INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

Zara: An Empathetic Interactive Virtual Agent

Pascale Fung, Anik Dey, Farhad Bin Siddique,
Ruixi Lin, Yang Yang, Wan Yan, Ricky Chan Ho Yin

Human Language Technology Center
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology, Hong Kong

pascale@ece.ust.hk, adey@connect.ust.hk, fsiddique@connect.ust.hk,
rlinab@connect.ust.hk, yyangag@connect.ust.hk, ywanad@connect.ust.hk, eehychan@ust.hk

Abstract

Zara, or 'Zara the Supergirl', is a virtual robot that can show empathy while interacting with an user, and at the end of a 5-10 minute conversation, it can give a personality analysis based on the user responses. It can display and share emotions with the aid of its built in sentiment analysis, facial and emotion recognition, and speech module. Being the first of its kind, it has successfully integrated an empathetic system along with the human emotion recognition and sharing, into an augmented human-robot interaction system. Zara was also displayed at the World Economic Forum held at Dalian in September 2015.

Index Terms: speech recognition, human-computer interaction, emotion recognition, empathy module

1. Introduction

Human-computer interactions, or simply put, speaking to machines has been a common phenomenon in people since the late 1990s. As humans get more used to interacting with voice and gesture controlled machines, they start to anticipate emotional feedback from the machines, and it becomes increasingly necessary for machines to recognise communication features such as intention, humor and sarcasm. This kind of communication can be made possible by having an empathy module in the machines that will enable it to extract emotional features from human interaction, and thereby decide an appropriate response. Even so, empathetic robots related research is still at the very beginning stage, but current methodologies that combine signal processing, sentiment analysis and machine learning algorithms, can enable robots to 'understand' human emotion [1].

Zara the Supergirl is a prototype system, that runs as a web program on a server in the cloud. It is presented on the screen as a virtual robot, with the aid of a cartoon character to display itself. As more data will be collected from the interactions with the different users, Zara can be incorporated with better and 'smarter' machine learning algorithms, thereby making it more empathetic. In addition, Zara can be, in the future, installed into a humanoid robot to give it a physical form.

2. System Design

2.1. Personality Recognition

In the current Zara system, questions are asked to the user from six different domains in order to assess the user's personality [2] and identify the type from the 16 different MBTI personality types[1]. Using sentiment analysis and emotion recognition on each user response, we calculate the score for each personality dimension (namely Introversion-Extraversion, Intuitive-Sensing, Thinking-Feeling, Judging-Perceiving), based on previous work done [3].

The transition from one state to another is controlled using the dialogue management system, that takes care of two different parts. One is to control the flow of the questions that Zara asks to the user and the user answers, and the other is the user initiated question(s) to Zara and handling such challenges.

2.2. Face and Speech Recognition

The program starts when a user face is detected, and a snapshot of the face is taken. The facial features are analysed and the algorithm tries to guess the gender and ethnicity of the user along with a confidence score.

For our speech recognition module, we use English audio data with 1385 hours from LDC corpora and public domain corpora for acoustic model training. We train our acoustic models by Kaldi speech recognition toolkit [4]. We train deep neural network (DNN) HMMs with 6 hidden layers. The DNN is initialized with stacked restricted Boltzmann machines (RBMs) which are pre-trained in a greedy layerwise fashion. Our text data contains 88.6M sentences. It comprises of acoustic training transcriptions, web crawled news and book data, Cantab filtering sentences on Google 1 billion word LM benchmark, weather domain queries, music domain queries and common chat queries. We train witten-bell smoothing interpolated trigram language model (LM) and CE based recurrent neural network (RNN) LM using the SRI-LM toolkit [5] and CUED-RNNLM toolkit [6] respectively. The ASR decoder performs search on weighted finite state transducer (WFST) graph for trigram LM and generates lattice, and then performs lattice rescoring with RNN LM. The ASR system achieves 7.6% word error rate on our clean speech test data.

2.3. Emotion recognition

2.3.1. Raw audio analysis

For the training of our speech audio emotion recognition we used around 200 hours of TED[2] audio data, that were labelled in 13 second frames into 11 mood categories. We set 5ms as the window size and 3.25ms as the moving step for the first convolutional layer. The regional max-pooling layer takes 40

[1] https://www.personalitypage.com/html/
[2] https://www.ted.com/talks

vectors each time, and the size of the second convolutional layer is 26, with moving step being 1. We then execute the maximum function over all the vectors of the outputs of the second pooling layer [7].

The first convolutional layer accepts a short period of audio as input. Then the model moves to convolute the adjacent period of audio with fixed overlap of last period. After going through the first convolutional layer, the data format is converted into a matrix. The next layer, max-pooling layer, is a form of non-linear down-sampling. It partitions the input matrix into a set of non-overlapping smaller matrices. For each sub-region, outputs the entry-wise maximum value in one dimension. The second max-pooling layer is to output the entry-wise maximum on the entire matrix instead of sub-regions. The embedding layer performs similar function as that of multi layer perceptron, which maps the vector into a probabilistic distribution over all the mood categories [8] [9].

2.3.2. Lexical features

Language understanding of lexical features is also used for the sentiment analysis in Zara. The user responses are matched against a pool of emotion lexicons in order to look for positive and negative keywords. The LIWC[3] dictionary is used, and the score is calculated based on the keywords found [10]. Furthermore, double negatives are dealt with by looking for *negate* words from LIWC, and the score is altered if necessary (for e.g, "I am not feeling well" will have a negative sentiment score, even though 'well' is a positive lexicon).

Also, if the length of a sentence is longer than five words, a 5-gram model is used and the final sentiment score is the total score across all the 5-grams. This is helpful in the case for longer sentences, where there is a higher chance of double negatives occurring. There are cases when there are too few or no LIWC keywords detected in the sentence, when it becomes difficult to calculate the correct sentiment. This is why we need the raw audio analysis to give a more comprehensive emotion score.

3. Handle challenging situations

There have been quite a few cases where a user challenges Zara by not answering a question directly. For example, 12.5% of the users challenged Zara by asking completely unrelated questions. In addition, 24.62% of users showed some kind of verbal challenges during the conversation, 37.5% of which were just trying to avoid the question by giving an irrelevant answer. The different challenging responses to Zara can be subdivided to six major categories, namely: 1. Seeking disclosure reciprocity, 2. Asking for clarification, 3. Avoidance of topic, 4. Deliberate challenge of Zara's ability, 5. Abusive language, 6. Garbage. Some of these categories are also common when humans converse with each other [11].

Avoiding the topic was most common, and in psychology it is seen as a way to deal with pressure, panic, discomfort or worry in most humans [12]. Some users directly were unwilling to continue, while others passively implied that they had the intention to not continue.

Abusive language, for example swearing, inappropriate or insulting remarks, were relatively rare. The cause behind these cases could be that some users were unable to build the trust with Zara and were slightly irritated or annoyed, rather than feeling happy and being excited about the conversation [13].

Although Zara is built to be empathetic with a kind and gentle personality, it is also given some witty traits. For example, if a user keeps using inappropriate language (e.g. swears at Zara), then the dialogue won't continue unless the user apologizes. Also, if Zara gets a general question that is irrelevant to the personality test (for e.g, "What is the weather in Hong Kong"), then the question will be answered from a general knowledge database using a search engine API[4].

4. Conclusion

Zara is a prototype system depicting an empathetic virtual robot that can understand and share human emotions. It is true that we are still in the primary stage of building empathetic and friendly robots and current robots like Zara will have flaws in them. Nevertheless, the crucial objective is to build robots that are more human like, so just like humans, they will not be perfect. Hopefully, if we succeed, future machines and robots will be able to 'get' us and empathise with humans, and instead of harming us in any way, they will be our caretakers, our teachers, and our companions.

5. References

[1] P. Fung, "Robots with heart," *Scientific American*, pp. 60–63, 2015.

[2] T. Polzehl, S. Moller, and F. Metze, "Automatically assessing personality from speech," *In Semantic Computing (ICSC), 2010 IEEE Fourth International Conference*, pp. 134–140, 2010.

[3] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, pp. 457–500, 2007.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[5] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *INTERSPEECH*, 2002.

[6] X. Chen, X. Liu, Y. Qian, M. J. Gales, P. Woodland *et al.*, "Cued-rnnlm–an open-source toolkit for efficient training and evaluation of recurrent neural network language models," 2016.

[7] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6964–6968, 2014.

[8] P. Golik, Z. Tuske, R. Schluter, and H. Ney, "Convolutional neural networks for acoustic modelling of raw time signal in lvcsr," *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] D. Palaz and R. Collobert, "Analysis of cnn based speech recognition system using raw speech as input," *Interspeech*, no. EPFL-CONF-210029, 2015.

[10] J. Pennebaker, R. Booth, R. Boyd, and M. Francis, "Linguistic inquiry and word count: Liwc2015," *Austin, TX: Pennebaker Conglomerates*, 2015.

[11] L. R. Wheeless and J. Grotz, "The measurement of trust and its relationship to self-disclosure," *Human Communication Research*, vol. 3, no. 3, pp. 250–257, 1977.

[12] S. Roth and L. Cohen, "Approach, avoidance, and coping with stress," *American psychologist*, vol. 41, no. 7, p. 813, 1986.

[13] T. Nomura, T. Uratani, T. Kanda, M. K., K. H., Y. Suehiro, and S. Yamada, "Why do children abuse robots?" *In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pp. 63–64, 2015.

[3] `http://liwc.wpengine.com/`

[4] `https://www.houndify.com//`