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

Aleksandar Andonov, Maria Schmidt, Jan Niehues, Alex Waibel

Karlsruhe Institute of Technology, Karlsruhe Germany
Email: aleksandar.andonov@student.kit.edu
firstname.lastname@kit.edu
Web: www.isl.anthropomatik.kit.edu

Abstract

Many goal-oriented spoken dialog systems lack a social component like small talk which often features in human-human communication. In this work we aim to alleviate part of this problem by generating sentences which have the goal to appeal to the user and increase the probability of a response. Such sentences are suitable to break the ice in the beginning of a conversation and are therefore referred to as "ice-breaking" throughout this paper. Furthermore, we use data from the Twitter account of the user in order to infer the users interests. By generating sentences about these interests we utilize the existence of homophily in social networks. A user study shows that the described system outperforms one which chooses interests at random. Furthermore, we note that 70% of the study participants would answer the system and continue talking on the same topic which was introduced by the generated sentence.

1 Introduction

The continuing development of spoken dialog systems (SDSs) has benefited the industry and society in various ways. Such systems have been deployed in various Human-Robot Interaction tasks,e.g., to give information about bus schedules or act as a therapist and measure stress levels (see [1] and [2]). Dialog systems however still lack the intelligence of real human-human conversations and often fail to grasp the context of the conversation if it was not modeled or trained beforehand. Moreover, a lot of systems lack the usual small talk which is a feature of nearly all human conversations.

Online social networks, on the other hand, 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 feature posts which develop into topics in human-human conversations. Recent statistics show that Facebook alone has around 1,5 billion monthly active users[1], which underlines the importance that such networks have gained in recent years.

These qualities of social networks and the relative ease with which data from social networks can be read and interpreted, make them a suitable source of information for dialog systems. By utilizing the available data in Twitter and applying state-of-the-art techniques to it, as described in Section 2 and 3, we create an innovative way of generating "ice-breaking" sentences for use in a SDS. Such sentences can be used at the beginning of a conversation as means to imitate the usual small talk in human-human conversations and so make existing SDSs more appealing to the user.

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

2 Background and Related Work

Twitter is seen as a combination of a microblogging service, a social network and a news platform [3]. Opposed to other social networks like Facebook, the "following" concept makes relations in Twitter unidirectional. The site features a well-defined markup syntax. User mentions are marked with a "@" and signal a relationship between the tweet and the mentioned user. Reposting a tweet from another user is marked with a preceding "RT" (for retweet) and hashtags which usually make up the keywords in the tweet, by a "#". Additionally, users can "favorite" tweets which they like and retrieve them from the favorites tab in their account page.

Online social networks have allowed researchers to explore homophily, which 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 (see [4] and [5]). Weng et al. also show in [6] that homophily exists 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.

Some users of online social networks are more influential than others and their posts are more likely to get shared by a large number of users as shown in [7]. A number of different algorithms to determine these users are proposed in the scientific literature. The K-core algorithm[8] designates users which are in the core of the network as more influential than others and uses k-decomposition to determine these. TwitterRank [6], dedicated to Twitter, ranks users similarly to PageRank [9], but takes into account the different topics in which the user is interested. Our system uses the In-Degree algorithm [6], which has been deployed by Twitter and others, to measure the influence of users.

Twitter data has been widely used in research, e.g., to measure the sentiments of a population [10] or to predict stock markets [11]. Furthermore, in [12] Bessho et al. propose a dialog system based on Twitter data. 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.

Social dialog systems have been studied in the scientific literature with one of the most prominent early examples being the rule-based ELIZA system [13] proposed by Weizenbaum in 1966. A more recent example is the Tick-Tock conversational agent [14] 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. In this work we aim to extend existing dialog systems by offering a social dialog component,

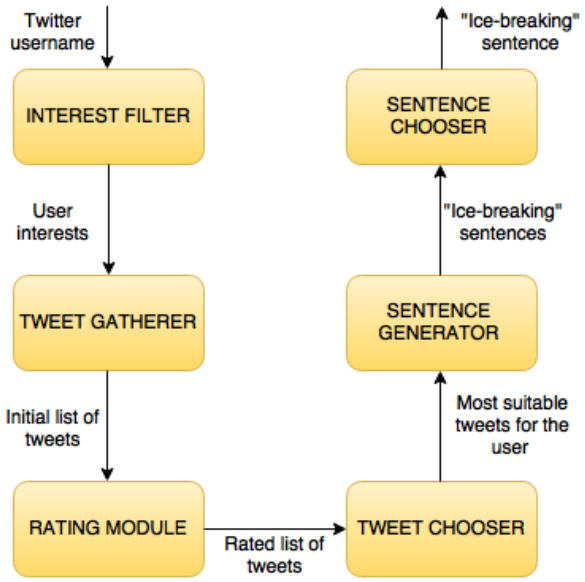

Figure 1: System Overview with the input and output of each system module

which can be used by existing systems to capture the attention of the user and to enter in a more natural conversation which features small talk. Additionally, the described system extends our existing dialog system [15].

"Ice-breaking" sentences can also be used in a dialog between a human and a robot assistant which proactively offers him help. This dialog situation is the topic of the SecondHands project[2] of which this work is part. Since a high degree of human-robot interaction is one of the main goals of the project, this system can be used to achieve a more natural and appealing dialog by generating ice-breaking sentences in the beginning.

3 Generating "Ice-Breaking" Sentences from Tweets

The architecture of the system features a modular design which enables us to easily make changes to the general flow of the system (see Figure 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 (*Interest Filter* module). Thereafter, tweets which are associated with these interests are gathered online and evaluated by the system (*Tweet Gatherer* and *Rating Module*). The *Tweet Chooser* module then selects a group of tweets based on this evaluation. These tweets are given to a Natural Language Generation (NLG) component which transforms them into sentences that fulfill the goals of an "ice-breaking" sentence (*Sentence Generator*). Eventually, the last module chooses the final output of the system from these sentences (*Sentence Chooser*).

3.1 Infering the User Interests

In order to deduce the interests of the target user, we analyze the list of users which the client (target user) follows. We try to classify each of these users into different categories. Then we deduce the user interests from these categories based on the frequency of their occurence among the users which the client follows.

First, every user is categorized as important or unimportant, whereby important users are defined as users with more than 50,000 followers. In order to determine the categories associated with the important users, we search 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 us to use the categorization system of Wikipedia. This is important since a number of categories 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, we use the categories assigned to each article about an important user as the interests which this user represents.

We then count how often each interest occurs among the list of important users followed by the client. The more often an interest occurs, the higher it is rated and the more likely it is to use it for the generation of sentences.

3.2 Gathering and Rating Tweets

We gather tweets for the ten highest rated interests which we inferred as described in the previous section. Lower ranked interests are hereby removed which 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.

We then select users which are followed by the client and which are associated with one of the inferred interests. The number of users added for a single interest must however not exceed a certain value (for the evaluation we have added a maximum of 5 users pro interest). This, on the one hand, prevents the list of being over-flooded by users from one interest and, 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, we gather the five most recent tweets for each of the selected users. The interest based on which the creator of a tweet was added to the possibly interesting users, is also added to the tweet as additional information. We, thus, assume that the tweet represents this interest.

3.3 Rating the Tweets

In order to select the most interesting tweets for the sentence generation and the presentation to the user, we assign each tweet a total score, which is calculated as a linear combination of three other separate scores.

The first one is the interest score (IS) which rates the tweets based only on the interest that they are associated with. Hereby, the more often the interest has occurred among the users followed by the client, the higher it is rated. Additionally, we calculate a user score (US) which measures the strength of the association between the client and the respective creator of the tweet. This is accomplished by checking the number of times a tweet was added to the favorites list of the client. The last score is the time score (TS) which ranks the tweets based on the time of their creation. Hereby, newer tweets are ranked higher

[2]More about SecondHands - `https://secondhands.eu/`

$$c_v^i * IS + c_v^u * US + c_v^t * TS = TOTAL\_SCORE$$

Figure 2: IS, US, TS stand respectively for the interest, user and time score. c_v^i, c_v^u and c_v^t stand respectively for the interest, user and time score weight.

since it is less probable that the user has seen them and since we assume that developing stories are more appealing to the majority of users.

Besides calculating comparable scores for each of the three stages, we also calculate a weight for each one of the scores. These weights represent the confidence in the estimated scores and are used as coefficients in the linear combination which results in the total score (see Figure 2). All three weights are calculated using the coefficient of variation (CV) which is defined as the ratio of the standard deviation to the mean.

3.4 Generating the Output

After ranking the possibly interesting tweets we select the highest rated 20 tweets and transform these into "ice-breaking" sentences which capture the attention of the client and encourage them to give a response to the system. The generation is rule-based and results 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 positive impact on the quality of the output. A more sophisticated statistical generator is, of course, a possible alternative.

Furthermore, we remove hashtags placed at the end of a sentence since these are mostly unconnected to the sentence, and retain hashtags at the middle and the beginning of a sentence which usually are 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"[3].

We also remove links from the tweet, but inform the user that a link was posted. The removal is needed since a lot of the tweets which feature a link, partially describe its contents and since the system is oriented to natural dialog which rarely features someone spelling a link.

4 Results and Discussion

In order to evaluate the system, we performed a user evaluation study in which participants could test the system online and then fill out a questionnaire consisting of 5-point Likert scale and yes-no questions. In order to compare the results, we set up a baseline system ("SystemX"), which randomly chooses a tweet which is then converted to a sentence by the same rule-based NLG component as used by the original system. The participants rated both systems, whereby the baseline system was presented to the participants in the same way as the original system. In order to avoid any bias based on the order in which the systems were tested, 50% of the participants tested the baseline system first and then the original system and the other 50% tested both systems in the opposite order. The study was text-based in order to concentrate on evaluating the system at hand. In conclusion, we checked if the differences in the results for both systems are significant by applying statistical tests on the results.

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

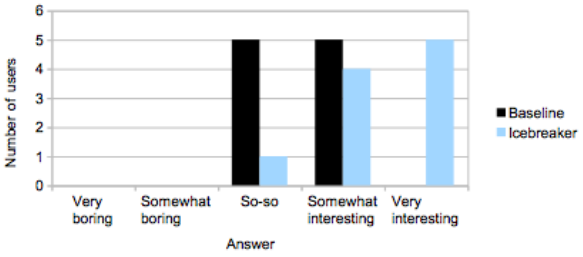

Figure 3: Distribution of the answers of the question "How interesting is the general topic of the ice-breaking sentence to you?". Icebreaker designates the original system and is in light blue, the baseline in black.

The gender of the participants was almost 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 since most of the participants were students.

4.1 Results about the General Interest in the Topic of the Sentence

The users were asked how interested they were in the general topic of the "ice-breaking" sentence which was presented to them. The baseline system achieved an average score of 3.5 while the original system achieved 4.4. (see Figure 3) The difference between these two means is nearly 1.0 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 ranged from -1.48 to -0.32 which points to a substantial difference between the means.

The perceived interest in the topic of the "ice-breaking" sentence seems to correlate with the perceived appeal. In 12 of the 20 questionnaire evaluations the interest in the topic and the appeal of a sentence were evaluated with the same value. The rest differed with the interest value always being one level higher than the perceived appeal.

4.2 Results about the Perceived Appeal of the Sentence

The users evaluated how appealing the generated "ice-breaking" sentence was. The average score for the baseline system was 3.3 while the same score for the original system was 3.8. The distribution of the answers for the two systems is presented in Figure 4. In order to compare these two means a paired t-test was performed on the data from the users who tested both systems. The results confirmed that a statistically significant difference exists and yielded a two-

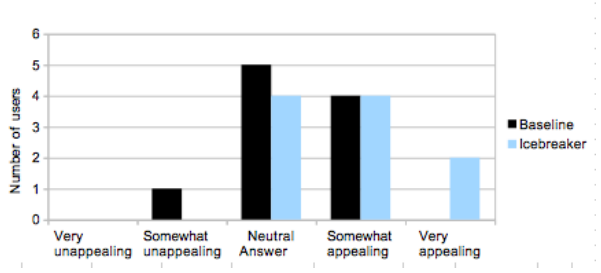

Figure 4: Distribution of the answers of the question "How appealing is the ice-breaking sentence to you?". Icebreaker designates the original system and is in light blue, the baseline in black.

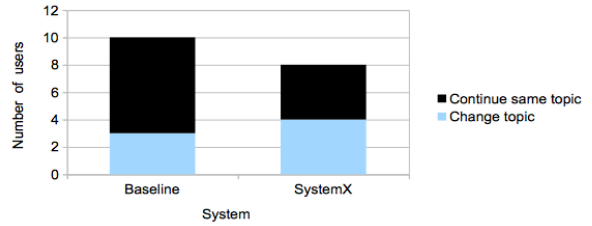

Figure 5: Distribution of the answers on the question if users would continue on the same topic or change it. This question was asked only if users have declareded that they would respond to the system.

tailed p-value of 0.0232. The 95% confidence interval for the difference between the original and baseline systems ranged from 0.14 to 1.42.

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 unfamiliar terms and people, makes a 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 to reply. The average appeal score given to an "ice-breaking" sentence which did not contain any unfamiliar terms or people was 4 ("Somewhat appealing"). Sentences with unfamiliar terms or people were rated on average with around 3.36 (near to "Neutral"). An unpaired t-test on the underlying data for these means yields a two-tailed p-value of 0.0338 which implies that the difference between the means is statistically significant. 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.

4.3 Results about Continuing the Conversation

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. 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 have the knowledge needed to understand the tweet. 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 5.

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". Such sentences can be, e.g., used to initiate a dialog between a robot assistant and a human in the SecondHands project. They should appeal to the user and increase the probability of a response. The evaluation of the system showed that it did outperform the baseline in generating sentences which were perceived as more appealing and additionally at choosing topics which interested the target user more. The perceived appeal of the generated sentences was rated highly as was the general interest in the topic of the sentences.

A 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. 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 or SDSs.

A more sophisticated statistical generator for the conversion of tweets to "ice-breaking" sentences is another possible topic for future work. One way to gather data for such a classifier would be to utilize different crowdsourcing platforms and present various tweets to the users asking them to form "ice-breaking" sentences involving these. Tweets, however, often have different representations with some tweets containing pictures, some being concise and containing only keywords and others describing events thoroughly. Thus, it is important to keep in mind that such differences in the representation can lead to a different degree of linguistic variance in the gathered data as shown in [16].

Finally, we note that we see further possibilities in utilizing information from social networks in the domain of SDSs and artificial intelligence in general. Such data could be used by 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 SDSs and making them more appealing to the user. Furthermore, it can also enhance these by adapting to the user even before a conversation has taken place.

Acknowledgements

This work has been conducted in the SecondHands project which has received funding from the European Union's Horizon 2020 Research and Innovation programme (call:H2020-ICT-2014-1, RIA) under grant agreement No 643950.

References

[1] A. Raux, B. Langner, D. Bohus, A. W. Black, and M. Eskenazi, "Letâ^{TM}s go public! taking a spoken dialog system to the real world," in *in Proc. of Interspeech 2005*, Citeseer, 2005.

[2] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1061–1068, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 591–600, ACM, 2010.

[4] J. A. Smith, M. McPherson, and L. Smith-Lovin, "Social distance in the united states sex, race, religion, age, and education homophily among confidants, 1985 to 2004," *American Sociological Review*, vol. 79, no. 3, pp. 432–456, 2014.

[5] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[6] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, (New York, NY, USA), pp. 261–270, ACM, 2010.

[7] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528, ACM, 2012.

[8] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web.," 1999.

[10] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series.," *ICWSM*, vol. 11, no. 122-129, pp. 1–2, 2010.

[11] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[12] F. Bessho, T. Harada, and Y. 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*, pp. 227–231, Association for Computational Linguistics, 2012.

[13] J. Weizenbaum, "Elizaâ"a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[14] Z. Yu, A. Papangelis, and A. Rudnicky, "Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness," in *2015 AAAI Spring Symposium Series*, 2015.

[15] M. Schmidt, J. Niehues, and A. Waibel, "Towards an open-domain social dialog system," in *Proceedings of the 7th International Workshop Series on Spoken Dialogue System Technology (IWSDS)*, 2016.

[16] M. Schmidt, M. Müller, M. Wagner, S. Stüker, A. Waibel, H. Hofmann, and S. Werner, "Evaluation of crowdsourced user input data for spoken dialog systems," in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 427, 2015.