Zara The Supergirl: An Empathetic Personality Recognition System

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 the Supergirl is an interactive system that, while having a conversation with a user, uses its built in sentiment analysis, emotion recognition, facial and speech recognition modules, to exhibit the human-like response of sharing emotions. In addition, at the end of a 5-10 minute conversation with the user, it can give a comprehensive personality analysis based on the user's interaction with Zara. This is a first prototype that has incorporated a full empathy module, the recognition and response of human emotions, into a spoken language interactive system that enhances human-robot understanding. Zara was shown at the World Economic Forum in Dalian in September 2015.

1 Introduction

"Sorry I didn't hear you" maybe the first empathetic utterance by a commercial machine. Since the late 1990s when the Boston company SpeechWorks International began providing their customer-service software to other numerous companies, which was programmed to use different phrases, people have gotten used to speaking to machines. As people interact more often by voice and gesture, they expect the machines to have more emotional intelligence, and understand other high level communication features such as humor, sarcasm and intention. In order to make such communication possible, the machines need an empathy module in them, which is a software system that can extract emotions from human speech and facial expressions, and can accordingly decide the correct response of the robot. Although research on empathetic robots is still in the primary stage, current methods involve using signal processing techniques, sentiment analysis and machine learning algorithms to make robots that can 'understand' human emotion (Fung, 2015).

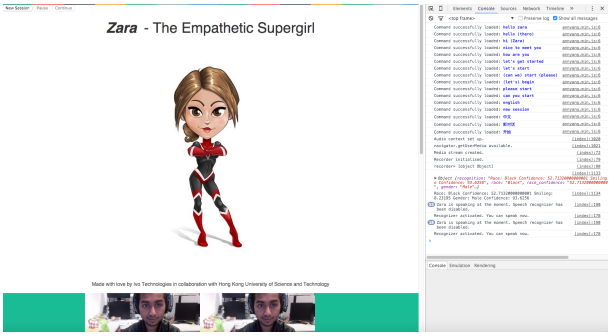

Figure 1: Screenshot of Zara

We propose Zara the Supergirl as a prototype system. It is a web program running on a server in the cloud. It is basically a virtual robot, with an animated cartoon character to present itself on the screen. Along the way it will get 'smarter' and more empathetic, by having machine learning algorithms, and gathering more data and learning from it. Later stage would involve installing the program into a humanoid robot, and therefore give Zara a physical body.

2 System Description

2.1 Design and Training

Zara's current task is a conversational MBTI personality assessor and we designed 6 categories of personality-assessing questions, each named as a 'state', in attempts to assess the user's personal-

Proceedings of NAACL-HLT 2016 (Demonstrations), pages 87–91,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

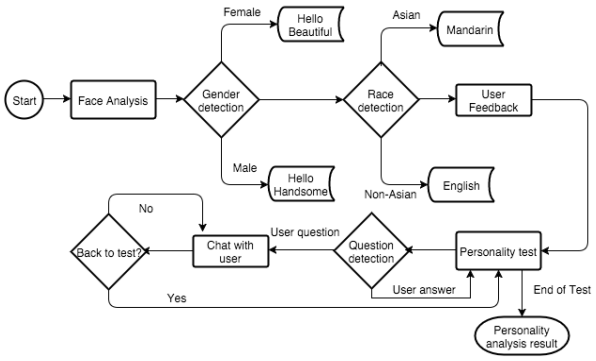

Figure 2: System state diagram

ity (Polzehl et al., 2010). These 6 states inquire about the user's earliest childhood memory, his or her last vacation, challenges at work, creative storytelling, friendship, and affinity toward human-robot conversations. Each state comprises of a series of questions, beginning with one opening inquiry with follow-up questions depending on the length of the user's preceding response. Each user is allocated 5-6 minutes to complete the personality assessment (approx. 1-2 minutes per question). The tests are conducted independently using url link rendered on a browser using built-in microphone and camera on Macs and PCs.

A dialog management system with different states is designed to control the flow of the conversation, which consists of one part machine-initiative questions from Zara and answers from human users, and another part user-initiative questions and challenges to Zara.

2.2 Facial and Speech Recognition

At the beginning of the conversation with the user, the program waits until a face is detected. The face recognition algorithm analyses the image captured by the computer's webcam to guess a possible gender and ethnicity.

For our speech recognition module, we use English audio data with 1385hrs from LDC corpora and public domain corpora for acoustic model training. We train our acoustic models by Kaldi speech recognition toolkit (Povey et al., 2011). 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. Cross-entropy (CE) criterion DNN training is first applied on the state alignments produced by discriminative trained GMM-HMMs. State alignment is then reproduced with DNN-HMMs, and DNN training with CE criterion is done again. Finally, sequence discriminative training on DNN-HMMs with state level minimum Bayes risk (sMBR) criterion is applied.

Our text data contains 88.6M sentences. It comprises 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 (Stolcke and others, 2002) and CUED-RNNLM toolkit (Chen et al., 2016) 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 decoder is designed for input audio data that is streamed from TCP/IP or HTTP protocol, and performs decoding in real time. The decoder supports simultaneous users by multiple threads and user queue. The ASR system achieves 7.6% word error rate on our clean speech test data.

2.3 Audio features for emotion recognition

The dataset we used for training speech emotion recognition is from the Emotional Prosody Speech and Transcripts, Linguistic Data Consortium (LDC) catalog number LDC2002S28[1] and ISBN 1-58563-237-6 (Liberman et al., 2002). The recordings contain audio and transcripts, which consist of professional actors reading a series of semantically neutral utterances (dates and numbers). There are 15 sentiments, 7 subjects and a total number of 2445 utterances. Each subject reads around 3,000 seconds.

We use openSMILE (Eyben et al., 2010) to extract features from LDC dataset. The features are calculated based on the INTERSPEECH 2009 Emotion Challenge feature set for emotion recognition. The final features are computed from a series of input frames and output a single static summary vector, e.g, the smooth methods, maximum and min-

[1] https://catalog.ldc.upenn.edu/LDC2002S28

Anxiety	Interest	Shame	Sadness	Pride	Elation
66.2	67.6	62.9	67.7	53.2	59.1
Neural	Despair	Hot anger	Disgust	Boredom	Happy
87.5	58.4	75.0	50.0	63.0	69.9
Panic	Contempt	Cold anger			
65.5	74.7	56.9			

Figure 3: Binary classification accuracy in percentage for each sentiment

imum value, mean value of the features from the frames (Liscombe et al., 2003).

For each sentiment, there are around 170 utterances. We implement an SVM learning method on the binary classification. We first construct a balanced dataset for each sentiment by choosing the same number of utterances per sentiment. Then we split the data into three categories, i.e. training, developing and testing parts in the ratio 6:2:2. We train the SVM with linear kernel and the maximum iteration time is 5,000. The development set is used to tune the number of iteration times. The model is chosen from the highest accuracy results from the development set (results given in figure 3).

2.4 Language understanding for sentiment analysis

In the first version of Zara, sentiment analysis is based on natural language understanding of lexical features. We look for keyword matches from a pool of positive and negative emotion lexicons from LIWC[2] dictionary. The positive lexicons have positive scores, and the negative lexicons have negative scores (Pennebaker et al., 2015). Moreover, when a *negate* word ('do not', 'cannot', etc) is present along with the emotion words, then the score is adjusted accordingly (for example, "I am not at all happy" would have a negative score, even though 'happy' is a positive emotion lexicon).

If there are more than five words in a sentence, then a n-gram model is used containing a number of 5 grams, which is then further analysed to give a total sentiment score across all the 5-grams. This tends to perform better than non n-gram methods in the case for long sentences.

2.5 Personality Analysis

We designed a set of personal questions in six different domains in order to classify user personality from among sixteen different MBTI personality types[3]. The original MBTI test questionnaire contains about 70 questions. We asked a group of training users to answer this questionnaire but also answer questions from Zara. The personality type generated by the MBTI questionnaire is used as the gold standard label for training the Zara system. Based on user answers to Zara's questions, scores are calculated in four dimensions (namely Introversion - Extroversion, Intuitive - Sensing, Thinking - Feeling, Judging - Perceiving).

Level	Introvert	Extravert
Conversational behaviour	Listen Less back-channel behaviour	Initiate conversation More back-channel behaviour
Topic selection	Self-focused Problem talk, dissatisfaction Strict selection Single topic Few semantic errors Few self-references	Not self-focused* Pleasure talk, agreement, compliment Think out loud* Many topics Many semantic errors Many self-references
Style	Formal Many hedges (tentative words)	Informal Few hedges (tentative words)
Syntax	Many nouns, adjectives, prepositions (explicit) Elaborated constructions Many words per sentence Many articles Many negations	Many verbs, adverbs, pronouns (implicit) Simple constructions* Few words per sentence Few articles Few negations
Lexicon	Correct Rich High diversity Many exclusive and inclusive words Few social words Few positive emotion words Many negative emotion words	Loose* Poor Low diversity Few exclusive and inclusive words Many social words Many positive emotion words Few negative emotion words
Speech	Received accent Slow speech rate Few disfluencies Many unfilled pauses Long response latency Quiet Low voice quality Non-nasal voice Low frequency variability	Local accent* High speech rate Many disfluencies* Few unfilled pauses Short response latency Loud High voice quality Nasal voice High frequency variability

Figure 4: Summary of identified language cues for extraversion and various production levels (Mairesse et al., 2007)

We use the output of the sentiment analysis from language and emotion recognition from speech as linguistic and speech cues to calculate the score for each personality dimension based on previous research (Mairesse et al., 2007). For each response, the individual score for each of the four dimensions is calculated and updated, and the final score in each dimension is the group average of all the responses.

3 Handling user challenges

The personality test consists mostly of machine-initiative questions from Zara and human answers. However, as described in the user analysis section below, there are scenarios where the user does not respond to questions from Zara directly. 24.62% of the users who tried Zara exhibited some form of verbal challenge in their responses during the dialogue conversation, of which 37.5% of users evade the questions with an irrelevant answer. 12.5% of

[2] http://liwc.wpengine.com/

[3] https://www.personalitypage.com/html/high-level.html

users challenged Zara's ability more directly with questions unrelated to the personality test.

Challenge here refers to user responses that were difficult to handle and impeded the flow of conversation with Zara. They include the following 6 types: 1. Seeking disclosure reciprocity; 2. Asking for clarification; 3. Avoidance of topic; 4. Deliberate challenge of Zara's ability; 5. Abusive language; 6. Garbage.

Several of the above categories can be observed in human-human interactions. For instance, seeking disclosure reciprocity is not uncommon in human conversations (Wheeless and Grotz, 1977).

Responses that revealed some form of avoidance of topic was the largest response group. Avoidance in psychology is viewed as a coping mechanism in response to stress, fear, discomfort, or anxiety (Roth and Cohen, 1986). In the dataset collected, two types of avoidance were observed. Users who actively avoid the topic specifically reveal their unwillingness to continue the conversation ("I dont want to talk about it", "I am in no mood to tell you a story Zara") while users who adopts a more passive strategy had the intent to discontinue the conversation implied ("Let's continue.", "Make it a quick one", "You know...").

Abusive language includes foul, obscene, culturally and socially inappropriate remarks and the like. Currently collected data revealed surprisingly few inappropriate comments such as "get lost now" and "None of your business". These challenges are comparatively mild. Owing to the context of Zara's role as a personality assessor, the reasons here for abuse could be the need to trust the robotic assessor and feeling of discomfort instead of the common enjoyment or group thinking reasons (Nomura et al., 2015).

Asking for clarification examples included "Can you repeat?" and "Can you say it again?". Clarification questions observed in this dataset are primarily non- reprise questions as a request to repeat a previous utterance (Purver, 2004).

Deliberate challenge of a robot's ability was also observed. This took the form of direct requests ("Can I change a topic?", "Why can't you speak English?" in the Chinese mode), or statements unrelated to the questions asked ("Which one is 72.1 percent?").

Zara is programmed with a gentle but witty personality to handle different user challenges. For example, when abusive language is repeatedly used against her, she would ask for an apology after expressing concern for the user's level of stress. If the user asks a general domain question unrelated to the personality test (e.g. "What is the population of Hong Kong"), Zara will try to entertain the question with an answer from a general knowledge database using a search engine API[4], much like Siri or Cortana. However, unlike these other systems, Zara will not chat indefinitely with the user but will remind the user of their task at hand, namely the personality test.

4 Future Work

We are also working on a second approach for the audio emotion recognition. This uses a deep neural network framework, with raw audio data from TED[5] audio database as training data. A total of around 200 hours of TED audio data was used, and was labelled in 13-second frames. This labelled data was used to train a binary classifier of 11 different mood categories.

An FFmpeg command-line software is used to extract the envelop of the raw audio input. Each value is a 16-bit integer. Since the sample rate is 8 kHz and each training sample is around 10 seconds in length, the input dimension should be 80,000. We set 5 ms (40 integers) as the window size and 3.25 ms (26 integers) as the moving step for the first convolutional layer. The regional max-pooling layer takes 40 vectors each time. The window size of the second convolutional layer is 26 which is slightly smaller than the first one. The moving step is 1. We execute the maximum function over all the vectors of the outputs of the second pooling layer.

There are two convolutional, two max-pooling and one embedding layers in the CNN model. 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, and the vector input is converted into a matrix. The next layer, max-pooling layer, is a form of non-linear down-sampling. It partitions the

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

[5] https://www.ted.com/talks

input matrix into a set of non-overlapping smaller matrices. For each sub-region, it 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, which outputs a vector. The embedding layer performs similar function as that of a multi layer perceptron, which maps the vector into a probabilistic distribution over all categories (Palaz and Collobert, 2015) (Golik et al., 2015).

5 Conclusion

We have demonstrated a prototype system of an empathetic virtual robot that can recognize user personalities from speech, language and facial cues. It is too early to say that the time of empathetic and friendly robots has arrived. We have so far developed only the most primary tools that future emotionally intelligent robots would need. The empathetic robots including Zara that are there currently, and the ones that will be there in the near future, might not be completely perfect. However, the most significant step is to make robots to be more human like in their interactions. This means it will have flaws, just like humans do. If this is done right, then future machines and robots will be empathetic and less likely to commit harm in their interactions with humans. They will be able to get us, understand our emotions, and more than anything, they will be our teachers, our caregivers, and our friends.

Acknowledgments

This work was funded by grant #16214415 of the Hong Kong Research Grants Council.

References

Xie Chen, Xunying Liu, Yanmin Qian, Mark JF Gales, Philip Woodland, et al. 2016. Cued-rnnlm–an open-source toolkit for efficient training and evaluation of recurrent neural network language models.

F. Eyben, M. Wollmer, and B. Schuller. 2010. opensmile - the munich versatile and fast open-source audio feature extractor. *ACM MM, Florence, Italy*, pages 1459–1462.

Pascale Fung. 2015. Robots with heart. *Scientific American*, pages 60–63.

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

Mark Liberman, Kelly Davis, Murray Grossman, Nii Martey, and John Bell. 2002. Emotional prosody speech and transcripts ldc2002s28. *Philadelphia:Linguistic Data Consortium*.

J. Liscombe, J. Venditti, and J.B. Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. *Columbia University Academic Commons*.

Francois Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, pages 457–500.

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

Dimitri Palaz and Ronan Collobert. 2015. Analysis of cnn based speech recognition system using raw speech as input. *Interspeech*, (EPFL-CONF-210029).

J.W. Pennebaker, R.J. Booth, R.L. Boyd, and M.E. Francis. 2015. Linguistic inquiry and word count: Liwc2015. *Austin, TX: Pennebaker Conglomerates*.

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

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society.

M. Purver. 2004. The theory and use of clarification requests in dialogue. *Unpublished doctoral dissertation, University of London*.

S. Roth and L.J. Cohen. 1986. Approach, avoidance, and coping with stress. *American psychologist*, 41(7):813.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

L. R. Wheeless and J. Grotz. 1977. The measurement of trust and its relationship to self-disclosure. *Human Communication Research*, 3(3):250–257.