hierarchical methods.* In this group of methods clusters are formed based on the distances between the elements, so the core idea is that the closer objects are more related to each other than the objects that are at the further distance.

- *Partitioning methods.* These methods start with some initial partitioning and further relocate the elements from cluster to cluster looking for the best combination of instances and clusters. Typically partitioning methods require from the user to know the number of clusters that have to be formed.

- *Distribution-Based methods.* This group of methods assume that the data is distributed according to some mathematical distribution, therefore the main task of density-based clustering methods is to find out the distribution parameters and clusters.

- *Model-based methods.* These methods aim at finding a fit between the data and some mathematical model.

- *Density-based methods.* In this group, clusters are defined as high-density areas and they do not a predefined number of the clusters as an input parameter [30].

K-means partitioning clustering algorithm is a commonly used algorithm, which is the most efficient in terms of execution time [50]. But its downside is that there a need to know the number of clusters to be formed. There are some techniques that allow to define a number of clusters, but due to the easier implementation, it was decided to use DBSCAN density-based algorithm which average run time complexity is $O(n \log n)$ [16]. Its main benefit is that there is no need to know the number of clusters, the algorithm defines them automatically. In 2014 This algorithm received Test of Time award at the KDD Conference that is given to outstanding papers which had a substantial impact on the data mining community [1]. It is the most widely used algorithm in its group. As long as clustering is an additional part of data analysis step and is not the main focus of the current work, DBSCAN is a sufficient clustering algorithm that can be used to explore the clustering possibilities for data analysis step.

Chapter 4

Approach

In this chapter, the actual approach that was used to build the data-to-text system is described. The insights, architecture and its each component are discussed separately.

4.1 Insigths

Based on the information obtained from the domain experts and the primary analysis of the provided data, 12 types of insights were defined. These 12 insights could be divided into six subgroups. These groups and insights which belong to each group are the following:

- insights related to the overall timespan
  - Insight that specifies the time period over which the subscriber has been using the service.
    Example: "Mr Smith has been using our service for almost 3.5 years."
- insights related to the number of calls to the call center
  - Insight that specifies how many times the subscriber has been in contact with the call center for the last n month(s). It also contains information about the number of the test calls.
    Example: "In the last two months Ms. X has been in contact with our call center eight times, one of which was test call."
- insights related to the reasons for the calls
  - Insight that specifies the most frequent reasons for calls during the whole period of the service subscription, where reasons for the calls refer to the case_description field in the class diagram presented in Figure 2.1.
    Example: "During his time using our service, the most frequent reason Mr. Smith has contacted us has been due to he has fallen down."
  - Insight that specifies reasons for calls for the past n months
    Example: "During that time Ms. Smith has requested assistance two times, fallen down two times."
- insights that compare particular subscriber with the related to him or her group of people; relation is defined based on some criteria
  - Insight that compares subscriber with other subscribers who belong to the same age group based on the number of falls
    Example: "Mr. Smith fell less than other people in his age group."

  - Insight that compares subscriber with other subscribers of the same gender and who belong to the same age group based on the number of falls
    Example: "Compared with other women in her age group, Ms. Smith has a tendency to fall more."
  - Insight that compares subscriber with other subscribers who live in the same type of area based on the number of falls
    Example: "Mr. Smith fell less than other people in rural area."

- insights related to the number of calls for some period of time

  - Insight that specifies the n months period with the biggest number of calls of the health case_type
    Example: "The two months period showing the most significant number of health related issues started on July 22, 2015."
  - Insight that compares number of calls for the past n months with the typical n months period
    Example: "Compared to the typical two months period, this last period showed a larger number of health related calls."
  - Insight that compares number of calls for the past n months with and the preceding n months period
    Example: "During the most recent period Ms. Smith has had more health-related contacts than during the preceding two-month period."

- insights related to the specific number of calls

  - Insight that that specifies the number of calls for the past n months compared to the average number
    Example: "Regarding the number of health related contacts, for the past two months Mr. Smith has faced 4 health-related issues while averaging one health related issues every two months."
  - Insight that specifies if the subscriber has more calls on some specific week days
    Example: "It looks like Ms. Smith has more cases on Saturday than on other days."

The types of the insights were defined alongside with the domain experts taking into the account the available data. There is a possibility to change the types of the insights and their content or to add a new one when needed. This means that the system is adjustable and whenever the requirements changed, there is a possibility to upgrade the system accordingly.

4.2 Architecture

To be able to generate a summary of the health situation of the PERS subscriber it was decided to build a data-to-text system, which would generate a natural language summary out of the raw data provided by the call center of the personal emergency response system. To make this system able to adapt to changes in the requirements and input data, it was decided to use the standard architecture of the data-to-text systems presented in Figure 4.1.

This architecture consists of four steps: data analysis, data interpretation, document planning and realization. First, the data goes as an input to the data analysis step where the data is being processed and analyzed with the purpose to get some meaningful information that will represent the input event log from the call center. The next step is the data interpretation which constructs messages based on the analyzed data. At the document planning phase, it is decided which insights should be shown in summary and the last step generates the final summary. Realization step generates the final summary that will be shown to the reader.

To support this architecture multiple modules were developed, they are:

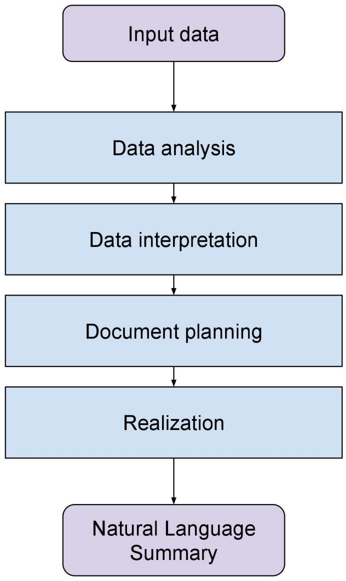

Figure 4.1: The architecture that was used to develop the data-to-text system.

- analytics module
- clustering module
- message generation module
- scoring module
- summary generation module

These modules are responsible for different parts of the architecture of the developed data-to-text system and they are mapped to the standard architecture pipeline in the following way, which is presented in Figure 4.2.

The summary generation process works as follows. The raw data goes as an input to the data analysis step which is performed using the analytics module alongside with the clustering module. Then the data interpretation step includes message generation module which builds messages based on the data analyzed and obtained in the previous step. The scoring module that represents the document planning step is responsible for evaluating the importance of each message by scoring them, which further will help to decide which sentences should be included in the summary. The last step decides which messages will be presented in the final summary based on the output of the previous step. Further, all the parts of the summary generation process will be discussed individually in more details. To be able to generate messages and perform data analysis, some rules defined by the domain experts are needed. These rules were expressed using insights which represent the message part of the summary generation process. The following insights were defined:

1. Insight that specifies the time period over which the subscriber has been using the service.
2. Insight that specifies how many times the subscriber has been in contact with the call center for the last n month. It also contains information about the number of the test calls.
3. Insight that specifies the most frequent reasons for calls during the whole period of the service subscription.

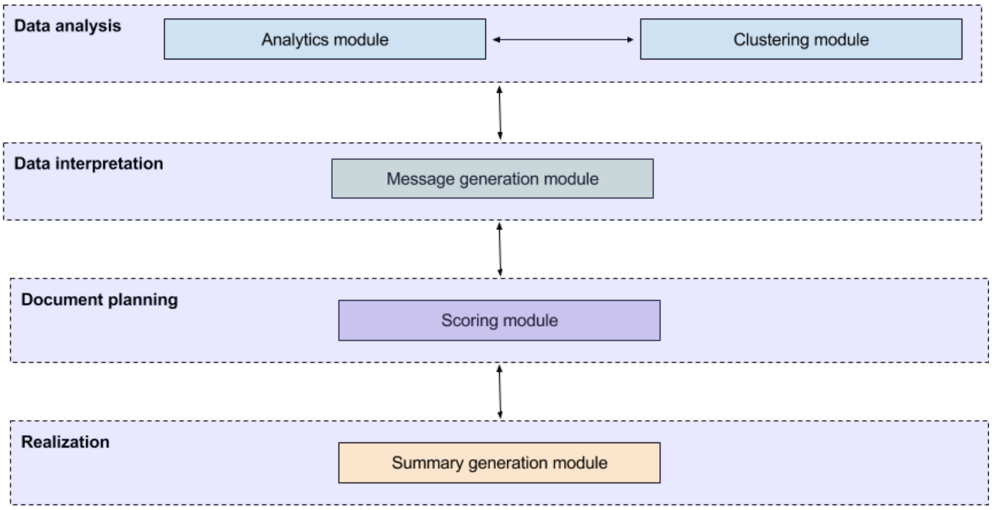

Figure 4.2: The mapping between the used approach and standard architecture of the data-to-text systems.

4. Insight that specifies the reasons for calls for the past n months.
5. Insight that compares subscriber with other subscribers who belong to the same age group based on the number of falls.
6. Insight that compares subscriber with other subscribers of the same gender and who belong to the same age group based on the number of falls.
7. Insight that compares subscriber with other subscribers who live in the same type of area based on the number of falls.
8. Insight that specifies the n months period with the biggest number of calls.
9. Insight that compares the number of calls for the past n months with the typical n months period, specifying if it is bigger, smaller or the same.
10. Insight that compares the number of calls for the past n months with and the preceding n months period, specifying if it is bigger, smaller or the same.
11. Insight that that specifies the number of calls for the past n months compared to the average number of health-related calls.
12. Insight that specifies if the subscriber has more calls on some specific week days.

The domain knowledge was also used during the scoring phase. Further on, the insights would be referred by a number, to simplify the description process.

Data analysis step was divided into two modules: analytics module and clustering module. The reasons for such a decision are the following: clustering is used by multiple insights such as #5, #6, and #7, therefore clustering should be defined as a separate entity which can be reused by multiple insights. Clustering also allows to get statistics on the clustered population, which further can be reused by different insights. As long as clustering provide additional functionality that can be reused by multiple insights it was decided to logically separate it from the analytics module which contains functionality only for each separate insight. This will allow to provide more clean and transparent structure, which would make easier further development and testing. Moreover, it would allow to change clustering logic separately from the insights analysis, for instance, it would be possible to use features and possibilities of other languages or other clustering algorithms. In general reduction of the coupling in the design provides more flexibility and it would be easier to develop the product further and will make the problem less complicated.

4.2.1 Analytics module

This is a very first step in the summary generation pipeline. Each insight has its own data analysis approach. The input data for the analysis step is the history of the calls of the subscriber to the call center. Insights have the following approaches to perform the data analysis:

1. Extract the time period for which subscriber is using the system.

2. Define the period for the data extraction should be performed, extract the number of the calls for the whole period of subscription and find the number of the test calls in that period. The length of the period is customizable and can be adjusted to the requirements and the current needs of the system.

3. Count the number of all call reasons, find the fraction of each call reason and compare it to the threshold value. This threshold defines the value of the fraction after which the call reasons are treated as frequent.

4. Define the period for which the data analysis should be performed and all the cases that have happened for this period. The length of the period is customizable and can be adjusted to the requirements and the current needs of the system.

5. The number of calls related to the falls of the particular subscriber is compared with the average number of the falls for the age group to which the subscriber belongs to. Age groups and the mean number of the fall-related calls are defined using clustering module. The outcome of the comparison is the linguistic comparator which specifies if the subscribe fell less, more or the same number of times compared to subscribers in his/her age group. The issue for which the comparison is performed could be changed.

6. This insight uses the same approach of data analysis as the previous one, but now the gender of the subscriber is taking into account so that he or she compared to the specific gender group. The outcome of the comparison is the linguistic comparator which specifies if the subscriber fell less, more or the same number of times compared to subscribers of the same gender and in the same age group. The issue for which the comparison is performed could be changed.

7. This insight is similar to the previous two, but now the location of the subscriber is taken into account. The subscriber belongs either to the rural or the urban area, so only subscribers from the same area and same age group are compared. The outcome of the comparison is the linguistic comparator which specifies if the subscriber fell less, more or the same number of times compared to the subscribers located in the same area and the same age group. The issue for which the comparison is performed could be altered.

8. Define the period length for which the analysis should be performed and then find the period of this length with the biggest number of health-related calls. The search is performed through the whole period of the user's subscription. The length of the period is customizable and can be adjusted to the requirements and the current needs of the system.

9. The number of calls for the past n months is compared to the average number of calls for the period of this length. The outcome of the comparison is the linguistic comparator which specifies if the number of calls in the last period was bigger, smaller or equal to the average number. The number n is chosen based on the requirements of the system and could be changed if needed.

10. The number of calls in the last n months is compared with the number of calls in the preceding n month. The outcome of the comparison is the linguistic comparator which specifies if the subscriber had more, less or same number of health-related contacts in the last period. Number n is chosen based on the requirements of the system and could be changed if needed.

11. The number of the calls for the past n months is compared to the average number of calls for the period of the same length. The outcome of this insight is two specific values. Number n is chosen based on the requirements of the system and could be changed if needed.

12. The number of events is calculated for each week day and then using the interquartile range (IQR) is used to find the outliers in the obtained statistics. The outcome is the day of the week to which the outlier value belongs.

Interquartile range

In the last insight, we define the weekday at which subscriber tends to have more health related events than on other days. This is done using the outlier detection and the interquartile range (IQR) [67]. First, the number of calls is counted for each weekday, so that we would have statistics of a number of calls for each weekday. Further, these seven values are used to detect the outlier values, in order to find if there is a weekday that has outlier number of calls. The approach of finding outliers is the following:

1. for the list of values find the first quartile $Q1$ and the third quartile $Q3$
2. find the IQR for the weekly statistics which equals to $Q3 - Q1$
3. find the outlier values. The outlier values are the values which are located above the $Q3 + 1.5 * IQR$

Weekdays with the outlier values will be specified in the insight message, as the days at which subscriber tends to have bigger number health-related calls.

4.2.2 Clustering module

It was decided to cluster subscribers based on their age. There were three parameters for subscribers: age, postal code and gender. Gender does not have to be clustered because it could take only two values, regarding the location it was interesting to see if there is a difference between rural and urban areas, therefore it was decided to cluster people by age. Moreover, at a different age, people tend to have different health-related issues. In the context of this thesis, it was more important to test the overall concept of subscribers clustering and to find out if it will be interesting and important insight for readers, rather than focus on the choice of some specific clustering method or parameter. Some of the insights compare the subscriber to other subscribers who belong to his or her age group. Clustering is used to define these age groups, meaning that after clustering is applied, subscribers of different ages are assigned to some particular group which contains ages of some range. To perform the clustering, the density based algorithm DBSCAN is applied. It is an unsupervised clustering algorithm which groups data points that are closely located. To be able to cluster we have to define two parameters: *epsilon* which is the maximum distance between two points when they will be considered in one cluster and a minimum number of data points to form a cluster [16]. These parameters are set by the system developer and could be changed according to the requirements. Example result of the DBSCAN clustering algorithm application can be seen in Figure 4.3.

Clustering module is also responsible for collecting statistics of the clustered population so that it can be further used during the analysis phase. It collects statistics about what is the average number of each health issue for each particular cluster.

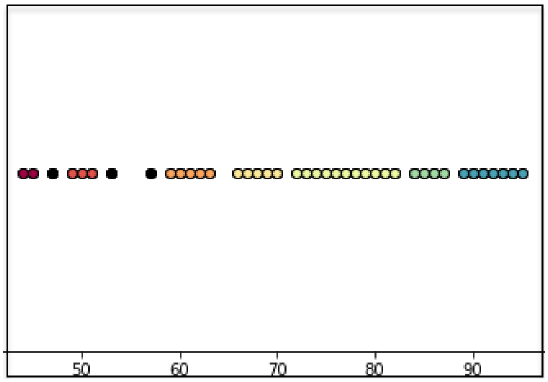

Figure 4.3: Example of the DBSCAN algorithm application.

Data limitations

During the exploratory analysis, it was discovered that few subscribers do not contain values in such fields as the date of birth, gender or postal code. As long as some insights compare subscriber to a group of similar subscribers, it will not be possible to generate these insights for the people with missing data. This will not be possible due to the fact that clustering and analysis techniques require this data, so they will be able to assign the subscriber to some cluster of subscribers. If the person is not assigned to any cluster, it is not possible to compare him with other people. In this case, the subscriber will not get these insights in the final summary.

Another issue that could be raised in future is the incorrect data. For instance date of birth of a subscriber can be unrealistic and claim that person was born 110 years ago. Although, this data could be correct in some rare cases. Such a situation could be a reason to double check the data with the service provider. The same is applicable for the other data fields with a similar situation. In case if the data was correct, but it appears to be an outlier, the insights that compare subscriber with similar subscriber will not be generated because there are no similar subscribers.

4.2.3 Message generation module

The task of this module is to generate messages based on the information that has been obtained during the previous step. For this purpose the template based approach is used, meaning that each insight has a predefined natural language template which is further will be supplemented with the information provided by the data analysis step as well as with the general data about the subscriber which is needed to make summary personalized.

At this step two additional messages are generated besides the insights. They are not related to the data analysis process, their purpose is to make the summary more pleasant to the readers so that they will have a feel of kind of a human touch and not just a machine-generated text.

There were defined three potential readers of the generated summaries, they are a subscriber of the PERS, caregiver, and a general practitioner. Each template is adjusted according to the specific user type to which it will be communicated. Templates that were used to generate text could be found in Appendix C. This module allows to modify the text of the message, add a new template and adjust it overall to the new requirements or new user group.

To support the workflow of the analytics and message generation module, the list of possible call reasons, their short description and score were separated from the system, so it would be able to easily modify the list, which includes the adding of the new health issues. This actions will not require to make changes in the developed system.

4.2.4 Scoring module

This module scores each insight on a scale from 0 to 1, where zero means the lowest priority of the insight and one represents the highest priority of the insight. Scoring is performed based on the individually defined rules. These scores allow to define the interestingness of the insight. Further on, these scores will be used in the realization phase to decided which insights will be shown in the final summary. During the scoring of the insights that includes comparators like "less", "more" the emotional aspect of the reader was taking into account. If the summary is shown to the subscriber, the insights that have comparator "more" will be scored less than insights with comparator "less". For the general practitioner and the caregiver, there will be no difference in scores because they can handle any information and moreover it is important for them to have all the information.

Scoring approach

Two messages that are intended to be shown as a first, greeting phrase and as the last closing phrase are constantly scored by 1, therefore there is a very high chance that they will be shown in all the final summaries. This is done with the intention to keep the structure of the summary similar to the usual conversations. Regarding the insights, they are scored in the following way:

1. It is redundant to show to the readers how long is the subscriber has been using the service if the summary would be generated every month or every two months. Therefore it was decided to show the time on the service insight if it was shown relatively long time ago, and the longer time ago it was shown, the higher score gets the summary. To get the score, the following function was defined:

$$\begin{cases} score = -\frac{1}{min} * (max - days \bmod max) + 1 & \text{, if } (days \bmod max) >= min \\ score = 0 & \text{, if } (days \bmod max) < min \end{cases} \tag{4.1}$$

   where *days* represent the number of days for which subscriber is using the service, *max* represents the number of days after which we prefer to show the summary and *min* represents a number of days after which score starts to increase. So we basically want to show the summary every *max* number of days and the score would be increasing while the period for which the insight was not shown will be increasing. An example of the formula application which was used in the system development is defined below:

$$score = -\frac{1}{90} * (180 - days \bmod 180) + 1 \tag{4.2}$$

   90 represents three months and 180 represents six months. After 90 days of showing this insight the score starts to increase, so at day 91 the score will be 0.01 and at day 179 it will be 0.99.

2. The score for the second insight consists of two scores: the *amount_score* and the *ratio_score*. The *amountscore* is represented by the following function:

$$\begin{cases} amount\_score = \frac{1}{coeff} * cases\_calls & \text{, if } cases\_calls <= coeff \\ amount\_score = 1 & \text{, if } cases\_calls > coeff \end{cases} \tag{4.3}$$

   where *coeff* stands for the number of calls at which the score will be equal to 1, so will reach the maximum value and the *cases_calls* represents the number of the health-related calls for the specified period. An example of such a function, which has been used in the system development is specified below:

$$score = \frac{1}{4} * cases\_calls \tag{4.4}$$

The *ratio_score* represents the ratio of the test calls to the number of the health-related calls:

$$\begin{cases} ratio\_score = 0 & \text{, if } test\_calls = 0 \text{ and } cases\_calls = 0 \\ ratio\_score = 1 & \text{, if } test\_calls = 0 \text{ and } cases\_calls \neq 0 \\ ratio\_score = \frac{1}{ratio} * \frac{cases\_calls}{test\_calls} & \text{, if } \frac{cases\_calls}{test\_calls} <= ratio \\ ratio\_score = 1 & \text{, if } \frac{cases\_calls}{test\_calls} > ratio \end{cases} \tag{4.5}$$

where *ratio* stands for the ratio of test calls to the health-related calls. For instance, in the developed system the ratio was equal to 2, meaning that if the number of health-related calls is more than two times bigger than the number of the test calls, then the *ratio_score* will be equal to 1. A specific example of the function could be found below:

$$\begin{cases} ratio\_score = 0 & \text{, if } test\_calls = 0 \text{ and } cases\_calls = 0 \\ ratio\_score = 1 & \text{, if } test\_calls = 0 \text{ and } cases\_calls \neq 0 \\ ratio\_score = \frac{1}{2} * \frac{cases\_calls}{test\_calls} & \text{, if } \frac{cases\_calls}{test\_calls} <= 2 \\ ratio\_score = 1 & \text{, if } \frac{cases\_calls}{test\_calls} > 2 \end{cases} \tag{4.6}$$

The final score is a combination of two mentioned above scores:

$$score = ratio\_score * amount\_score \tag{4.7}$$

3. The insight about the frequent reasons of calls treated as an interesting insight that should always be presented in summary, therefore its score is constant and equals to 1.

4. A score of this insights consists of two scores: uniqueness score and number score. Number score depends on the ratio of the number of cases that happened in the previous period to the number of health-related for the last period. The more events happened in the last period, the higher score will be. In case if previous period didn't contain any health-related calls, the score will be equal to a number of calls in the last period:

$$\begin{cases} number\_score = \frac{1}{2} * \frac{RP}{PP} & \text{, if } PP \neq 0 \\ number\_score = \frac{1}{2} * RP & \text{, if } PP = 0 \end{cases} \tag{4.8}$$

where RP is a number of health-related calls for the most recent period and PP is a number of health-related calls for the previous period.

The uniqueness score represents how unique are the calls for the last period. If they are unique compared to the previous period than it is more important to mention them in the final summary. In this context, unique cases mean that cases didn't happen in the previous period.

$$uniqueness\_score = \frac{3}{4} * UC \tag{4.9}$$

where UC stands for unique calls for the last period.

The final score consist of both of the scores:

$$total\_score = \frac{1}{2} * (number\_score + uniqueness\_score) \tag{4.10}$$

If the total score got value more than 1 then it is rounded up to 1.

5. The score for this insights depends on the obtained linguistic comparator and the reader type. The insight for the caregiver and general practitioner get score 0.8 in all the cases, and if the summary is shown to the subscriber, the score will be 0.8 if the comparator is "less" otherwise it will be 0.

6. The score of this insight is always equal to 1, as it was chosen to be a representative insight among three insights that compare a subscriber to the population, therefore it will be shown in a number of summaries.

7. In this case, the situation is similar to the insight #5, the score depends on the linguistic comparator and the reader type. The insight for the caregiver and general practitioner gets score 0.9 with all linguistic comparators, and if the summary is shown to the subscriber, the score will be 0.9 if the comparator is "less" otherwise it will be 0.

8. The score for this insight depends on how long time ago this period has taken place. The closer it is to the last record in the database the higher score it gets. So more recent periods are scored higher.

$$score = \frac{1}{2} * (\frac{first\_date - start\_date}{days\_in\_system} + \frac{first\_date - end\_date}{days\_in\_system}) \tag{4.11}$$

   where $first\_date$ is the date when the first call has occurred, $days\_in\_system$ is the number of days for which user is using the system, $start\_date$ is the start day of the period with the biggest number of health-related calls and the end_date is the last day of that period.

9. Score for this insight depends on the linguistic comparator that was generated and the reader type. The insight is shown to the caregiver and general practitioner, in any case, so it always has score 1 and if the summary is shown to the subscriber, the score will be equal to 1 if the linguistic comparator is equal to "less" otherwise the score will be 0.

10. In this case, the score depends on the obtained linguistic comparator and the reader type. The insight is shown to the caregiver and general practitioner in any case, so it always has score 1 and if the summary is shown to the subscriber, the score will be equal to 1 if the comparator is "less" and in the other case, the score will be 0.

11. The score for this insight depends on the ratio of the number of the health-related calls for the past period to the average number of health-related calls during the typical period.

$$\begin{cases} score = \frac{1}{2} * \frac{last\_period}{average\_period} & \text{, if } \frac{cases\_calls}{test\_calls} <= 2 \\ score = 1 & \text{, if } \frac{last\_period}{average\_period} > 2 \end{cases} \tag{4.12}$$

   This means that the insights are more interesting if for the last period the subscriber had an unusually bigger number of health-related issues than he has in average, which could be pretty important to mention because it could be a signal that something bad is happening.

12. If a weekday with an outlier value was found the insight is scored 1, otherwise 0. This is a binary scoring due to the fact that here we have two possible situations: either we find some unusual weekday or not. If we find some pattern, it is pretty important to inform the reader about it, if not there is no need to mention it.

$$\begin{cases} score = 1 & \text{, if a pattern was found} \\ score = 0 & \text{, if a pattern was not found} \end{cases} \tag{4.13}$$

inital phrase	0.8
#1	0.7
#2	0.9
#3	0.8
#4	0.9
#5	0.8
#6	0.9
#7	0.8
#8	0.8
#9	0.8
#10	0.8
#11	0.1
#12	0.1
final phrase	0.8

Table 4.1: Importance scores for the insights.

4.2.5 Summary generation module

After the messages were generated and scored the last step in the data-to-text system architecture should be performed. This module is responsible for the generation of the final text. All the messages with their scores are sent to this module.

Apart of the individual scoring, there is another scoring layer at the summary generation module. This step contains scores for each insight that was generated in the system. These scores represent the overall importance of the insights. It sets the final score to the insight, which will allow to make a final decision on which insights should be shown to the reader. Tuning of this scores will allow to specify the priority for the insights. The score could take values between 0 and 1, where 1 is the highest score. The score is also set for the non-insights that could be present in the text. In this case, they are an initial and final phrase. The scores which were chosen for the insights could be seen in Table 4.1.

The final score is calculated using the following formula:

$$total\_score = \frac{1}{2} * (importance\_score + individual\_score) \tag{4.14}$$

One more parameter that is needed for the summary generation is the number of sentences of which the summary will consist of. It is a variable that could be changed by the developer of the system. For this system, the parameter was set to ten which means that the final summary will contain ten sentences. The selection of the summary content depends on the final score of the messages and the number of sentences that should be present, so the first n messages with highest scores will be included in the summary.

Each generated summary will follow a particular structure. Based on the input data, domain knowledge, and generated insights it was decided that the summary that will be shown to the users of the system should consist of the following blocks:

- introduction clause
- analysis of the whole subscription period
- analysis of the specified period
- overall findings
- final clause

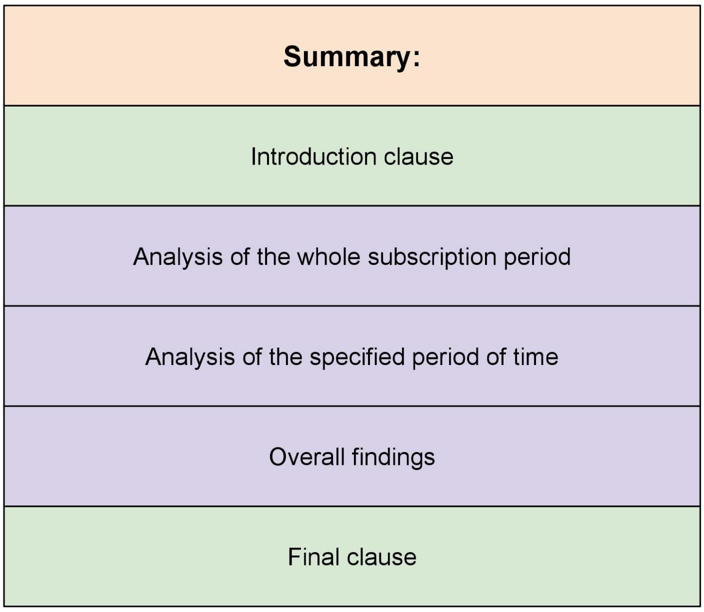

Figure 4.4: The structure of a summary.

The structure of the automatically generated summaries is presented in Figure 4.4.

In this work, it was the design decision to follow the presented structure as long as texts generated by data-to-text systems do not follow some particular text structure. It was decided to choose this particular structure so that a reader will get more human touch with the introduction and the final clauses. After the introduction, they will be shown the analysis of the overall subscription period, so they will get a more general overview, which will be followed by the text that focuses on the analysis some particular period of last n-months. In addition to that information, they will get insights to the overall findings, which, for instance, could include some interesting and relevant patterns of their subscription period, comparison with other similar subscribers, etc. The structure of the summary could be changed and adjusted to the new or updated requirements if needed.

4.3 Remark

This architecture and approach were used with the intention to make the system easy to change and adapt to new requirements and health-related issues that can be inserted into the system. A scoring approach, messages content, insights, in general, can be changed without the need to radically change the whole system, but only by changing the logic related only to the new insights.

Chapter 5

Implementation

5.1 Architecture

Data-to-text systems mostly follow similar architecture that was described in the background chapter. During development of this particular system, the same principles and ideas were kept in mind while developing the summarization tool. This means that the architecture of the developed system follows the same text creation pipeline. The architecture could be found in Figure 5.1.

Summarization software consists of 7 parts:

- analytics module
- clustering module
- message generation module
- scoring module
- summary generation module
- graphical user interface (GUI) module
- tools module

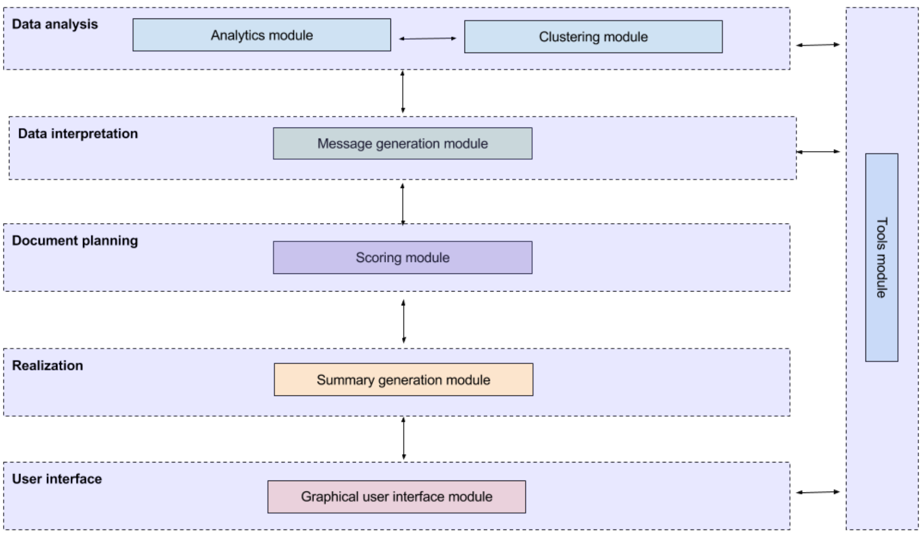

Figure 5.1: Architecture of the developed system.

The system works as follows. First, the GUI module presents the main window to the user where the user has to choose parameters for the summary generation, which are the subscriber/user id and the reader type. After the parameters were chosen, they are sent through the summary generation and message generation modules, where the last one obtains user info based on his id and send this data to analytics module. The analytics module alongside with the clustering module starts to analyze the data. The analyzed data is sent to the message module. At the next step, messages are generated based the obtained data. Afterwards, all the messages get scored using the scoring module. The messages and their scores used as an input data for the summary generation module which is responsible for the generation of the final text. In the end, GUI module shows the final text to the reader. Tools module assists other modules in performing minor technical tasks such as converting a string to an integer or retrieving data from a file. Representation of this process could be found in Figure 5.2. Further, each module will be discussed separately.

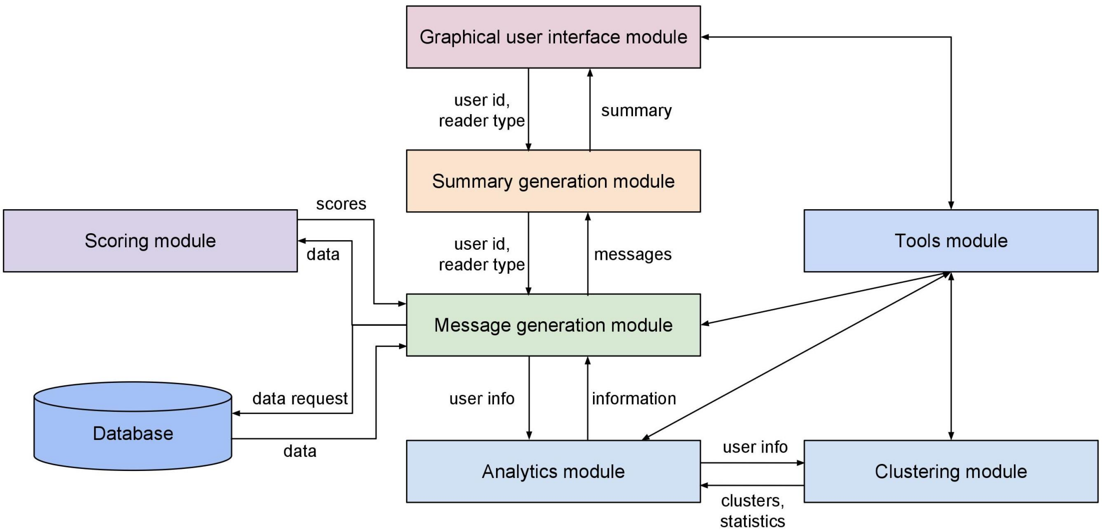

Figure 5.2: Diagram that represents the interconnections and data flow between the modules.

5.1.1 Analytics module

This is the module that is responsible for the logic of the program. It analyses the raw input data and as output provides data about:

- how long is the person uses the service
- amount of contact times and how many of them were test contacts
- frequent reasons for call
- health-related issues that happened recently
- compares two periods of time
- compares subscriber with other subscribers of the same age group

This module produces that further will be used as an input to the message generation module. The module works as follows. Using GUI user chooses for which subscriber the text should be generated, then the analytics module extracts log data of that person and analyzes it for each insight separately. To support the analysis and make the system scalable, a separate file that contains health-related issues with their textual description on individual weightings are recorded. This makes it easy to update the list of possible health issues, even for the person that doesn't have engineering education, because the data stored in a spreadsheet file. The example of this file can be found in Figure 5.3.

	A	B	C
1	case_name	description	weight
2	Request Assistance	requested assistance	0.6
3	Deceased	Mr. X has died on	1
4	Anxious	felt anxiety	0.6
5	Seizure	had seizure	1
6	Broken Bone	had a broken bone	0.9
7	Breathing Problem	faced a breathing problem	1
8	Fell	fallen down	1
9	Stroke/Numbness	had a stroke	1
10	Nausea/Vomiting	had nausea possibly with vomiting	0.8
11	Illness	been ill	0.7
12	Unconscious	been unconcious	0.9
13	Bleeding/Injury	had bleeding or injury	1
14	Confused	been confused	0.7

Figure 5.3: Separate file with the possible health-related issues, their textual description and weighting.

Analytics module consists of the following functions:

- time_on_service[1]
- contacts_number
- frequent_calls
- recently_happened
- calls_per_period
- compare_people
- periods
- day_of_week
- last_average
- last_previous

time_on_service is a function which allows to find out for how long the subscriber is using the system. As an input, it gets subscriber data and as output, it returns the number of days for which subscriber is using the service. It simply takes the difference between the first and the last record in the event log of the subscriber.

contacts_num is a function that allows to find out how many times did the subscriber contacted the call center and how many of those calls were test calls for some particular period of time. As an input, it takes user data and returns a tuple with the following elements: period, a total number of contacts, a number of test contacts. It is needed to specify a period for which the statistics should be collected. Then for this period, all the contacts are counted.

frequent_cases is a function which finds the most frequent reasons for calls for the whole period of subscription. As an input, it takes user data and returns a string which contains the textual description of all the frequent cases concatenated with 'and'. The function works as follows, first, it counts how many times each health issue has happened with the subscriber. Then we look for an issue that has happened the biggest number of times and divide all the other numbers by this maximum value. This allows to normalize data and to get values between 0 and 1, where 1 corresponds to the most frequent health issue. To find the most frequent reasons for calls we also introduce a threshold which allows to decide which calls would be treated as frequent. At this

[1]all function names are faked

point, the threshold equals to 0.75, but it could be changed to any another value. This means that all the cases that have value more than or equal to 0.75 will be treated as frequent. After the frequent cases were selected, we generate their textual description using file which contains the textual representation of health issues.

happened_recently is a function that selects calls that have happened in the specified period of time, counting from the most recent day, so for example what has happened for the last two months. An input data is the event log of the calls and the output is a string which contains a description of health issues for that period, a list of those cases and the list of the cases for the previous period of the same length. These data would be required for the scoring step.

calls_per_period compares the number of calls that happened for the last period with the previous period of equal size, so it allows to detect whether for the most recent period number of calls has increased or reduced. The period consists of a specific number of months. Its output is the coefficient which specifies the number of months and linguistic comparator "less", "more" or "same amount of". We divide the event log for periods of 30 days (approximately one month, the period should not be very precise). Then we specify the coefficient, which means a number of periods we want to analyze, f.e. if coeff equals 2, it means that we want to compare last two months with previous two months. Then we count a number of events in each period and compare them. Comparison returns one of the three values "more", "less" or "same amount of" specifying that the first period had less, more or same amount of events as the second period.

compare_people is a function that compares one particular subscriber with other subscribers from the same age group and who share other specified characteristics. First, it clusters people by their age using clustering module. Then it counts the statistics of falls for the particular person and compares it to the average number of falls for his/her age group. After the comparison, we get the comparator which specifies if the person falls more, less or same amount of as an average person in a cluster. There are two additional functions that perform similar tasks, but one of them takes into account the gender of the person and the second takes into account the place of living, more precisely if it is rural or urban.

5.1.2 Cluster_people module

This module clusters subscribers by their age using DBSCAN algorithm and afterwards calculates the statistic about each cluster, specifying the average number of health issues per cluster. Module contains two main functions *clustering* and *get_statistic*.The clustering part was implemented using the implementation of DBSCAN from the scikit-learn, open source Python library [5]. As an input, it takes a list of ages of all the subscribers that need to be clustered. To define the density necessary to form a cluster we set two parameters: *eps* and *min_samples*. Using experiments, it was found that the most suitable values for these parameters in the particular case are the following:

- eps = 1
- min_samples = 2

The output of the DBSCAN is the dictionary in which label of the cluster is a key and value is a list of ages which are located in the cluster. Some subscribers who do not have indicated age, therefore they could not be assigned to any cluster, therefore it is not possible to get insights that involve clustering for them.

get_statistics is a function that collects statistic for each cluster. It counts the average number of cases that happen per cluster.

5.1.3 Message generation module

This module is responsible for the generation of the messages. It consists of 9 functions:

- initial_phrase
- time_in_service

- contacts_number
- frequent_calls
- recently_happened
- events_in_period
- compare_people
- random_quote
- final_phrase

As long as there are multiple use cases for the generated text, messages are generated separately for each of them, so in the end, the final summary would be addressed to the person, caregiver or the doctor. Each function calls corresponding functions from the analytics module and uses the returned result as an input for the template text. Messages are basically templates that are filled with the appropriate information received from analytics module.

5.1.4 Scoring module

Scoring of the insights that were obtained in the previous step is performed according to rules specified in the approach chapter. At this step, each insight represented as message receives a score. Further, these insights with the scores are sent to the summary generation module.

5.1.5 Summary generation module

This module is responsible for the generation of the final text. As an input, it gets all the messages and their scores from the previous modules and as a result, it generates a final text. This module requires a number of sentences that will be present in the final text and importance scores for each insight. After combining two scores, the individual score of each message and the importance score, it selects the n insights which will be present in the final summary. Then it aggregates them in one summary which is ready to be shown to the user.

5.1.6 Tools module

This is a module that contains functions that assist other modules of the system in performing some tasks, like converting a string to date, get the subscribers' data to appropriate format, etc.

The module contains the following functions:

import this method takes as an input CSV file, which contains cases name, description and weight and converts it to a dictionary, which simplifies further manipulation with the cases.

string2date converts string which contains date and time to a datetime object.

age_int converts string which contains age of a person to an int. Some profiles do not contain age of a subscriber, and if this is a case, the method will return -1.

days_string takes a number of days as an input and returns its representation in natural language format. The example of the output: more than two years.

ids get the id from the file path. It uses regular expressions to perform this task.

5.1.7 GUI module

Graphical user interface was developed using Python package Tkinter [2], which is Pythons standard Graphical User interface package. The screenshot of the interface could be found in the Figure 5.4.

The program has one window, which contains two drop-down menus, one button and the main text box where the generated summary will be shown. The first drop-down menu provides available use-cases: subscriber him or herself, caregiver and general practitioner. Next drop-down

Figure 5.4: Graphical user interface.

menu contains a list of ids of subscribers for whom the summary will be generated if to choose one of the ids. They follow by the button run, which starts the summarization process. At the right part of the window the main text box is located, where the summary is shown.

To get the summary user actions should be the following:

1. From the first drop down menu select who will read the summary. It could be either the person him/herself, caregiver or the general practitioner
2. From the second drop down menu choose id of the person for whom the summary should be generated
3. Hit the run button

After the completion of the above steps in the text box would be generated a summary for a chosen person. The whole process is presented in Figures 5.5 - 5.7.

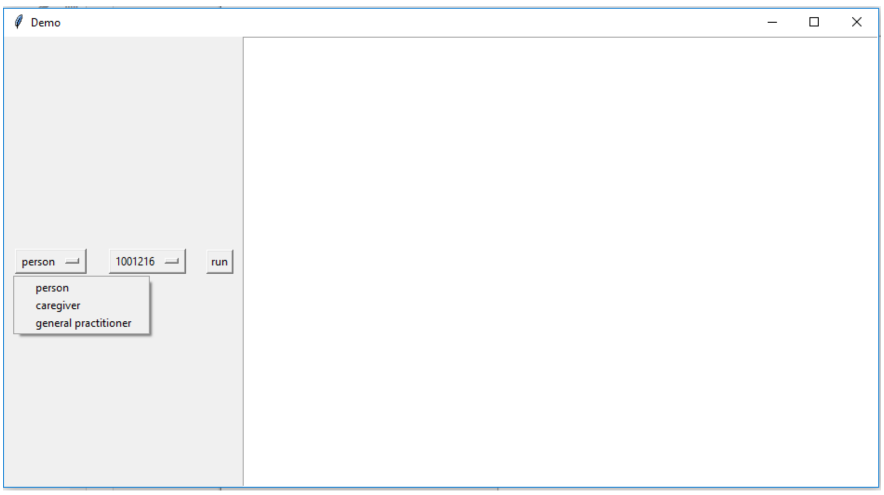

Figure 5.5: Step 1.

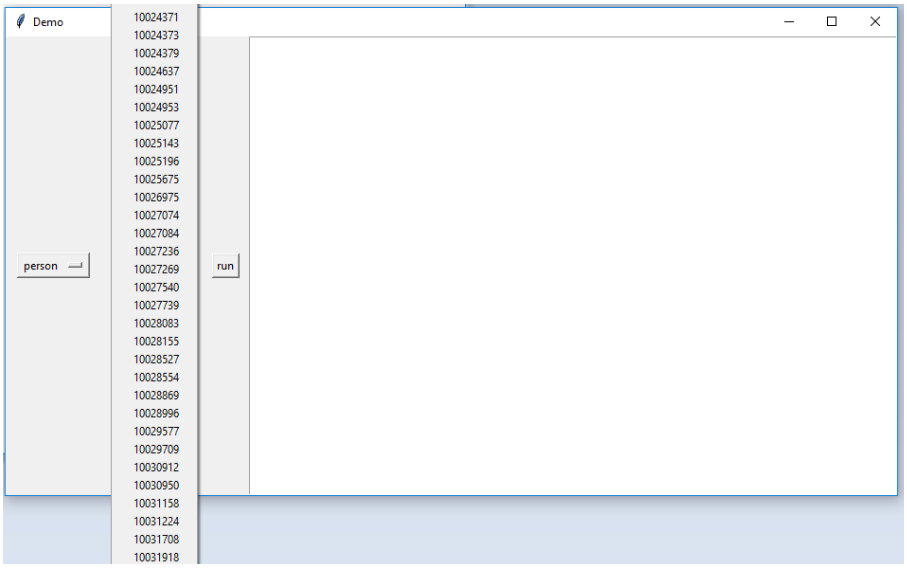

Figure 5.6: Step 2.

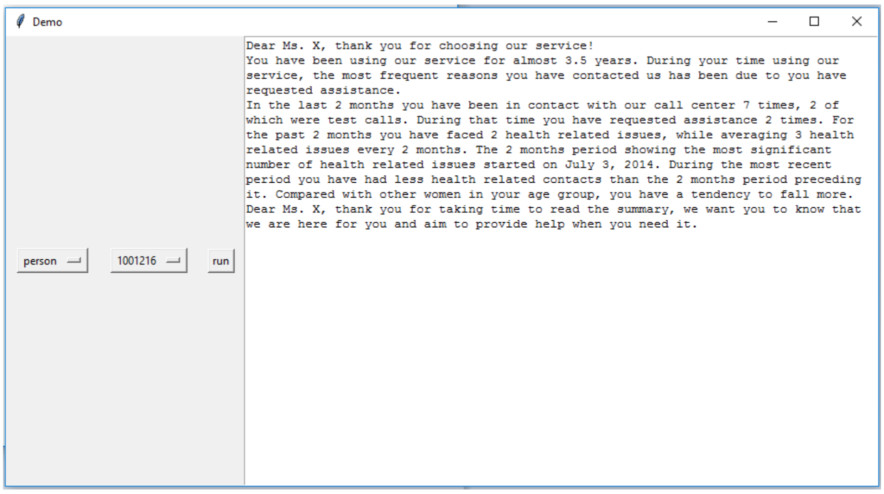

Figure 5.7: Step 3.

Chapter 6

Evaluation

After the summary was generated, the next step would be to evaluate it. As long as the intention of this textual summary is to bring additional value to the customers and to help their caregivers to get familiar with the health situation of the subscriber, it is important to understand if it is easy to understand the summary for the potential readers and how much additional value does it bring, as well as to get an open feedback. Answers to these question will help to understand the advantages and the downsides of the automatically generated summary, which will allow to improve the user experience and the quality of the textual summary.

The general aspects that can be evaluated are the following [18]:

- accuracy
- fluency and intelligibility
- task evaluation

Accuracy describes to which extent the generated summary represents initial data; fluency and intelligibility represents the quality and the readability of the generated text, this could include such metrics as coherence, stylistic appropriateness, syntactic correctness; task evaluation allows to assess how well has the system performed the task, this metric could be evaluated using multiple methods and questions.

There are multiple ways available for evaluation of the quality of the text that was generated by the data-to-text system. The system developed during this master thesis can be classified as a data-to-text system, therefore methods which are used to evaluate data-to-text systems are applicable to the developed system as well. All these methods could be classified into three main groups [47]:

- evaluation of the impact of the generated text on the summary readers

  This approach includes a range of various methods such as physocological experiments or randomized controlled clinical trials [49] as well as the other techniques from the application domain.

- evaluation of the quality of the summary made by the human expert

  This type of evaluation could be performed in multiple ways. Human experts may be asked to rate the automatically generated text and specific aspects of the text using some scale such as Likert scale or asked to give an open feedback on the text or ask to compare two texts, where one is automatically generated and the other one is manually written.

- comparison of the automatically generated summary with the human-generated summary using automatic metrics

  This evaluation could be done using such metrics as BLEU [42], ROUGE [37] or METEOR [32] which tend to compute the n-gram overlap of the two texts.

First and third types of evaluation are not suitable in the context of the current master thesis and the chosen development approach, therefore the assessment of the generated text would be done using the second type of evaluation, which involves human experts judgment.

The first group of evaluation methods which includes psychological experiments and trials could give a vital information regarding the influence of the generated summary on the readers of the summary, especially elderly people, which could be a good source that allows to understand how well the information is presented. But in the context of this master thesis, this approach was not considered as viable, due to the fact that it was not possible to reach the elderly people who use the service and this type of evaluation is pretty complex and complicated to perform.

The third type of the evaluation, which involves automatic methods such as BLEU, ROUGE and METEOR also is not suitable for the assessment of the generated text in the context of this master thesis. BLEU and other similar metrics compare the automatically generated text with the manually created text, which is treated as kind of golden standard text [47]. To be more specific, these metrics compare n-grams in the automatically created text with the n-grams in a 'golden standard' text and counts the number of their matches [42]. Basically, BLEU is checking the word to word match of two text, so according to BLEU the closer automatically generated text to the manually created the better it is. METEOR was created taking into account these weaknesses of the BLEU and it looks not only for exact word match but also for the morphological variants of the words [32]. The way these metrics work is the actual reason why it is not possible to use them to evaluate the generated text. They assume the existence of the golden standard text, but there is no such a text for the provided data and task. Multiple different texts could be created out of the available data and they will depend on the requirements and the domain experts working with the data. Therefore it is not possible to use this kind of metrics. Moreover, NLG community suggests that evaluation results obtained using automatic metrics could be meaningless because it can not capture the impact of the text on the readers, but only the match of two texts [47].

Taking into account the described downsides of the two types of methods as well as the desirable outcome of the evaluation, it was decided to perform the evaluation involving human experts, which is currently the most popular method to evaluate NLG systems [47]. The intention of the evaluation is to get feedback from the potential readers on the various aspects of the automatically generated text.

6.1 Experimental Setup

To collect feedback from the human experts with the different backgrounds it was decided to use the crowdsourcing service Amazon's Mechanical Turk (MTurk), which allows to collect feedback from a big number of responders in a relatively small amount of time. Multiple studies suggest that Mechanical Turk allows to obtain data of high quality and it is as reliable as the data obtained using traditional methods of feedback collection [44] [11].

The overall idea of the experiment is to show the generated summaries to the human experts from the Mechanical Turk and to get feedback regarding the usefulness, clarity and the willingness to use the service as well as to get the free comments using open questions. To perform this, ten texts generated by the developed system were chosen. Half of them were chosen randomly and another half was chosen manually. Random selection was done with the intention to avoid biased results and the manual selection allows to show the whole range of possibilities of the system and the various insights that were possibly not included during the random selection. Anyway, all the texts are generated using template sentences, therefore all of them are similar and the content is mostly the same, apart from some minor changes, which means that even two texts will give a very clear understanding of what kind of summaries does the system generate. One example text that would be shown to the human experts could be found in Figure 6.1.

To be able to participate in the experiment participant should meet the following criteria:

- they are between 40 and 65 years old

  This age range was chosen due to the fact that in this age people have parents which most

Dear caregiver, below we provide you a summary of the person's health situation.

Ms. X has been using our service for almost 3.5 years. During her time using our service, the most frequent reasons Ms. X has contacted us has been due to she has fallen down.

- During that time Ms. X has fallen down 1 time, been in the hospital 1 time.
- Regarding the number of health related contacts, for the past 2 months Ms. X has faced 2 health related issues, while averaging 1 health related issues every 2 months.
- Compared to the typical 2 months period, this last period showed a larger number of health related calls.
- During the most recent period Ms. X has had more health related contacts than the 2 months period preceding it.

Compared with the other women in her age group, Ms. X have a tendency to fall more. It looks like Ms. X has more cases on Saturday than on other days.

Dear caregiver, thank you for taking care of our client.

Figure 6.1: One of the texts that would be shown to the human experts.

probably face some health related challenges, therefore human experts in this age most likely have some experience and knowledge that will help them to understand the presented text and situation, as well as to give feedback. Moreover, people of this age are potential users of the PERS and the developed system.

- they are based in the US

  As long as the PERS for which the system in this master thesis was developed is currently launched in the US and provides service for the US citizens, it was decided to involve human experts who are based in the US

- they have an MTurk-approval rate of 95%, which means that at least 95% of the prior tasks they completed on MTurk were of acceptable quality

To get the feedback in a quick and convenient way the questionnaire was developed. At the very beginning of the questionnaire could be found an introduction to the evaluation which they will perform. In the introduction participants are asked to imagine the situation in which they will use personal emergency response system and are getting familiar with the PERS and how does it work, then they are told about the additional service, that was developed during this master thesis. After the explanation, they are provided with the example of the automatically generated text and then asked to do the following:

Q1: to summarize presented text in one sentence

Q2: to rate the following four expressions on a five-point Likert scale - [Strongly disagree, Disagree, Not sure, Agree, Strongly agree]:

  Q2.1: It was easy to understand the text

  Q2.2: The summary would help me to get familiar with the persons health status

  Q2.3: It would be useful to get such a summary from time to time, for example, every two months

  Q2.4: I would agree to pay an additional fee for such type of a service, which summarizes the health status of my elderly relative

Q3: to leave an open comment about the text and to specify what is the overall impression regarding the text

Q4: to tell if there is something additional that the reader want to see

Q5: to tell if there is something that is not worth mentioning

Q6: to tell if they have any personal experience of taking care of an elderly family member

The example of the full questionnaire can be found in Appendix A.

Due to unexpected problems with the funding of the study, it was not possible to use MTurk anymore, therefore evaluation was performed internally within Philips and only the answers from Philips employees are used in the evaluation. The content of the questionnaire didn't change since the previous setup. The actual questionnaire was created using the Philips internal tool for questionnaires creation and management. Afterwards, it was distributed via internal groups of volunteers and via emails to the departments were researchers that take part in this study work. There was no obligation to take part in the questionnaire, it was upon employees to decide if they want to participate. The questionnaire was anonymous and the same requirements for the age were held, therefore only employees that are between 40 and 65 years old could participate, which was mentioned in the study description.

6.2 Results

In total 17 people have participated in the survey. They gave answers to all the open questions and rated all four expressions using the Likert scale. In the following section we will analyze the answers that used a Likert scale and in section 6.2.2 we will discuss the free-text answers.

6.2.1 Answers to the questions that used Likert scale

There were four questions where survey participants were asked to rate four expressions based on the five-point Likert scale were the choices were "Strongly disagree", "Disagree", "Not sure", "Agree" and "Strongly agree", participants had to choose one option per expression. The detailed number of answers could be found in Table 6.1 with corresponding pie chart for each expression in Figure 6.2.

The first expression that should be rated claims that "It was easy to understand the text". 10 out of 17 responders, which is almost 60% of all the responders, mention that they disagree or strongly disagree with this statement, 4 out of 17 responders, which is almost 25% of responders, agreed with the statement. This shows that final texts have some downsides which make it more difficult to understand the summary. Comments on the free text questions could give some insights regarding the issues in the texts and what are the possible improvements.

	Strongly disagree	Disagree	Not sure	Agree	Strongly agree
Q2.1: It was easy to understand the text	3	7	3	4	0
Q2.2: The summary would help me to get information about the current health status of the person.	0	2	4	11	0
Q2.3: It would be useful to get such a summary from time to time, for example, every two months	0	2	4	8	3
Q2.4: I would be willing to pay an additional fee for a service, which summarizes the health status of my elderly relative	0	3	10	4	0

Table 6.1: Responses to questions that used Likert scale.

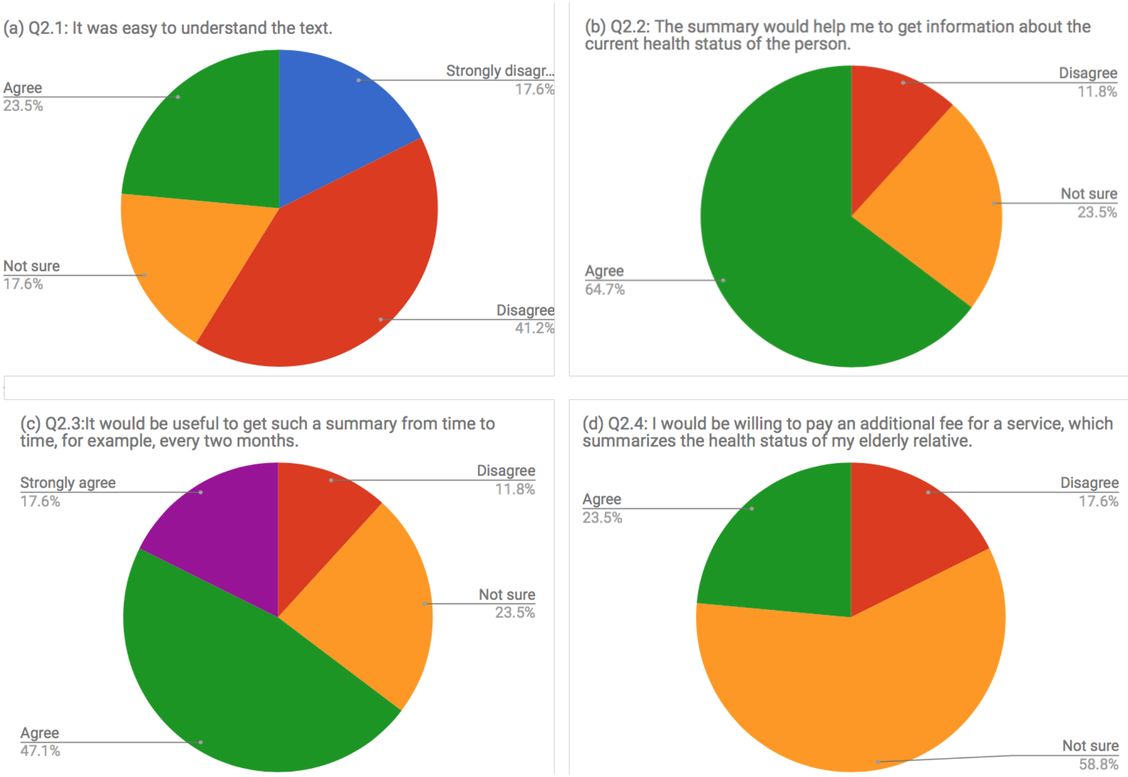

Figure 6.2: Pie-chart representation of responses to the questions that use Likert scale.

The second expression stated that "The summary would help me to get information about the current health status of the person". Despite the fact that it was not easy to understand the text, most of the responders agree that such type of summary would help them to understand the current health status of the subscriber and only two out of 17 responders disagree with this statement.

Next expression claimed that "It would be useful to get such a summary from time to time, for example, every two months.". 11 out of 17 responders agree or strongly agree with this claim, while only two disagree. Four responders were not sure if it would be useful to get such a summary from time to time. This means that readers see potential benefits of this service which is a good sign for the further improvements and development.

The last claim was "I would be willing to pay an additional fee for a service, which summarizes the health status of my elderly relative". The majority of the responders were not sure if they would be ready to pay for such a service, while four responders agree to pay and three disagree. This shows that readers were not convinced that this service is worth additional money and some improvements should be made to make the service more appealing in the eyes of the potential customers.

6.2.2 Analysis of free-text answers

Based on the four open questions we were able to collect quite a lot of free-text feedback. Most of the comments were related to the text readability and others were related to the content of the summaries. The overall impression is that responders were mostly focused on the quality and readability of the text and not on the information that was delivered using the text. This issues could be related to the not sufficient quality of the generated text, which causes the shift of the readers' focus.

Comments related to the text itself and its quality mostly contained remarks that the generated text is too long, not intuitive, contains repetitive parts, lacks natural flow, and needs some improvement, but they mention that in general, it is readable. Suggestions from readers were that it should take less time to read the text, the summary could include bullet points, illustrations or little tables which would shorten the text and would make it easier to perceive the information in general. The overall suggestion was to work more on text, to make it more emotive and coherent.

Regarding the content of the texts, responders mentioned that they would like to see more details about the health issues that the subscriber had, such as the specific time of the issue, how serious it was, how long did it last and what actions were taken. Responders also mention that they are missing more actionable and recent information. Few responders indicate that some insights could be removed from the final text, but there is a lack of explanation regarding why they think that some insights should be eliminated. For instance, two responders mentioned that insight that highlights the n-month period with the biggest number of events could be removed. The insight that was mentioned in one of these comments is the following: "The two-months period showing the most significant number of health related issues started on March 12, 2016". Based on the other comments there could be made an assumption, that this information is pretty old compared to the current date and may be not as useful as more recent information.

6.2.3 Discussion

The overall feedback states that such summarization service is a promising idea and it could be useful to have such a tool, although many people are not sure if they are ready to pay for it. This could be explained by the feedback on summaries. Responders mentioned that summaries are readable, but still have to be improved, so it will be easier to comprehend the overall message. This is especially important if the text will be shown to the actual PERS subscribers, who are mostly elderly people, due to the fact that for them it would be even more complicated to read and understand a pretty complex text. Responders also have expressed the willingness to see some additional information which means that insights could be changed. In general, there is a clear interest in such a service and there is room for improvement. As long as only Philips employees took part in the evaluation, the results could be biased.

6.3 Analysis of the insights scores

Another additional analysis that could be performed is the analysis of the scores dynamic. It could be analyzed if the insights' scores change over the time, meaning that the scores that are calculated using some logic are not static over the whole period and the summary would contain different insights.

Insights number two, three, eight and eleven which provide information regarding the number of calls, issues that happened recently comparison of two periods and shows the period with the biggest number of events respectively are scored using functions that were described in section 4.1.4. Other insights have static scores that are not changing over the period. Therefore it was decided to explore the scores of insights that have to change over the time to see if they really change. Taking three random subscribers out of the experiment that was conducted before, we analyze scores for the whole period of their subscription checking the score for each particular insights at every period of time. One unit of time was taken equal to two months. Figures 6.3 - 6.5 represent how the insights scores are changing over the time. X axis is the time periods, where 0 is the start of the subscription and one unit is equal to two months, and the Y axis represents the score which can be between 0 and 1. It can be seen that the scores are not static.

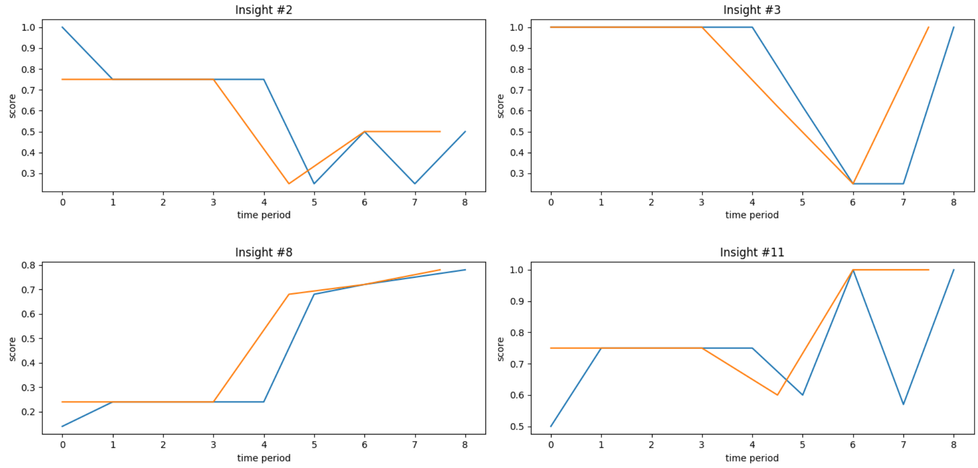

Figure 6.3: Scores for the four insights for user 1.

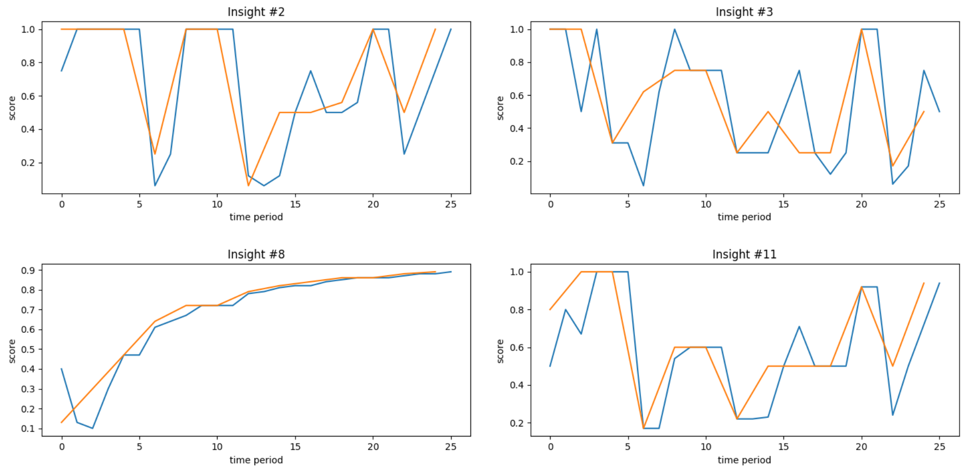

Figure 6.4: Scores for the four insights for user 2.

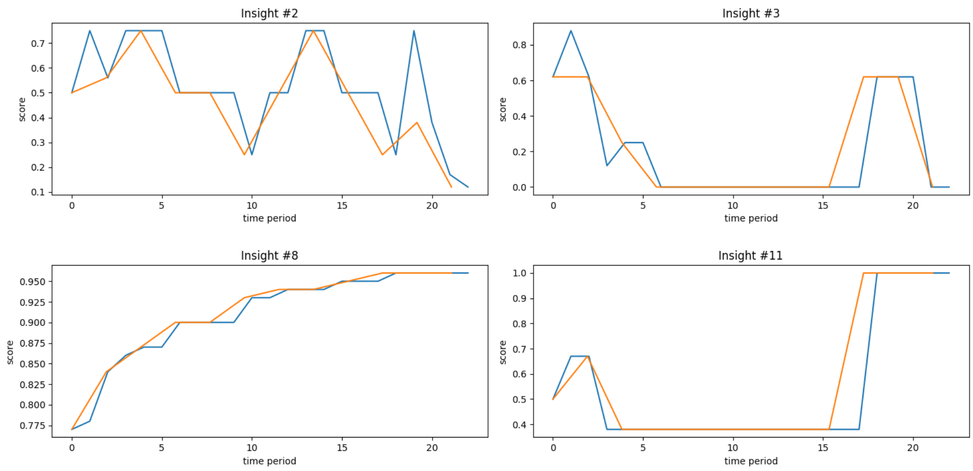

Figure 6.5: Scores for the four insights for user 3.

Chapter 7

Conclusions

This thesis discusses the possibility of summarization of the event logs provided by the call-centers of Personal Emergency Response Systems. The need for the data summarisation has evolved due to the rising costs of healthcare and increasing amount of information produced. The goal of the thesis is to find out if it is possible to generate a useful summary out of the PERS data. This was formulated in the following research question:

- *How to generate a useful natural language summary out of the event logs provided by the call-center of Personal Emergency Response Systems?*

This question was subdivided into the subquestions which were answered throughout the thesis. Further in the chapter, the summary of the results and the further research opportunities are described.

7.1 Summary of contributions

The answer to the main research question consists of the answers for the three subquestions. Therefore the answers for the subquestions are presented below.

The first research subquestion which is *"How to generate a natural language representation of the data generated by Personal Emergency Response Systems?"* was answered in chapters 2 and 3. Chapter 2 introduces the environment in which the summary should be generated. It describes the Personal Emergency Response Systems in general, the data that will be used to generate the summary, user groups which will actually read the natural language summary and the concept of the insights and the summary, describing the overall idea of the automatically generated summary. Chapter 3 describes approaches that allow to generate the linguistic summaries. Several architectures and approaches were introduced to describe the process of the natural language summary generation. Most of them follow similar process of summary generation which contains such steps as data analysis, which allows to analyse the data and to extract patterns and interesting findings, data interpretation step at which the extracted data is formed into messages or new and more high-level patterns could be found, document planning step, during which the decision regarding the events to be shown is made and the implementation step which is responsible for the generation of the final text. In this chapter were also described some data analysis techniques that could be used during the data analysis step. Further, the existing data-to-text systems were reviewed, which allowed to learn their summary generation process.

The answer for the second research subquestion *"What should be the architecture of the system that will generate natural language summary?"* was described in chapter 3 and 4. To be able to answer this question the existing data-to-text architectures were described in chapter 3. The chosen architecture follows the standard data-to-text architecture which contains four steps: data analysis, data interpretation, document planning, and realization. Based on this architecture five modules were developed: analytics module and clustering module which represent data analysis step, message generation module which represents the data interpretation step, scoring module which represents the document planning step and the summary generation step which represent the last step in the summary generation process. Each module is described in more details in chapter 4.

The third research subquestion *"How to decide which information should be present in the generated summary?"* is answered in chapters 3 and 4. Basically, the decision on which information should be included is made during the data analysis and document planning step. First, the analysis of raw data from the call center is performed and after that, it is scored according to the user defined rules. These scores reflect the interestingness of the processed data and insight. Afterwards, the obtained scores allow to decide which insights should be included in the final summary. In healthcare domain, the analysis mostly performed using the user-defined rules which allows to detect abnormal values and patterns. The rules which are used in the thesis and described in more details in chapter 4. Chapter 3 also mentions additional common data analysis techniques that can be used to improve the data analysis process. It was decided to use clustering technique which allowed to compare one subscriber to other similar subscribers.

The fourth research subquestion *"How to evaluate the summary that was generated from data generated by Personal Emergency Response Systems?"* was answered in chapter 6, where various evaluation techniques were presented. To evaluate the results of the current thesis the human experts evaluation was chosen. To perform this evaluation, we have created a questionnaire which presents the generated text and asks multiple questions that allow to collect feedback on the text. The evaluation was performed internally in Philips. The overall idea was supported by the responders, but they mention that more work on the actual text should be done. Additionally, the scores for some particular insights were examined to see if they are changing over the time.

7.2 Further steps

During the evaluation step it was found that generated texts have some downsides that make it more complicated for the readers to understand the text, therefore the first improvement step should the revision of the texts that are being generated. This would help to make it easier to read and comprehend the text for the readers. This is especially important for the elderly readers. This improvement is a vital change for the further development of the system, due to the fact that readers first focus their attention on the grammar and structure of the text and only after go to the content. If a text would contain mistakes or it will be hard to comprehend it, readers could start to doubt if the service provider can generate a quality summary. The improvement of the text generation should be performed at the document planning and realization steps of the architecture. Another step towards the step improvement could be done by adding bullet point, simple table or graphics.

It was also found that readers were interested in some additional information like more detailed description of the case; some information seemed as repetitive; some of the insights were redundant. Therefore, there is a need to revise the insights that are generated currently. Many readers mentioned that they are more interested in recent information, thus the first step could be to focus more on the recent period.

Regarding the data analysis step, there will always be room for an improvement as long as there are multiple techniques that can be applied. Currently, an initial improvement that could be implemented is the frequent items mining, to explore if there are some interrelated health issues. But this could be a challenging task due to a pretty small concentration of events. This technique would also require the domain knowledge to see if the mined data makes sense.

Bibliography

[1] 2014 sikdd test of time award winners. http://www.kdd.org/awards/view/2014-sikdd-test-of-time-award-winners. Accessed: 2017-06-27.

[2] 25.1. tkinter python interface to tcl/tk. https://docs.python.org/3.6/library/tkinter.html. Accessed: 2017-06-30.

[3] Arria platform. https://www.arria.com/platform/. Accessed: 2017-06-20.

[4] Nlg neonatal reporting. http://info.arria.com/neonatal-care-report-use-case-download. Accessed: 2017-06-20.

[5] sklearn.cluster.dbscan. http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html. Accessed: 2017-06-30.

[6] Personal emergency response systems (pers): A research brief. http://www.strategyr.com/MarketResearch/Personal_Emergency_Response_Systems_PERS_Market_Trends.asp, 2017. Accessed: Accessed 12-July-2017.

[7] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[8] Alert1. Alert1. https://www.alert-1.com/, 2017. Accessed 12-July-2017.

[9] Arria. *TECHNICAL OVERVIEW: The Arria NLG Engine*, 2014.

[10] Roberto J Bayardo Jr. Efficiently mining long patterns from databases. *ACM Sigmod Record*, 27(2):85–93, 1998.

[11] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[12] José Coch. Interactive generation and knowledge administration in multimeteo. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 300–303, 1998.

[13] Fiorella De Rosis and Floriana Grasso. Affective natural language generation. In *Affective interactions*, pages 204–218. Springer, 2000.

[14] Koenraad De Smedt, Helmut Horacek, and Michael Zock. Architectures for natural language generation: Problems and perspectives. In *Trends in Natural Language Generation An Artificial Intelligence Perspective*, pages 17–46. Springer, 1996.

[15] Dursun Delen. *Real-world data mining: applied business analytics and decision making.* FT Press, 2014.

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[17] Joshua C. Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, and Dean F. Sittig. Summarization of clinical information: A conceptual model. *J. of Biomedical Informatics*, 44(4):688–699, August 2011.

[18] Joshua C. Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, and Dean F. Sittig. Summarization of clinical information: A conceptual model. *J. of Biomedical Informatics*, 44(4):688–699, August 2011.

[19] Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. From data to text in the neonatal intensive care unit: Using nlg technology for decision support and information management. *AI Commun.*, 22(3):153–186, August 2009.

[20] Eli Goldberg, Norbert Driedger, and Richard I. Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and Their Applications*, 9(2):45–53, April 1994.

[21] GreatCall. GreatCall: Cell Phones, Medical Alert & Safety for Seniors. `https://www.greatcall.com`, 2017. Accessed 12-July-2017.

[22] Medical Guardian. Medical Guardian: Medical Alert Systems - Senior Monitoring Devices. `https://www.medicalguardian.com`, 2017. Accessed 12-July-2017.

[23] Catalina Hallett, Richard Power, and Donia Scott. Summarisation and visualisation of e-health data repositories. 2006.

[24] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[25] James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse. *Artif. Intell. Med.*, 56(3):157–172, November 2012.

[26] James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, Cindy Sykes, and Dave Westwater. Bt-nurse: Computer generation of natural language shift summaries from complex heterogeneous medical data. *Journal of the American Medical Informatics Association*, 18(5):621–624, 2011.

[27] Kathryn Hyer and Linda Rudick. The effectiveness of personal emergency response systems in meeting the safety monitoring needs of home care clients. *Journal of Nursing Administration*, 24(6):39–44, 1994.

[28] Janusz Kacprzyk, Anna Wilbik, and Sławomir Zadrozny. Linguistic summarization of trends: a fuzzy logic based approach. In *Proceedings of the 11th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 2166–2172, 2006.

[29] HJ Kaiser. Health care costs: Key information on health care costs and their impact. *Kaiser Family Foundation*, 2012.

[30] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[31] Karen Kukich. Design of a knowledge-based report generator. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 145–150. Association for Computational Linguistics, 1983.

[32] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[33] Anna S Law, Yvonne Freer, Jim Hunter, Robert H Logie, Neil McIntosh, and John 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*, 19(3):183–194, 2005.

[34] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets.* Cambridge university press, 2014.

[35] LifeFone. LifeFone: Medical Alert System & Medical Alarm. `https://www.lifefone.com/`, 2017. Accessed 12-July-2017.

[36] Philips Lifeline. Medical Alert Systems & Aging in Place Solutions — Philips Lifeline. `https://www.lifeline.philips.com`, 2017. Accessed 15-April-2017.

[37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2014.

[38] Saad Mahamood and Ehud Reiter. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21. Association for Computational Linguistics, 2011.

[39] Dee McGonigle and Kathleen Mastrian. *Nursing informatics and the foundation of knowledge.* Jones & Bartlett Publishers, 2014.

[40] World Health organisation. Frequently asked questions. `http://www.who.int/suggestions/faq/en/`, 2017. Accessed 23-May-2017.

[41] World Health Organization et al. Preamble to the constitution of the world health organization, as adopted by the international health conference, new york, pp. 19-22 (june 1946): signed on 22 july 1946 by the representatives of 61 states,(official records of the world health organization, no. 2 p. 100) and enterd into force on 7 april 1948. *http://www. who. int/-abouwho/en/definition. html*, 1948.

[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[43] François Portet, Ehud Reiter, Jim Hunter, and Somayajulu Sripada. Automatic generation of textual summaries from neonatal intensive care data. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 227–236. Springer, 2007.

[44] David G. Rand. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299:172 – 179, 2012. Evolution of Cooperation.

[45] Ehud Reiter. Has a consensus nl generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 163–170. Association for Computational Linguistics, 1994.

[46] Ehud Reiter. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 97–104, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[47] Ehud Reiter and Anja Belz. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Comput. Linguist.*, 35(4):529–558, December 2009.

[48] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems.* Cambridge University Press, New York, NY, USA, 2000.

[49] Ehud Reiter, Roma Robertson, and Liesl M. Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artif. Intell.*, 144(1-2):41–58, March 2003.

[50] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[51] Robert E Roush, Thomas A Teasdale, Jane N Murphy, and M Stella Kirk. Impact of a personal emergency response system on hospital utilization by community-residing elders. *Southern medical journal*, 88(9):917–922, 1995.

[52] Deb K Roy. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 16(3):353–385, 2002.

[53] EIT Digital Master School. Software and Service Architectures (SSA). `https://masterschool.eitdigital.eu/programmes/ssa/`, 2017. Accessed 14-April-2017.

[54] Somayajulu Sripada, Ehud Reiter, and Ian Davy. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10, 2003.

[55] Randi Stokke. The personal emergency response system as a technology innovation in primary health care services: an integrative review. *Journal of medical Internet research*, 18(7), 2016.

[56] Department of Economic United Nations and Population Division Social Affairs, editors. *World Population Ageing 2015 (ST/ESA/SER.A/390).* United Nations Copyright, New York, 2015.

[57] Anna Wilbik, James M Keller, and Gregory Lynn Alexander. Linguistic summarization of sensor data for eldercare. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 2595–2599. IEEE, 2011.

[58] Ronald R. Yager. A new approach to the summarization of data. *Information Sciences (New York)*, 28:69–86, 1982.

[59] Ronald R Yager. Quantified propositions in a linguistic logic. *International Journal of Man-Machine Studies*, 19(2):195–227, 1983.

[60] Ronald R. Yager, Kenneth M. Ford, and Alberto J. Cañas. *An approach to the linguistic summarization of data*, pages 456–468. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991.

[61] Yanbin Ye and Chia-Chu Chiang. A parallel apriori algorithm for frequent itemsets mining. In *Software Engineering Research, Management and Applications, 2006. Fourth International Conference on*, pages 87–94. IEEE, 2006.

[62] Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(01):25–49, 2007.

[63] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[64] Lotfi A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst.*, 90(2):111–127, September 1997.

[65] Lotfi A Zadeh. A prototype-centered approach to adding deduction capability to search engines-the concept of protoform. In *Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium*, volume 1, pages 2–3. IEEE, 2002.

[66] Lotfi Asker Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1(1):3–28, 1978.

[67] Sergio Zani, Marco Riani, and Aldo Corbellini. Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, 28(3):257–270, 1998.

Appendix A

Example of the questionnaire

This appendix contains an example of the questionnaire. The introduction to the questionnaire and questions will remain the same for all the questionnaires but the text will be randomly changed from questionaire to questionnaire.

A.1 Questionnaire

Dear participant, thank you for participating in the survey. Below you can find a quick introduction to the questionnaire.

Imagine that you are taking care of your elderly family member who lives independently, but not in the same household with you. You want to stay updated about his or her overall health status. By the health status, we mean the ability of the person to live independently.

Your relative uses a personal emergency response system (PERS) that allows seniors and people with special needs to connect to a call centre when they need help or assistance. The user can call either by himself or herself, or the system can make calls automatically.

There are three types of calls:

1. test calls aimed to test the proper work of the system
2. health related calls
3. neutral calls all other calls, which are not related to the health issues.

One way to keep caregivers aware of the health and wellness of the elderly person is to provide them regularly a written summary of the recent service history. The focus of this study is to assess the requirements of such summarization system and we are asking your opinion of several aspects of the summary.

Below you will find an example of the automatically generated text, which you will be asked to evaluate using guiding questions to this text. The text is generated about your relative that is using PERS service.

Dear caregiver, below we provide you a summary of the person's health situation. Ms. Smith has been using our service for almost 3.5 years. During her time using our service, the most frequent reasons Ms. Smith has contacted us has been due to she has fallen down. During that time Ms. Smith has fallen down 1 time, been in the hospital 1 time. Regarding the number of health related contacts, for the past 2 months Ms. Smith has faced 2 health related issues, while averaging 1 health related issues every 2 months. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Ms. Smith has had more health related contacts than the 2 months period preceding it. Compared with other women in her age group, Ms. Smith have a tendency to fall more. It looks like Ms.

Smith has more cases on Saturday than on other days. Dear caregiver, thank you for taking care of our client.

---

1. Please, summarise the automatically generated text in one sentence.

   Please rate the following expressions:

2. It was easy to understand the text Strongly disagree, Disagree, Not sure, Agree, Strongly Agree

3. The summary would help me to get information about the current health status of the person Strongly disagree, Disagree, Not sure, Agree, Strongly Agree

4. It would be useful to get such a summary from time to time, for example every two months. Strongly disagree, Disagree, Not sure, Agree, Strongly Agree

5. I would agree to pay additional fee for a service, which summarizes the health status of my elderly relative. Strongly disagree, Disagree, Not sure, Agree, Strongly Agree

6. Please, leave an open comment about the text. Specify what is your overall impression regarding the text.

7. If there is something additional that you want to see?

8. If there is something that is not worse mentioning?

9. Do you have any personal experience of taking care of an elderly family member?

Appendix B

Texts

In the appendix ten texts that were chosen for the evaluation are presented. Half of them were chosen manually and another half randomly. Table B.1 contains the selection details.

Text number	Selection type	Selection reason
1	random	none
2	random	none
3	random	none
4	random	none
5	random	none
6	manual	Requested assistance 3 times
7	manual	5 events for the past month, while averaging 2 events
8	manual	Had pretty big numbet of events (8)
9	manual	Requested assistance and had more events on Tuesdays
10	manual	Fall less, has been in the hospital

Table B.1: Text and their selection type and criteria.

B.1 Randomly selected texts

1

Dear caregiver, below we provide you a summary of the health situation of Mr. Smith. He has been using our service for more than 3 years. During his time using our service, the most frequent reasons Mr. Smith has contacted us has been due to he has fallen down.

In the last 2 months, Mr. Smith has been in contact with our call center 2 times. Regarding the number of health-related contacts, for the past 2 months, Mr. Smith has faced 1 health related issue while averaging 1 health related issue every 2 months. The 2 months period showing the most significant number of health related issues started on November 22, 2014. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Mr. Smith has had more health-related contacts than the 2 months period preceding it.

Compared with other men in his age group, Mr. Smith has a tendency to fall less.

Dear caregiver, thank you for taking care of our client.

2

Dear caregiver, below we provide you a summary of the health situation of Mr. Smith. He has been using our service for almost 3.5 years. During his time using our service, the most frequent reasons Mr. Smith has contacted us has been due to he has fallen down.

Regarding the number of health-related contacts, for the past 2 months, Mr. Smith has faced 1 health related issue while averaging 0 health related issues every 2 months. The 2 months period showing the most significant number of health related issues started on January 20, 2014. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Mr. Smith has had more health-related contacts than the 2 months period preceding it.

Compared with other men in his age group, Mr. Smith has a tendency to fall less. Mr. Smith fell less than other people in rural areas.

Dear caregiver, thank you for taking care of our client.

3

Dear caregiver, below we provide you a summary of the health situation of Mr. Smith.

He has been using our service for almost 3.5 years. During his time using our service, the most frequent reasons Mr. Smith has contacted us has been due to he has fallen down. During that time Mr. Smith has faced a breathing problem 1 time.

Regarding the number of health-related contacts, for the past 2 months, Mr. Smith has faced 1 health related issue, while averaging 1 health related issues every 2 months. The 2 months period showing the most significant number of health related issues started on May 27, 2014. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Mr. Smith has had more health-related contacts than the 2 months period preceding it.

Mr. Smith fell less than other people in rural areas.

Dear caregiver, thank you for taking care of our client.

4

Dear caregiver, below we provide you a summary of the health situation of Ms. Smith. She has been using our service for almost 3.5 years. During her time using our service, the most frequent reasons Ms. Smith has contacted us has been due to she has fallen down.

In the last 2 months, Ms. Smith has been in contact with our call center 5 times. During that time Ms. Smith has fallen down 4 times. Regarding the number of health-related contacts, for the past 2 months, Ms. Smith has faced 4 health-related issues, while averaging 2 health related issues every 2 months.

Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Ms. Smith has had more health-related contacts than the 2 months period preceding it.

Compared with other women in her age group, Ms. Smith has a tendency to fall less.

Dear caregiver, thank you for taking care of our client.

5

Dear caregiver, below we provide you a summary of the health situation of Ms. Smith. During her time using our service, the most frequent reasons Ms. Smith has contacted us has been due to she has fallen down.

In the last 2 months, Ms. Smith has been in contact with our call center 4 times, 1 of which was test call. Regarding the number of health-related contacts, for the past 2 months, Ms. Smith has faced 4 health-related issues, while averaging 2 health related issues every 2 months. The 2 months period showing the most significant number of health related issues started on November 28, 2015. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Ms. Smith has had more health-related contacts than the 2 months period preceding it.

Compared with other women in her age group, Ms. Smith has a tendency to fall more. It looks like Ms. Smith has more health related calls on Tuesdays than on other days.

Dear caregiver, thank you for taking care of our client.

B.2 Manually selected texts

6

Dear caregiver, below we provide you a summary of the health situation of Ms. Smith. During her time using our service, the most frequent reasons Ms. Smith has contacted us has been due to she has fallen down.

In the last 2 months, Ms. Smith has been in contact with our call center 2 times. During that time Ms. Smith has faced a breathing problem 2 times. Regarding the number of health-related contacts, for the past 2 months, Ms. Smith has faced 2 health-related issues, while averaging 1 health related issue every 2 months. The 2 months period showing the most significant number of health related issues started on August 16, 2014. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Ms. Smith has had more health-related contacts than the 2 months period preceding it.

It looks like Ms. Smith has more cases on Friday than on other days.

Dear caregiver, thank you for taking care of our client.

7 Dear caregiver, below we provide you a summary of the health situation of Ms. Smith. She has been using our service for almost 3.5 years. During her time using our service, the most frequent reasons Ms. Smith has contacted us has been due to she has fallen down.

During that time Ms. Smith has requested assistance 3 times, fallen down 2 times. Regarding the number of health-related contacts, for the past 2 months, Ms. Smith has faced 5 health-related issues, while averaging 1 health related issue every 2 months. The 2 months period showing the most significant number of health related issues started on March 16, 2016. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Ms. Smith has had more health-related contacts than the 2 months period preceding it.

Compared with other women in her age group, Ms. Smith has a tendency to fall less.

Dear caregiver, thank you for taking care of our client.

8

Dear caregiver, below we provide you a summary of the health situation of Ms. Smith. During her time using our service, the most frequent reasons Ms. Smith has contacted us has been due to she has fallen down.

In the last 2 months, Ms. Smith has been in contact with our call center 8 times, 1 of which was test call. During that time Ms. Smith has fallen down 6 times, requested assistance 1 time. Regarding the number of health-related contacts, for the past 2 months, Ms. Smith has faced 8 health-related issues, while averaging 2 health related issues every 2 months. The 2 months period showing the most significant number of health related issues started on December 22, 2015. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Ms. Smith has had more health-related contacts than the 2 months period preceding it.

Compared with other women in her age group, Ms. Smith has a tendency to fall more. It looks like Ms. Smith has more cases on Sunday than on other days.

9

Dear caregiver, below we provide you a summary of the health situation of Mr. Smith. He has been using our service for almost 3.5 years. During his time using our service, the most frequent reasons Mr. Smith has contacted us has been due to he has fallen down.

During that time Mr. Smith has requested assistance 2 times. Regarding the number of health-related contacts, for the past 2 months, Mr. Smith has faced 2 health-related issues, while averaging 1 health related issues every 2 months. The 2 months period showing the most significant number of health related issue started on November 12, 2015. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Mr. Smith has had more health-related contacts than the 2 months period preceding it.

It looks like Mr. Smith has more cases on Tuesday than on other days.

Dear caregiver, thank you for taking care of our client.

10

Dear caregiver, below we provide you a summary of the health situation of Ms. Smith. During her time using our service, the most frequent reasons Ms. Smith has contacted us has been due to she has fallen down.

In the last 2 months, Ms. Smith has been in contact with our call center 4 times, 1 of which was test call. During that time Ms. Smith has fallen down 1 time, been in the hospital 1 time. Regarding the number of health-related contacts, for the past 2 months, Ms. Smith has faced 2 health-related issues, while averaging 0 health related issues every 2 months. The 2 months period showing the most significant number of health related issues started on March 12, 2016. Compared to the typical 2 months period, this last period showed a larger number of health related calls. During the most recent period, Ms. Smith has had more health-related contacts than the 2 months period preceding it.

Compared with other women in her age group, Ms. Smith has a tendency to fall less.

Dear caregiver, thank you for taking care of our client.

Appendix C

Text templates

- subscriber
  1. "You have been using our service for " + period length + ". "
  2. "In the last n months you have been in contact with our call center m times k of which were test calls."
  3. "During your time using our service, the most frequent reasons you have contacted us has been due to " + reasons + ". "
  4. "During that time you had " + recent_calls + ". "
  5. "You fell more/less than other people in your age group."
  6. "Compared with other men/women in your age group, you have a tendency to fall less/more."
  7. "You fell less/more than other people in urban/rural area. "
  8. " The n months period showing the most significant number of health related issues started on " + date + ". "
  9. "Compared to the typical n period, this last period showed a smaller/bigger/equal number of health related calls."
  10. "During the most recent period you have had less/more health related contacts than during the preceding n-months period. "
  11. "For the past n you have faced m health related issues, while averaging k health related issues every n months."
  12. "It looks like you have more health related calls on " + week day + " than on other days. "
  13. Final phrase: "Dear Ms/Mr X, thank you for taking time to read the summary."
  14. Random quote in the end of the summary: "we want you to know that we are here for you and aim to provide help when you need it.", "we would like to continue helping you to live safely and independently at home for the years to come.", "we want you to know that we are here to help you to continue living safely and independently. ", "we are ready to keep providing help to ensure your safely at home. "
- caregiver
  1. Ms/Mr X + " has been using our service for " + period length + ". "
  2. "In the last n months Mr/Ms X has been in contact with our call center m times k of which were test calls."

3. "During her/his time using our service, the most frequent reasons Ms/Mr X has contacted us has been due to " + reasons + ". "
4. "During that time Mr/Ms X had " + recent_calls + ". "
5. "Mr/Ms X fell more/less than other people in his/her age group."
6. "Compared with other men/women in his/her age group Mr/Ms X has a tendency to fall less/more."
7. "Mr/Ms X fell less/more than other people in urban/rural area."
8. "The n period showing the most significant number of health related issues started on " + date + ". "
9. "Compared to the typical n period, this last period showed a smaller/bigger/equal number of health related calls."
10. "During the most recent period Mr/Ms X has had less/more events than during the preceding n-months period. "
11. "Regarding the number of health related contacts, for the past n Mr/Ms X has faced m health related issues, while averaging k health related issues every n months. "
12. 'It looks like Mr/Ms X has more health related calls on " + week day + " than on other days. "
13. Final phrase: "Dear caregiver, thank you for taking care of our client."

- general practitioner

1. Ms/Mr X + " has been using our service for " + period length + ". "
2. "In the last n months Mr/Ms X has been in contact with our call center m times k of which were test calls."
3. "During her/his time using our service, the most frequent reasons Ms/Mr X has contacted us has been due to " + reasons + ". "
4. "During that time Mr/Ms X had " + recent_calls + ". "
5. "Mr/Ms X fell more/less than other people in his/her age group."
6. "Compared with other men/women in his/her age group Mr/Ms X has a tendency to fall less/more."
7. "Mr/Ms X fell less/more than other people in urban/rural area."
8. "The n period showing the most significant number of health related issues started on " + date + ". "
9. "Compared to the typical n period, this last period showed a smaller/bigger/equal number of health related calls."
10. "During the most recent period Mr/Ms X has had less/more events than during the preceding n-months period. "
11. "Regarding the number of health related contacts, for the past n Mr/Ms X has faced m health related issues, while averaging k health related issues every n months. "
12. 'It looks like Mr/Ms X has more health related calls on " + week day + " than on other days. "
13. Final phrase: "Dear doctor, thank you for taking care of our client."