Journal of Biomedical Informatics 51 (2014) 41–48

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Patients' involvement in e-health services quality assessment: A system for the automatic interpretation of SMS-based patients' feedback

Stefania Rubrichi [a,*], Andrea Battistotti [b], Silvana Quaglini [a]

[a] Laboratory for Biomedical Informatics, Deparment of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[b] Azienda Ospedaliera Fatebenefratelli e Oftalmico, Milano, Italy

ARTICLE INFO

Article history:
Received 10 August 2013
Accepted 3 March 2014
Available online 12 March 2014

Keywords:
e-Health
Patients' feedback
Health service assessment
SMS
Information extraction
Conditional random fields

ABSTRACT

Purpose: Effective communication between patients and health services providers is a key aspect for optimizing and maintaining these services. This work describes a system for the automatic evaluation of users' perception of the quality of SmsCup, a reminder system for outpatient visits based on short message service (SMS). The final purpose is the creation of a *closed-loop control system* for the outpatient service, where patients' complaints and comments represent a feedback that can be used for a better implementation of the service itself.

Methods: SmsCup was adopted since about eight years by an Italian healthcare organization, with very good results in reducing the no-show (missing visits) phenomenon. During these years, a number of citizens, even if not required, sent a message back, with comments about the service. The automatic interpretation of the content of those SMS may be useful for monitoring and improving service performances.Yet, due to the complex nature of SMS language, their interpretation represents an ongoing challenge. The proposed system uses conditional random fields as the information extraction method for classifying messages into several semantic categories. The categories refer to appreciation of the service or complaints of various types. Then, the system analyzes the extracted content and provides feedback to the service providers, making them learning and acting on this basis.

Results: At each step, the content of the messages reveals the actual state of the service as well as the efficacy of corrective actions previously undertaken. Our evaluations showed that: (i) the SMS classification system has achieved good overall performance with an average F1-measure and an overall accuracy of about 92%; (ii) the notification of the patients' feedbacks to service providers showed a positive impact on service functioning.

Conclusions: Our study proposed an interactive patient-centered system for continuous monitoring of the service quality. It has demonstrated the feasibility of a tool for the analysis and notification of the patients' feedback on their service experiences, which would support a more regular access to the service.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Among e-health applications [1,2], those based on short message service (SMS) are emerging as effective methods for health promotion. As a matter of fact, SMS messages are widely available and accessible, allowing for reaching individuals of all socioeconomic status. Moreover, they are asynchronous, that is they can be accessed at a time that suits an individual.

Mobile phone applications in healthcare setting cover several clinical areas and focus on improving processes or outcome of care. In this work we focus on the specific problem of patient's non-attendance or no-shows. A no-show is a missed visit, i.e. a visit that has been scheduled but not respected by a patient, without any notice from him. This phenomenon is common in every healthcare organization that delivers services on a scheduling basis, and may have various causes: the most frequent is that patients simply forget the appointment, or cannot attend it due to a sudden disease, or a last minute business, and forget to notify the doctors [3]. No-shows are a serious problem for both healthcare organizations and patients since they reduce the efficiency and quality of care delivery. Every such event causes waste of resources and time: a planned visit, while not executed, still entails a fixed cost that is not reimbursed by the national healthcare system. Equally, clinicians' time, which could have been used to serve other patients, is lost and waiting lists extend. On the other hand, from patients'

* Corresponding author. Address: via Ferrata 1, 27100 Pavia, Italy. Fax: +39 0382 525638.

E-mail address: stefania.rubrichi@unipv.it (S. Rubrichi).

http://dx.doi.org/10.1016/j.jbi.2014.03.003
1532-0464/© 2014 Elsevier Inc. All rights reserved.

perspective, a no-show might result in reduction of care quality because it can affect the health of patients who may need treatments.

Since forgetfulness is the major cause of no-show, a reminder system should alleviate the problem. In 2005, such a system, called SmsCup, has been implemented at the healthcare organization "Azienda Ospedaliera di Pavia" (from here on "AO-Pavia"), achieving very good results (the no-show rate decreased from 8% to 4.2%) [4,5]. The patient is not required to answer the reminder, but someone does, for different reasons: to thank for the service, to notify a possible error (e.g., he does not remember having booked any visit), to ask for further details, etc. After collecting a significant number of answers, we decided to analyze them, in order to evaluate patients' perception of the service. Moreover, this analysis should help to individuate possible system weaknesses and consequently spur corrective actions. As a matter of fact, any reminder that a patient perceives as an error, reveals an organizational bug or malpractice. The purpose of this study is to exploit patients' responses to reminders for improving the quality of service. The patients' SMSs, in fact, may be used for a double purpose. First, they might point out incorrect behaviors, of both patients and health care professionals, that inhibit the effective use of the service. For instance, elderly patients often give to the organization the mobile number of a relative (such as a son, a grandson, etc.), getting him/her confused when receiving the reminder. Just as, front-office operators sometime do not ask for the mobile number at all or, in front of a first refusal, they do not pursue the patient's consensus, by explaining them the usefulness of the service. Second, analyzing patients' SMSs from time to time, we can measure the effects of system improvements made on the basis of the SMSs themselves (e.g., observing a decreasing number of SMS claiming for errors, indicates a service improvement).

In this work, we present a system, based on natural language processing (NLP) techniques for SMS classification, that supports the healthcare organization in the outpatient service improvement.

2. Background and related work

2.1. Involving patients in service development

A better understanding of patients' preferences, needs and values is becoming an important issue in modern healthcare, especially in view of the increasing attention to patient-centered care [6]. Some studies have examined the effect of several forms of patients' involvement on healthcare service provision (i.e., improved patient-provider electronic communication, range of consultations, patient forums, interviews with service users), across a range of settings [7–10]. They recognized the great potential that such initiatives can have on services improvement, but asserted that it is too early to make strong conclusion about the impact on health outcomes and quality. A more recent literature review [11] explored the impact of ICT on patients' satisfaction. Despite the absence of clear evidence of positive impact, the authors found a widespread awareness of the need of incorporating patients' perspective into the care delivery and suggested the inclusion of patients' satisfaction as a strategic component of quality in medical informatics. Thus, after focusing on communication within and among healthcare organizations [12], over the past decade, ICT has increasingly considered patient-health providers communication, with the aim of making care more patient-centered.

2.2. Mobile devices for healthcare

Interventions involving cell phone found in scientific literature employ both cell phone voice and SMS technologies, and cover a variety of health areas such as diabetes [11], smoking cessation [13], HIV/AIDS [14], asthma [15], hypertension [16], physical activity [17], orthodontics [18], hepatitis vaccinations [19], stress management [20], physical disability [21], dialysis [22] and general outpatient clinics [23,24]. A number of studies have assessed the effectiveness of different systems [4,5,11,13–20]. As a result, they showed how such interventions have brought positive impact in term of health outcomes (e.g., compliance with medication taking and smoking cessation) and care processes (e.g., lower number of failed appointments and quicker diagnosis and treatment). For more details, research on the use of cell phones is well described in some comprehensive reviews by Krishna et al. [25] and by Hasvold et al. [26]. These systems, however, mostly rely on predefined reminders to be sent to patients according to some clinical conditions, but do not consider to receive a feedback from the patient in the form of an SMS text to analyze. Since this is the focus of our work indeed, in the next section we illustrate existing NLP methods for extracting information from SMS.

2.3. Information extraction from SMS text

Information extraction (IE) techniques have become an invaluable resource for searching about a particular topic in electronic archives of scientific literature [27,28], and for enriching the content and the utility of electronic clinical systems [29]. Excellent efforts have been documented in the literature on IE from textual biomedical documents [27–32], and its subsequent application in summarization, case finding, decision-support, or statistical analysis tasks. As well, the automatic analysis of SMS text through NLP techniques could allow for properly accessing and processing the SMS text and thus for deducing its syntactic and semantic structure. However, despite the growing significance of SMS as a means for the delivery of healthcare information, to date little has been published on NLP approach specific to SMS in the domain of medicine. We are aware of only one publication [33] about a system designed for extracting specific information from patients' informal SMS on medication management.

Some works exist about normalization of text message more in general [34–36]. As a matter of fact, the SMS language is far from standard: users are creating a novice language, overlooking orthographic and syntactic rules with a great emphasis on compressions (e.g., ad hoc abbreviation and acronyms due to space restrictions), and written representations of the sounds, such as "r" instead of "are". Another common phenomenon in SMS is represented by emoticons, such as :-(, :-) and ;-). All these aspects contribute to make NLP analysis of SMS an ongoing challenge.

3. Methods

3.1. Functional architecture

Fig. 1 shows the system we developed on top of SmsCup. This system allows for preparing, every working day, an SMS package to be sent to patients. The software retrieves data from the database hosted by the AO-Pavia. Then, SMSjob, a commercial gateway, sends the SMS package. The system sends this simple message three days before the scheduled date: *"The Healthcare Company of Pavia reminds you the visit of dd/mm. If you want to cancel it, call free 800448800. Thank you for the cooperation"*.

As mentioned in the introduction, patients receiving an SMS are not supposed to answer, but someone (about 0.5–1%) does. Considering from 120,000 to 150,000 SMSs sent/year, we receive about 600–750 replies/year. It is difficult to provide a more precise figure because some of the replies are empty or apparently unrelated to the service (due to patients' mistakes).

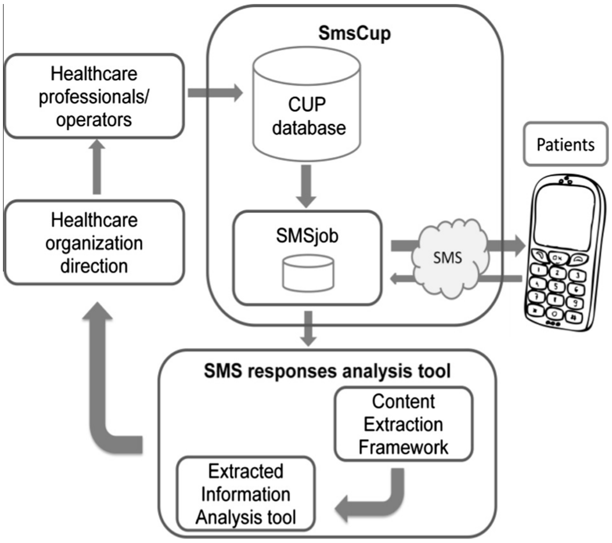
SmsCup
Healthcare professionals/ operators
CUP database
Patients
Healthcare organization direction
SMSjob
SMS
SMS responses analysis tool
Content Extraction Framework
Extracted Information Analysis tool

Fig. 1. Schema of the SmsCup system and of the responses analysis tool. AO-Pavia front-office or call center operators collect visit-booking data and patients' mobile phone numbers, and store them in CUP database. SMSjob gateway retrieves booking data, sends SMS messages and (occasionally) receives replies. CRF extracts information from patients' replies, which are analyzed, and then a feedback is returned to the AO-Pavia responsible administrators (direction). According to the reported issues, they take opportune corrective actions.

The core of the system, shown in Fig. 1, is the SMS response analysis tool, which relies on the conditional random fields (CRF) algorithm (see next paragraph) for retrieving meaningful content from SMS. Such approach entails the classification of text or portion of text into a number of application-dependent semantic categories that we have defined in the form of hierarchy of concepts. It integrates the semantics via engineered features, which describe both the semantic and syntactic peculiarities of SMS content. Afterwards, it discriminates between semantically interesting and uninteresting content through the automatic adaptation of such features. This extraction framework is then used to evaluate outpatient service and plan its improvements. Basically, the SMS statistics are reported back to the healthcare direction. They took up interventions to accordingly improve the service.

For the purposes of this paper, we performed a baseline statistics, and another one after one month, in order to inquire about effects of interventions.

3.2. SMS content extraction

Our approach for IE consists of five major steps: (1) semantic model for the representation of SMS content, in order to find out the concepts to be extracted; (2) preprocess of the SMS text; (3) hand annotation, according to the previously developed conceptual model; (4) definition of a set of binary features that express some descriptive characteristics of the data and conversion of the stream of words into features; (5) process of the data through the CRF.

These steps are described in the following paragraphs.

3.2.1. The semantic representation of SMS

Typically, a system for IE is based upon specific domain knowledge. Therefore, we read the messages, in order to interpret their meaning and to understand the patients' motivations to send them, with the purpose of identifying the underlying semantic classes that are appropriate to ensure comprehensive coverage of the concepts described in the SMS that are valuable for our purpose. The derived classification, shown in the Box 1 below, includes four major semantic categories, or types. Each category has been further characterized at one or more sub-levels.

Box 1 Concepts hierarchy.

1. Service perception
1.1. Appreciation
1.1.1. Confirmation
1.1.2. Thanks
1.2. Disappoint
1.2.1. Scorn
2. Further detail
2.1. Ward/time
2.2. Type of visit
3. Further request
3.1. Cancellation/Shifting
3.2. Problem with toll number
4. Error notification
4.1. Misunderstanding
4.1.1. City misunderstanding
4.2. Actual error
4.2.1. Surprise
4.2.2. Repeated error
4.2.3. Live elsewhere
4.2.4. Visit already done
4.2.5. Visit already cancelled
4.2.6. Visit already shifted

We provide a brief explanation of each category.

1. *Service perception* deals with the overall judgment of the service. We identified two levels of perception: *Appreciation and Disappoint.* Appreciation messages express suitability and satisfaction with the service. We distinguish between *Confirmation*, that classifies messages where patients confirm the appointment (e.g. "I confirm the visit of Jan 26th 2011, thank you"), and *Thanks*, for messages that display gratitude and approval (e.g. "Very good initiative, thanks for reminding me!"). Disappoint messages have negative content, and in particular a *Scorn content*, where the patient seems bitterly disappointed with the service because either he gets the reminder by mistake or he dislikes the service at all (e.g. "Do not send any other message and save public money").
2. *Further detail* considers the completeness of the information received. We grouped this kind of messages under two types: *Ward/time*, where the patient asks for more details about ward and/or time schedule (e.g. "I confirm the visit, could I have more information about ward and time?") and T*ype of visit*, where either the patient just asks for more information (e.g., "Could you please remind me what this is about?"), or he seems surprised, but nevertheless he considers the eventuality of having forgotten the appointment (e.g., "I do not remember booking any visit, what is this about?"). In both of cases there is an explicit request for visit details. Note that these requests are probably due to the fact that the reminder was formulated as concisely as possible, to comply with privacy regulation (in case the message is read by another person than the patient to whom it was addressed).
3. *Further request*, covering demands for further operations such as cancellation or shifting. Despite the reminder suggests to contact the toll number for canceling/moving the visit, some patients try to do it via SMS. We distinguish between *Cancellation/Shifting* (e.g. "I intend to move the appointment") and *Problem with the toll number* (e.g. "I cannot call the toll number, please cancel the visit").

4. *Error notification*, where the patient's SMS reports issues attributable to error. Messages belonging to this category reveal, more than the others, critical aspects of the system, which would require more careful analysis and interventions. We made a first distinction between *misunderstanding* and *actual errors*. Misunderstanding means that patients perceive the reminder as erroneous, but it is just a misunderstanding due to the lack of some details, for example *City misunderstanding*: actually, the reminder does not precise the city or structure of the visit but just says "Healthcare Agency of Pavia". This can generate a misunderstanding, especially when the visit is in cities belonging to the AO-Pavia, but other than Pavia. Italian healthcare agencies, in fact, are local organizations that provide services in specific areas, including several cities, in structures closer to where people lives (e.g. "Dear healthcare agency, the reservation is for the hospital of Voghera, not the hospital of Pavia"). Actual error is instead the semantic category that refers to SMS that reveal *real* errors. We identified five types of *errors*: *Surprise*, when the patient definitely asserts that he has never booked any visit and thus an error has been made (e.g. "This is surely a mistake, I never booked any visit"). This could happen for several reasons: mobile numbers are not accurately provided or collected or a fake visit has been booked (this revealing a severe malpractice); *Repeated error*, when a patient points out that it is not the first time he receives a reminder by mistake (e.g. "Every week I receive this kind of messages but I repeat I NEVER BOOKED ANY VISIT"); *Live elsewhere*, when a patient specifies that he lives in cities or regions that do not belong to the AO-Pavia (e.g. "... But I live in Tuscany"); *Visit already cancelled/shifted/done*, when a patient informs that his visit has been already cancelled, shifted or done. In these last three cases an error occurred and it is attributable to healthcare operators who do not timely update the electronic agenda when an exception occurs.

The most specific semantic types available in the hierarchy are assigned to the SMS. Each SMS message can be tagged with one or more semantic types.

3.2.2. Preprocessing of SMS

As long as the label prediction is on a word-by-word basis, and decisions are made for one message at a time, the first stage of our extraction algorithm consists in splitting SMS texts into tokens. All the messages were passed through a pre-processing engine, which tokenized them, by automatically detecting tokens (i.e. words) boundary. We used white spaces to determine token boundaries.

In addition, we considered a normalization step that mainly includes removing all punctuation and fixing orthographic errors. Several exceptions, in fact, occurred within the texts: this is because SMS messages were written by persons from different walks of life, and SMS language is more error-prone in general. In order to account for them, we employed an orthographic corrector, which automatically fixed such mistakes. Incidentally, SMS text usually contains abbreviations (e.g., for the Italian language: «ki» → «chi», «cmq» → «comunque», «xk» → «perché»). We implemented a Java class to purposely manage them.

3.2.3. Hand annotation of SMSs

We collected two sets of SMS messages: one for training and testing the classifier, the other one for evaluating SmsCup progresses. The first set includes responses received from July 2005 to July 2011, the second one, messages gotten from September 2011 till July 2012. Among the about 7000 SMS messages, 3000 have been used for the analysis. The remaining ones were not usable because either they were empty or clearly uncorrelated to the service.

The gold standard was generated by manual annotation of the SMS corpus. Corpus annotation is the practice of adding to a corpus interpretative linguistic information (part-of-speech –POS- tags, syntactic structure, co-references, etc.) or information derived from domain knowledge (i.e. semantic annotation). Because of the complex and ambiguous nature of language, it is necessary to undertake manual work in order to obtain an optimally tagged corpus. Two biomedical engineers performed the annotation process together. They availed themselves from an SmsCup operator to interpret particularly unclear or ambiguous messages. One fundamental problem in corpus annotation is the definition of what constitutes an entity to be tagged. Semantic descriptions must be well defined and easy to understand by the domain expert who annotates the text: to this aim we exploited the conceptual model of information conveyed in the SMS messages, previously illustrated, and we used the following fifteen semantic labels: *Thanks, Confirmation*, *Scorn, Ward/Time*, *TypeVisit*, *Cancellation/Shifting*, *CityMisunderstanding*, *Surprise*, *RepeatedError*, *LiveElsewhere*, *AlreadyShifted*, *AlreadyDone*, *AlreadyCancelled*, *TollNumber*, *None*. The label None has been given to indicate elements that are not relevant for this research.

Some examples of hand-annotated SMSs are:

1. $\langle$Very good initiative$\rangle_{\textbf{Thanks},}$ $\langle$I confirm$\rangle_{\textbf{Confirmation}}$ the visit of 26th gen 2011 could I have more information about $\langle$ward and time$\rangle_{\textbf{WardTime}}$?"
2. Dear healthcare agency, the reservation is for the $\langle$hospital of Voghera$\rangle_{\textbf{CityMisunderstanding},}$ $\langle$not the hospital of Pavia$\rangle_{\textbf{CityMisunderstanding}}$"
3. $\langle$Don't send$\rangle_{\textbf{Scorn}}$ any other messages and $\langle$save public money$\rangle_{\textbf{Scorn}}$!"
4. Every week I receive this kind of messages but $\langle$I repeat I NEVER BOOKED$\rangle_{\textbf{RepeatedError}}$ ANY VISIT".

3.2.4. Feature definition and text-to-feature conversion

The feature construction process aims at capturing the salient characteristics of each token (i.e. a word) in order to help the system to predict its semantic label. When defining feature functions, we construct a set of real-valued features $b(x, i)$, for the observed word, that express some characteristics of the empirical distribution of the training data. The feature produces a numerical value. As an example, we can define a simple feature, that takes into account if a word is capitalized:

$$b(x,i) = \begin{cases} 1 & \text{if the word } x \text{ in the position } i \text{ in the sentence is capitalized} \\ 0 & \text{otherwise} \end{cases}$$

Feature definition is a critical stage for the success of feature-based statistical models such as CRF. We compiled four types of features to describe the data: (1) lexical features, (2) morphosyntactic features, (3) semantic features, and (4) hybrid features.

Lexical Features cover word-related features such as: the identity of words, that is the vocabulary obtained from the training data; the identity of the 3 words before; the identity of the 3 words after; the lemmas of these words. About the number of words before and after, it is clear that the more context words analyzed, the better and more precise the results are. However, widening the context window quickly leads to an explosion of computational and statistical complexity. For our experiments, we estimated a suitable window size of $[-3,3]$.

Morphosyntactic Features include the POS of all the words, and the negation feature which indicates if a negation is detected (e.g., not, no) at a limited distance ($[-4;4]$) of words before and after the considered one. Here, the window length has been chosen so that the most part of negative constructs in the Italian language can be detected. We used TreeTagger [37] to perform POS tagging of the corpus sentences.

Semantic Features use external semantic resources, related to a specific task. The first class of semantic features indicates whether a specific "trigger" word occurs in the SMS text (e.g., the word "toll" may be useful for identifying problems with the toll free number). After browsing our corpus we manually collected a list of trigger words associated to different semantic areas (e.g. change, cancellation). Every time a text token, or its lemma, occurs in such a list, the feature is activated, indicating that the token belongs to the specific semantic area. The second class relies on a particular semantic resource that is the lists of all Italian cities and regions. In several messages, mention is made of Italian cities or regions to report an error (e.g., "You definitely got the wrong number because I live in *Sicily*"). We found that this type of feature can provide good clues for identifying such error notifications.

Hybrid Features take into account a combination of semantic and lexical/morphosyntactic properties of tokens. For instance, with reference to moving an appointment, we found that adding tense information to the semantic is useful for distinguishing a shifting request (e.g., "I *intend* to move the appointment to the next week") from a shifting notice (e.g., "I *have already moved* the visit"). To this end, we created a binary feature, which considers the list of verbs indicating shifting in conjunction with tense aspects, in particular present and past tense, respectively.

Within the CRF model, each token is represented by the corresponding label and the set of active features. Then, stream of tokens has been automatically converted to features using a Java class: every time a text word takes on the value of a feature, the feature will become active for that token.

3.2.5. Data process

As mentioned, for the purpose of this work, we employed the CRFs algorithm, which belongs to the family of *supervised machine learning* approaches [38]. In *supervised learning*, one chooses a classifier based on a training set of examples for which the correct labels are available. The aim is to select the classifier that *generalizes* well to unseen examples. More formally, we are given a set $S = (x^{(i)}, y^{(i)})_{i=1}^{N}$ of SMS $x \in X$ together with the corresponding correct labels $y \in Y$ from the tag set. A classifier h is a mapping from the input space X to the output domain Y. It performs well if it correctly predicts the labels y of some new and unknown SMS x.

CRFs predict the labels of words by using large number of interdependent features, such as those described in the previous paragraph. This can be seen as a way to "capture" the hidden patterns of labels and features, and "learn" what the likely output might be, given these patterns. Our system uses the MALLET [39] implementation of CRFs.

4. Results

4.1. Classification algorithm performance

We randomly split the first set of SMS messages into two sets: one for training ($n = 1768$) and one for testing ($n = 443$). We measure the performance of our model on the individual labels using the standard evaluation metrics for machine learning algorithms, i.e. Recall, Precision and F1-measure [40]. In dealing with multi-label classification, we combined the performance results of the different labels, both computing their arithmetic mean, so giving equal weight to each of the labels (macro-averaged) and computing a weighted mean, using for each label the number of times it occurs in the dataset (micro-averaged). Macro-averaged metrics are often dominated by the performance on rare labels, while micro-averaged metrics are dominated by the performance on frequent labels. The two ways of measuring performance are hence complementary, and both are informative.

Our experiments show that the CRF, with carefully designed features, can classify the SMS content with an overall accuracy of around 92%. Table 1 reports the recognition scores for the different labels and the overall ones.

As expected, labels whose training examples are scarce suffer from relatively low performance (e.g. *RepeatedError* and *AlreadyShifted,* that are hardest to extract). On the other hand, some other labels such as *CityMisunderstanding* and *AlreadyCancelled*, although rare, perform better. Such labels, in fact, can rely on more dedicated and informative features, which is an important factor that contributes to the good performance.

In order to estimate the variability of the overall performance, we ran the models on 40 trials. In each trial, the SMSs in the new training set are randomly sampled from the original dataset. Fig. 2 shows the micro-averaged Recall, Precision and F1-measure.

Fig. 3 shows the variability of the F1-measure calculated on the individual labels. As expected the labels *Confirmation*, *Thanks*, *None* and *Surprise* present a very low variability.

4.2. Exploiting SMS classification for the reminder service evaluation

We used the classifier to examine the SMS content and thus to evaluate the SmsCup service. To this aim, we chose to consider just the responses received between August 2008 and July 2011 for a total of 1481. Responses from the first three years were employed as training set, but not in the evaluation, in such a way to let the system enough time to establish itself and the patients to become familiar with the service.

Table 1
Performance results of the classifier.

Label	N_{train}	N_{test}	Precision (%)	Recall (%)	F_1-measure (%)
Cancellation/Shifting	191	50	93.02	80.00	86.02
Thanks	1100	267	97.72	96.25	96.98
None	5759	1540	91.70	96.95	94.26
Confirmation	1097	306	95.89	91.50	93.65
Surprise	1147	261	87.50	88.51	88.00
LiveElsewhere	80	48	93.55	60.42	73.42
VisitType	139	38	86.67	68.42	76.47
AlreadyShifted	21	13	100.00	38.46	55.56
AlreadyDone	27	7	71.43	71.43	71.43
CityMisunderstanding	60	27	100.00	66.67	80.00
RepeatedError	48	11	100.00	36.36	53.33
TollNumber	48	11	100.00	100.00	100.00
AlreadyCancelled	55	15	85.71	80.00	82.76
Scorn	106	13	100.00	53.85	70.00
Ward/Time	85	19	88.89	42.11	57.14
Overall(Micro-average)	–	–	92.50	92.38	92.04
Overall(Macro-average)	–	–	92.81	71.39	78.60

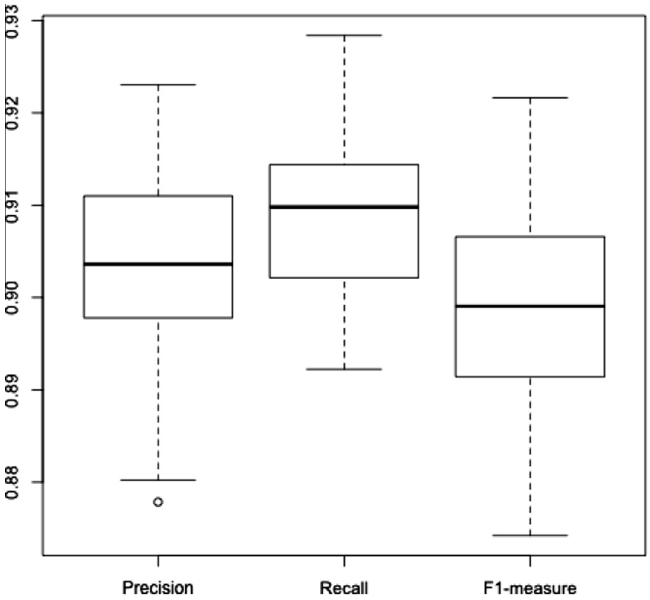

Fig. 2. Overall performance, in terms of micro-averaged Precision, Recall and F1-measure, estimated on 40 trials. Boxplots show median, interquartile range, and range.

Initially, messages were processed by the CRF and their content categorized according to the hierarchy of concepts described above. Fig. 4 shows the message distribution according to main content categories and subcategories.

The total percentage exceeds 100% because single messages may include a combination of conceptual categories. For instance, when confirming an appointment, patients may ask for more information about ward and time (e.g., "I confirm the visit, could you remind me the time?") or may point out the city where the visit should take place (e.g., "I confirm the visit but at the hospital of Vigevano".). In these cases, we counted the messages twice, once under *Appreciation* category and the other one under *Further Information* and *Error Notification* categories, respectively. Also, when deleting a visit, patients may produce as motivation that the visit has been already made, cancelled or moved (e.g., "I wish to cancel the appointment because the visit has already been made"). Similarly, we considered such responses as both *Further Request* and *Error Notification* messages. On the contrary, messages including several content categories that refer to the same macro-category were considered once. For example, in reporting an error the patient may add that he lives elsewhere or that it is not the first time it happens (e.g., Surely there must be a mistake, I live in Tuscany). In our distribution, messages of this type have been counted once as *Error Notification* messages.

Overall, our analysis showed that the service is delivered consistently and patients are satisfied with it. Actually, about the 70% of SMS responses have positive content since they were assigned to the *Appreciation* category of the conceptual hierarchy. A fairly small percentage of messages contained request for more detailed information (4.32%) such as ward time and type of visit, or for specific actions (8.71%) such as visit shifting or cancellation. Only 0.27% expresses definite dislike for the service.

However, despite the majority of responses concerned service suitability and patients endorsements, nearly one fourth (22%) were classified as *error notifications*, ranging from *Misunderstanding* to *Actual Errors*, with a sharp superiority of the *Surprise* category (about the 18%) over the other ones. This result reveals that there are some aspects that appear to damage the service efficiency and quality, and which should therefore be considered for improvement. In particular, reminders sent by mistake (i.e. *sending error*) have been identified as main cause for patient dissatisfaction. Among error notifications, the content category *Surprise* is the most frequent one and, in addition, the category *Scorn* occurs often concurrently with *Surprise*.

The results of our content analysis were reported back to the SmsCup providers. They focused on the weak areas and undertook some improvement plans to better meet the needs detected in patients' messages. They made health professionals, as well as front-office personnel, aware of such defaults and invited them to pay careful attention when booking visits and collecting phone numbers.

One month after, we collected patient responses and analyzed them for evaluating the effects of interventions implemented as consequence of the first results. In this phase, we considered just *sending error* issue. Specifically, we plotted, year by year, the total percentage of *Surprise* and *Type Visit* messages gotten from August 2008 until June 2012. These two content categories, in fact, are the

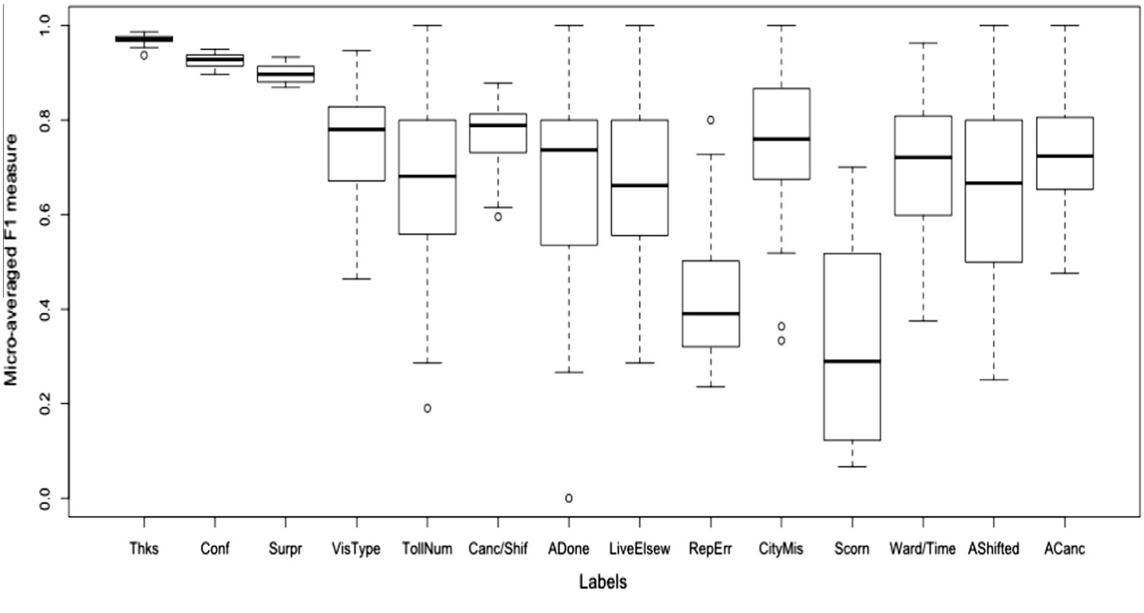

Fig. 3. Performance results on individual labels, in terms of micro-averaged F1-measure, estimated on 40 trials. Labels are abbreviations of those reported in Table 1.

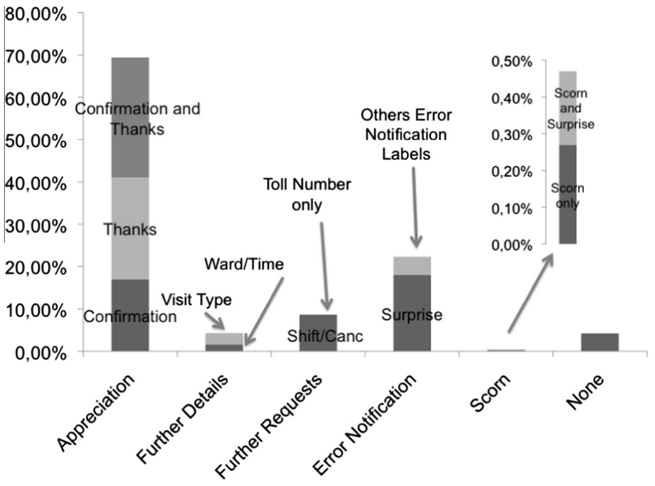

Fig. 4. Distribution of messages by main content categories. Bars show a further labeling according to subcategories. Scorn bar has been zoomed for visualization purposes.

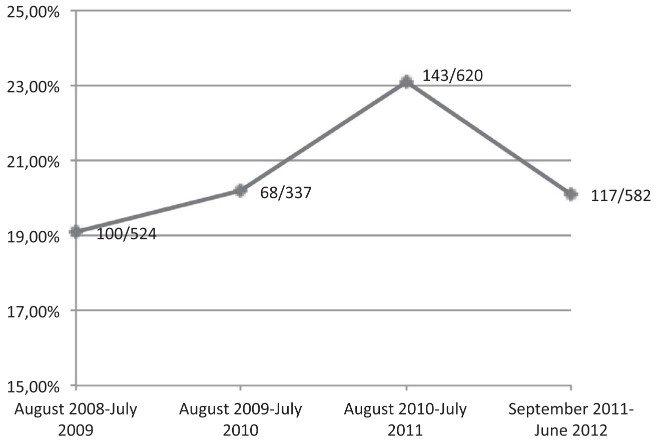

Fig. 5. Annual rate of *Surprise* and *Type Visit* messages gotten from August 2008 until June 2012.

ones that more reflect sending errors. Fig. 5 shows that, even if not markedly, these interventions produced an inversion of the upward trend. Thus, it may be argued that there is a tendency towards a positive impact, but also that corrective actions must be continued.

5. Discussion

Despite searching for a rigorous patient-centered evaluation framework seems to be an important future task in health IT, current literature does not provide any major indication about optimal techniques to perform it. A study by Akesson et al. [41] reviewed existing studies on patients' experience with health IT and categorized them according to their main focus as well as the method of evaluation. Although the study is more focused on the description of the consumers' experience, we noticed that the methods used for assessing such experiences were usually structured questionnaires and interviews on quality of life. The patient is thus explicitly asked to assess the service according to several measures, such as confidence with the service, impact on health outcomes or quality of life. In this work, on the contrary, we have proposed an indirect evaluation method, i.e. we explored the possibility of measuring the patient's perception of the quality of a reminder service, through automatic interpretation of the his SMSs. The patient, by spontaneously sending back the message, is not aware of providing the developer with feedbacks on the service. In this way, the evaluation results should be more actual and less dependent on confounding factors such as user motivation induced by the interviewers.

Other previous works have reported on content analysis of electronic messages between patients and health providers [42–44]. Unlike our study, the majority of these works has mostly been limited to a manual analysis of message content and has concerned health issues of several clinical settings (e.g. pediatrics, internal medicine, primary care). They have evaluated the content of patients' e-mails, to assess the benefit of web-based communication services. Although such studies are hardly comparable with our one, we nevertheless report on them as proof of the increasing interest in considering patients' perceptions to enhance and promote the performance of health services. On the whole, consistently with our study, they suggested that e-health can lead to safer care by improving communication between patients and health providers.

A work by Stenner et al. [33] presented an IE system that extracts medication information form patients' SMS messages, as part of a mobile-phone based medication management system. They describe a similar system that extracts information from patient-generated SMS texts, following a different approach based on custom lexicons and existing knowledge sources to extract medical information. Their method achieves an F1-measure comparable with our one. Yet, authors have focused on the accuracy of the extraction system without examining in depth the impact of such extractions on the medication management system.

On the contrary, our evaluation has demonstrated that analysis of patients' feedback helps providers considering strategies to improve the system. Moreover, the same analysis allowed domain experts (i.e., healthcare administrators) to identify anomalies in the outpatient management workflow. Such anomalies were related to visit booking performed by healthcare personnel who either had no *official* role to do it, or did it without following the correct protocol, with the result of non-traceability of the reservation. After implementation of corrective measures, further message analysis showed a trend towards improvement (Fig. 5).

Owing to the great success of SmsCup service in reducing the rate of no-show, in July 2012 it was extended to the whole Lombardy region and its management centralized at regional level instead of single hospital trusts. Exploiting our results about reminder interpretation problems (see categories 2.1 and 4.1.1 in Box 1), the text of the message has been reformulated in such a way to include the city where the visit will be held as well as the name of the hospital and time schedule. In light of these changes, it would be interesting to examine the evolution of these message categories as a further test of the classifier.

Eventually, we report some limitations of our study. Despite the very good overall accuracy, if we consider the individual labels, the results highly depend on the size of training examples as well as features definition, with some labels that suffer from low performances and high variability. This is an important open question for future work, which should aim to improve results of such labels, so that they too can be considered for system monitoring.

About the definition of labels, a possible improvement could be the reference to standard ontologies and terminologies [45].

Another limitation is related to the automation of the feedback reporting. In our work, the SMS classification has been manually reported and discussed with the domain experts. In the future, we could implement an automatic reporting tool integrated with the SMS responses analysis tool, in such a way that healthcare administrators can periodically and autonomously monitor the system.

6. Conclusions

In short, we have presented an example of electronic engagement of patients in the quality assessment and management of an e-health service. We developed an NLP framework for simultaneous identification of multiple semantic entities, in patients' SMS messages generated in response to a reminder system. The extracted information is then used to identify and fix possible defaults that threaten the effective and appropriate functioning of the service. Our empirical evaluation shows that the classifier achieves high overall accuracy. The results and the ready adaptability of the approach we have adopted show that our system is suitable for the classification of SMS content and thus it can be used from here onto monitor the users' satisfaction and system performances in time.

Acknowledgments

We gratefully thank Miriam Ferrari and Arianna Milanesi for extracting SMS from the legacy system of the "Azienda Ospedaliera della Provincia di Pavia", StéphanieEku-Ndam for the first manual tagging of SMSs, and Enrico Cuoco for fruitful discussion about dropout issue.

References

[1] Eng TR. The eHealth landscape: a terrain map of emerging information and communication technologies in health and health care. Springer; 2001. p. 9.
[2] Atkinson NL, Gold RS. The promise and challenge of eHealth interventions. Am J Health Behav 2002;26:494–503.
[3] Husain-Gambles M, Neal RD, Dempsey O, et al. Missed appointments in primary care: questionnaire and focus group study of health professionals. Br J Gen Pract 2004;54:108–13.
[4] Battistotti A, Quaglini S, Cuoco E. Contrastare il dropout ambulatoriale: il sistema dell'Azienda ospedaliera di Pavia. Mecosan 2007;16(62):119–34.
[5] Battistotti A, Quaglini S, Cuoco E. Reducing dropouts in outpatient care through an SMS-based system. Stud Health Technol Inform 2006;124:935–40.
[6] Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: The National Academics Press; 2001.
[7] McGeady D, Kujala J, Ilvonen K. The impact of patient–physician web messaging on healthcare service provision. Int J Med Inform 2008;77(1):17–23.
[8] Crawford MJ, Rutter D, Manley C, Weaver T, Bhui K, Fulop K, et al. Systematic review of involving patients in the planning and development of health care. BMJ 2002;325(7375):1263.
[9] Hiidenhovi H, Nojonen K, Laippala P. Measurement of outpatients' views of service quality in a Finnish university hospital. J Adv Nurs 2002;38(1):59–67.
[10] Ye J, Rust G, Fry-Johnson Y, Strothers H. E-mail in patient–provider communication: a systematic review. Patient Educ Couns 2010;80(2):266–73.
[11] Rozenblum R, Donzé J, Hockey PM, Guzdar E, Labuzetta MA, Zimlichman E, et al. The impact of medical informatics on patient satisfaction: a USA-based literature review. Int J Med Inform 2013;82(3):141–58.
[12] Marchetti D, Lanzola G, Stefanelli M. An AI-based approach to support communication in health care organizations. Lect Notes Comput Sci 2001;2101:384–94.
[13] Bramley D, Riddell T, Whittaker R, et al. Smoking cessation using mobile phone text messaging is as effective in Maori as non-Maori. N Z Med J 2005;118:U1494.
[14] Andrade AS, McGruder HF, Wu AW, et al. A programmable prompting device improves adherence to highly active antiretroviral therapy in HIV-infected subjects with memory impairment. Clin Infect Dis 2005;41:875–82.
[15] Ostojic V, Cvoriscec B, Ostojic SB, Reznikoff D, Stipic-Markovic A, Tudjman Z. Improving asthma control through telemedicine: a study of short-message service. Telemed J E-health 2005;11:28–35.
[16] Kim HS, Song MS. Technological intervention for obese patients with type 2 diabetes. Appl Nurs Res 2008;21:84–9.
[17] Hurling R, Catt M, Boni MD, Fairley BW, Hurst T, Murray P, et al. Using Internet and mobile phone technology to deliver an automated physical activity program: randomized controlled trial. J Med Internet Res 2007;9:7.
[18] Bos A, Hoogstraten J, Prahl-Andersen B. Failed appointments in an orthodontic clinic. Am J Orthod Dentofacial Orthop 2005;127:355–7.
[19] Vilella A, Bayas JM, Diaz MT, et al. The role of mobile phones in improving vaccination rates in travelers. Prev Med 2004;38:503–9.
[20] Riva G, Preziosa A, Grassi A, Villani D. Stress management using UMTS cellular phones: a controlled trial. Stud Health Technol Inform 2006;119:461–3.
[21] Nguyen T, Garrett R, Downing A, Walker L, Hobbs D. Telecommunications access: matching available technologies to people with physical disabilities. Aust Phys Eng Sci Med 2006;29:87–97.
[22] Azzini I, Falavigna D, Gretter R, Lanzola G, Orlandi M. First steps toward an adaptive spoken dialogue system in medical domain. Interspeech 2001:1327–30.
[23] Downer SR, Meara JG, DaCosta AC. Use of SMS text messaging to improve outpatient attendance. Med J Aust 2005;183:366–8.
[24] Capozzi D, Lanzola G. An agent-based architecture for home care monitoring and education of chronic patients. COMPENG'10; 2010. p. 138–140.
[25] Krishna S, Boren SA, Balas EA. Healthcare via cell phones: a systematic review. Telemed E-health 2009;15(3):231–40.
[26] Hasvold PE, Wootton R. Use of telephone and SMS reminders to improve attendance at hospital appointments: a systematic review. J Telemed Telecare 2011;17:358–64.
[27] Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2002;18:1124–32.
[28] Hanisch D, Fluck J, Mevissen HT, Zimmer R. Playing biology's name game: Identifying protein names in scienfific text. In: Proceedings of the symp biocomput; 2003. p. 403–14.
[29] Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 2006:589–99.
[30] Rubrichi S, Quaglini S, Spengler A, Russo P, Gallinari P. A system for the extraction and representation of summary of product characteristics content. Artif Intell Med 2013;57(2):145–54.
[31] Friedman C, Johnson SB, Forman B, Stanner J. Architectural requirements for a multipurpose natural language processor in the clinical environment. In: Proceedings of the annu symp comput appl med care; 1995. p. 347–51.
[32] Aronson A. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of the AMIA annu symp; 2001. p. 17–21.
[33] Stenner SP, Johnson KB, Denny JC. PASTE: patient-centered SMS text tagging in a medication management system. J Am Med Inform Assoc 2012;19(3):368–74.
[34] Investigating SMS Text Normalization using Statistical Machine Translation. <http://nlp.stanford.edu/courses/cs224n/2009/fp/27.pdf> (accessed April 2013).
[35] Aw A, Zhang M, Xiao J, et al. A phrase-based statistical model for SMS text normalization. In: Proceedings of the COLING/ACL on main conference poster sessions. Sydney, Australia: Association for, Computational Linguistics; 2006.
[36] Byun J, Lee SW, Song YI, Rim HC. Two phase model for SMS text messages refinement. In: The 23rd AAAI conference on artificial intelligence (AAAI 2008) workshop 04 – enhanced messaging, Chicago, Illinois, USA, AAAI; 2008.
[37] Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods in language processing; 1994. p. 44–9.
[38] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the international conference on machine learning; 2001.
[39] McCallum A. Mallet: a machine learning for language toolkit. Tech rep; 2002. <http://mallet.cs.umass.edu>.
[40] Van Rijsbergen C. Information retrieval. 2nd ed. London, UK: Butterworths; 1979.
[41] Akesson KM, Saveman BI, Nilsson G. Health care consumers' experiences of information communication technology – a summary of literature. Int J Med Inform 2007;76(9):633–45.
[42] White CB, Moyer CA, Stern DT, Katz SJ. A content analysis of e-mail communication between patients and their providers: patients get the message. J Am Med Inform Assoc 2004;11(4):260–7.
[43] Sitting DF. Results of a content analysis of electronic messages (email) sent between patients and their physicians. BMC Med Inform Decis Mak 2003;3:11.
[44] Anand SG, Feldman MJ, Geller DS, Bisbee A, Bauchner H. A content analysis of e-mail communication between primary care providers and parents. Pediatrics 2005;115(5):1283–8.
[45] Falasconi S, Lanzola G, Stefanelli M. Ontology and terminology servers in agent-based health-care information systems. Methods Inf Med 1997;36(1):30–43.