Moderator Assistant: a Natural Language Generation-based Intervention to Support Mental Health via Social Media

As online mental health support groups become popular they require more from volunteers and trained moderators who help their users through ‘interventions’ i.e. responding to questions and providing support. We present a system that supports such human interventions using Natural Language Generation (NLG) techniques. The system generates draft responses aimed at reducing moderators’ workload, and improving their efficacy. NLG and human interventions were compared through the ratings of 35 psychology interns. The NLG-based system was capable of generating messages that are grammatically correct with clear language. The system needs improvement, however moderators can already use it as draft responses.

Keywords: Mental health, online support groups, interventions, social media, NLG

Introduction

Mental health problems are known to cause disability, decrease productivity, and reduce overall quality of life. The World Health Report (2001) states that one in four people worldwide will meet criteria for a mental disorder at some point during their life. According to the Australian Bureau of Statistics (2007), almost half (45%) of Australians aged 16 to 85 years experience a mental disorder at some stage. Depression and anxiety are the most prevalent mental disorders in Australia and elsewhere. Depression alone is predicted to be one of the world’s largest health problems by 2020 (Murray & Lopez, 1996). Despite high prevalence rates, the diagnosis and treatment of mental disorders has long been neglected, especially in rural populations where access to quality care is limited (Burns et al., 2010; Clarke & Yarborough, 2013; Strecher, 2007). Moreover, people are often reluctant to seek help, with only 13% of males and 31% of females aged 16 to 24 years with a mental health problem accessing a clinical service (Slade et al., 2009). In many cases, the lack of available trained mental health professionals, as well as the intensive time and cost needed for treatments, allow for only a minority of people experiencing problems to be treated and supported (disease, 2008; Doherty, Coyle, & Sharry, 2012). Strong stigmatising attitudes and beliefs towards mental health disorders are other key factors that have resulted in a wide treatment gap and reluctance in the help-seeking process (Clarke & Yarborough, 2013; Henderson, Evans-Lacko, & Thornicroft, 2013).

Internet-based interventions have the potential to jump many of the traditional barriers when accessing and receiving mental health treatment. The anonymous nature of Internet-based interventions has been found to increase participants utilisation of self-help options (Ybarra & Eaton, 2005). Furthermore, web-based interventions provide an alternative to face-to-face patient care (Currell et al., 2000) while also eliminating travel and treatment waiting times, increasing treatment accessibility and flexibility, reducing overall cost, and, perhaps most importantly, increasing access to mental healthcare (Doherty, Coyle, & Sharry, 2012). This has allowed structured interventions models, such as computerized/Internet-based cognitive behaviour therapy (CBT) to receive a lot of attention over the years (Christensen, Griffiths, & Jorm, 2004; Spek et al., 2007). A number of randomised studies have particularly investigated the effects of Internet-
based interventions on depression and anxiety related disorders (Spek et al., 2007). The Internet-based CBT approaches have proven to be effective, especially with therapist support.

One of the most promising aspects of Internet-based tools and interventions is the widespread availability of online communities and peer support groups enabling people in distress to identify with others with similar needs and problems, share feelings and information, provide and receive advice, and develop a sense of community. Online peer support groups are becoming increasingly popular on social networking websites such as Facebook as well as for organisations such as ReachOut.com in Australia. Some of these support groups are moderated by trained young people or allied mental health staff (e.g. ReachOut.com), giving people the opportunity to receive help from professionals and use resources developed by experts. However, as such communities keep growing, the amount of work required of the moderators continues to increase, ultimately making quality support unsustainable (Xxx & Xxx, 2013).

One way to address this problem would be to automate the generation of interventions (e.g. posts or email responses) using computer programs. This would require detecting a problem (i.e. making a diagnosis) and generating an appropriate text that would be useful to the help-seeking individual. While it is technically demanding to generate human-quality feedback even in the simplest application, this challenge may be insurmountable in the context of complex mental health issues. A possible solution would be to augment the abilities of human moderators, helping them reach out to more people (i.e. help-seekers), more effectively and efficiently (Xxx & Xxx, 2014). This could be done using Natural Language Processing and Generation tools that filter, sort posts and generate draft responses that the moderators could then use and subsequently track the impact of their feedback.

Currently, templates are used to generate standardized responses; however, their value is limited as the content tends to be simplistic, static, rigid, repetitive, and only partially-appropriate for the target user. Within the health domain, personalization has been considered critical to patient-centered care and a number of studies have used Natural Language Generation (NLG). NLG is a subfield of artificial intelligence and computational linguistics which primarily focuses on producing human-like text from non-linguistic data with specific communicative goals (Reiter, Dale, & Feng, 2000). To date, NLG has generally been used for the authoring and personalization of webpages containing patient education materials. DiMarco et al. (2007) called this Information Therapy describing a system with personalized preoperative information, including resources that would typically be presented in a series of brochures discussing various surgical procedures. The system had a collection of reusable texts, each annotated with linguistic and formatting information, that the NLG tools automatically drew from to select, assemble and tailor the reader-specific pieces of text. Numerous studies have shown that NLG systems are able to produce dynamic human-like, individualised sentence structures suitable to various contexts (see proceedings of the International Natural Language Generation Conference for examples). Moreover, NLG systems can generate tailored and meaningful interventions by combining psychological strategies (Van Bilsen, 2013; Wiemer-Hastings et al., 2004) and techniques applied by moderators in peer support groups.

The aim of this project is to develop a Natural Language Generation Service (NLGS) that will create draft responses (i.e. interventions) to social media posts using input from a mental health knowledge base. The interventions can then be edited by

moderators and delivered to individuals through social networks or online support groups.

A first step in this project involved generating interventions in response to posts related to two mental health conditions: depression and anxiety. The conditions were chosen as they are the most common forms of mental illness in Australia and elsewhere. A sample of posts (n=25) collected from various mental health support groups/forums were used as the basis for generating interventions. These reflect the typical posts received by moderators of online support groups. We then asked a senior moderator from a youth mental health organisation in Australia, and three mental health professionals to write responses (i.e. interventions) for the same posts. Finally, both human and system interventions were rated by University psychology students/ interns (n=35) using quality measures designed specifically for the study. In this paper we evaluate the following in detail:

1. The quality of NLGS interventions as responses to posts on depression, anxiety or both.
2. The quality of NLGS compared to human-generated interventions.

Our study contributes the first evaluation of a NLG system in a mental health application. The system is novel in that it is being developed to support a human moderator by providing a draft intervention, rather than fully automating the response process. This approach, where the technology augments human capabilities, is particularly useful in contexts where those providing feedback might not have expertise in clinical psychology (something the system can help with) but have useful personal experiences they can share (something the computer cannot).

Natural Language Generation

The Internet helps deliver early interventions to at risk, help-seeking individuals and brings together people with shared health problems. Internet-based interventions can help large populations with minimal time, effort and cost through self-help programs (e.g. web-based) and minimal-contact therapy settings (e.g. emails, phone calls) (Barak et al., 2008; Doherty, Coyle, & Sharry, 2012; Spaulding et al., 2010). The field has grown and some structured frameworks and taxonomies for research of computer-mediated and Internet-based interventions have been developed (Barak, 2009; Barak & Grohol, 2011). Computerized interventions using different modalities such as online chat (Dowling & Rickwood, 2013), relational agents (Bickmore & Gruber, 2010), and interactive graphical exercises (Coyle et al., 2007; Doherty, Coyle, & Sharry, 2012) have been investigated. These approaches can be suitable for engaging users through human-human or human-computer dialogue and interactivity.

Text-based interventions, such as those employed in this study can be used in synchronous communication (e.g. the online chat) or asynchronous communication (e.g. emails and discussion forums), providing supportive messages with suggested activities or resources based on the problem(s) identified. Text-based interventions can be in the form of fixed responses targeting the overall community (e.g. via webpages), but a more nuanced and dynamic approach is to generate personalised text interventions to provide ad-hoc messages in human-like natural language structure (Reiter, Dale, & Feng, 2000). Text generation approaches in the form of natural language have been used in a variety
of applications each focusing on a particular problem. Back in the 60’s the system called ELIZA (Weizenbaum, 1966) was one of the first that emulated a Rogerian psychologist through dialogue and certain types of conversation (e.g. psychological issues). ELIZA was an early development but was a source of inspiration for programmers and developer in Artificial Intelligence that attempted such type of human-computer interaction.

New developments in human-computer interaction nowadays allow for much more sophisticated interfaces. In particular, Affective Computing (Picard, 1997) can make text generation systems more natural (Dockrey, 2007). For example, automated conversational coaches (Hoque et al., 2013) and robots (Breazeal, 2003) have been developed that aim of provide a variety of proto-social responses (e.g. simulating affects) by detecting natural social cues (e.g. speech, gaze, posture, facial expressions, etc.). Some applications have aimed to help crisis counsellors by analysing psychological and emotional patterns through text-based platforms (e.g. chat, SMS). For example, Fathom (Dinakar et al., 2015) is a natural language interface that makes use of machine learning approaches and probabilistic graphical models to extract and visualize psychological and emotional patterns in patients (e.g. during calls with counsellor). The statistics and visualization then allows the counsellor to respond accordingly. As for text generation, NLG based systems like PyschoGen (Dockrey, 2007) have been proposed that generates responses based on emulated mental/ emotional states.

By considering the psychological and emotional factors, NLG approaches would be suitable for automatically generating interventions that express empathy and compassion along with the client-centric health information and resources. This can be ideal for mental health clinicians, where information about a specific patient can be presented in the form of a report or as part of structured interventions. For moderators in online support groups, such information that can be used for quickly customizing and replying would greatly reduce their workload.

Even though the concept of text generation was developed much earlier (Appelt, 1985; McKeown, 1992), the field of NLG only started to mature in the late 1990s when new comprehensive structures of NLG systems suitable for real-word applications were proposed (Reiter, 1999; Reiter, Dale, & Feng, 2000). Following this, several NLG systems were developed for a growing number of applications (Gatt & Reiter, 2009; Reiter, 1999; Varges et al., 2012). At the end of 1990s, Reiter and Dale (2000) wrote “Building Natural Language Generation Systems”, the first book to provide a comprehensive overview of the tasks involved in building a NLG system.

A number of NLG frameworks that facilitate the development of new systems have been created, including SimpleNLG (Gatt & Reiter, 2009). Others focus on a single application, like SemScribe (Varges et al., 2012) which produces clinical reports from medical observations entered into a structured entry form, and BabyTalk (Portet et al., 2007), which provides support to medical professionals to make decisions based on large amounts of information. In recent years, researchers have started to apply NLG techniques to provide personalised health information for individual patients (DiMarco et al., 2007). For example, some attempts have been made to generate letters tailored for smokers using a NLG system called STOP (Reiter, Dale, & Feng, 2000; Reiter, Robertson, & Osman, 2003). However, these first steps aiming to offer personalised interventions in physical health have yet to be achieved in mental health applications.
Tailored Information Systems using NLG

Tailored patient information systems produce personalised medical information and/or advice (Reiter & Osman, 1997). The information can be patient-centric by providing information about an individual’s health condition or diagnosis, or doctor-centric by providing patient reports to doctors. Tailored systems provide more appropriate information relevant to each individual and therefore are more effective (Bental, Cawsey, & Jones, 1999). Evaluations of tailored information systems provide evidence that they may improve the quality and effectiveness of personalized texts.

SemScribe (Varges et al., 2012) is a system that automates the process of generating medical reports (particularly in cardiology), in natural language based on individual medical observations. By using NLG for a fully automatic mapping between non-linguistic input and linguistic output, it enables the doctor to get the corresponding medical report immediately after they enter observations (Faulstich et al., 2011).

The Baby Talk project (Portet et al., 2007) was developed to present clear summaries of medical data about sick babies in a neonatal intensive care unit. The data included physiological signals (e.g. heart rate, blood pressure), patient related notes, and laboratory test results. BT-45, the first Baby Talk system was able to generate written summaries of 45 minutes of clinical data by combining techniques from intelligent signal processing and NLG. An experiment showed that BT-45 texts were as effective for decision support as conventional visualisations (Portet et al., 2007).

Not all studies have shown improvements. STOP (Reiter, Dale, & Feng, 2000) is another NLG system that generates short tailored smoking cessation letters based on users’ responses to a four-page smoking questionnaire. A clinical trial showed that STOP was not effective as recipients of a tailored letter were less likely to stop smoking compared to recipients of a non-tailored letter (Reiter, Robertson, & Osman, 2003).

Generic Architecture for NLG

There are several possible architectures for NLG systems, but the one proposed by Reiter and Dale (Reiter, Dale, & Feng, 2000) is broadly compatible with most applications. In this architecture, three components are connected together into a pipeline. More specifically, a Document Planner determines the content and structure of a document. A Microplanner decides how to communicate the content and structure chosen by the Document Planner. This involves choosing words and syntactic structures. A Surface Realiser maps the abstract representations used by the Microplanner into an actual text. Message, Document Plan and Text Specification represent the input and output of each component.

Moderator Assistant: NLG Service for Mental Health Interventions

We have adopted the Reiter and Dale (2000) architecture as part of our mental health intervention module for the Moderator Assistant (MA) (Xxx & Xxx, 2013). The MA system is able to retrieve all incoming posts from nominated social media groups/forums using their Application Programming Interface (API). A triage module of the MA system, which implements a text classifier using NLP and machine learning techniques, is responsible for identifying mental health categories (e.g. depression, anxiety) from social media posts. This module also retrieves the timestamp of the post, name of the person, and other details that can be used as input by the NLG component.
The interventions generated by NLG can then be administered by moderator and posted back as comments to corresponding posts using the API. The overview of our NLG architecture is shown in Figure 1.

Figure 1: Overview of the NLG Architecture for Mental Health Interventions

The first step in this architecture is **Content Determination** where *Messages* are instantiated. Each *Message* represents a chunk of data that can be grouped together to express a specific meaning. The second step is **Document Structuring**, where the *Messages* are combined into *Document Plan* using schema and heuristic algorithms in order to group different kinds of *Messages* together in a logical order. This represents a tree structure with *Messages* as terminal nodes and *Discourse Relation* as internal nodes.

Although *Document Plan* groups *Messages* together, it does not specify how the information inside a *Message* should be structured. Therefore, the domain model expressed inside *Messages* need to be mapped into words that make sense. The third step is **Lexicalisation and Aggregation**, where words and syntactic structures are chosen to communicate the information in the *Document Plan*. This is a very important part of providing mental health intervention through this NLG architecture. The meaning of the information needs to be expressed correctly as inappropriate feedback may have a negative impact on the user.

Templates were used in **Content Determination**, which were retrieved from mental health professionals as well as by extracting the some common feedback/comments from Livejournal, Facebook, and ReachOut.com posts. These templates are mostly formed in complete sentences, therefore, the resulting *Messages* also consist of well-structured sentences. Only the other *Messages*, such as greetings
with name, need to be refined in the **Lexicalisation and Aggregation** stage. The resulting document from this step is a **Proto-phrase Text Specification**.

The **Proto-phrase Text Specification** can be used as the input to the **Surface Realisation** directly. It can also be refined in step four, **Referring Expression Generation**, where the symbolic names of entities are replaced by the semantic content of noun phrase referring expressions. The output of this stage is the **Text Specification**, which contains all information needed, as well as the message structure and the sentence structure.

The **Lexicalisation and Aggregation** and **Referring Expression Generation** steps do not affect the NLG process for this version of the NLGS architecture because the **Text Specification** has exactly the same structure as the **Document Plan**. Therefore, the **Lexicalisation and Aggregation** and **Referring Expression Generation** steps are not implemented in this version of the NLGS architecture. The **Text Specification** contains all the necessary information, which is then passed to **Surface Realiser**. This converts the **Text Specification** into real text from the abstract representations. The system will then produce the intended feedback. The following sections give the details of different parts of the NLGS architecture.

**Defining Messages**

A **Message** is essentially a form of particular configuration of domain elements, and it may contain different levels of information for each particular system (Reiter, Dale, & Feng, 2000). In order to define the message, we need to analyse the indented output that is to be generated as part of the intervention. Analysing several examples of real-world interventions from our dataset, we identified the following four types of messages that appear in social media interventions for mental health:

- **Greeting the person posting** (**Greeting Message**)
- **Comforting the person experiencing mental health problems** (**Comforting Message**)
- **Suggestions to the person experiencing mental health problems** (**Suggestion Message**)
- **Encouragement to the person experiencing mental health problems** (**Encouragement Message**)

Four types of messages were constructed for the intervention. Figure 2 shows how messages from the four types are grouped together to form an intervention.

![Figure 2: Example intervention divided by Messages](image-url)
Content Determination

In the *Content Determination* phrase, the system instantiates *Messages* designed in the previous section using post-related (e.g. mental health category) information extracted from social media posts and other information (e.g. current time). The NLGS system implements the content determination logic inside a group of ‘Feedback Generator’ classes using the `generateFeedback` method. The ‘Intervention Generator’ class handles the overall NLG generation tasks, including the content determination task (Figure 3). According to the figure, the *Greeting Message* is generated based on current timestamp. The `generalFeedbackGenerator` generates the *Message* that is suitable for any type of mental health categories whereas the specific mental health category feedback generators (e.g. `DepressionFeedbackGenerator`, `AnxietyFeedbackGenerator`) create *Messages* based on the mental health categories detected in the social media post. If no mental health category can be identified from the post then the `unknowCategoryFeedbackGenerator` is triggered. Finally, the *Messages* are combined into a `List<Message>` object. Currently, these feedback generators cannot generate a more personalised feedback *Message* due to the limitations of the information extracted from posts.

Figure 3: Overall *Content Determination* Logic

Greetings Generator

The greeting contains two parts, the first part is generated using the current timestamp (Table 1), and the second part is generated randomly (Table 2). These two are then combined to form the output Greeting Message.

Table 1: Greeting based on the current time.

<table>
<thead>
<tr>
<th>Current Time</th>
<th>Greeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>0am-6am</td>
<td>Hi.</td>
</tr>
<tr>
<td>6am-12 noon</td>
<td>Good morning</td>
</tr>
<tr>
<td>12 noon-18pm</td>
<td>Good afternoon</td>
</tr>
<tr>
<td>18pm-24pm</td>
<td>Good evening</td>
</tr>
</tbody>
</table>

Table 2: Random Greeting

<table>
<thead>
<tr>
<th>How are you?</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>How are you doing?</td>
<td></td>
</tr>
</tbody>
</table>
General Feedback Generator

The general feedback does not relate to a specific mental health problem. It contains text suitable for any type of mental health category. It can generate random Comforting Messages, Suggestion Messages, and Encouragement Messages based on the templates (i.e. knowledge base) provided by mental health professionals. Other than that, it also contains feedback providing suggestions based on the posting behaviour (e.g. the time when the post was submitted). For example, it will generate feedback similar to the following if the person posted late at night:

“It seems that you post really late; Healthy sleep habits can make a big difference in your quality of life. Make whatever adjustments you need to sleep 7-8 hours/night. Respect your need for sleep, and trust me, many other things will just fall in place.”

Depression and Anxiety Feedback Generator

This feedback generator produces feedback suitable for a specific mental health issue (e.g. depression, anxiety). It can generate random Comforting Messages, Suggestion Messages, and Encouragement Messages under its mental health category based on the template provided by mental health professionals. As the MA (Xxx & Xxx, 2013) builds on the NLP component, which is responsible for extracting personalised information from the original post, the feedback generator can be improved with new data or features about the user or post.

Document Plan, Document Structuring and Realiser

All the Messages are retrieved from the Content Determination and then separated into different Message lists according to the Message type (i.e. Greeting, Comforting, Suggestion, Encouragement).

Each Document Plan is a node in a tree structure, containing a parent (also a Document Plan), a topic (the information carrying document plan), and constituents. The constituents contains the children document plans and the discourse relationship (e.g. Sequence, Contrast, Elaboration) between them. Each node in the tree contains a complete Document Plan for each Message that are already in the form of surface text.

In order to instantiate the Document Plan, both schema and heuristic algorithms are used in the Document Structuring phrase. In this process, all Document Plans that contain same types of Messages are grouped together into a higher level Document Plan. Finally, according to the order of different types of Messages, the final Document Plan is constructed.

The Realiser constructs the final intervention by traversing the Document Plan tree using post-order traversal. This is achieved by combining all the Document Plan contents (i.e. the node of the tree) together.
**Design and Methods**

**Evaluation Study**

Augmentation is the process of supporting the moderator (as opposed to automation where the computer accomplishes tasks normally done by the human). The quality of the augmentation is related to the quality of the texts automatically generated, which we evaluate by measuring the variation of the output texts and their appropriateness in relation to the corresponding mental health problem and post. This section details the process of selecting sample posts, assessing for variation (Jaccard distance) in the generated texts, and evaluating the quality of the measures, ratings and the overall evaluation process.

**Pilot Evaluation**

As a way of testing the system and the quality of text generation, we performed a pilot evaluation for NLGS in the context of responding to depression and anxiety posts, where three mental health professionals rated the NLGS interventions along with the human interventions (Xxx & Xxx, 2015). Both sources of interventions were randomized and then presented for the rating procedure. Despite variations in rating scores, results showed that the quality of the interventions generated by NLGS for depression and anxiety were satisfactory in relation to the early development and nature of the application. As part of an extended evaluation, 35 University psychology students/interns rated the NLGS in order to provide a broader sense of quality of the interventions. The following section describes the extended evaluation.

**Main Evaluation**

This section presents the main evaluation for NLGS in the context of depression and anxiety. In order to evaluate the performance of NLGS, 25 social network posts related to depression and anxiety were chosen. These two categories were chose because the end user organization (ie. ReachOut) found them the most critical categories in a triage system.

With those 25 posts as input, we generated 25 corresponding interventions using NLGS. Three clinicians (two psychologists, one psychiatrist) and a trained moderator separately wrote responses (i.e. interventions) for the 25 posts. The clinicians are experts in the field of mental health and are collaborating closely with the project. The moderator is a senior staff in ReachOut who has a lot of experience in supporting young people though their forums. We hypothesized that the two groups (clinicians versus moderator) would generate two different types of interventions each with their own qualities. In order to simulate the environment in which they may be responding to users, the original posts were presented to the clinicians and the moderator with the respective categories using Google Blogger and interventions were collected as comments.

All the interventions were rated by participants as described in section ‘Rating Interventions’. The project was approved by The University of Xxxxxx Human Research Ethics Committee.
Selecting Posts for Intervention

We collected sample posts from two online peer-supported groups (Livejournal, Facebook) as well as one online, moderated health support group (ReachOut.com). The author’s name (i.e. username) and identifying information were removed from each post. Initially two psychologists and a psychiatrist selected 90 posts out of 4,583 that were classified under depression, anxiety and 14 other mental health related categories (e.g., self-harm, suicide, drug/alcohol use, bullying/violence, medication/treatment, psychosis, bipolar, eating disorder, personality disorder, sleep, accessing help, positive emotion, self-care, etc.) These posts were used as gold standards for training the participants and were assumed to be best examples of the total 4,583 posts. Of the 90 posts, a total of 25 related to depression and anxiety or both were randomly selected. The final distribution was: seven of depression, eight of anxiety and 10 combined (contained both depression and anxiety) posts. The clinicians and the moderator had to read the individual posts and write corresponding responses as part of the human interventions; hence the total number of 25 posts allowed a reasonable workload for this task.

NLGS Interventions and Measure of Variations

The sample 25 posts were used as input for NLGS to generate 25 matching interventions. The NLGS interventions are intended to be dynamic; therefore, it is useful to evaluate the variation of the output text to avoid the repetitious nature of the interventions such as when responding to posts indicating similar mental health problems to the same recipient within a short period. By measuring the dissimilarity between interventions that NLGS generated using the Jaccard distance, we are able to identify the variation in the 25 NLGS interventions. Jaccard distance is obtained by subtracting the Jaccard similarity coefficient from 1. In this context, the dissimilarity is defined as the difference in the number of the union and the intersection of words in sentences divided by the number of the union of the words in the sentences.

The NLGS interventions have an average of 0.79 Jaccard dissimilarity, which indicates that the system is able to generate interventions with good variation. With that being said, since the interventions all relate to a specific mental health topic, the variation is not extremely high as some keywords repeatedly appeared under the same topics. The average Jaccard dissimilarity for the seven depression interventions, eight anxiety interventions, and 10 combined interventions are 0.71, 0.67, and 0.68, respectively.

Quality Measures

In order to rate the interventions, quality measures were developed specifically for the project by research staff at the Xxxx. These measures were then used to rate the 75 interventions (25 NLGS interventions, 25 moderator interventions, and 25 mental health professional interventions). The following questions in Table 3 were asked to measure quality of the interventions.

Table 3: Quality Measure questions and response type.
<table>
<thead>
<tr>
<th>Questions</th>
<th>Response Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>The intervention is grammatically correct (grammatical)</td>
<td>Likert scales: Strongly Disagree (1), Disagree (2), Neither (3), Agree (4), Strongly Agree (5)</td>
</tr>
<tr>
<td>The language used in the intervention is clear and unambiguous (clarity)</td>
<td>Likert scales: Strongly Disagree (1), Disagree (2), Neither (3), Agree (4), Strongly Agree (5)</td>
</tr>
<tr>
<td>The intervention is appropriate (appropriateness)</td>
<td>Likert scales: Strongly Disagree (1), Disagree (2), Neither (3), Agree (4), Strongly Agree (5)</td>
</tr>
<tr>
<td>The intervention provides the recipient with useful advice (usefulness)</td>
<td>Likert scales: Strongly Disagree (1), Disagree (2), Neither (3), Agree (4), Strongly Agree (5)</td>
</tr>
<tr>
<td>The intervention is likely to encourage the recipient to take positive steps towards enhancing their mental health and wellbeing (positive reinforcement)</td>
<td>Likert scales: Strongly Disagree (1), Disagree (2), Neither (3), Agree (4), Strongly Agree (5)</td>
</tr>
<tr>
<td>What is your overall rating of the intervention? (overall)</td>
<td>Likert scales: Very poor (1), Poor (2), Average (3), Good (4), Excellent (5)</td>
</tr>
<tr>
<td>In your opinion, was this intervention machine-generated?</td>
<td>Discrete: YES, NO, Don’t know</td>
</tr>
<tr>
<td>Do you have any comments regarding this intervention?</td>
<td>Comment box</td>
</tr>
</tbody>
</table>

**Rating Interventions**

The participants who rated the interventions (human and NLGS) were aged from 18 to 27 years and were mostly undergraduate students from first year to fourth year pursuing
Psychology or an equivalent university degree. The cohort of raters was considered informed and interested enough in mental health issues, yet not expert psychologists. This is a representative sample of the human moderators who do this job, both in age, interest and prior knowledge of mental health first aid.

A total of 44 psychology students from a variety of universities in Australia were recruited and allocated the rating task; however, only 35 interns completed the rating task. In order to facilitate the rating process, a rating system was developed in-house and was explained to the raters before they started the task. The rating system presented a form for collecting demographic information followed by the rating task. Each participant rated a total of 50 interventions (i.e. 25 NLGS interventions and 25 human interventions). A comprehensive face-to-face training was provided by a psychiatrist who described the experiment and the rating procedure. Participants were asked to complete as many ratings as possible during the training and all issues (e.g. questions and confusions) were resolved through discussion. They completed the remaining task over a period of one week.

To explore if the order of presentation had an effect on the results, the participants were randomly allocated to one of four groups. Each group either started with human interventions (trained professional moderator or mental health professional) or the NLGS one. In the second stage they annotated the other type: NLGS for the former and either trained professional moderator (ReachOut) or mental health professional (Clinician), for the latter (see Table 4). Initially, all 44 interns were divided into the four groups equally. The participants were not informed about the group allocated to them as well as the order of presentation.

<table>
<thead>
<tr>
<th>Group</th>
<th>First Part Interventions (25)</th>
<th>Second Part Interventions (25)</th>
<th>Num. of Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>Clinician</td>
<td>NLGS</td>
<td>9</td>
</tr>
<tr>
<td>RM</td>
<td>ReachOut</td>
<td>NLGS</td>
<td>11</td>
</tr>
<tr>
<td>MC</td>
<td>NLGS</td>
<td>Clinician</td>
<td>8</td>
</tr>
<tr>
<td>MR</td>
<td>NLGS</td>
<td>ReachOut</td>
<td>7</td>
</tr>
</tbody>
</table>

Hypotheses

We hypothesised that rating scores change over time and that the quality of interventions would be perceived as good initially but drop towards the end. We believed that when raters see many of the system generated interventions in a short period of time, they may start to find them less interesting.

While comparing NLGS ratings with human ratings, we hypothesised that rating scores would change based on the order of presentation. More specifically, we propose that the system-generated interventions would be rater higher if the raters saw the human intervention responses after the NLGS interventions and vice versa. We believe that when raters see the human interventions in the first order, they may find the NLGS interventions less appealing.

Data Analyses

The Likert scale for first six questions in Table 3 was converted to 1.00-5.00 values.
Then the average and standard deviation (SD) scores were calculated over the participants \(n = 35\) for the following scenarios.

1. NLGS interventions for three categories individually: Depression, Anxiety, and both (Depression and Anxiety).
2. The first, middle, and last proportions of NLGS interventions according to order of presentation.
3. Human and NLGS interventions individually for all categories combined.
4. Human and NLGS interventions individually according to order of presentation.

The percentage of NLGS that received rating above 2.00 and 3.00 were calculated individually for the three categories and combined (i.e. overall) over all participants. This was used to report the proportion of the interventions receiving high rating scores (above 2.00 and 3.00).

The one-tailed t-test \((p<0.05)\) was conducted for reporting significant differences in rating scores for scenarios (3) and (4).

**Results**

This section presents the outcome of the NLGS interventions using the rating scores. The results are presented for the following:

1. The quality of NLGS interventions.
2. The quality of NLGS in comparison to human generated interventions.

**Overall quality of NLGS interventions**

Firstly, we report the quality of NLGS interventions (Figure 4) based on the average and standard deviation (SD) for three types of intervention categories: Depression, Anxiety, and both (Dep&Anx). Grammatical and Clarity (Q1 and Q2) received average scores slightly above 4.00 for all three types of intervention categories. Both Depression and Anxiety for Appropriateness (Q3), Positive Reinforcement (Q5) and Dep&Anx for Usefulness (Q4) received average scores slightly above 3.00. As for all other cases the average score was slightly below 3.00. This indicates that the interventions were grammatically correct and clear, but average in relation to delivering interventions that are appropriate to the corresponding post, providing useful resources, and offering advice to improve general wellbeing. The SD is lower for Grammatical and Clarity compared to the other questions for the three categories, indicating that the raters had more agreement in scoring Grammatical and Clarity, where as opinions were more varied in relation to other questions.
The majority of the NLGS interventions received ratings above 2.00 (Table 5). Over 90% of the ratings scored above 2.00 for *Grammatical* and *Clarity* (Q1 and Q2), whereas the remaining questions scored 60-70%. The result is similar for achieving ratings above 3.00 for *Grammatical* and *Clarity* (Q1 and Q2), however only 30-50% for * Appropriateness* (Q3), *Usefulness* (Q4), *Positive Reinforcement* (Q5), and *Overall* (Q6) ratings (Table 6).

Table 5: Proportion of interventions receiving rating above *Disagree* (2.00)

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>0.93</td>
<td>0.93</td>
<td>0.63</td>
<td>0.73</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.95</td>
<td>0.98</td>
<td>0.63</td>
<td>0.71</td>
<td>0.68</td>
<td>0.63</td>
</tr>
<tr>
<td>Dep&amp;Anx</td>
<td>0.98</td>
<td>0.97</td>
<td>0.56</td>
<td>0.63</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td>Overall</td>
<td>0.96</td>
<td>0.96</td>
<td>0.60</td>
<td>0.68</td>
<td>0.65</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 6: Proportion of interventions receiving rating above *Neutral* (3.00)

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>0.89</td>
<td>0.87</td>
<td>0.47</td>
<td>0.52</td>
<td>0.44</td>
<td>0.30</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.89</td>
<td>0.91</td>
<td>0.45</td>
<td>0.54</td>
<td>0.46</td>
<td>0.40</td>
</tr>
<tr>
<td>Dep&amp;Anx</td>
<td>0.96</td>
<td>0.91</td>
<td>0.35</td>
<td>0.39</td>
<td>0.36</td>
<td>0.29</td>
</tr>
<tr>
<td>Overall</td>
<td>0.92</td>
<td>0.90</td>
<td>0.42</td>
<td>0.48</td>
<td>0.41</td>
<td>0.33</td>
</tr>
</tbody>
</table>

In Figure 5, we present the average scores for the first, middle, and last proportion of the interventions in the time-series over all raters for the three categories. According to the results the first and middle five interventions received higher rating scores compared to the last five interventions for * Appropriateness* (Q3), *Usefulness* (Q4), *Positive Reinforcement* (Q5), and *Overall* (Q6). The ratings for *Grammatical* and *Clarity* were consistent for the first, middle, and last five interventions.
Quality of NLGS Interventions vs. Human Interventions

We also compare the rating scores (i.e. performance) of NLGS intervention with human intervention. Table 7 gives the average and SD scores over all raters for the NLGS and human interventions. *Grammatical* (Q1) and *Clarity* (Q2) have similar rating scores for both NLGS and human interventions, however the average rating is above 4.00 for all questions for human interventions. The standard deviations indicate higher variations in ratings for NLGS (except of *Grammatical* and *Clarity*). The difference in the scores of NLGS and human interventions were significant ($p<0.05$) for all questions except *Grammatical* (one-tailed $t(1748)=1.19$, $p=0.12$).

Table 7: Comparing interventions average (SD) ratings (NLGS vs. Human)

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLGS</td>
<td>4.25</td>
<td>4.18</td>
<td>2.98</td>
<td>3.19</td>
<td>3.07</td>
<td>2.89</td>
</tr>
<tr>
<td></td>
<td>(0.76)</td>
<td>(0.72)</td>
<td>(1.24)</td>
<td>(1.16)</td>
<td>(1.15)</td>
<td>(1.14)</td>
</tr>
<tr>
<td>Human</td>
<td>4.29</td>
<td>4.47</td>
<td>4.33</td>
<td>4.31</td>
<td>4.18</td>
<td>4.14</td>
</tr>
<tr>
<td></td>
<td>(0.93)</td>
<td>(0.64)</td>
<td>(0.77)</td>
<td>(0.75)</td>
<td>(0.83)</td>
<td>(0.85)</td>
</tr>
</tbody>
</table>

The authorship (i.e. human or computer) was recognisable as shown in the answers to Q7, “in your opinion, was this intervention machine-generated?”. We find that 76% of the NLGS interventions were correctly recognised by raters to be machine generated. Table 8 gives the percentage scores for the individual intervention types. As for the human interventions, 82% were assumed by raters to be generated by humans.

Table 8: Opinion about Q7 (is machine generated?)

<table>
<thead>
<tr>
<th></th>
<th>Q7</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLGS</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>80 %</td>
</tr>
<tr>
<td>Anxiety</td>
<td>71 %</td>
</tr>
<tr>
<td>Dep&amp;Anx</td>
<td>79 %</td>
</tr>
<tr>
<td>Overall</td>
<td>76 %</td>
</tr>
</tbody>
</table>

In Figure 6, we present data for the average (and SD) scores for the NLGS interventions based on the order of presentation. *MFirst* represents when raters were...
presented with the NLGS interventions first and then human ratings and MSecond represents the opposite. According to the results MFirst received slightly higher average rating scores compared to MSecond for Appropriateness (Q3), Usefulness (Q4), Positive Reinforcement (Q5), and Overall (Q6). The ratings for Grammatical and Clarity (Q1 and Q2) were opposite. This indicates that for Grammatical and Clarity the ratings were good for NLGS due to having rich text contents generated by the system. However, the raters have a slightly lower perception of the quality of NLGS interventions after seeing human interventions that were tailored for the user as well as providing support and other useful resources. The difference in the scores of MFirst and MSecond were significant (Table 9).

![Figure 6: Average rating scores for NLGS based on presentation order (Machine vs. Human)](image)

<table>
<thead>
<tr>
<th>Question</th>
<th>T-test score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical (Q1)</td>
<td>t (873)=3.31, p&lt;0.05</td>
</tr>
<tr>
<td>Clarity (Q2)</td>
<td>t (873)=1.65, p=0.05</td>
</tr>
<tr>
<td>Appropriateness (Q3)</td>
<td>t (873)=1.87, p&lt;0.05</td>
</tr>
<tr>
<td>Usefulness (Q4)</td>
<td>t (873)=3.35, p&lt;0.05</td>
</tr>
<tr>
<td>Positive Reinforcement (Q5)</td>
<td>t (873)=2.78, p&lt;0.05</td>
</tr>
<tr>
<td>Overall (Q6)</td>
<td>t (873)=2.78, p&lt;0.05</td>
</tr>
</tbody>
</table>

**Discussion and Conclusions**

This paper presented NLGS, an NLG-based system for generating interventions for supporting mental health. Despite the work on NLG-based implementations in other domains (Portet et al., 2007; Reiter, 1999; Varges et al., 2012), its use in mental health, particularly for interventions, is still very novel. Previous work, such as PsychoGen (Dockrey, 2007) have attempted NLG approaches in the mental health domain but are limited in terms of generating long /complicated natural language output and lacks a proper evaluation of its quality.
The positive findings of our study suggest that the system is capable of generating natural language interventions in this domain. More specifically, the NLGS produced intervention-based responses that were clear and grammatically correct. Although the aim of NLGS is not to replace a human moderator, this system could potentially be very useful for providing moderators with draft responses which would reduce their workload, even if those responses requiring editing, and allow them to meet increasing demand. While questions remain as to the ability of the NLGS in relation to the generation of personalized messages, in practice and in the context of a sensitive area such as mental health, personalized messages are always encouraged to come from humans.

The Moderator Assistant system (including the NLGS component) is being deployed in an Australian mental health organization, and we are evaluating time saving and other benefits that moderators may find. One approach in the real world is to automate the process of detecting concerning contents and generating corresponding responses. Despite this being a very cost and time effective solution, it is unsuitable for sensitive issues like mental health and its support. Instead, human moderators can administer contents detected (e.g. keyword-based, NLP) and generated (e.g. NLG) by machines. For example, the moderators in ReachOut.com have to ensure that they collect, read, and understand the contents posted by the community and then respond with resources (e.g. links) and personal experiences related to the concerns. The Moderator Assistant system aims to help detect some of the issues the moderators listen out for and provide template responses for the corresponding concerns for them to administer and use. The NLGS will provide the support for the later.

**Limitations**

This study has two important limitations. The first, related to its ecological validity, is that the way moderators and end users perceive the quality of posts (both human or NLG generated) would be different in a real life situation to what we have been able to do here. Second, we have not attempted to evaluate the impact that the interventions have on health outcomes. The differences in perceived quality may or may not have significant impact on the way the interventions help end-users. This is a common problem, the health impact of human generated interventions in peer-support groups are often not measured directly.

As part of future work, other mental health categories (e.g. self-harm, suicide) and personalised information (e.g. age, sentiment, cognitive processing, etc.) from social media posts will be extracted for the NLGS input in order to address a broader range of mental health problems and to improve the quality of the personalised messages. Furthermore, the currently system only uses the mental health categories (e.g. depression, anxiety) for individual posts to generate the interventions. Any previous responses or dialogue between the moderator and the help-seeker should be considered as part of future work by storing historical information/keywords in the NLG knowledge base.

As for evaluating the quality of interventions, the procedure presented in this paper is based on a small sample size (i.e. raters) with quality measure questions developed specifically for this study. The questions can be revised in future studies for reporting the quality of interventions as the enhancement of NLGS progresses. Despite
the limitations, the evaluation presented in this paper provide good insight into the capability of the NLGS for generating natural language responses in the mental health domain.

References


Clarke, G., & Yarborough, B. J. (2013). Evaluating the promise of health IT to enhance/expand the reach of mental health services. General Hospital Psychiatry, 35(4), 339-344.


