Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2069–2077
Marseille, 11–16 May 2020

A Term Extraction Approach to Survey Analysis in Health Care

Cécile Robin, Mona Isazad Mashinchi, Fatemeh Ahmadi Zeleti, Adegboyega Ojo, Paul Buitelaar
Insight Centre for Data Analytics, NUI Galway
Galway, Ireland
{cecile.robin, mona.isazad, fatemeh.ahmadizeleti, adegboyega.ojo, paul.buitelaar }@insight-centre.org

Abstract

The voice of the customer has for a long time been a key focus of businesses in all domains. It has received a lot of attention from the research community in Natural Language Processing (NLP) resulting in many approaches to analysing customers feedback ((aspect-based) sentiment analysis, topic modeling, etc.). In the health domain, public and private bodies are increasingly prioritising patient engagement for assessing the quality of the service given at each stage of the care. Patient and customer satisfaction analysis relate in many ways. In the domain of health particularly, a more precise and insightful analysis is needed to help practitioners locate potential issues and plan actions accordingly. We introduce here an approach to patient experience with the analysis of free text questions from the 2017 Irish National Inpatient Survey campaign using term extraction as a means to highlight important and insightful subject matters raised by patients. We evaluate the results by mapping them to a manually constructed framework following the Activity, Resource, Context (ARC) methodology (Ordenes et al., 2014) and specific to the health care environment, and compare our results against manual annotations done on the full 2017 dataset based on those categories.

Keywords: health care, patient experience, term extraction, natural language processing, ARC framework, patient engagement, evaluation methodology

1. Introduction

Patient engagement became recently a focus in Ireland with the creation three years ago of the National Care Experience Programme[1] in partnership between the Health Information and Quality Authority (HIQA), the Health Service Executive (HSE) and the Department of Health. The programme is responsible for the National Inpatient Experience Survey (NIES) conducted for the first time in 2017 in 39 public hospitals from six hospital groups, and is now in its third edition. The first edition of the survey collected feedback of patients who spent at least 24 hours in a public hospital during the month of May 2017, and retrieved 13,000 responses. The questions were mainly of multiple choice type, except three open-ended questions at the end asking responders to provide more details on the positive or negative experiences they had, as well as suggestions for improvement. The survey was designed to assess patients' experience during their stay and to identify particular areas in need of improvement. The aim for health care organizations is to be able to act on patient feedback, and create solutions for better service in terms of quality and safety within hospital care.

An analysis of the close-ended survey questions was previously carried out, and a project with the Insight Centre for Data Analytics[2] (Galway, Ireland) was established to analyse and provide findings in the comments written in natural language from the open-ended questions of the 2017 and 2018 editions of the survey. From this work resulted several outcomes, among which the creation of a framework specific to the field of health and hospital care for classifying the different complaints and compliment into categories. Those categories are divided into three types following the Activity, Resource and Context (ARC) methodology defined in Ordenes et al. (2014) (see section 2.1. for more details on the approach). From the project was also produced a dataset of comments manually annotated with the ARC framework categories mentioned above.

The contribution provided in this paper is twofold. We first introduce an intermediate level term extraction approach to patient experience analysis, with the claim that this level of granularity can help in identifying important and particular aspects of complaints and compliments, and make them more easily actionable than typical broader level terms. Second, we describe a novel approach to term extraction evaluation by testing our hypothesis through a mapping to the ARC framework. All experiments in this study were performed on the 2017 dataset of answers, as it was the only fully annotated corpus available at the time of the study.

We first give an overview of the current and previous work done in the field of survey analysis, customer feedback and patient engagement. We then explain our hypothesis of using intermediate-level term extraction in the context of patient engagement, followed by the term extraction methodology and the experiment itself. We describe next our approach to term extraction evaluation for the purpose of this task. We finally provide results and a discussion on the term extraction experiments and the outcomes of the evaluation task.

2. Related Work

2.1. Automatic Approaches to Survey Analysis and Customer Feedback

Obtaining customer feedback through surveys is an efficient way to gauge people's reactions and appreciations to a product, a service, or measure the popularity of a brand/company/organization. The structured feedback provides a quick quantitative insight to customer's general feeling. However, the so-called Voice Of the Customer is precious for any organization, and one might want to capture the precise elements and aspects of their complaints or compliments in order to detect the major problems to be addressed (or things to pursue). This is one limitation of tick-

[1] https://yourexperience.ie/about/who-we-are/
[2] https://www.insight-centre.org/

box type of questions from surveys, along with the potential bias conveyed by the questions and the options to choose from. Therefore, it is not uncommon to find at the end of a questionnaire one or more open questions, seeking for general feedback or other comments to be made (O'Cathain and Thomas, 2004). The data retrieved from those sections represents a mine of information, and care must be taken in making the most of it. A qualitative analysis of this data is needed, however the cost and effort to manually analyze the replies makes automatic tools an advantageous choice. Natural Language Processing (NLP) techniques using text mining and topic modeling in particular are the most popular approaches for this task.

Espinoza et al. (2018) describe a semi-automatic approach based on automatic text clustering using distributional semantics models. However interpreting those clusters is not straightforward, and they consist of single word terms which can be too broad or vague to allow a correct interpretation. For instance "location" for a hotel review can relate to its distance to the city centre, or the atmosphere of the area (dodgy, lively...). Their method also still requires human effort to manually rework the clusters. Ordenes et al. (2014) explore a holistic approach using a framework of Activities, Resources and Contexts (ARC) to describe each customer compliment or complaint through those three key components. They build a text mining model based on linguistic patterns inspired by data they previously manually annotated using the ARC framework. However, the model they developed is specific to the type of service in which their research was tested thus not directly applicable to our domain of focus.

In the health care area specifically, Maramba et al. (2015) investigated web-based tools to approach patient feedback analysis. Text clouds are widely used within the health care community for surveys and forum messages analysis. In their study they compare three different Web applications for text cloud generation. They find difficult to estimate the significance of the clouds because of the loss of information that results from the dissociation from the context, and express worries about missing out on nuances. They reckon more sophisticated technologies from the NLP domain are needed for a more informative insight into the data.

2.2. Approaches to Patient Engagement

In general, a more valuable information for improvement is provided to practitioners by considering patient perspectives and feedback across the care continuum, rather than looking at specific services in isolation from a clinical or hospital management viewpoint (Cunningham and Wells, 2017)).

Previous experiences from patient survey analysis have shown that a quantitative analysis can highlight the general feeling of patients, but it can be limited in terms of giving adequately detailed explanations of problems which matter to them. In the first Scottish Cancer Patient Experience Survey (SCPES), in the goal to understanding patients' experiences of care, Cunningham, and Wells (2017) implemented a thematic analysis on all free-text comments provided by participants through the survey. This way of collecting patients experience provided policymakers with a more in-depth insight into particular issues within the healthcare system. Also, the qualitative analysis of comments had positive impacts on the public, as patients felt that their needs were understood which consequently improved their confidence in the healthcare system.

In another study looking to provide more insights into the experience of patients with cancer in the London National Health Service (NHS) trusts, researchers have employed a framework analysis approach to process patients comments. They designed a thematic framework used to analyse 15403 comments from over 6500 patients. Their results showed mainly positive comments, while those the negative ones were more related to the quality of care, with a focus on poor communication, inadequate care, and waiting times in outpatient departments (Wiseman et al., 2015). The qualitative analysis targets providers to be more informed about the areas of care in need of improvement.

3. Term extraction for Patient Engagement

Similar to the UK, in Ireland for the first time, the NIES explored the experiences of adult patients admitted to Irish acute hospitals in May 2017 and 2018. The majority of the questions in the questionnaire were in Likert-scale format, however three open-ended questions were included in the survey, which enabled patients to describe their experience in more detail. A framework analysis of the free-text data allowed researchers to categorize participants comments into 23 themes. These themes were further grouped into ten themes, to give planners a better overview of the patients experience. Free text questions provide an actionable insight into the experiences of patients. Such feedback gives healthcare organizations an opportunity for double-loop learning, allowing them to revisit some of the assumptions underlying their services (Reddick et al., 2017).

As presented in the related work, most approaches to free text question analysis make use of text mining and term extraction techniques. When annotated data is not available for machine learning algorithms, frequency-based techniques are commonly used to assess the importance of a term and most often rely on single words, thus often failing in effectively representing and detecting the particular aspects of the complaints and compliments. We use in this study an approach to term extraction (further described in 4.2.1.) which we claim provides a better insight to the patient experience than more typical term extraction techniques. By targeting terms at an intermediate level of specificity, we allow to cover important information regarding the claims of the patients while preserving the generality needed to highlight major themes. This method would therefore reduce the cost of future manual work, while preserving the authenticity of the data and the main outcomes that comes from it.

4. Term Extraction Experiment

4.1. Dataset

The dataset is split into two corpora (one for positive comments and one for negative comments), each containing individual replies to the following open questions of the National Patient Survey 2017 questionnaire:

- Q59 Was there anything particularly good about your hospital care?
- Q60 Was there anything that could be improved?

together with a set of metadata including information such as the type of admission, gender of responder, age group etc. (if available). In order to ensure that the comments do not allow to identify particular authors or practitioners, some elements have been carefully anonymized before they were shared with us. All anonymized pieces of text were replaced by the type of information anonymized in square brackets, such as patients' or doctors' names, addresses, specific types of condition, dates, etc. (eg. "[Doctor name]"). Some more details on the number of responses are given below in table 1.

Question	Nb answers to Q (%)
Q59 (Positive)	$9,254$ (67.52%)
Q60 (Negative)	$8,130$ (59.32%)
Total questionnaires collected	$13,705(100\%)$

Table 1: Dataset statistics

After a first glance at the data, it appeared that many survey responders did not answer one or any of the open questions, or provided an invalid or unusable reply for our analysis. We identified those noisy entries in our dataset and filtered them out, in particular empty fields, variations of "NaN"/ "null"/"none", or also "everything" (which is not specifying any specific aspect of improvement or quality therefore discarded), etc. Table 2 presents statistical measures of the comments in the dataset: average length (in character), standard deviation (SD), percentiles and the maximum length (in character).

Question	Avg.	SD	25%	50%	75%	Max
Positive	96	112	33	66	127	3497
Negative	129	200	23	74	168	3714

Table 2: Some statistics on the comments from the dataset

4.2. Term Extraction

We adopt a term extraction strategy for tackling the detection of particular aspects of concerns raised by patients, as shown to be the preferred approach in previous literature for similar or related cases. However, we overcome limitations of previous studies by adapting our approach towards deeper and more explainable results with terms of a higher degree of specificity needed for this study. For instance, we want to reduce as much as possible human effort and interaction with the system for time and cost purpose. We also want a system that could be reused for future surveys (the 2019 campaign is currently being closed), and that is not dependent of external resources for reasons that will be explained more in detail below.

Our term extraction methodology makes use of features provided by the open source knowledge extraction tool Saffron[3]. The flexibility of this tool allows us to specify different elements of the term extraction process, and tailor the system to suit our needs better. The general approach follows commonly used steps of term extraction as described below, and the custom features chosen are given in more detail in 4.2.2..

4.2.1. Approach

Candidate Term Selection: The first step in the term extraction process is to select all potential terms - candidate terms - from the dataset which are not necessarily domain specific yet. For this, we focus on noun phrases as the unit carrying most of the information, following well-known characteristics such as limitations on authorized POS tags and patterns of tags to be considered a noun phrase.

In addition, a minimum and maximum length, and a minimum number of appearance within the corpus can be specified. Saffron uses Apache OpenNLP models[4] for the linguistic analysis (lemmatization, POS tagging).

Scoring: The second step makes use of scoring functions to calculate the relatedness between the noun phrase and the domain. Several functions following different approaches are available to choose from in Saffron. Some of them rely essentially on occurrence frequency, some on reference corpora, etc. (see (Astrakhantsev, 2018) for a detailed review of different types of scoring).

Ranking and Filtering: These scores obtained are then used to rank the candidate terms by relevance. An individual scoring function can be chosen, or a few of them combined with a voting algorithm approach to aggregate them (Zhang et al., 2008). A threshold can be specified by selecting top N terms, or terms observing a minimum scoring value, to obtain the final list of terms selected for the task ordered from the most relevant to the least relevant for the domain of the dataset.

4.2.2. Settings

In this study, we are looking for intermediate level terms which, as opposed to high level terms, capture more precise facets on the complaints and compliments and help to provide a deeper analysis. With this in mind, we defined the following settings for our experiment. We allow terms of minimum two and maximum five words, as terms that are longer in length provide more information than single words terms which are very generic (Bordea et al., 2013) (Maramba et al., 2015). We only consider terms that appear at least twice in the corpus to remove noise and reduce irrelevance. We also select the OpenNLP Perceptron model for the POS tagging as we found that it gave better accuracy than the Maxent model (default in Saffron) for the texts we had. We keep the default settings of Saffron for the noun phrase detection specification, following the recommendations of Bordea et al. (2013). Finally, we set the limit of maximum terms to generate to 1000, so that we focus on extracting fewer but more pertinent terms over a less meaningful bigger quantity.

[3] https://github.com/insight-centre/saffron

[4] http://opennlp.sourceforge.net/models-1.5/

The second step is the selection of the scoring function. As Zhang et al. (2008) observed in their comparison of term extraction methods, there is no single Automatic Term Extraction (ATE) algorithm that consistently performs well in all domains. They propose in Zhang et al. (2018) an additional layer aiming at improving a wide range of ATE algorithms, by relying on external resources from the domain corpus. However, our experiment corpus is quite uncommon in that the domain is at the crossroads of everyday life concepts (waiting, meals, etc.) and the medical environment (doctors, wards, A&E), without being scientifically medical (it is unlikely to find many technical terms on drugs, enzymes, etc.). Medical resources are therefore not relevant here, and using data from previous patient experience studies in other countries might bias our findings towards problems experienced by the health care systems of those countries. In this study, we choose to use the ComboBasic scoring function. This method derives from Bordea et al. (2013)'s original methodology, a variant of C-value (Frantzi et al., 2000), however it takes into account the embeddedness of a term within, and also as part of other terms to calculate its importance in the corpus. This promotes terms that are used to create more specific terms, and also which are using a more generic term in their composition, therefore that are from a more intermediate level, which suits our intention. In addition, this method has the benefit of not using external corpora either from or outside the domain.

5. Evaluation Procedure

5.1. Approaches

5.1.1. Traditional Approaches

Traditionally, term extraction evaluations are performed against data manually annotated by either experts in the specific domain, like the GENIA corpus (Kim et al., 2003) used by Zhang et al. (2018), or often by non-experts trained to annotate for the specific task, such as the Patent dataset used as evaluation by Judea et al. (2014), or in the SemEval 2017 Task 10 (Augenstein et al., 2017). Most of the time, those gold standards are taken as a base to calculate the F1 score and average precision, based on the exact matching of terms. For many methods extracting single word terms, it is easier to evaluate the precision based on exact matching. It also avoids the difficulty of deciding how to account for terms that are "partially" right.

Multi-word terms, even though more informative, create more challenges for the evaluation than single words as the length and structure of the term can vary a lot between different methods, and there can be disagreements between annotators on where to set the limits. Whether to include some modifiers in a term or not can for example be more or less relevant depending on the degree of specificity wanted, which depends on the domain and the use case. In the Aspect Term Extraction sub task of SemEval 2014 Task 4 (Pontiki et al., 2014), despite the strict guidelines given for the task of manual creation of the gold standard, the authors acknowledge disagreements between the annotators on deciding the exact boundaries of the terms (whether to include conjunctions, some adjectives, etc.)

Astrakhantsev (2018) and Šajatović et al. (2019), who both allow for multiword term extraction, rely on exact matching to evaluate their method with the average precision at K. They thus do not take any consideration for partial or close match. Zhang et al. (2008) evaluate different term extraction algorithms by asking three judges to annotate 300 candidate terms with the guidelines to "mark those they believed to be terms one would expect to encounter when reading texts about animals", which allows some flexibility on the term length for the calculation of the precision.

5.1.2. Term - Framework Mapping Approach to Evaluation

Taking into account previous experiences on this type of task, we propose here a novel evaluation framework to measure the quality of the terms extracted. We compare them against a gold standard, not at the term level but instead by mapping the terms to higher categories from a domain specific terminology to which we have access (the ARC framework), and use it as a point of comparison.

One main motivation for the particular study described in this paper is to get a meaningful and representative insight into the (good and bad) experiences of patients by using automatic techniques, saving on time and the effort of manual work for future campaigns. The 2017 edition of the NIES campaign was the first one of its kind in Ireland and no similar data was ever previously gathered in the country. The goal of the overall project in which this study integrates is to report on the findings from the content of the comments, therefore a manual annotation of the dataset based on the ARC framework (delved into more details in 5.2.1.) was performed and provided as part of the outcomes of the project. This richly annotated data was made available to us within the boundaries of the project, which we could use to evaluate our automatic method and create a baseline for future NIES campaigns.

In order to evaluate the quality of our system, we wish to compare the terms extracted by our automatic approach with the text segments manually annotated with the framework. However, those segments can be very variable in length and linguistic unit types it can contain (a single adjective, a verb and a complement, a whole sentence, etc.) as opposed to the restricted characteristics used to define a candidate term during the automatically process. It is therefore not possible to assess the quality of the terms extracted using direct string matching with this resource. Instead, we propose a novel approach for term extraction evaluation, making use of the higher level categories from the manually constructed ARC based framework. This validation is performed in several steps.

We first map the automatically extracted terms with the categories from the ARC framework. For this, we use a similarity metric (Jaccard coefficient) to match our terms with the text segments manually annotated from the comments in both the negative and positive corpora. We then map the terms with their corresponding framework categories using the manual annotations as a bridge to connect the two. Second, we count the occurrences of each category in both the manual and the ARC mapped dataset, and we rank them based on their absolute frequency. We therefore obtain two ranked lists, one directly deriving from the gold standard

of manual annotations, and one created through a mapping technique from automatically extracted terms. An illustration of the evaluation is presented below in figure 1.
This method thus compares the top results of the automatic approach and of the manual one at the framework category level. It allows us to concentrate more on the quality of the topics raised (and by extension refer back to the aspects of complaints and compliments conveyed in the more specific terms) rather than on the precise form of the term, which is more adapted to the type of outcome we want to provide here.

5.2. Evaluation Dataset

As we described previously, these experiments are part of a bigger project involving a deep qualitative report on the results of the open ended questions. As part of it, a framework was created, and the dataset of comments manually annotated with these categories specifications, divided into three types.

5.2.1. The ARC Approach

The framework described here is based on the ARC approach, which includes three value-creating types. It was constructed with the initial aim of helping to systematically organize the large number of patients' textual comments into manageable and meaningful chunks of information. We describe the three types of the ARC framework in more detail below.
Activity - Services which consist of activities between patients and healthcare providers. *Activities* are lower level concepts used to illustrate specific care patients receive under each stage of care. Eg. *"Cleaning (Ward)"*
Resource - Resources are typically the entities with which (or whom) the users interact to realize their goals. To ensure consistency, the *resource* aspect of the ARC framework was developed based on a terminology, SNOMED[5], a global standard for health terms. Eg. *"Nurse"*, *"Wheelchair"*
Context - Patients' evaluation of their healthcare service experience is provided in a specific context identified by the patient. The main contextual elements affecting the patient experience are situational and personal factors. Thus, *context* involves activities beyond the direct control of the service provider. Eg. *"Long stay in ward"*, *"Staff under pressure"*
Both the *activity* and *context* types were developed by two domain specialists using available literature and the NIES official reports on the close ended questions. A second round of review for agreement and refinement was done with a third contributor. In total, the framework contains 32 categories for the *activity* type while the *resource* type contains 242 categories, and the *context* 256.

5.2.2. Dataset Annotation

The annotation of the dataset was performed by three coders trained for this task. They were asked to annotate any text segment that would refer or induce a particular framework category, without restrictions on length or linguistic unit. The annotators were in constant interaction with each other during the task in order to maintain a consistency of decisions and insure the quality of the annotations. 10% of the comments were commonly annotated by the three participants in order to determine the agreement. The calculated average Cohen's Kappa (two by two) for the inter-rater reliability was of 0.66 for the 2017 dataset over all open ended questions, which is commonly considered a substantial agreement. In total, the final dataset used for evaluation in this study comprises 54095 annotated text segments in the negative comments and 24791 in the positive comments. Annotations referring to the most generic category (that is *all*), used when the reply does not give a particular opinion on an aspect of the care, were discarded from our study. Table 3 presents the top 5 annotation categories found for the *activity* type in the negative comments.

Activity	**Occurrences**
Patient Care on the Ward	3018
Providing facilities (Ward)	1322
Staff Management (Ward)	1276
Meal and Catering (Ward)	1166
Patient Care in Emergency	975

Table 3: Top 5 *activity* framework categories from the manually annotated dataset

More than 96% of the annotations were found to cover half of all framework categories and above 82% covered a quarter of available categories. This distribution demonstrates that most concerns and compliments of patients are localised around a few themes of importance only.

5.3. Term - Taxonomy Mapping

We use the manually annotated corpus as a means to map the terms we extracted with the more generic framework categories, and explore if the automatic system reveals the same or similar important themes to address as the manual annotations highlights. We want to evaluate how much the limited amount of automatic extracted terms can be as representative as the exhaustive manual annotations. We therefore take the two lists of 1000 terms we extracted for each of the negative and positive corpus, then look for instances of them in the comments and verify whether they match a corresponding annotated text segments or not.
Since the annotators were not given limitations in the length of text to select in the manual tagging task, the annotated text segments are very variable, and exact matches between the manual and the automatic procedure are therefore very limited. In the negative corpus, among the 54095 manually annotated text segments, only $1,237$ instances of the automatic terms found an exact match (approx. 2.3% of the total annotations, eg. "quality of food"), and 771 out of $24,790$ in the positive corpus (approx. 3% of the total annotations, eg. "good experience"). We also explore the possibility to account for close and related matches, by considering a valid association any term that loosely conveys the concept of the text segments (eg. "was on trolley in corridor" and "trolley in corridor", "private medical details discussed" and "medical detail").

[5] www.snomed.org

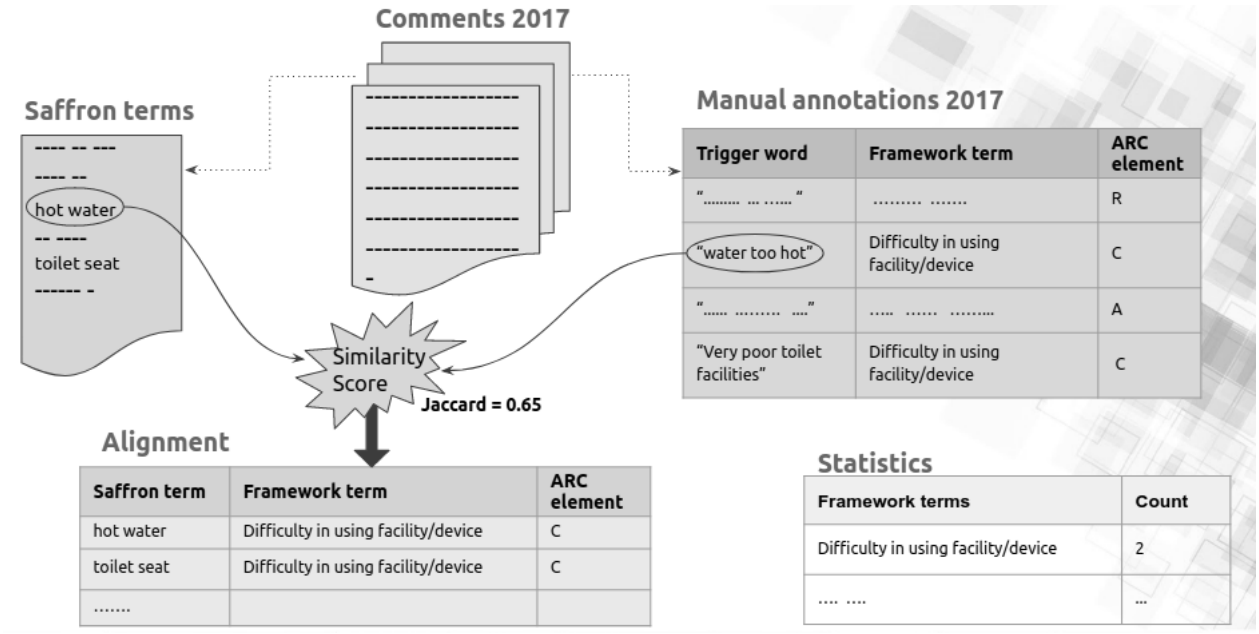

Figure 1: Mapping process for the creation of a higher level evaluation dataset

We first match the extracted terms with the manually annotated text segments. For this task, we measure the string similarity between each pair by calculating the Jaccard similarity index. We retained matching pairs above a manually set threshold of 0.25, under which the quality of the approximate matching significantly dropped.
After the similarity matching process, we found that the instances of terms covered 9.4% of the annotations in the negative comments, and 12.3% in the positive comments. See the breakdown per *activity*, *resource* and *context* in table 4.

Sentiment	All	Activity	Resource	Context
Negative	9.4%	9.7%	6.2%	14.5%
Positive	12.3%	7.4%	20.1%	7.2%

Table 4: Percentage of manual annotations covered by the automatic terms for each ARC type

Finally, we aggregate the terms at the framework level and create the ranking by number of occurrence of each category, for each ARC type. Our evaluation will compare this ranked list with the ranking from the manual annotations, for each one of the negative and positive corpus, and for each ARC element.

6. Results and Discussion

6.1. Term Extraction Experiment Results

In table 5, we present the top 15 terms ranked by score resulting from the experiments described in the previous section.
We observe some overlap between terms extracted from positive and negative comments, for example *nursing staff*, *stay in hospital*, *member of staff* and *time in hospital*. Those terms tend to relate more to resources with which

Positive Term	Negative Term
stay in hospital	nursing staff
nursing staff	private health insurance
good care	other patient
level of care	lack of communication
excellent care	health insurance
member of staff	stay in hospital
great care	private health
care from doctor	time in hospital
good experience	lot of pain
hospital staff	member of staff
hospital care	length of time
first time in hospital	cup of tea
members of staff	waiting time
medical staff	lot of pressure
time in hospital	family member

Table 5: Top 15 Terms from positive and negative comments

patients are typically confronted, therefore the experience can typically go towards one sentiment or another. In total, out of the terms extracted for each corpus about 30% were overlapping. Some unique terms found in the negative comments included *private health insurance*, *lack of communication*, *lot of pain* or *lot of pressure*, and in the positive comments *care from doctor*, *friendliness of staff* or else *excellent care*. We can see that these intermediate-level terms give some meaningful information about the different aspects of service appreciated by patients or to focus on for improvement, as opposed to the level that could be achieved with single word terms, very likely to miss important aspects on topics that are crucial for patients. We also notice that our method generated several terms em-

bedded into each other, such as *private health insurance*, *health insurance*, *private health*. This allows to put an emphasis on all terms which belong to a broader concept about health insurance: "I have private health insurance and was in a public ward", "Despite having all private health details", "I never gave my health insurance number".

This qualitative analysis of the terms extracted with our method proves how a fully automatic approach can provide an informative representation of patients' experience and we can already have at a first glance an idea of typical problems to be addressed, or things that patients particularly appreciate. In the following section, we will evaluate our results against the manually annotated corpus (following the procedure described in 5.1.2.) and provide a more quantitative analysis of the results.

6.2. Framework Mapping Experiment Results

After performing the mapping between the terms and ARC framework as described in the previous section, we compare the ARC-based ranked lists obtained for each of the automatic and manual analysis. Table 6 presents an example of an extract of the ranked list obtained for the *activity* type in positive comments.

Rk	Manually extracted Framework Terms	Automatically Generated Terms
1	Patient Care on the Ward	Patient Care on the Ward
2	Meal and Catering (Ward)	Patient treatment
3	Patient treatment	Meal and Catering (Ward)
4	Cleaning (Ward)	Providing facilities (Ward)
5	Providing facilities (Ward)	Patient Care in Emergency
6	Communication / Information Exchange with Patient (Ward)	Cleaning (Ward)
7	Patient Care in Emergency	Communication / Information Exchange with Patient (Ward)
8	Surgery / procedure (treatment)	Staff Management (Ward)
9	Outpatient	Communication / Information Exchange with Patient (Treatment)
10	Communication / Information Exchange with Patient (Treatment)	Admission

Table 6: Top 10 *activity* framework categories for positive comments

6.3. Evaluation

We perform two types of tests to measure the quality of our approach: first, the recall at k with k=10 and k=20 as we are interested in the most important topics for patients, and second we calculate the Spearman correlation rank, based on the top 20 framework categories for each ARC type (which we called *Spearman 20*), and for the whole list of categories (*Spearman*). This way, we want to compare the lists of results while accounting for the ranking, and test on both the most important categories, and on all of them. The evaluation results are presented in the table 7 for positive comments and 8 for negative comments.

ARC	Recall 10	Recall 20	Spear. 20	Spear.
Act.	80%	85%	0.79	0.70
Res.	80%	70%	0.45	-0.13
Cont.	70%	75%	0.41	0.16

Table 7: ARC framework mapping evaluation for positive comments

ARC	Recall 10	Recall 20	Spear. 20	Spear.
Act.	90%	90%	0.77	0.89
Res.	70%	80%	0.62	-0.26
Cont.	50%	65%	0.4	0.34

Table 8: ARC framework mapping evaluation for negative comments

The recall on k=10 and k=20 show high results for all types, with over 70% in most cases except for the type *context* in negative comments. The best recall and Spearman correlation coefficient scores were achieved for the *activity* type in both the negative and positive comments with between 80% and 90% of recall, and 0.7 to 0.9 of Spearman correlation. The lowest recall was obtained for the *context* in the negative comments. This can be explained by the fact that patients typically tend to be highly precise in describing aspects that negatively impacted on their experience. In contrast, the positive comments are overall less specific, therefore covering less categories from the framework. (e.g. "Just want to say I was treated with the height of respect"). For example, the quite specific negative comment "Bathrooms in [Ward Type] were dirty come afternoon time." will never appear as an equivalent positive version in the corpus. As for the first ranked categories, the automatic method matched the manual one for *category* and *context*. Indeed "*Patient care on the ward"* was the top ranked *activity* in both negative and positive comments, *"Received care"* was the highest ranked *context* in positive comments, and *"Long waiting time"* in negative comments. As for *resource*, the manual annotation resulted in "Nurse" as first rank category in both negative and positive comments, while the automatic algorithm identified the more generic "Staff" (with "Nurse staff" ranked in second position).

The Spearman coefficient is a measure that accounts for the ranking itself, which is important for us as we want

to isolate major issues shared by many patients. Calculated on the top 20 categories only, the correlation is quite strong with around 0.4 for the *resource* and *context* types in the positive comments and the *context* type in the negative comments (commonly interpreted as a "moderate" correlation), and above 0.6 for the *activity* type in the positive and negative comments and the *resource* type in the negative comments (commonly interpreted as a "strong" correlation). The figures drop to a "weak" /"very weak" correlation for the *resource* and *context* types when the Spearman coefficient is calculated over the totality of categories. This distinction with the *activity* type score is very likely to be explained by the much higher number of categories available for those types, as opposed to the limited number of *activity* categories. The annotated text segments are therefore spread over many more categories and the ranking is much more difficult to match between the manual and automatic approaches. However, as we noticed in section 5.2.2., the vast majority of themes expressed by patients are using 25% of all available categories. The top categories are thus the most meaningful and relevant ones to take into account in this study, and the automatic approach presented here proved to achieve good results in covering them.

7. Conclusion

Patient experience analysis is a delicate subject of study. Patients, health care organizations and related authorities all eagerly await the outcomes of such campaigns which assess the quality of the services provided and which is crucial for shaping the plan of action defined for future months or years. Some useful preliminary information can be obtained from close-ended questionnaire questions, but it is really when people are allowed to express themselves freely that the data becomes rich, with details on contexts, resources, etc. As in every domain, a manual analysis and annotation of such data requires available specialists or trained people for the task, making it a long and costly process to repeat. Natural Language Processing techniques are a natural choice when it comes to automatically extracting information and summarizing big quantities of text, while leveraging time and cost. However, for such a sensitive domain, one has to be careful about not loosing pieces of information in the process.

In order to provide an automatic analysis which does not loose information on the important details from patients feedback and which can highlight the major themes expressed, we performed an extraction of terms at an intermediate level of specificity. The experiments showed that despite not being as exhaustive as the manually annotated dataset is, the evaluation performed through a mapping technique over more general categories has proven this method to be successful in bringing up key aspects of subjects that matter to patients.

This approach of automatic processing could therefore serve future campaigns to assess patient satisfaction more quickly and at low cost without loosing information, and to track success or failure of health care entities in addressing problems through time.

8. Acknowledgements

This work presented in this paper was partially supported by the Patient Centred Service Improvement (PaCSI Project) and by Science Foundation Ireland grant 12/RC/2289_2 (Insight).

9. Bibliographical References

Astrakhantsev, N. (2018). Atr4s: toolkit with state-of-the-art automatic terms recognition methods in scala. *Language Resources and Evaluation*, 52(3):853–872, Sep.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., and McCallum, A. (2017). SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August. Association for Computational Linguistics.

Bordea, G., Buitelaar, P., and Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*, 01.

Cunningham, M. and Wells, M. (2017). Qualitative analysis of 6961 free-text comments from the first national cancer patient experience survey in scotland. *BMJ open*, 7(6):e015726.

Espinoza, F., Hamfors, O., Karlgren, J., Olsson, F., Persson, P., Hamberg, L., and Sahlgren, M. (2018). Analysis of open answers to survey questions through interactive clustering and theme extraction. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 317–320, New York, NY, USA. ACM.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, Aug.

Judea, A., Schütze, H., and Bruegmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 290–300, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182, 07.

Maramba, I., Davey, A., Elliott, M. N., Roberts, M., Roland, M., Brown, F., Burt, J., Boiko, O., and Campbell, J. (2015). Web-based textual analysis of free-text patient experience comments from a survey in primary care. *JMIR medical informatics*, 3(2):e20–e20, May.

O'Cathain, A. and Thomas, K. J. (2004). "any other comments?" open questions on questionnaires - a bane or a bonus to research? *BMC Medical Research Methodology*, 4(1):25.

Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T., and Zaki, M. (2014). Analyzing customer experience feedback using text mining: A linguistics-based approach. *Journal of Service Research*, 17(3):278–295.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Reddick, C. G., Chatfield, A. T., and Ojo, A. (2017). A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government facebook use. *Government Information Quarterly*, 34(1):110–125.

Šajatović, A., Buljan, M., Šnajder, J., and Dalbelo Bašić, B. (2019). Evaluating automatic term extraction methods on individual documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 149–154, Florence, Italy, August. Association for Computational Linguistics.

Wiseman, T., Lucas, G., Sangha, A., Randolph, A., Stapleton, S., Pattison, N., O'Gara, G., Harris, K., Pritchard-Jones, K., and Dolan, S. (2015). Insights into the experiences of patients with cancer in london: framework analysis of free-text data from the national cancer patient experience survey 2012/2013 from the two london integrated cancer systems. *BMJ open*, 5(10):e007792.

Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Zhang, Z., Petrak, J., and Maynard, D. (2018). Adapted textrank for term extraction: A generic method of improving automatic term extraction algorithms. In *SEMANTICS*.