Karlsruhe Institute of Technology

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Using Tweets as ”Ice-Breaking” Sentences in a Social Dialog System

Bachelor’s Thesis of

Aleksandar Andonov

at the Department of Informatics
Institute for Anthropomatics and Robotics (IAR)

Reviewer:	Prof. Dr. Alexander Waibel
Second reviewer:	Dr. Sebastian Stücker
Advisor:	M.A. Maria Schmidt

08. July 2015 – 07. November 2015

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, 07.11.2015

. . . .

(Aleksandar Andonov)

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text. I have followed the respectively valid KIT statutes for safeguarding good scientific practice.

Karlsruhe, 07.11.2015

. . . .

(Aleksandar Andonov)

Abstract

Data from social networks has been utilized in many domains, a prominent example being advertising. This data covers a huge number of topics, but data about a specific user often relates to the interests of its author. Since all active social network users produce data, they also reveal their interests and views on topics which they care about.

In this work a system to extract these interests and use them to create “ice-breaking sentences” is implemented by using Twitter as the online social network of choice. “Ice-breaking” sentences should appeal to the user at the beginning of the dialog and increase the probability of a response from the user. Thus, their goal is to “break the ice” with the user. Many dialog systems could profit from implementing a system like the one proposed in this work in order to gain sympathy with the user by utilizing the widely researched existence of homophily between people with the same interests.

The system is evaluated through a user study which shows that it is better in generating “ice-breaking” sentences than a system which chooses interests randomly. Additionally we note the high mean scores for the perceived appeal of the sentences and the interest in the general topic of the sentences. Additionally, we note that 70% of the test users would answer the system and continue talking about the same topic which was introduced by the “ice-breaking” sentence.

Zusammenfassung

Daten von sozialen Netzwerken werden in vielen verschiedenen Bereichen angewendet. Diese Daten decken eine große Anzahl an Themen ab und stellen die persönlichen Interessen des Autors dar. Da alle aktiven Benutzer von sozialen Netzwerken Daten generieren, ist es möglich ihre Interessen und Einstellungen zu für sie interessanten Themen auszuspüren.

Im Rahmen von dieser These wird ein System implementiert, das Sätze generiert, die das Eis in einem Gespräch brechen sollten. Das bedeutet, dass die Sätze für den Benutzer am Anfang des Gesprächs interessant sein sollten und die Wahrscheinlichkeit für eine Antwort steigern. Ein solches System würde vielen Dialogsystemen von Nutzen sein da mit einem solchen System diese dem Benutzer sympathischer sein würden. Das basiert auf der Existenz von Homophilie zwischen Menschen, die die gleichen Interessen haben.

Das System wurde durch eine Benutzerstudie evaluiert, die zeigt, dass das System höher bewertete Sätze generiert als ein anderes System, das Interessen zufällig auswählt. Die Studie zeigt darüber hinaus, dass die Benutzer die Sätze interessant fanden. Dabei hätten 70% der Befragten auf den Satz geantwortet und wären bei dem selben Thema geblieben.

Acknowledgments

This thesis was partially written at *Carnegie Mellon University* in Pittsburgh, USA. This extremely positive and enriching experience was possible thanks to the InterACT Exchange Program and was funded with an IGEL - ”Informatik GrEnzenLos” scholarship. I would therefore like to thank all people involved and supproting these programs.

Additionally, I would like to thank my advisor at the KIT, Maria Schmidt, for her lasting commitment to this work and all the flexibility shown during my stay in the USA. I am also very grateful to Professor Alex Waibel for his ideas regarding the direction of this work and his support concerning the exchange program. I would furthermore like to thank all people who helped me during my stay at Carnegie Mellon, especially Jae Cho and Susanne Burger who always had good tips on where to hear an exciting talk. Finally and most importantly I would like to thank my parents for their support, commitment and love.

Contents

1. Introduction

Thousands of years ago the first humans started using language as means to communicate with each other, share their ideas and discuss ways to cooperate and improve their daily lives. Since these ancient times, the means of communication have evolved further. The invention of symbols, alphabets and writing systems (5000 years ago) enabled communication between people which are separated in space and time. The advancements in telecommunication like the invention of the telegraph and telephone allowed further for real-time long-distance communication. The recent upsurge in the use of smartphones, tablets and other mobile and Internet-connected devices made the long-distance real-time communication even more natural since people are no longer bound to a specific place in order to use these services (in contrast to the traditional corded telephone).

Dialog systems represent another step in the evolution of communication. The continuing development of such systems benefits the industry and society in various ways. For example, such systems can be deployed to give information about bus schedules or act as a therapist and measure stress levels (see [20] and [6]). Dialog systems however still lack the intelligence of real human-human conversations and often fail to grasp the context of the conversation if this context was not modeled or trained beforehand. Moreover a lot of systems lack the usual small talk which is feature of nearly all human conversations.

Online social networks like Facebook and Twitter are also part of the long-going change in the way we communicate. These networks allow us to receive information about topics in which we are interested and from people that we find important (e.g., friends, celebrities). They also enable us to determine the preferences of a new or even a potential acquaintance and often feature posts which develop into topics in human-human conversations (e.g., Trump posts in the USA). Recent statistics show that Facebook alone has around 1,5 billion monthly active users[1], which underlines the popularity that such networks have gained in recent years.

These qualities of social networks and the relative ease with which social networks' data can be read and interpreted, make them a suitable source of information for dialog systems. For example, pages which were liked by a user on Facebook often represent topics in which this user is interested. Such topics would be suitable for a small talk conversation since they are more likely to capture the attention of the user (since the user has showed interest in them). Furthermore, the concept of tagging oneself at certain locations gives additional information where the user was before the dialog with the system.

[1] Data from Statista - http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Sentences featuring these locations could also interesting in a small talk conversation.

This thesis therefore demonstrates how to utilize some of the available data in a social network (Twitter) and use it to create sentences on topics which are appealing to the user and which increase the probability of the user responding to the system. Such sentences can be used at the beginning of a conversation as means to imitate the usual small talk which features in human-human conversations and make the system more appealing to the user.

Additionally, the widely researched phenomenon of homophily states that people sharing the same interests or values, are more likely to become friends. This furthermore motivates the use of the system described in this thesis since the system creates sentences by taking the target user's interests into account. Thus, the system acts as if it has the same interests as the target user and is therefore more likely to be received in a positive manner. Other dialog systems (e.g., goal-oriented) can hence make use of the system proposed in this thesis in order to engage the user in a more natural way.

The next Chapter 2 presents related work and basic notions from the domains of sociology, dialog systems and recommendation systems which have influenced this work. Chapter 3 then gives deeper insight into the implementation of the system created in the scope of this thesis. Chapter 4 presents the user study conducted in order to evaluate the system and discusses its result. Finally Chapter 5 presents the conclusions from this thesis and possible future work.

2. Background and Related Work

The notion of using data from a social network about a specific user in order to start a dialog with them, relates mainly to three general research aress: social networks and sociology, dialog systems and recommendation systems. The relations between these areas and this thesis are presented in the next sections as well as concrete examples of research which influenced this work.

2.1. Social Networks and Sociology

Social networks and their analysis are concepts which predate modern online social networks like Facebook and Twitter, but there is no consensus in the scientific community on the exact time when the term social network was used for the first time. It is however evident that the term became popular in the 1930s and is examined in the work of Jacob L. Moreno (1934)[17]. It is difficult to give an exact definition of what a social network is either, but a widespread definition is that it represents a social structure consisting of a set of social actors and the dyadic relationships between them. [1]

Online social networks thus represent an implementation of this concept on the Internet and offer the users the ability to define their own personal social networks. This enabled the scientific community to conduct large-scale research on the effects which occur in social networks with better precision than before the advancement of the internet-based social networks. The main reason for this is the availability of very precise, relatively cheap and objective data, in comparison to the manual and often expensive collection of subjective data that largely dominated the research before. For example, if a researchers wanted to examine the social circle of a specific person before the rise of the online social networks, they had to first ask the specific user to define this circle and then interview all people from the circle in order to get accurate data about their interests and preferences. In an online social network all this effort can be replaced with just examining the data already available in the social network.

The next subsections will examine three effects which are widely studied within online social networks: information diffusion (Subsection 2.1.1), social influence (Subsection 2.1.2) and homophily (Subsection 2.1.3). The relationship between these effects and this work are also explained.

[1] Wikipedia - https://en.wikipedia.org/wiki/Social_network

2.1.1. Information Diffusion

Information diffusion in a social network is a term which describes the spread of information within this network. Online social networks often offer tools like shares in Facebook and retweets in Twitter to facilitate this concept. Additionally, studies show that such networks increase the flow of diversified opinions and new information to their users[1] and establish themselves as major factors in socially important situations like the 2010 Arab Spring.[9]

Two main types of models are used to analyze the information diffusion in online social networks: explanatory and predictive models. Explanatory models are given the time of arrival of new information to each user and based on that, create a graph which shows how each user was influenced by the others. Such models are used to explain how a piece of information was propagated through the network. Predictive models, on the other hand, try to predict how a piece of information is going to spread through a network given the network itself and information on the spread of previous pieces of information. [7]

The notion of information diffusion relates to the topic of this thesis since a successful "ice-breaking" sentence would have to contain information which is interesting for the user and which they would therefore eventually diffuse (e.g., retweet on Twitter). This work does not use a large-scale model of the whole online social network (Twitter), but rather examines only a part which contains the target user and their following list.

2.1.2. Social Influence

Social influence is a concept which is closely related to information diffusion and describes the fact that an action taken by a single person in a social network, can influence others to perform similar actions. A large-scale study involving data from 253 million Facebook users gives an example for the effects of social influence. It shows that the probability of sharing a certain URL increases with the number of friends which have shared this URL. [1] This suggests that the probability of a user sharing a certain URL is influenced by the amount of friends which have shared this link already.

A main problem in determining the effects of social influence in online social networks is distinguishing it from social selection. The latter increases the probability of people with similar interests to become friends in the first place, whereas the former causes people who are already befriended to adopt the same preferences or perform the same actions as their friends.[14] The distinction between the two is problematic since both result in people having the same interests or performing the same actions. Efforts to solve this problem involve the use of data in addition to the one provided from the online social network like a paper by Lewis et al.[14] suggests. Taking this approach, the study from the paper involved 1640 college students and combined data from Facebook and academic and housing data from the college. The study examined if friendships on Facebook influence the music preferences of the students. The results of the study suggested that friends had the same interests not because they were influencing one another, but because their shared interests were one of the reasons they became friends.[14]

Some users have more social influence than others and their posts are more likely to get shared by a large number of users. Determining which users are more influential is crucial to applications such as marketing and recommending. This task is also important to the system implemented in the scope of this work since the generated sentences have to be interesting to the user and such sentences are often written by influential users in the network.

A number of different algorithms to determine the influential users are proposed in the scientific literature. The K-core algorithm[12] designates users which are in the core of the network as more influential than others and uses k-decomposition to determine these.[25] TwitterRank, dedicated to Twitter, ranks users similarly to PageRank, but takes into account the different topics in which the user is interested. This thesis uses the In-Degree algorithm, deployed by Twitter and others, to measure the influence of users.[25]

2.1.3. Homophily

Homophily designates the phenomenon that people with similar interests are more likely to associate with each other. The scientific literature offers a large amount of research on homophily which proves that religious, racial and socio-economical homophily exist.[22][16] Weng et al also show that homophily is present in Twitter and that users who are in a following relationship are more likely to share the same interests than users who are not related.[25]

The system implemented in the scope of this work, relies on these findings and assumes that people who are interested in a certain topic are more likely to engage with a system which discusses this topic. Therefore the sentences generated by this system concern topics which are interesting to the target user.

2.2. Twitter

Twitter is a combination of a microblogging service, a social network and a news platform[13] and as of October 2015 has reached more than 300 million monthly active users.[2] Twitter implements relations between its users by using the following principle which states that if you follow a user, you will get all of their future tweets in your feed and that you can follow anyone without asking for special permissions (except when the user is private). Opposed to other social networks like Facebook, the following concept makes relations in Twitter unidirectional. Users of the site can write posts which are limited to 140 characters and will be posted to the feeds of all their followers.

[2] Data about number of users from Statista - http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Hashtags are keywords in a post and have gained popularity as a concept after they were introduced in Twitter. The site features a well-defined markup syntax, where other users are marked with a preceding ”@” (user mentions), retweets (posting a post from another user) with a ”RT” and hashtags with ”#”. Additionally, users can ”favorite” tweets when they like them and can retrieve them from the favorites tab in their account page.

The topology which Twitter offers is similar to that of other social networks, but with the difference, as mentioned, that relations between users are unidirectional. Additionally, Kwak et al. show that the degree of separation in Twitter is lower in comparison to other social networks and support the findings of other studies (Weng et al.[25]) that homophily exists on Twitter.[13] These findings support the dualistic nature of Twitter which is viewed, on the one hand, as a social networks and, on the other, as an information source and news media.

Because of these interesting features, Twitter has been widely used in research, e.g., to measure the sentiments of a population[18] or to predict stock markets.[3] Furthermore, Bessho et al. propose a dialog system based on Twitter data.[2] The system creates replies by comparing the input to a list of utterance-reply pairs, whereby the input is compared to the utterance. The response of the system is then the utterance-reply pair with the highest similarity score, but if this score is below a certain threshold, a real-time crowdsourcing platform is used instead.

The system implemented in the scope of this thesis uses Twitter data to form “ice-breaking” sentences for a number of reasons. First, it offers a less restrictive API than other alternatives like Facebook and public user content can be accessed without seeking approval from the user. Second, the limit on the length of posts provides for short and informative sentences which can be converted into “ice-breaking” sentences with less effort and excludes long texts which are not suitable for that particular task. Furthermore, as studies have shown [13], a lot of Twitter users use the platform as a source of information about news and current events which makes posts from the network very suitable for “ice-breaking” sentences since new and interesting information can be presented to the user in these sentences.

However, the use of Twitter has also its downsides. The limit on the length of posts leads a lot of users to write grammatically incorrect sentences and use a lot of abbreviations. Further observations have shown an extensive use of links to explain a post with more characters which is especially true for accounts belonging to news providers.

2.3. Dialog Systems

Dialog is a common feature of human communication, but it possesses a large number of different definitions. In the scope of this work, dialog is defined as a conversation between two or more actors[3] (also called interlocutors). A dialog system is thus a system

[3] Merriam-Webster - http://www.merriam-webster.com/dictionary/dialogue

which can be an actor in a dialog. One example for the increasing popularity of such systems is their advancement in the consumer market and include Apple's Siri, Microsoft's Cortana or IBM's Watson.

The next sections of this chapter will examine different types of dialog systems: spoken and text-based dialog systems, goal-oriented and social dialog systems. The relationship between these systems and the system implemented in the scope of this thesis are also explained.

2.3.1. Text-Based Dialog Systems

Text-based dialog systems (TBDS) are systems which are able to engage in a dialog with a user utilizing text as a communication modality. Such systems have different designs, but the common problems which they face can be categorized into Natural Language Understanding (NLU), Dialog Management (DM) and Natural Language Generation (NLG). The next paragraphs of this section describe these problems and common ways to solve them.

The first problem which a TBDS faces is to extract the system-relevant information from the textual user input. This problem is called NLU and can be solved by, e.g., constructing simple rule-based grammars to parse relevant information.

Once all relevant information is extracted, the system must analyze it and decide which information should be contained in the response. This problem is called DM and one way to solve it would be to define a set of rules for every possible input from the NLU module. An example of a system which follows this notion is ELIZA, proposed by Weizenbaum in 1966.[24] This system first searches the user input for predefined keywords. If no keywords are found, the system outputs a context-free sentence or an earlier output. If keywords are found, the system chooses a rule to transform the user input into a response sentence, whereby the choice of this rule depends on the keywords found in the input.

The last problem which the system faces is the NLG problem which has the goal to create natural sentences from the information output of the DM module. One example of a strategy for this problem would be, for example, to define a fixed number of sentences corresponding to each type of information.

A popular example of TBDS are chatbots which aim at texting with a human in such a natural way that the human fails to distinguish them from other human users. A popular chatbot is Cleverbot[4] which learns from previous conversations and replies to a given input in the same way as a human has answered the question before. Thus, with the growing number of conversations Cleverbot also gets cleverer and has more options for a response. More recently, Vinyals et al. describe a chatbot using variants of Long Short

[4] Cleverbot - http://www.cleverbot.com

Term Memory (LSTM) recurrent neural networks which can harness the wisdom of large and noisier data sets. [23]

2.3.2. Spoken Dialog Systems

Spoken dialog systems (SDS) are similar to TBDS, but use speech instead of text as their primary modality. In consequence such systems face the additional problems of automatic speech recognition (ASR) and text-to-speech (TTS) synthesis. An overview of a typical architecture of such systems can be seen in Figure 2.1.

The ASR problem occurs at the very beginning of the processing chain and consists of assigning textual representation to an acoustic signal. A popular way to achieve this is to extract important features from the acoustic signal and use an acoustic and a language model to determine the most probable textual representation of the acoustic signal. Once this is done, the textual representation can be used as an input to a TBDS.

The TBDS then produces textual output and the system faces the additional problem of TTS synthesis. One way to solve this problem would be to record the sounds of the different phonemes or biphones and then build the whole spoken output from these smaller pieces. This method is called concatenative speech synthesis. [10]

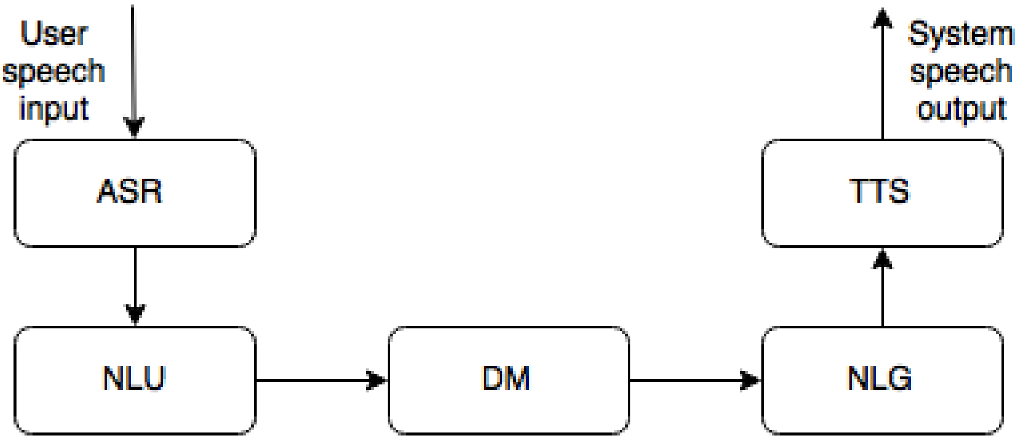


Figure 2.1.: A typical architecture pattern for a SDS

2.3.3. Goal-Oriented Dialog Systems

Goal-oriented dialog systems are designed to fulfill a certain goal. Therefore their ability to handle situations which are not directly connected with this goal is often limited or non-existent.

A possible design for such systems is one which resembles a finite-state machine where the final state represents the goal which the system has to accomplish. In order to reach this state, the system must first go through other states which represent subgoals. The state transition function is then a function which maps from the set of goal-relevant words and expressions to the set of states (subgoals) defined by the system.

Goal-oriented dialog systems have been widely researched and deployed for everyday tasks. In particular, there is a large number of goal-oriented TBDS which are widely used when a GUI is not available. An example for a different type of goal-oriented dialog system is Let's Go proposed by Raux et al. This system is a SDS and has the goal to provide schedule information about the Port Authority of Allegheny County's buses out of the working hours of human telephone operators.[20][19] Let's Go was deployed actively in March 2005 and has received 40 to 60 calls each night with an average success rate of around 40% with the main difficulties faced by the system laying in the automatic speech recognition but also in recovery and error strategies.[20]

State-of-the-art systems utilize statistical models with one popular choice being Partially Observable Markov Decision Processes (POMDPs). These models allow representing the uncertainty which exists in SDS since different possibilities for the current dialog state are analyzed. Additionally they simplify error recovery strategies since it is always possible to switch to a different hypothesis about the current dialog state. [26]

2.3.4. Social Dialog Systems

Social dialog systems describe systems which do not have a fixed goal, but rather aim to engage the user in a natural dialog. In this sense such systems can also be viewed as goal-oriented with the goal being to reproduce small talk and its features which are present in everyday human-human conversations.

Such systems have been studied in the scientific literature with one of the most prominent early examples being the rule-based ELIZA system proposed by Weizenbaum in 1966.[24] A more recent example is the Tick-Tock conversational agent which tries to keep the user in the conversation for as long as possible. This is accomplished by measuring the user engagement with the current topic and deploying different strategies depending on the level of engagement.[27]

Social dialog systems are especially closely related with the system implemented in the scope of this work since “ice-breaking” sentences are also a feature present in everyday small talk. In addition the system has also the goal to engage the target user and raise the probability for a response from them which is also one of the key goals of social dialog systems.

2.4. Recommendation Systems

Recommendation systems are defined in Wikipedia as systems which are a subclass of information filtering systems and seek to predict the ’rating’ or ’preference’ that users would give to an item.[21] Such systems are especially useful for Internet platforms which have a huge number of data which means that users are overflown with data which might not be interesting for them at all. Schwartz and Ward indeed point out that in a situation when users have too much options to choose from, they might feel suppressed and miserable. (Chapter 6 in [15])

In consequence of this problem, a large number of popular Internet platforms which manage big numbers of data and have to present them to a user make use of recommendation systems. Examples include electronic commerce websites like Amazon where the recommended items should be the ones which are most likely to be sold to a particular user and video sharing platforms like YouTube and Vine where the recommended videos should be the ones which are most likely to keep the user entertained.

As it is evident from the examples given, recommendation systems fulfill different purposes depending on the platform where they are deployed. Ricci et al. distinguish five major goals for recommendation systems: increasing the number of items sold, diversifying the items sold, increasing user satisfaction, increasing user fidelity and having a better understanding of what the user wants. Recommendation systems often try to optimize the fulfillment of more than one of these goals while trying to fulfill the goals which the user has when using the system.[21]

The recommendation system deployed by the video-sharing website YouTube, for example, tries to increase user satisfaction with the system by providing users with possibly interesting video recommendations and sparing them the time to search through the enormous database. At the same time the system has to also offer diversified content so that the users can also find videos which are completely new to them.

Furthermore, the system must be designed by taking features of the underlaying platform in mind. For example, YouTube features short and noisy interactions and two different types of data: content and user data. Content data is formed by the video metadata and the video input wheres the user data is further split into explicit and implicit data. Explicit user data is formed by actions like favoriting a video or writing a comment, whereas implicit user data is a result of the user interacting with a video like watching a portion of a video. By examining the data the system deployed by YouTube constructs a set of possibly interesting videos and ranks these in three different stages: video quality, user specificity and diversification. The highest ranked videos are then shown to the user.[5]

The system implemented in the scope of this work has some similarities with recommendation systems and the one deployed by YouTube. In particular, the selection of a tweet which should be interesting to a target user is in its core a recommendation task. The design of the system is also similar to the one used by YouTube since tweets are also ranked and selected based on their overall qualities (time when they were created), user specificity (how well do they suit the interests of the user) and diversification (choosing only a number of tweets for every interest).

3. Implementation

This chapter gives an overview of the system which was implemented in the scope of this thesis. Given a Twitter username it generates sentences which appeal to this specific user and can therefore serve as "ice-breaking" sentences in a dialog.

The first section describes the abstract architecture of the whole system and does not go into any detail about the separate components. The remaining sections of this chapter are then dedicated to describing these modules and their abstract function. The implementation of each of these in the described system is also explained. Furthermore, Section A.1.1 in the Appendix shows and describes an important code example from the system.

3.1. Overview

The architecture features a modular design which enables to easily make changes to the general flow of the system (see Figure 3.1), introduce new algorithms and mix existing ones together in new ways.

The only input needed by the system is the username of a Twitter user. Given that the system deduces the interests of the user from their Twitter profile (Section 3.2 goes into more detail how this is done). Thereafter tweets which are associated with these interests are gathered online and evaluated by the system (respectively Section 3.3 and Section 3.4). Another algorithm then chooses a group of tweets based on this evaluation (Section 3.5). These tweets are given to a NLG component which transforms them into sentences that fulfill the goals of an ”ice-breaking” sentence (Section 3.6). Eventually the last algorithm chooses the final output of the system from these sentences (Section 3.7).

3.2. Interest Filter (IF)

The Interest Filter (IF) module has the task to extract the interests of the user from their Twitter account. The output of this module is a list of interests and additionally a non-negative score for each interest. Higher scores are hereby interpreted as being more important to the target user (client). The output of this module is subsequently given to the next module, the Tweet Gatherer, as input.

The interface of this module is implemented by the Basic Interest Filter. In order to deduce the interests of the target user, the list of users which the client follows (the following list of the client), is analyzed following the logic presented in Figure 3.2.

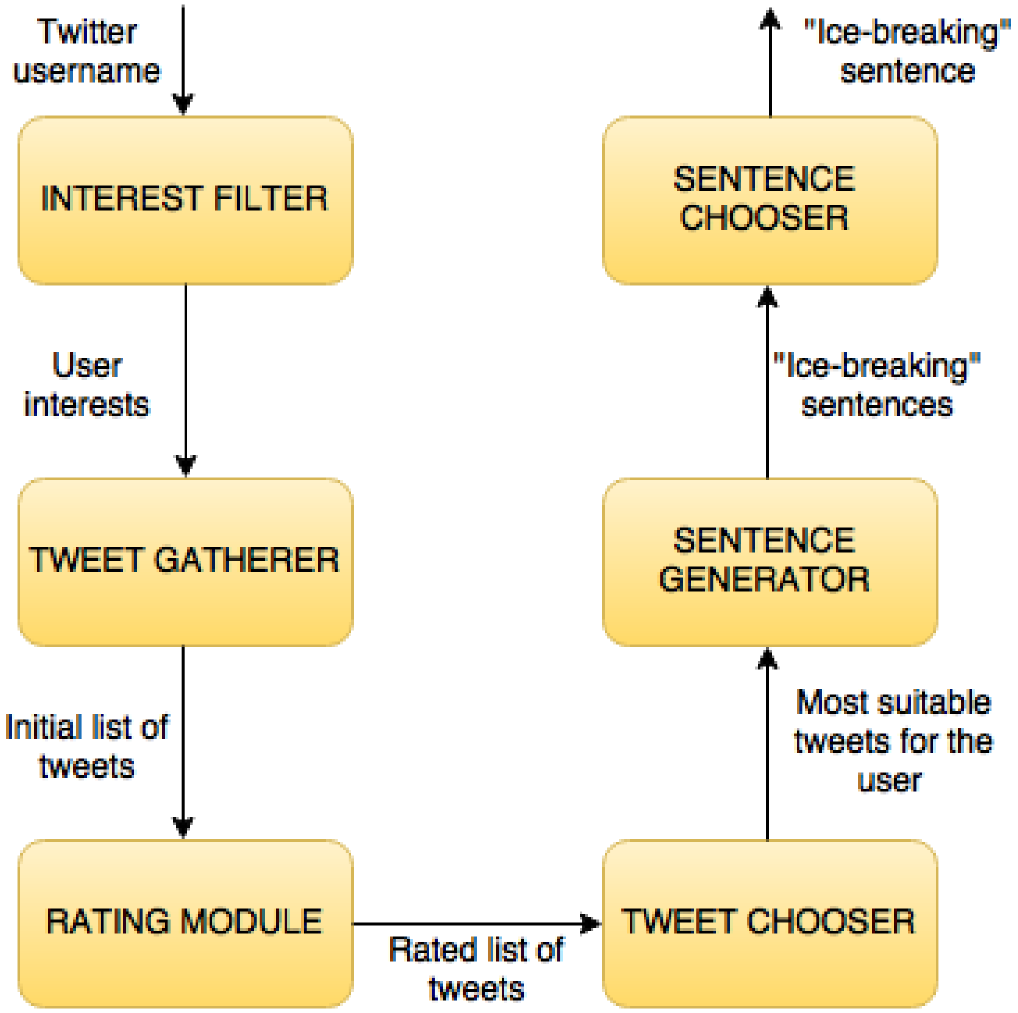
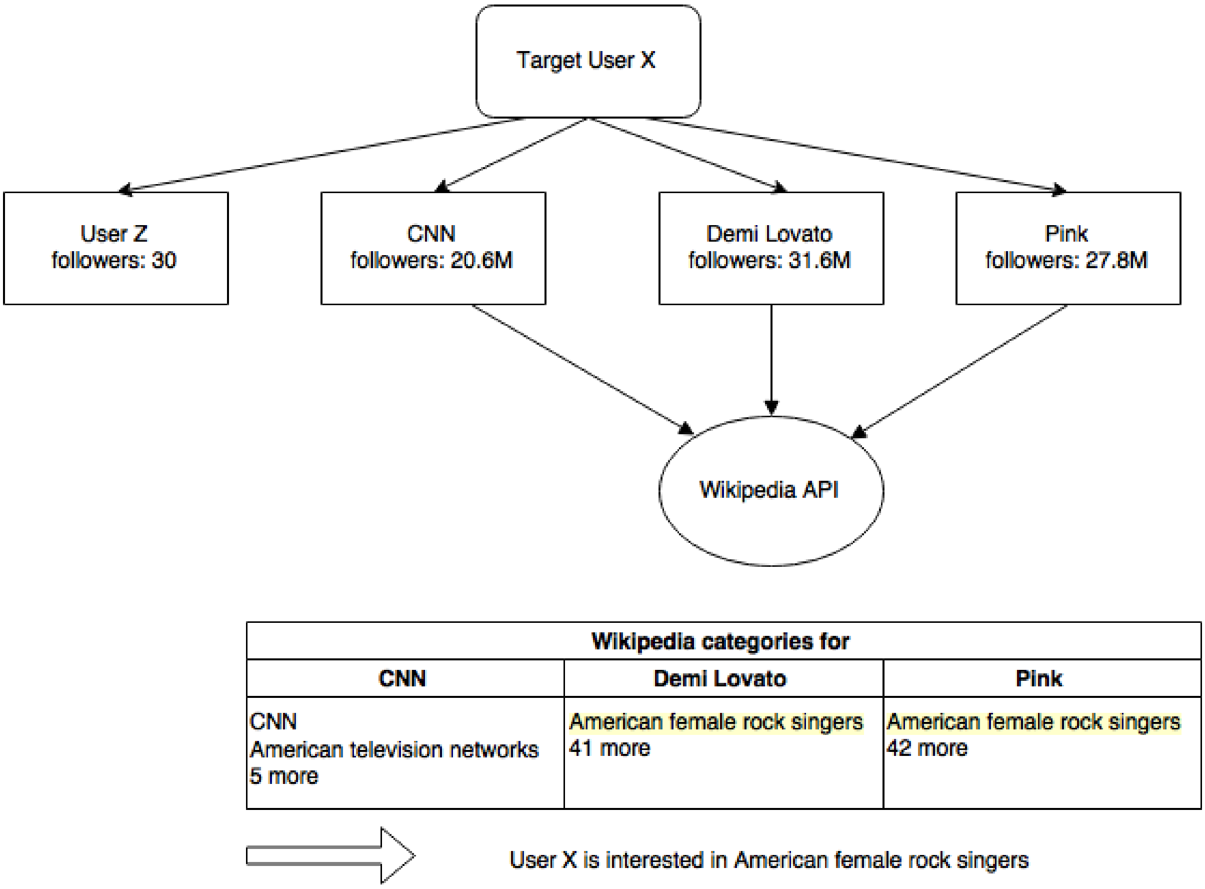


Figure 3.1.: System Overview

Unlike Facebook's differentiation between pages and users, Twitter's API offers no such categorization. This raises the problem of determining which users can be easily mapped to interests using data from Wikipedia. Such users would, for example, stand for businesses, brands, organizations, etc. and usually have a large number of followers.

Based on these observations every user in the following list is categorized as important or unimportant whereby important users are defined as users with more than 50,000 followers. It is assumed that because of their popularity among Twitter users, the important users represent and stand for some specific interests.

In order to determine these interests, the system searches Wikipedia for an article about each important user. If such an article exists, it serves the system in two different ways. First, it confirms the hypothesis that the user is indeed an important person/organization/place, who/which stands for a specific array of interests. Second, it allows the use of the categorization system of Wikipedia. This is important since a number of categories



Wikipedia categories for		
CNN	Demi Lovato	Pink
CNN American television networks 5 more	American female rock singers 41 more	American female rock singers 42 more

Figure 3.2.: The target user follows UserZ, CNN, Demi Lovato and Pink. Since CNN, Demi Lovato and Pink have more than 50 000 followers they are considered popular by the system. Therefore the searches for their respective articles on Wikipedia and associates them with the categories these articles have. These categories are treated as interests. The category ”American female rock singers” is assigned to both the article about Pink and Demi Lovato. Therefore it is more probable that the user is interested in American female rock singers and the IF module ranks this interest higher than the others.

like Demography or Sport describe interests. It is therefore possible to use the name of a category as the name of an interest.

Following this observation the IF uses the categories assigned to each article about an important Twitter user as the interests which this user represents. Thus, every important user is now associated with the categories assigned to their respective Wikipedia article. These categories are hereby treated as the interests which the user represents (see Figure 3.2 for an example).

At this stage every important user is associated with their own list of interests. These separate lists are merged together, whereby interests which occur multiple times are added only once and the information about the number of their occurrences is stored separately. This newly formed list represents the first part of this module's output: the list of the client's interests. Therefore the module now focuses on calculating the second part of the output: the non-negative score for each interest.

This score is calculated by the Basic Interest Filter as the number of users an interest has been assigned to. This follows the thought that if a client follows, e.g., a great number of IT companies, they are then genuinely interested in the IT industry. The Basic Interests Filter would reflect this by returning higher scores for IT industry related interests since they will be assigned to more users from the client's following list.

The interest filtering phase featured two major problems. Both were related to the unique microculture which Wikipedia represents.

The first problem was the different depth of categorization which is caused by the fact that everyone can create categories and assign them to articles in Wikipedia. Some categories are therefore very abstract and do not represent interests at all, e.g., *Living people* or *1984 births*, while others are too specific and feature only a small number of articles, e.g., *RAAF commands*[1]. Since large abstract categories are assigned to a large number of articles, they were consistently rated higher by the Basic IF. This caused the system to mainly choose very general categories like *Living people* which did not represent the interests of the user. In order to overcome this problem the system now performs an additional check and filters out categories with more than 1000 and less than 10 entries. These numbers were chosen only based on observations of possibly interesting Wikipedia categories and subjective judgement about how well the IF functions with them.

The second encountered problem was the ambiguity of the Wikipedia search results since, e.g., multiple people can have the same name. Therefore the results of the search could feature a number of possible articles. The implementation does not offer any particular strategy to resolve this ambiguity. It rather relies on the fact that if a client follows a large number of users and if false classification does occur, then the interests assigned to the misclassified user will be outliers compared to rest of the client's interests. The Basic

[1] RAAF - Royal Australian Air Force

IF will therefore calculate a lower score for them and they will be at the bottom of the ranked list of the client's interests.

3.3. Tweet Gatherer (TG)

The Tweet Gatherer (TG) module has the task to collect tweets about the interests determined by the IF. Thus, the output of this module is a list of tweets which is then used for the creation of "ice-breaking" sentences.

This module is implemented by the system as the Basic Tweet Gatherer (BTG). The size of the gathered data can be controlled via three variables: *INTEREST_COUNT*, *USERS_PRO_INTEREST*, *TWEETS_PRO_USER*.

The list of interests generated by the IF is first truncated so that its maximum size is not larger than the value of the *INTEREST_COUNT* variable. Lower ranked interests are hereby removed from the list. This results in less noise in the data since outliers caused by categorization misclassification in Wikipedia are removed. Additionally, this also counteracts the misclassification due to ambiguous Wikipedia search (see previous Section 3.2).

In the next step a list of possibly interesting users is created. The users are chosen from the following list of the client and represent the interests defined by their respective Wikipedia article's categories. For each interest in the truncated list the users representing this interest are added if they were not added already before. The number of users added from a single interest must however not exceed the value of the *USERS_PRO_INTEREST* variable. The addition of this variable offers additional flexibility to the system. On the one hand, it can prevent the list of being "over-flooded" by users from one interest, but on the other, ensures flexibility as such "over-flooding" may be favored, e.g., when the list of interests is truncated to a very small size.

Eventually the output list of tweets is populated by adding tweets from each possibly interesting user. Hereby a maximum of *TWEETS_PRO_USER* tweets are added to the list for each user. The interest because of which the creator of a tweet was added to the possibly interesting users is also added to the tweet as additional information. It is said that the tweet represents this interest.

The three control variables described above also enable controlling the size of the output list. In particular, they set the following upper bound for the length of the returned list:

$$INTEREST\_COUNT * USERS\_PRO\_INTEREST * TWEETS\_PRO\_USER \tag{3.1}$$

3.4. Rating Module (RM)

The Rating Module (RM) has the task to rate the tweets collected by the TG. The output of this module is thus a rated list of tweets, which are then used by a separate Tweet Chooser module to choose the best matching tweets.

This module is implemented as the Basic Tweet Evaluator (BTE). The BTE algorithm rates the tweets in three stages: interest, user and tweet stage. It additionally calculates weights for the score of each stage. A total score is then calculated based on the weights and the separate stage scores.

All three scores are calculated on a scale between 1 and 16, whereas more suitable tweets should receive higher scores. The choice of these numbers is arbitrary and mainly reflects the choice to define eight time intervals. This topic is discussed further in Subsection 3.4.3.

3.4.1. Interest Score (IS)

The first stage in which the tweets are rated is the interest stage. In this stage the interest stage score is calculated for each tweet added by the TG. As explained in the previous Section 3.3 each such tweet is assigned to an interest. The interests score rates the tweets based only on this interest and it is therefore sufficient to calculate the score for each such interest. This score represents the strength of the association between the target user and the specific interest.

The interests are already ranked by the IF module. The algorithm takes advantage of this and just maps this ranking to the 1 to 16 scale. This is accomplished by dividing the sorted list of interests in 16 equally large parts. Thereby the interests which were at the bottom of the sorted list form the first part and are given an interest score of 1. The following parts are then progressively given larger scores until the last part, containing the highest ranked interests, is given the score of 16. Thereby after all interest in a part were rated with X, the interests from the next part are rated with X+1.

3.4.2. User Score (US)

The second stage of the ranking algorithm is the user stage. In this stage the tweets are assigned a user stage score which ranks them based on their creator. This score measures the strength of the association between the client and the respective creator of the tweet.

In order to analyze this relationship the favorites list of the client is analyzed (more on favorites lists in Section 2.2). For each creator the algorithm checks the number of times a tweet was added to the favorites list of the client and saves these numbers as the initial user scores. The maximum and minimum initial scores define an interval which is divided into 16 equally large intervals. These intervals define the mapping of the initial scores to the 1 to 16 scale. For example, if tweets from a creator B were added X times to the favorites list and X is in the Y-the interval, then the user score of creator B is Y (whereby Y is between 1 and 16).

3.4.3. Time/Tweet Score (TS)

The last stage in which tweets are ranked is the tweet stage. In this stage tweets are evaluated based on their features irrespective of the target user. The concrete implementation of this module used by the system only ranks the tweets based on the time they were created. Thus, the tweet score is in this case synonymous with the time score. Hereby, newer tweets are ranked higher since it is less probable that the user has seen them and since it is assumed that developing stories are more appealing to the majority of users.

The time score is calculated based on hard-coded predefined boundaries which are presented in Table 3.1. A total of eight intervals are defined, whereas the newest tweets get the maximum score of 16. With each boundary passed, the score which tweets are assigned, gets decremented by two until tweets older than 1 day all get rated with 2.

As mentioned in the beginning, the choice of a scale from 1 to 16 reflects the choice of the eight time intervals. Since the score gets decremented by two each time a boundary between two intervals is passed and the lowest possible score is 2, the maximum score is 16. The choice to define exactly eight time intervals is arbitrary as is the decrementation by two. These values are, of course, therefore not optimal and better values possibly exist. Finding these would however involve measuring subjective values like the appeal of the ”ice-breaking” sentence for every chosen parameter combination. This would require big amounts of data on how well systems with different parameters perform which is currently not present. Since gathering such information would go beyond the scope of this work, the system currently largely relies on the expert knowledge of the author of this thesis about social networks.

Time Score	From	To
16	0 sec	15 min
14	15 min	30 min
12	30 min	1h
10	1h	3h
8	3h	6h
6	6h	12h
4	12h	24h
2	24h	∞

Table 3.1.: Time intervals with corresponding time scores

3.4.4. Weights Calculation

Besides calculating comparable scores for each of the three stages, the BTE algorithm also calculates a weight for each one of the scores. These weights represent the confidence in the estimated scores and are used to calculate the total score (see Equation 3.3).

All three weight are calculated using the coefficient of variation (CV) (see Equation 3.2) which is defined as the ratio of the standard deviation to the mean. As a standardized measure of dispersion the CV is used to find out if the data is well dispersed.

$$c_v = \frac{\overbrace{\sigma}^{\text{Standard deviation}}}{\underbrace{\mu}_{\text{mean}}} \tag{3.2}$$

A high dispersion would mean that the mapping from the absolute values to the scale of 1 to 16, did create artificially small differences in the scores. For example, interest X which has occurred 100 times could be rated with 16 and an interest Y, which has occurred only 15 times with 15. This can, for example, happen if the system has found 16 different interests with all other interests besides X and Y having less than 15 occurrences. In a situation like this the interests will be ranked by the Basic IF module based on the number of occurrences they have. Thus, X will be ranked highest and Y second highest. When the RM then splits the 16 interests into 16 equally large sets during the interest stage, each set will contain exactly one interest. Since X was ranked highest by the IF module, it will be in the set with the highest interest score which is 16. Y will be in the set with the second highest interest score since all sets contain only one interest and Y was the second highest ranked interest by the IF module.

In a situation like the one described above, the system will suffer from a loss of information and X and Y will have nearly the same interest score which does not reflect the huge difference in the absolute number of occurrences. This loss could be partially impeded when using the CV as weight since high dispersion of the data would result in a higher CV and thus the difference between the interest score of X and Y will be larger than 1.

On the other hand, in case of low dispersion of the absolute data the mapping could create artificially big differences in the scores, although these did not exist in the absolute data. For example, interest X which has occurred 10 times could be ranked with 16 and interest Y which has occurred 8 times could be ranked with 1. This could, e.g., happen if the IF has found 16 different interests, whereby X has 10 occurrences, 14 other interests have 9 occurrences and Y has 8 occurrences. Similarly to the example above, the RM will then split the interests into 16 sets each one of which contains exactly one interest. In a situation like this X will have an interests score of 16 since no other interest has more or at least 10 occurrences and Y will have an interests score of 1 since no other interests has less or at most 8 occurrences.

In the case like the one described above the loss of the information form the absolute number of occurrences will result in an artificially small difference. In this case the low dispersion of the absolute data would cause the CV to be low, which would mean that all interest scores are deflated and have a smaller share in calculating the total score of the tweets. Thus, the estimates of the interest score would be deemed more unreliable.

The absolute data which is used to calculate the CV for the IS is the data which the IF used to rank the interests. In the case of the Basic IF this is the number of users from the following list an interest has been assigned to. The set containing this data for every interest found by the IF module is then formed. The mean and standard deviation of the data from this set are then determined which enables for the calculation of the CV. The resulting coefficient is then used as the interest score weight (c_v^i).

Before being mapped to the 1 to 16 scale, the US was represented by the number of times a tweet from each creator was added to the following list of the client. The mean and standard deviation of the set containing this data for each user is calculated in order to determine the CV. The resulting coefficient is the user score weight (c_v^u).

The calculation of the time score weight (TSW, c_v^t) differs from the calculation of the other types of weights. Instead of using the absolute values before their mapping to the 1 to 16 scale, the mapped time scores are used. This exception can be made, because the boundaries between the different score values are predefined and invariant to the absolute values of the data (unlike the US and IS). It prevents that huge time differences in old data have an impact on the TSW and ensures that differences in the time distribution of the tweets are weighted higher.

For example, if a user tweets once every month, their tweets would have very high dispersion and therefore the TSW would be very high when using absolute data. Most of the tweets have however a big chance of being irrelevant since the topic which they cover could have since evolved. This problem is solved when using the mapped scores since most of this user's tweets would be rated with 1 and the data would feature a relatively low dispersion and therefore low TSW.

The system would, on the other hand, get a list of tweets from the same period if a user tweets very often. This would lead to a low dispersion of the data and a low weight of the TS if absolute data is used. Again, this problem is solved by using the mapped scores since these would be more diverse and therefore feature a high dispersion.

Finally, the total sore is calculated by weighting each score with its respective weight and then summing all scores (see Equation 3.3). A higher total score indicates a higher chance that the tweet is interesting for the user.

$$c_v^i * IS + c_v^u * US + c_v^t * TS = TOTAL\_SCORE \tag{3.3}$$

3.5. Tweet Chooser (TC)

The Tweet Chooser (TC) module has the task to choose the tweets which will be transformed into sentences and eventually presented to the user. This module makes it possible to implement different strategies, e.g., choose different topics or introduce a random element. It uses the evaluated list of tweets which the RM generated and outputs a list of tweets which are to be transformed into sentences.

This module is implemented by the system as the TopXTweetChooser. As the name suggests, this algorithm forms the output list by choosing the X tweets with the highest overall score from the list of rated tweets generated by the RM. The system implemented in the scope of this work uses a X of 20.

3.6. Sentence Generator (SG)

The Sentence Generator (SG) module has the task to transform tweets into ”ice-breaking” sentences which capture the attention of the target user and encourage them to give a response to the system. This module thus serves as a natural language generation (NLG) module with the output being a list of natural sentences.

This module is implemented by the system as the Rule Based Sentence Generator (RBSG) which is, as the name hints, a rule-based NLG module. The rules which the generator applies are mostly simple and result in concatenating phrases like “Hey, did you know that” or ”Oh, X just tweeted” to the text of the tweet. Such generation is, of course, limited, but still has a feeble impact on the quality of the output. A more sophisticated statistical generator is of course a possible alternative given that enough data is present, but would exceed the boundaries of this thesis.

Furthermore, the RBSG removes hashtags placed at the end of a sentence since these are mostly not connected to the sentence before and retains hashtags at the middle and the beginning of a sentence since those are usually just used as words within the sentence. The hashtags at the end of a sentence usually also point to the topic of the tweet. These are therefore sometimes used by the generator to give information to the user about the topic of the tweet. For example, “Hey, NASA just tweeted about JourneyToMars and said that ’Lettuce’ tell you how veggies growing on @Space_Station will help on our JourneyToMars. A link was posted too!” was created from the original tweet ”’Lettuce’ tell you how veggies growing on @Space_Station will help on our #JourneyToMars: http://go.nasa.gov/1J2Qm9a”[2].

The RBSG also detects links and removes these from the sentence, but informs the user that a link was posted with the tweet. This step is taken since a lot of the tweets which feature a link, actually partially describe its contents and since the system is oriented to natural dialog which rarely features someone spelling a link.

[2] Original tweet - https://twitter.com/NASA/status/630424063509458945

3.7. Final Output

The last module, the Sentence Chooser (SC), analyzes the output of the SG and chooses the sentence which will be presented to the target user. This module offers additional flexibility to the system and allows, for example, for the detection of errors which occurred during the NLG. The output of this module is the output of the whole system.

The implementation of this module used by the system currently just chooses the sentence which was generated from the highest rated tweet. This is the output of the system. Further reading on the implementation of the system and a code example are presented in Section A.1.1 in the Appendix.

4. Evaluation

In order to evaluate the system implemented in the scope of this work, a user evaluation study was performed. Since it is not possible to determine if a sentence is really a good ”ice-breaking” one fully automatically and since the system is targeted at human users, this kind of evaluation is considered appropriate.

The next Section 4.1 gives an overview of the predefined goals for the user study. Section 4.2 describes the design of the study and Section 4.3 explains how this design was implemented. Section 4.4 gives an overview of the results which are then discussed in Section 4.5. Finally Section 4.6 summarizes the evaluation process.

4.1. Evaluation Goals

The user evaluation study has the main goal to measure if and how well the system described in this thesis achieves its main goal of creating “ice-breaking” sentences. In the scope of this work the definition of an ”ice-breaking” sentence is a sentence which is appealing to the user and increases the probability of a response. Therefore this study aims at measuring the appeal and the likelihood of a response from the user.

Apart from evaluating if and how well the system achieves its main goal, the study also aims at evaluating the performance of the separate modules and module groups involved in the implementation. This ensures that even if the system fails its main goal in the study, an analysis of the possible reasons can be performed. It also serves as a possibility to determine where the bottleneck of the system lays in means of user satisfaction.

The study additionally compares the system implemented in this work with a baseline system since the two goals described above are subjective and it is not easy to determine what values would mean that the system did well or not. A comparison with a different system however yields additional insight into these problems as it can be determined if the system performed better or worse than the baseline system.

Besides these goals, the evaluation study also seeks to determine if the use of personal information by the system raises security and privacy concerns in its users. This is an important issue for the system since it is designed not to ask any information or permission from any user besides the Twitter username of the target user. As a subgoal of this goal, the study also strives to gain some insight into how users envision the use of a system like the one proposed in this thesis.

4.2. Study design

The design of the user study started after the evaluation goals for the study had been defined. It was decided to create a live test study in which participants could test the system online and then fill out a questionnaire in which they could rate the performance of the system. An alternative approach, which was also considered, was to present the users with examples of the system output and ask them to rate the system based on these. The live system approach has however multiple advantages over the alternative. One of these is, for example, that in a live system test the users will evaluate an "ice-breaking" sentence which is tailored exactly for their interests and would not need to imagine being interested in things they have no interest in.

The decision to execute a live study however raises the question of how to present the system to the user. Since the study had to be done remotely, it was decided to design an Internet website on which the system could be displayed and at which the user could interact with it. The user interface for the interaction was text-based and was designed to resemble a chat. After the system generated an "ice-breaking" sentence it would ask the user to visit the questionnaire and fill it out.

Since the user study was performed in Germany and it was evident that a large number of participants in the study would not use Twitter actively, an additional feature to create a "dummy" user account was added to the system. This feature was designed to resemble a real Twitter account and the users had to decide on a username and choose other accounts to follow. The data collected during the creation of these accounts was not stored after the end of the user interaction in order to satisfy any privacy concerns from the participants.

Since the study had also the goal to compare the system described in this thesis with a baseline system, a search for such a system was performed. As such a system did not exist, to the best knowledge of the author of this thesis, a baseline system was designed. This system was named SystemX and features randomly choosing a tweet which is then converted to an "ice-breaking" sentence by the same NLG component as in the original system. The baseline system was presented to the user with a textual user interface and in an Internet website featuring the same design as the original system in order to avoid any bias based on the user interface.

Because the number of participants for the study was expected to be low, it was decided to ask all participants to test and evaluate both systems. Since it was not clear if any learning effects affect the systems and if the order of testing the systems would influence the users, it was decided to ask half of the participants to first test the baseline and then the original system and ask the other half to do the opposite. This way it is also possible to study if there are any learning effects affecting the system and how the different order of testing influences the results.

After taking the decision to perform a live study and designing the modules for the original and baseline system, it was time to design the most important third component of the study: the questionnaire. It had the main task to gather information from the participants which would help towards fulfilling the study's goals defined earlier (see Section 4.1). The questionnaire's design followed some of the guidelines provided by previous research like the ones from Hone and Graham [8].

First, as a good scientific practice, the participants were asked to fill in some demographic information like gender and age. Additionally, this part featured questions about which system was used and if this system was the first one used by the user. This was necessary in order to determine the order in which the systems were tested and also served as a reminder to the user as to which system exactly they are evaluating.

The second part of the questionnaire then focused on evaluating the IF, TG and TE modules. The users were asked how interested they were in the general topic of the "ice-breaking" sentence which was presented to them. This evaluated the performance of the IF module. Additionally, this part of the survey contained questions on whether the "ice-breaking" sentence was understood by the users and whether it contained any terms or people which were unfamiliar to them. These questions enabled to evaluate the TG module of the original system and compare it against the random choice made by the baseline system. The final question in this chapter asked the users if the information in the sentence was new to them. This allowed for studying if any correlation between new information and the perceived appeal of a sentence existed and additionally was aimed at verifying the use of the time score in the evaluation scheme.

The next section contained questions which evaluated the remaining modules and the system as a whole. The first question was for the users to evaluate how appealing the generated "ice-breaking" sentence was which was also one of the main goals of this study. Then, users were asked how natural or artificial the generated sentence seemed to them and could optionally describe what they found artificial about the sentence. This question was aimed at evaluating the performance of the NLG module used by both systems.

The next question asked the users whether they agreed with the statement that they felt their privacy threatened by the system. The last question of this section asked if they would continue the conversation with the system based on the "ice-breaking" sentence. If they answered with ”Yes", they could optionally answer if they were going to change the topic or continue on the same topic and if they answered with ”No", they could additionally enter reasons why they did not want to reply to the system. Both these questions aimed at fulfilling the goals defined in Section 4.1.

The last section of the questionnaire featured a comments sections which was optional. This section also asked the users what they think about the idea in general and what problems and advantages they saw in human-machine communication. As a help to the user, the questionnaire provided some example questions to think on like if the user would be willing to share information with such a system if it could be their friend (trust) or what part of the implementation did they like the least.

The ”dummy” feature and website for the original system, the baseline system and its website and the questionnaire formed the three main components of the user study. These were designed as to fulfill the evaluations goals defined in the previous section.

4.3. Implementation

The websites for the original and baseline systems were both designed and implemented using the Django framework[1]. They largely reused the design of the already existing social dialog website of the KIT[2] and featured only minor changes like different heading and first utterance by the system. Both websites[34] were hosted by the KIT and additionally needed to use the systems which were written in Java. The gap between the Python-based Django framework and the Java systems was bridged using the ZeroC Ice middleware[5] which provided for smooth communication between the website and the systems.

The ”dummy” feature of the original system was implemented as a text-based dialog where the user is first asked to choose a username. Then, the user has to choose users to follow from four popular Twitter users. These users are from different categories and the target user has to enter the number of the users in order to follow them. If the target user rejects two users from the same category consecutively one after the other, the system closes this category and chooses a new one. Once a category is closed, accounts from this category are never again proposed to the target user. An example dialog can be seen in Figure 4.1.

Once the study participant responds to the system, it asks them again to choose from another four users and this continues until they follow at least 15 users. The choice of the users to present follows the same logic as described in the previous paragraph. The number of users to follow is based on the consideration that Twitter proposes new users to follow 20 accounts. The system asks for only 15 users, because the websites lack the graphical user interface of Twitter employed to follow users and its extensive database, which make following a user on the survey’s websites much more time-consuming. After the user has reached 15 followers, their account is initialized and the Twitter users they have chosen are used just as those belonging to a real Twitter account in order to generate the “ice-breaking” sentence. Once this has been accomplished, the ”ice-breaking” sentence is sent to the website via ZeroC Ice and is displayed to the user together with a message to switch to the questionnaire and rate the system.

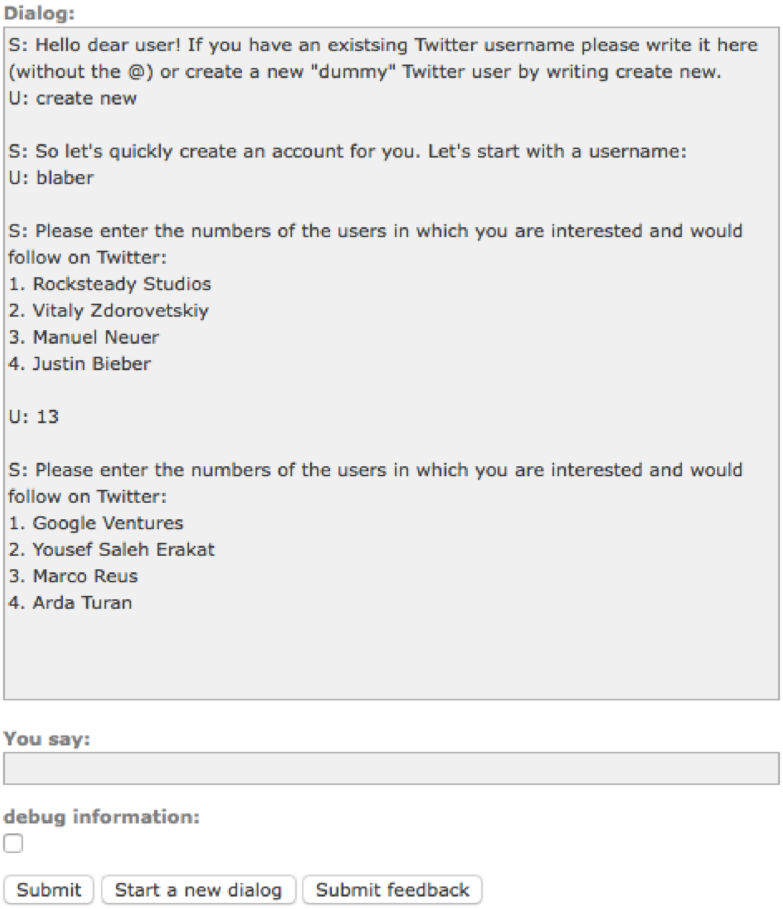
[1] Django framework - https://www.djangoproject.com

[2] Social dialog website at the KIT - http://socialdialog.ira.uka.de

[3] Baseline system’s website - http://socialdialog.ira.uka.de/icebreaker/systemX/

[4] Original systems’s website - http://socialdialog.ira.uka.de/icebreaker/

[5] ZeroC website - https://zeroc.com/index.html



Dialog:

S: Hello dear user! If you have an existsing Twitter username please write it here (without the @) or create a new "dummy" Twitter user by writing create new.
U: create new

S: So let's quickly create an account for you. Let's start with a username:
U: blaber

S: Please enter the numbers of the users in which you are interested and would follow on Twitter:
1. Rocksteady Studios
2. Vitaly Zdorovetskiy
3. Manuel Neuer
4. Justin Bieber

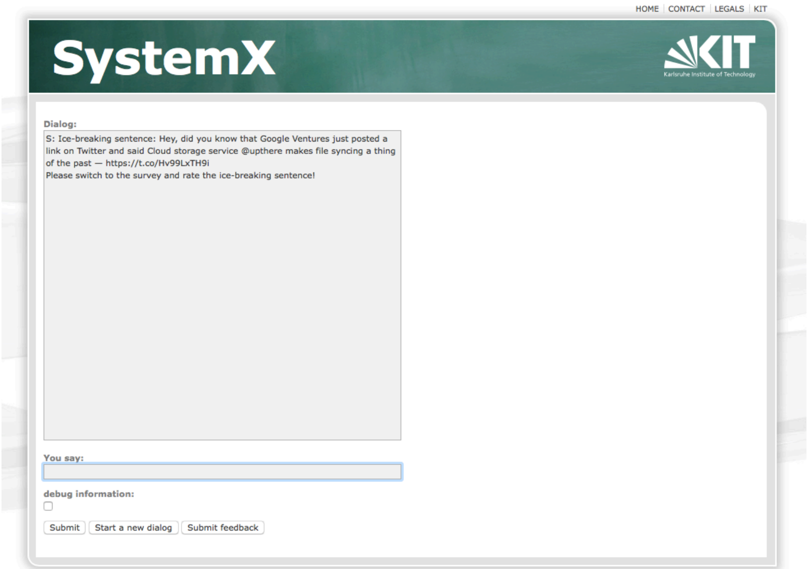
U: 13

S: Please enter the numbers of the users in which you are interested and would follow on Twitter:
1. Google Ventures
2. Yousef Saleh Erakat
3. Marco Reus
4. Arda Turan

You say:

debug information:
□

Figure 4.1.: The text box embedded within the website featuring an example dialog in order to create a ”dummy” account. The user wrote their utterances in the text box and could hit enter or the submit button to send them to the system. The response was then presented in the text box as illustrated.


HOME | CONTACT | LEGALS | KIT
SystemX
KIT
Karlsruhe Institute of Technology

Dialog:

S: Ice-breaking sentence: Hey, did you know that Google Ventures just posted a link on Twitter and said Cloud storage service @upthere makes file syncing a thing of the past — https://t.co/Hv99LxTH9i
Please switch to the survey and rate the ice-breaking sentence!

You say:

debug information:
□

Submit
Start a new dialog
Submit feedback

Figure 4.2.: The website where the baseline system (SystemX) was embedded. Users were presented with an ”ice-breaking” sentence when the website was opened. The orginal (Icebreaker) system featured the same design with a different title and dialog logic.

In order to present accounts from different interest domains to the target user, the study needed a database of such accounts with corresponding category relationships. For this reason such a database was created automatically before the start of the study. This was done by saving the categories which were used to create an “ice-breaking” sentence and the accounts belonging to these categories for a number of manually picked users. Through the manual choice of these users it was ensured that these users were interested in different domains (e.g., cars and racing, news and news websites, stand-up comedy).

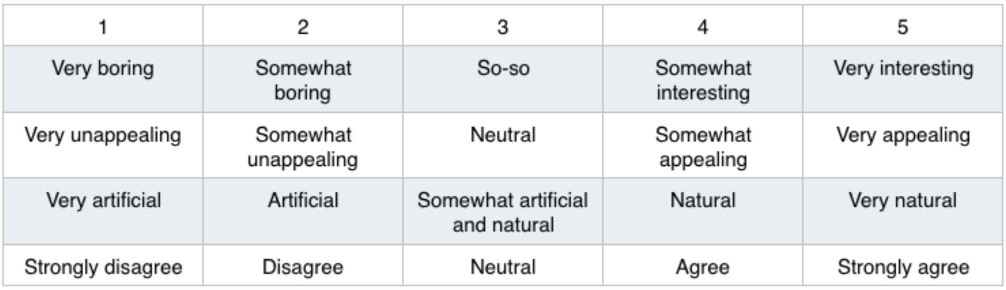
The baseline system also makes use of this database by choosing a random interest and then a random user associated with this interest. Then it selects a random tweet from the last *TWEETS_PRO_USER* tweets which this user has posted and uses the NLG module of the original system in order to generate an “ice-breaking” sentence from this tweet. The sentence is then sent to the website of the baseline system and displayed to the user. After this the user is urged to switch to the questionnaire and rate the system. Figure 4.2 illustrates the design of the website where the interaction with the baseline system took place. Section A.1.2 in the Appendix shows and explains part of the code used for the baseline system.

The questionnaire was created using LimeSurvey[6] and was accessible for everybody on the Internet. The questions in the survey had different structure with the most popular being Yes/No questions and questions using the 5-point Likert scale. Each Likert-scale question featured its own label set since the goals of the survey measured different properties of the system. The full survey can be found in the Appendix A.2.

[6] LimeSurvey website - https://www.limesurvey.org/en/

4.4. Results

The results presented in this section on questions using the Likert scale are analyzed by assigning numerical values to the scale answers. Negative answers are thereby assigned smaller numbers. The exact assignment can be found in Figure 4.3.



1	2	3	4	5
Very boring	Somewhat boring	So-so	Somewhat interesting	Very interesting
Very unappealing	Somewhat unappealing	Neutral	Somewhat appealing	Very appealing
Very artificial	Artificial	Somewhat artificial and natural	Natural	Very natural
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Figure 4.3.: Rules used for assigning numerical values to answers

The gender of the participants was very evenly split with six female and five male participants. The average age of the participants was 25 years and the median was 24. The age distribution was concentrated in the range between 21 and 27 years with only one participant out this age range.

Nine participants tested both systems while the other two each tested one system with the first one testing the baseline and the other one testing the original system. Five of the participants who tested both systems tested the original system first and then the baseline with the remaining four testing in the opposite order.

Out of the 10 participants who tested the original system only one tested it with a personal Twitter account. This test however was unsuccessful, because the Twitter account was private and required additional permissions from the owner.

The participants rated the baseline system in terms of appeal of the generated “icebreaking” sentence on average with a score of 3.3 (using the assignments defined in Figure 4.3). The same score for the original system was 3.8. The distribution of the answers for the two systems is presented in Figure 4.4.

All users who tested the original system declared that they would continue the conversation with the system. Out of these 70% (7) reported that they would continue on the same topic, whereas 30% (3) reported that they would change it. From the users testing the baseline system 20% (2) would end the conversation with the system. They reported that they either had no idea how to continue the conversation or that they did not know the person about which the tweet was. From the people who would continue the conversation, 50% (4) would change the topic and 50% (4) would continue on the same topic. These results are also presented graphically in Figure 4.6.

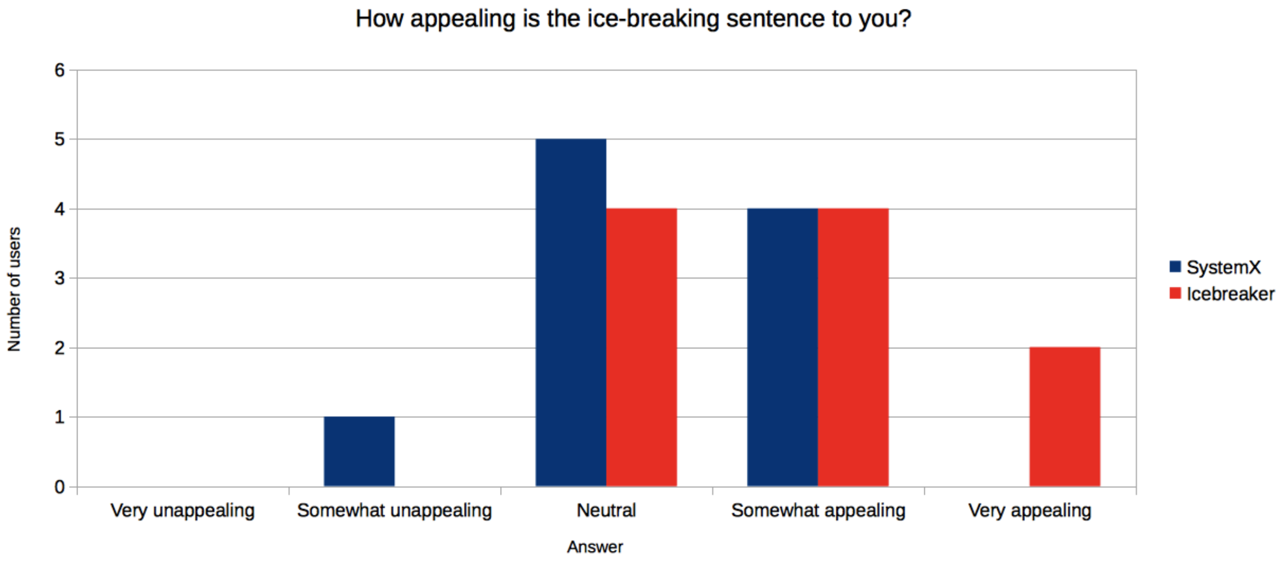


Figure 4.4.: Distribution of the answers of the question ”How appealing is the ice-breaking sentence to you?”

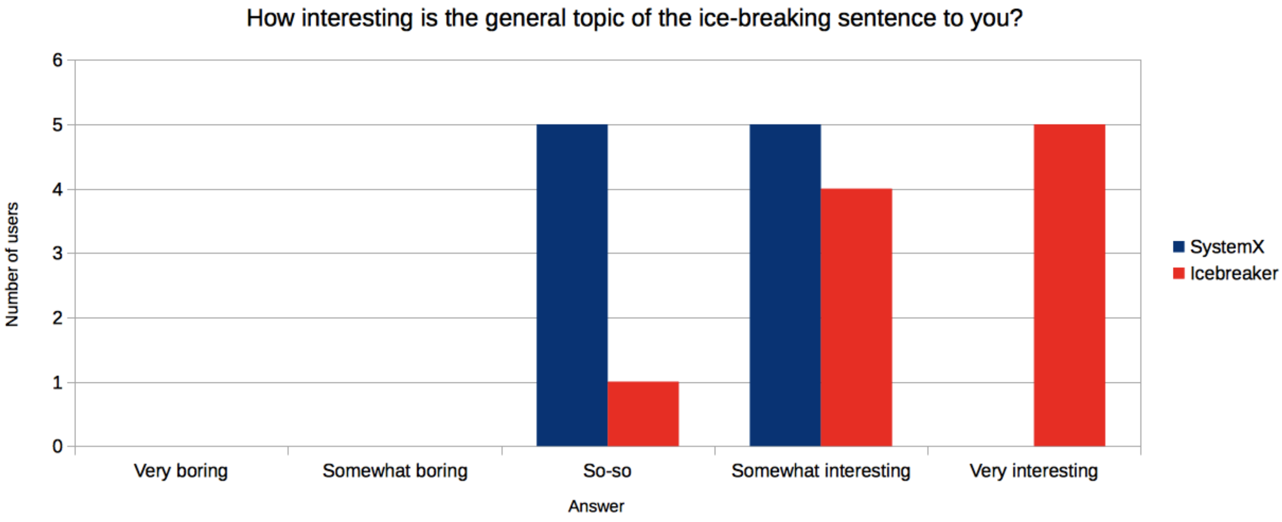


Figure 4.5.: Distribution of the answers of the question ”How interesting is the general topic of the ice-breaking sentence to you?”

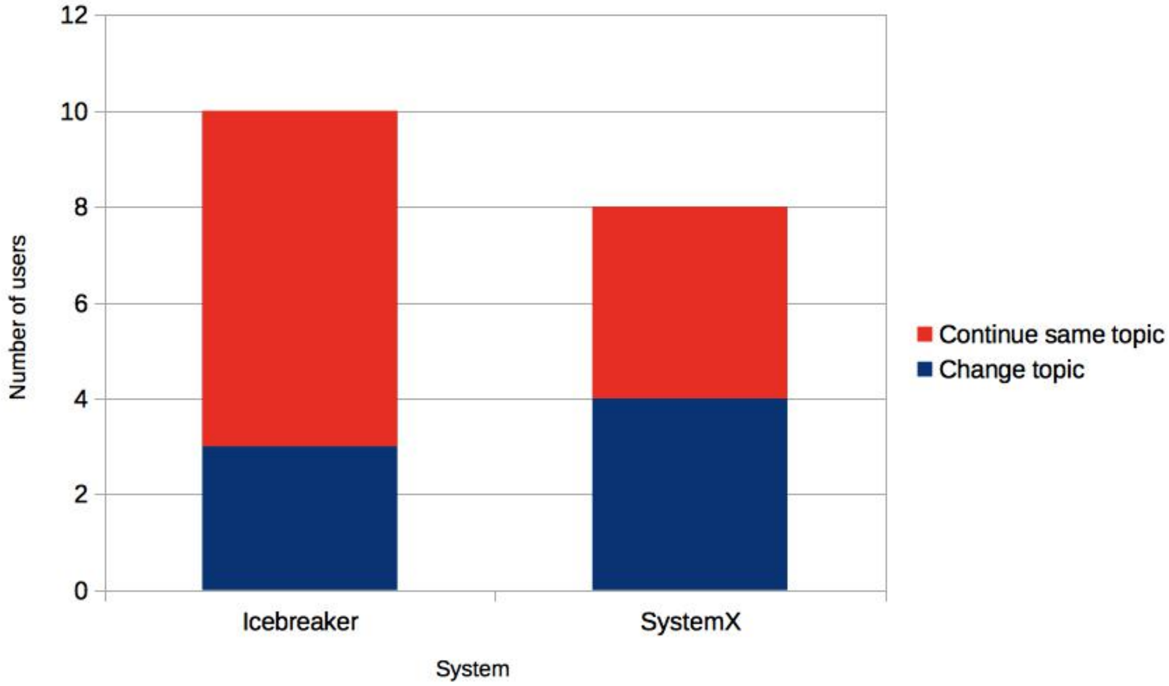


Figure 4.6.: Distribution of the answers on the question if users would continue on the same topic or change it. This question was asked to users only if they declared that they would respond to the system.

The participants who tested the baseline system rated their interest in the general topic of the "ice-breaking" sentence at 3.5 (the conversion to numerical values can be found in Figure 4.3). For the original system this value was 4.4. The exact distribution of the answers can be seen in figure 4.5.

Six participants who tested the original system reported that the generated "ice-breaking" sentence did not contain any terms or people they are unfamiliar with while the other four reported the opposite. The evaluation of the same question for the baseline system produced different results. Seven participants reported terms and people unknown to them and three reported the opposite.

All participants who tested the original system reported that they understood the "ice-breaking" sentence. This was however not the case for the participants testing the baseline system with two reporting not to have understood the question and the remaining eight reporting the opposite. These two participants also reported that the sentence did contain people or terms they are not familiar with. Additionally, all participants who tested any of the two systems reported that the information from the "ice-breaking" sentence is new to them.

Since the same NLG module was used for both the baseline and the original system, we combine the results for both. The average result which the participants gave to the “ice-breaking” sentence in terms of naturalness is 3.5 (using the assignments defined in Figure 4.3). The mode and median of the answers were both the answer ”Natural”. The reasons why the sentences were not seen as natural were different. The most common one was that the rule-based approach for the NLG did not work for every kind of tweet and that the tweets contained a big number of grammatical errors which were transferred to the ”ice-breaking” sentence.

The majority of the testers of both systems disagreed with the statement that they felt their security threatened by the system. The respective average scores were 1.8 for the baseline system and 1.9 for the original system which also was most closely to the answer “Disagree” (using the assignments defined in Figure 4.3). A detailed view of the distribution of the answers can be found in Figure 4.7.

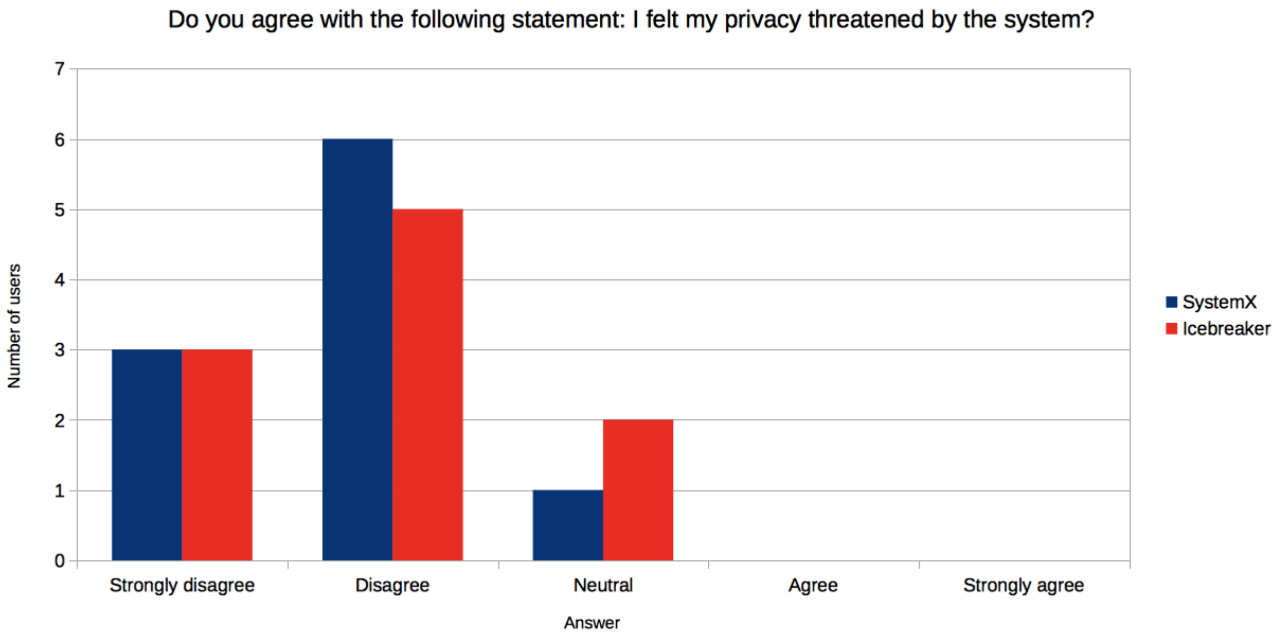
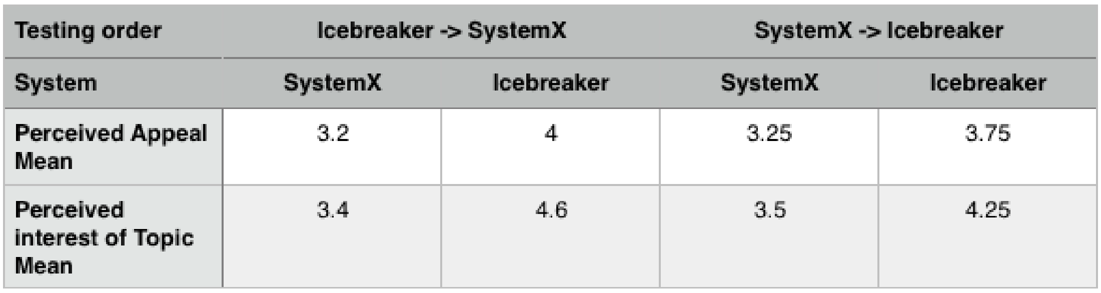


Figure 4.7.: Distribution of the answers of the question ”Do you agree with the following statement: I felt my privacy threatened by the system?”.

4.5. Discussion

The order of testing the baseline (SystemX) and the original (Icebreaker) system did not seem to have a significant effect on the answers of the users. Figure 4.8 presents the average results of the appeal of the system and how interesting the general topic of the generated sentence was. Small tendencies can be observed in this table like that the “ice-breaking” sentences generated by the baseline system were rated higher in terms of appeal when this system was tested first. These tendencies could however not be proven as statistically significant. Therefore in the rest of the evaluation the order in which the systems were tested is ignored and the two groups are mixed together. Any learning effects should at least be partially counteracted by the different order of testing.



Testing order	Icebreaker -> SystemX		SystemX -> Icebreaker	
System	SystemX	Icebreaker	SystemX	Icebreaker
Perceived Appeal Mean	3.2	4	3.25	3.75
Perceived interest of Topic Mean	3.4	4.6	3.5	4.25

Figure 4.8.: Mean scores for the perceived appeal and interest in the general topic of the sentence depending on the order in which the two systems were tested.

Furthermore, the analysis conducted in in this section should to always be interpreted cautiously since most of the study participants were computer science students or with university background in computer science. Moreover, a large number of the participants were personally contacted and urged to take part in the study. Since some of them were personally acquainted with the author of this thesis, they also had some background knowledge on the topic of the work and their personal relations may have influenced them to rate the system higher (although it was explicitly asked to avoid this).

The participants rated both the baseline and the original system relatively high in terms of the appeal of the "ice-breaking" sentence. There was only one user who rated one of the systems as "somewhat unappealing" and none who rated them as "very unappealing" (See Figure 4.4). The average score for the baseline system was 3.3 and for the original system 3.8.

In order to compare these two means a paired t-test was performed on the data from the users who tested both systems (nine in total). The conversion to numerical values is defined in Figure 4.3. The results confirmed that a statistically significant difference is existent in the data with a two-tailed P value of 0.0232. The 95% confidence interval for the difference between the original and baseline systems was from 0.14 to 1.42. This suggests that the original system would be on average more appealing than the baseline system.

The appeal of a sentence could be influenced by a number of factors. One possible correlation hinted by the data is that the lack of terms and people unfamiliar to the user, makes an "ice-breaking" sentence more appealing to the user. This could be the case since the user would possibly understand the sentence better and would have enough background information in order to form an interesting reply. The average appeal score given to an "ice-breaking" sentence which did not contain any unfamiliar terms or people was 4

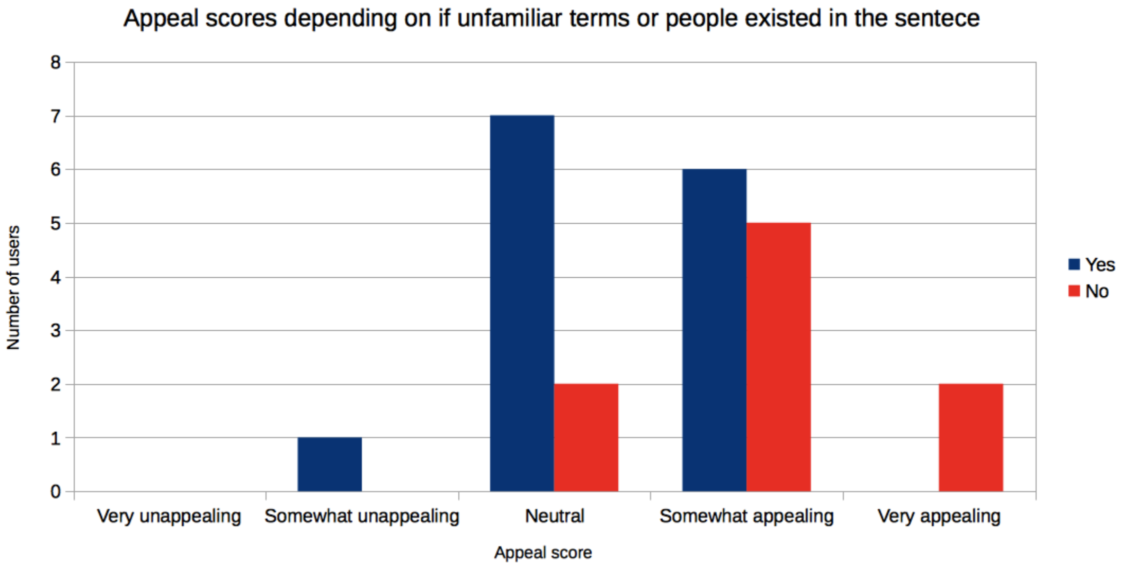


Figure 4.9.: Appeal scores depending on if unfamiliar terms or people existed n the sentence

(Somewhat appealing). Sentences with unfamiliar terms or people were rated on average with around 3.36 (more near to neutral). An unpaired t-test on the underlying data for these means (for exact distribution see Figure 4.9) yields a two-tailed *P* value of 0.0338 which implies that the difference between the means is statistically significant. The 95% confidence interval for the difference between the mean of the “Yes”-group and the mean of the “No”-group was from -1.23 to -0.05. This correlation could also explain the difference in the perceived appeal of the sentences generated by the original and the baseline systems since the sentences from the original system contained less terms or people that are unfamiliar to the target user.

The data also suggests that a significant factor to not understanding the “ice-breaking” sentence is the existence of unfamiliar terms or people. In both cases where a sentence was not understood, users have reported that the sentence contained such terms or people. Both occasions on which the sentence was not understood occurred while testing SystemX. The users rated the appeal of the sentence as “Neutral” which is below the mean score for SystemX.

The “ice-breaking” sentences generated by the original system have also largely fulfilled their goal to engage the user in a conversation since all users using the original system reported that they would continue the conversation. The data also suggests a correlation between the perceived appeal and the willingness to change the topic. Users who wished to change the subject rated the appeal of the system on average with 3.29 (more near ”Neutral”) while the ones who wished to remain at the topic rated it with 3.81 (more near ”Somewhat appealing”). The difference in the data however was not sufficient to show that this difference was statistically significant. An unpaired t-test yielded a two-tailed *P* value of 0.1628 and a 95% confidence interval from -1.30 to 0.24.

The data does however give a clear signal of a correlation between the decision to change or remain at a topic and the perceived interest of the topic. Users who decided to change the topic rated the general topic of the “ice-breaking” sentence with 3.57 (between ”So-so” and ”Somewhat interesting”). Users who remained at the same topic rated it with 4.36 (between ”Somewhat interesting” and ”Very interesting”). An unpaired t-test yielded a two-tailed *P* value of 0.0186 which shows that the difference in the means is statistically significant. The 95% confidence interval for this difference was from -1.43 and -0.15. This shows that when the topic of the conversation is changed, the topic of the ”ice-breaking” sentence is perceived as less interesting.

Another observation derived from the data is that the users who tested the baseline system reported that they are less interested in the general topic of the sentence than the ones who have tested the original system. The average score of the interest in the general topic of the sentences generated by the baseline was 3.5 (between “So-so” and “Somewhat interesting”). The same score was 4.4 (between “Somewhat interesting” and “Very interesting”) for sentences generated by the original system. The difference between these two means is nearly one point and was proven as statistically significant by an unpaired t-test on the data. The two-tailed *P* value was 0.0044 and the 95% confidence interval was from -1.48 to -0.32 which points to a substantial difference between the means. Additionally, a paired t-test was performed on data from the users who have rated both systems. The test furthermore showed that the difference was significant and yielded a two-tailed *P* value of 0.0278.

These results can be interpreted as an indication that the IF module chooses the interests significantly better than the baseline which chooses random ones. However, it is to be noted that the baseline chooses the interests randomly only from a subset of all possible interests (as explained in section 4.2). This could of course influence the result. In this study however all users except for one used the ”dummy” feature of the original system. Since the ”dummy” feature and the baseline system share the same database, the users of the original system and the baseline chose from the same pool of users and interests. The influence of not having all possible Twitter users is therefore very insubstantial when interpreting the results.

The perceived interest in the topic of the “ice-breaking” sentence seems to correlate with the perceived appeal. In 12 of all of the 20 questionnaire evaluations the interest in the topic and the appeal of a sentence are evaluated with the same value (after conversion to numerical values illustrated in Figure 4.3). The rest differ with the interest value always being one level higher than the perceived appeal of the sentence. This correlation could also further explain the difference in the perceived appeal of the baseline and the original system. The substantially higher values of the perceived interest in the topic of the sentences generated by the original system could be the reason for the perceived higher appeal of these sentences. A more detailed analysis of this trend would however require more data than was available from the survey.

The mean score for the naturalness of the sentences was 3.5 which lays between “Somewhat natural and artificial” and “Natural”. These results reflect the downsides of the use of a rule-based NLG module for this task. Since the mode and median were “Natural”, the lower mean score indicates that the generator failed at some inputs. The reason for this lays in the variance of the grammatical construction of the tweets used to form the sentences and sometimes in their grammatical incorrectness (as reported by study participants). There seems to be a small correlation between the naturalness of the sentences and their perceived appeal. However, because of the small number of participants which rated the sentences as artificial (only two), further analysis is not performed. Both of these participant have however rated the perceived appeal with 3 (Neutral) which is lower than the means for both systems.

The data also shows that the perceived threat to privacy was rather low. As can be seen in the figure on privacy (Figure 4.7), the answers for both systems barely had different distributions. This is rather surprising since the baseline system did not at all use the personal information of the participants, whereas the original system did. This trend could lead to the conclusion that the users have a personal view of their privacy limits and judge systems based on the threat they perceive based on these views rather than on facts.

In the comments sections the situation on privacy was very mixed with some users saying they would not want to share personal information with a system even if it could be their friend. This was also connected with the issue of trusting such a system as it could misuse personal data. A number of users saw application for an extended version of the proposed system as an artificial friend in the medical domain or in lonely times. Another idea was that the system described in this work could improve the initiative of existing chatbots like Cleverbot.

4.6. Summary

The study shows that the system implemented in the scope of this work does generate “ice-breaking” sentences which are perceived as more appealing than those created by the baseline system. Furthermore, the general topic of a sentence generated by the original system was perceived as more interesting than the topic of one generated by the baseline. The system has also succeeded in engaging the user in a conversation with all users wanting to respond and a large number of them wanting to continue on the same topic. Thus, the “ice-breaking” function of the original system does seem to function better than the baseline.

It is also to be noted that a more sophisticated NLG module than the rule-based one used in this work may be needed since it failed to produce natural sentences on a couple of occasions during the study. Privacy remains a topic of mixed feelings with most users not seeing the system implemented in this work as threatening. However, their views were very mixed when it comes to sharing more information.

5. Conclusion

This work proposes a system which generates sentences targeted at a specific user by utilizing information from their public Twitter profile. These sentences are suitable for the beginning of a conversations and are thus called “ice-breaking”. They should appeal to the user and increase the probability of a response. The notion of creating such sentences is mainly related to the domains of sociology, recommendation systems and social dialog system.

The system described in this work was evaluated through a user study where participants could test the system on an Internet website and then fill out a questionnaire. Additionally, they tested a baseline system which chooses topics and tweets randomly. The proposed system did outperform the baseline in generating sentences which were perceived as more appealing and additionally at choosing topics which interested the target users more. The perceived appeal of the generated sentences was rated highly as was the general interest in the topic of the sentences.

However, we note that the NLG component of the system (RBSG) could perform better since users, on average, perceived the sentences as somewhat artificial. This issue could be faced by either further extending the rule-based system with additional rules or by deploying a statistical model which could bring more diversity and flexibility in the sentence generation.

A more general goal for future work would be to extend the current system into a real dialog system. A system like this could utilize the information from previous dialogs and, when needed, more information about the user found in social networks like location or personal posts. This could, for example, be used when the topic of the conversation needs to be changed, because the engagement between the user and the system is lower than a certain threshold.

It is additionally of interest to analyze when an ”ice-breaking” sentence was successful and strategies to either change the topic gracefully or continue on the same topic by taking into account the context and the content of previous utterances and available social networks’ data. Such systems may rely on statistical models and machine learning techniques such as neural networks.

Finally, we note that we see further possibilities in utilizing data from social networks in the domain of dialog systems and artificial intelligence in general. Such information could, for example, be used by social dialog systems in a number of situations since it often features in small talk conversations. Apart from that, it is useful for personalizing existing goal-oriented dialog systems and making them more appealing to the user. Dialog systems can also adapt to the user even before they had a conversation with them by analyzing the data about this user present in social networks. We believe that these and other innovations would be a further step in the evolution in the ways we communicate.

Bibliography

[1] Eytan Bakshy et al. “The role of social networks in information diffusion”. In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 519–528.

[2] Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. “Dialog system using real-time crowdsourcing and Twitter large-scale corpus”. In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics. 2012, pp. 227–231.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. “Twitter mood predicts the stock market”. In: *Journal of Computational Science* 2.1 (2011), pp. 1–8.

[4] Christy MK Cheung, Pui-Yee Chiu, and Matthew KO Lee. “Online social networks: why do students use Facebook?” In: *Computers in Human Behavior* 27.4 (2011), pp. 1337–1343.

[5] James Davidson et al. “The YouTube video recommendation system”. In: *Proceedings of the fourth ACM conference on Recommender systems*. ACM. 2010, pp. 293–296.

[6] David DeVault et al. “SimSensei Kiosk: A virtual human interviewer for health-care decision support”. In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2014, pp. 1061–1068.

[7] Adrien Guille et al. “Information diffusion in online social networks: A survey”. In: *ACM SIGMOD Record* 42.2 (2013), pp. 17–28.

[8] Kate S Hone and Robert Graham. “Towards a tool for the subjective assessment of speech system interfaces (SASSI)”. In: *Natural Language Engineering* 6.3&4 (2000), pp. 287–303.

[9] Philip N Howard et al. “Opening closed regimes: what was the role of social media during the Arab Spring?” In: *Available at SSRN 2595096* (2011).

[10] Andrew J Hunt and Alan W Black. “Unit selection in a concatenative speech synthesis system using a large speech database”. In: *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. Vol. 1. IEEE. 1996, pp. 373–376.

[11] Akshay Java et al. “Why we twitter: understanding microblogging usage and communities”. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM. 2007, pp. 56–65.

[12] Maksim Kitsak et al. “Identification of influential spreaders in complex networks”. In: *Nature Physics* 6.11 (2010), pp. 888–893.

[13] Haewoon Kwak et al. “What is Twitter, a Social Network or a News Media?” In: *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10. Raleigh, North Carolina, USA: ACM, 2010, pp. 591–600. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772751. URL: http://doi.acm.org/10.1145/1772690.1772751.

[14] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. “Social selection and peer influence in an online social network”. In: *Proceedings of the National Academy of Sciences* 109.1 (2012), pp. 68–72.

[15] P Alex Linley, Stephen Joseph, and Martin EP Seligman. *Positive psychology in practice*. Wiley Online Library, 2004.

[16] Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a Feather: Homophily in Social Networks”. In: *Annual Review of Sociology* 27.1 (2001), pp. 415–444. DOI: 10.1146/annurev.soc.27.1.415. URL: http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.415.

[17] Jacob Levy Moreno. “Who shall survive?: A new approach to the problem of human interrelations.” In: (1934).

[18] Brendan O’Connor et al. “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.” In: *ICWSM* 11.122-129 (2010), pp. 1–2.

[19] Antoine Raux et al. “Doing research on a deployed spoken dialogue system: one year of let’s go! experience.” In: *INTERSPEECH*. 2006.

[20] Antoine Raux et al. “Let’s go public! taking a spoken dialog system to the real world”. In: *in Proc. of Interspeech 2005*. Citeseer. 2005.

[21] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[22] Jeffrey A Smith, Miller McPherson, and Lynn Smith-Lovin. “Social Distance in the United States Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004”. In: *American Sociological Review* 79.3 (2014), pp. 432–456.

[23] Oriol Vinyals and Quoc Le. “A neural conversational model”. In: *arXiv preprint arXiv:1506.05869* (2015).

[24] Joseph Weizenbaum. “ELIZA—a computer program for the study of natural language communication between man and machine”. In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[25] Jianshu Weng et al. “TwitterRank: Finding Topic-sensitive Influential Twitterers”. In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. WSDM ’10. New York, New York, USA: ACM, 2010, pp. 261–270. ISBN: 978-1-60558-889-6. DOI: 10.1145/1718487.1718520. URL: http://doi.acm.org/10.1145/1718487.1718520.

[26] Steve Young et al. “The hidden information state model: A practical framework for POMDP-based spoken dialogue management”. In: *Computer Speech & Language* 24.2 (2010), pp. 150–174.

[27] Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. “TickTock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness”. In: *2015 AAAI Spring Symposium Series*. 2015.

A. Appendix

A.1. Code examples

A.1.1. Basic Icebreaker

The following code example shows how an *Icebreaker* object which generates ”ice-breaking” sentences is built. The class utilizes six modules as described in Chapter 3. These use an underlying scheme which defines how these are connected. The one used in this code example, *IceBreakerSchemeExtended*, ensures that the modules are connected as illustrated in Figure 3.1.

```
package dialnet.mercury.tweetFilter.basic;

import java.io.IOException;

import oauth.signpost.exception.OAuthCommunicationException;
import oauth.signpost.exception.
    OAuthExpectationFailedException;
import oauth.signpost.exception.OAuthMessageSignerException;

import org.apache.http.client.ClientProtocolException;

import dialnet.mercury.data.TwitterComExceptionHttp;
import dialnet.mercury.data.
    TwitterComExceptionWithNoTwitterCode;
import dialnet.mercury.data.
    TwitterComExceptionWithTwitterCode;
import dialnet.mercury.tweetFilter.Database;
import dialnet.mercury.tweetFilter.Evaluator;
import dialnet.mercury.tweetFilter.Gatherer;
import dialnet.mercury.tweetFilter.IceBreaker;
import dialnet.mercury.tweetFilter.IceBreakerSchemeExtended;
import dialnet.mercury.tweetFilter.InterestFilter;
import dialnet.mercury.tweetFilter.SentenceChooser;
import dialnet.mercury.tweetFilter.SentenceGenerator;
import dialnet.mercury.tweetFilter.TweetChooser;

public class BasicIceBreaker implements IceBreaker{
        private InterestFilter iF;
        private Gatherer tG;
        private Evaluator tE;
        private TweetChooser tC;
        private SentenceGenerator sG;
        private SentenceChooser sC;
        IceBreakerSchemeExtended scheme;

        public BasicIceBreaker() {
                iF = new BasicInterestFilter();
                tG = new BasicTweetGatherer();
                tE = new BasicTweetEvaluator();
                tC = new TopXTweetChooser(20);
                sG = new RuleBasedSentenceGenerator();
                sC = new FirstChoice();
                scheme = new IceBreakerSchemeExtended(iF, tG
                    , tE, tC, sG, sC);
        }

        @Override
        public String generateIceBreaker(String username,
            String lang) throws OAuthMessageSignerException,
            OAuthExpectationFailedException,
            OAuthCommunicationException,
            ClientProtocolException,
            TwitterComExceptionWithTwitterCode,
            TwitterComExceptionWithNoTwitterCode,
            TwitterComExceptionHttp, IOException {
                return scheme.generateIceBreaker(username,
                    lang);
        }

        public Database getDB() {
                return scheme.getDB();
        }
}
```

A.1.2. Baseline Icebreaker

The following code example describes the *Icebreaker* object used by the baseline system. The *generateIceBreaker* method no longer uses an underlying scheme since this type of Icebreaker only needs a *SentenceGenerator* (NLG module). Interests and users are loaded from the database used by the ”dummy” system (*UserInit*). Random interests and a user assigned to this interest are chosen. The Twitter API is then called and a list of *TWEETS_PRO_USER* tweets is obtained. A random tweet from this list is chosen and transformed into an ”ice-breaking” sentence by the *SentenceGenerator* object (NLG module). This sentence is returned by the system with a message to switch to the survey and rate the system.

```java
package dialnet.mercury.tweetFilter.basic;

import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import java.util.Random;

import oauth.signpost.exception.OAuthCommunicationException;
import oauth.signpost.exception.
    OAuthExpectationFailedException;
import oauth.signpost.exception.OAuthMessageSignerException;

import org.apache.http.client.ClientProtocolException;

import dialnet.mercury.communication.TwitterCom;
import dialnet.mercury.data.Tweet;
import dialnet.mercury.data.TwitterComExceptionHttp;
import dialnet.mercury.data.
    TwitterComExceptionWithNoTwitterCode;
import dialnet.mercury.data.
    TwitterComExceptionWithTwitterCode;
import dialnet.mercury.data.TwitterUser;
import dialnet.mercury.tweetFilter.IceBreaker;
import dialnet.mercury.tweetFilter.SentenceGenerator;
import dialnet.mercury.user.UserInit;

public class BaselineIceBreaker implements IceBreaker {
        private SentenceGenerator sg;
        private List<String> interests;

        public BaselineIceBreaker() {
                sg = new RuleBasedSentenceGenerator();
                interests = UserInit.loadInterests();
        }

        @Override
        public String generateIceBreaker(String username,
            String lang) throws OAuthMessageSignerException,
            OAuthExpectationFailedException,
            OAuthCommunicationException,
            ClientProtocolException,
            TwitterComExceptionWithTwitterCode,
            TwitterComExceptionWithNoTwitterCode,
            TwitterComExceptionHttp, IOException
                            {
                Random rand = new Random();
                String interest = interests.get(rand.nextInt
                    (interests.size()));
                ArrayList<TwitterUser> users = UserInit.
                    loadUsers(interest);
                TwitterUser user = users.get(rand.nextInt(
                    users.size()));
                Tweet[] tweets = TwitterCom.getTweets(user.
                    getScreen_name(), BasicControlObject.
                    getInstance().getTWEETS_PRO_USER());
                Tweet tweet = tweets[rand.nextInt(tweets.
                    length)];
                tweet.setCreator(user);
                String sentence = sg.generateSentence(tweet)
                    ;
                String output = "Ice-breaking␣sentence:␣" +
                    sentence;
                output = output + "\nPlease␣switch␣to␣the␣
                    survey␣and␣rate␣the␣ice-breaking␣sentence
                    !";
                return output;
        }

}
```

A.2. Survey questions

The following pages present all questions and possible answers in the survey questionnaire. Conditions on which answers were displayed are presented as well.

Twitter-based Ice-breaker System Evaluation

Imagine you are talking to a robot or a person you have just met. In many cases the person will try to "break the ice" and try to find common topics with you. Automated systems however often are limited in this domain to just greeting you. Therefore the system you just tested tries to behave more naturally and create "ice-breaking" sentences which should capture your attention and appeal to you.

Dear Visitor,

we are extremely glad to welcome you to this survey! We need your help in evaluating a new module for our social dialog system. Given a Twitter username the new module generates sentences which appeal to the given Twitter user and are suitable for starting a conversation (to break the ice, thus also called ice-breaker sentences).

Thank you for your time and commitment!

Aleksandar

There are 19 questions in this survey

Demographics and general questions

[]Gender *

Please choose **only one** of the following:

○ Female

○ Male

[]How old are you? *

Only an integer value may be entered in this field.

Please write your answer here:

□

[]Which system did you just test? *

Please choose **only one** of the following:

○ SystemX

○ Icebreaker

You can check the title on the page of the system which states the name of the system. If you received an email with instructions, it should additionally say which link corresponds to which system.

[]Did you use your Twitter account in the process? *

Only answer this question if the following conditions are met:
Answer was 'Icebreaker ' at question '3 [DEMO3]' (Which system did you just test?)

Please choose **only one** of the following:

○ Yes

○ No

[]How frequently do you use Twitter? *

Only answer this question if the following conditions are met:
Answer was 'Yes' at question '4 [DEMO31]' (Did you use your Twitter account in the process?)

Please choose **only one** of the following:

○ On a daily basis

○ On a weekly basis

○ On a monthly basis

○ I have an account, but don't use it

[]If you are willing to let us further analyze how your account and the answers you gave relate, enter your username below:

Only answer this question if the following conditions are met:
Answer was 'Yes' at question '4 [DEMO31]' (Did you use your Twitter account in the process?)

Please write your answer here:

This only gives us the possibility to look at your publicly visible information.

[]Is the system you selected the first one you tested? *

Please choose **only one** of the following:

○ Yes

○ No

If you have already evaluated one of the two systems, select 'yes'.

Icebreaker general questions

[]How interesting is the general topic of the ice-breaking sentence to you? *

Please choose the appropriate response for each item:

Very boring	Somewhat boring	So-so	Somewhat interesting	Very interesting
○	○	○	○	○

For example, if you are interested in the automobile industry, you would probably rate something like "Autobild just tweeted that Toyota will recall multiple models due to airbag malfunctions." as very interesting.

[]Did the sentence contain any terms or people you are not familiar with? *

Please choose **only one** of the following:

○ Yes

○ No

[]Did you understand the sentence? *

Please choose **only one** of the following:

○ Yes

○ No

If the sentence made sense to you, please answer with yes.

[]Was the information in the sentence new to you? *

Please choose **only one** of the following:

○ Yes

○ No

Icebreaker questions

[]How appealing is the "ice-breaking" sentence to you? *

Please choose the appropriate response for each item:

Very unappealing	Somewhat unappealing	Neutral	Somewhat appealing	Very appealing
○	○	○	○	○

A highly appealing sentence is one which would capture your attention and lead to a response on your side.

[]How natural does the "ice-breaking" sentence seem to you? *

Please choose the appropriate response for each item:

Very artificial	Artificial	Somewhat natural and artificial	Natural	Very natural
○	○	○	○	○

Natural sentences occur in everyday conversations between humans. You should be able to think of everyday situations where you can hear such sentences.

[]Why did the sentence seem artificial?

Only answer this question if the following conditions are met:
Answer was 'Somewhat natural and artificial' *or* 'Very artificial' *or* 'Artificial' at question '13 [IB7]' (How natural does the "ice-breaking" sentence seem to you? ( ))

Please write your answer here:

[]Do you agree with the following statement: I felt my privacy threatened by the system. *

Please choose the appropriate response for each item:

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

[]Would you continue the conversation with the system based on the generated "ice-breaking" sentence? *

Please choose **only one** of the following:

○ Yes

○ No

[]Would you continue discussing the same topic as in the "ice-breaking" sentence or would you try to change the topic?

Only answer this question if the following conditions are met:
Answer was 'Yes' at question '16 [IB5]' (Would you continue the conversation with the system based on the generated "ice-breaking" sentence?)

Please choose **only one** of the following:

○ Continue on the same topic

○ Change topic

[]Why would you end the conversation with the system?

Only answer this question if the following conditions are met:
Answer was 'No' at question '16 [IB5]' (Would you continue the conversation with the system based on the generated "ice-breaking" sentence?)

Please write your answer here:

Comments

[]

Here you can leave any further comments and ideas. As a help you can think about the following questions:

- **Did you like the idea?**
- **Were there issues with the implementation which you didn't like at all?**
- **Would you be willing to share information with such a system if it could be your friend?**
- **If possible, would you have a robot as a friend?**

You don't have to answer all of these questions or restrict yourself to them, but every sentence is valuable and appreciated.

Please write your answer here:

List of Figures