Zara: A Virtual Interactive Dialogue System Incorporating Emotion, Sentiment and Personality Recognition

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang,
Dario Bertero, Wan Yan, Ricky Chan Ho Yin, Chien-Sheng Wu
Human Language Technology Center
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology, Hong Kong
pascale@ece.ust.hk,
[adey, fsiddique, rlinab, yyangag, dbertero, ywanad]@connect.ust.hk,
eehychan@ust.hk, b01901045@ntu.edu.tw

Abstract

Zara, or 'Zara the Supergirl' is a virtual robot, that can exhibit empathy while interacting with an user, with the aid of its built in facial and emotion recognition, sentiment analysis, and speech module. At the end of the 5-10 minute conversation, Zara can give a personality analysis of the user based on all the user utterances. We have also implemented a real-time emotion recognition, using a CNN model that detects emotion from raw audio without feature extraction, and have achieved an average of 65.7% accuracy on six different emotion classes, which is an impressive 4.5% improvement from the conventional feature based SVM classification. Also, we have described a CNN based sentiment analysis module trained using out-of-domain data, that recognizes sentiment from the speech recognition transcript, which has a 74.8 F-measure when tested on human-machine dialogues.

1 Introduction

As the availability of interactive dialogue systems is on a rise, people are getting more accustomed to talking to machines. Modern systems are equipped with better statistical and machine learning modules in order to help them get better over time. People have started expecting the machines to understand different aspects of dialogues, like intent, humor, sarcasm, etc. We want the system to connect with us more, by recognising our emotions. This requires machines to have an empathy module in them, that will enable them to give more emotional responses during the interaction with users (Fung, 2015).

We have developed a prototype system that is a web program that can be rendered on a browser, and is a virtual robot with a cartoon character to represent itself (Fung et al., 2015). It can converse with a user by asking a few questions related to the user's personal experiences, and can give a personality analysis based on the responses after a 5-10 minute conversation. At each round of interaction, the response to the user utterance is chosen based on the emotion and sentiment recognition results, some examples are shown below:

Zara: *How was your last vacation?*

User: *I went on a vacation last month and it was pretty bad, I lost all my luggage.*

Response: *That doesn't sound so good. Hope your next vacation will be a good one.*

User: *My last vacation was amazing, I loved it!*

Response: *That sounds great. I would like to travel with you.*

Conventional methods of emotion recognition require feature engineering (Schuller et al., 2009; Schuller et al., 2010), which is too slow for a task like this, and so cannot be used in interactive dialogue systems. Therefore, we use a Convolutional Neural Network (CNN) model that bypasses the feature extraction and extracts emotion from raw-audio in real-time.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 278–281, Osaka, Japan, December 11-17 2016.

2 System Description

2.1 Design

The main task of our system right now is the assessment of MBTI personality at the end of the conversation with the user (Polzehl et al., 2010). We have designed 6 unique classes of questions asking the user about their childhood memories, last vacation, work challenges, creativity in telling a story, companionship, and also their opinion on human-robot interactions. Each class is termed as a 'state' and each state consists of an opening question and other follow up questions, depending on the user response. Zara can be used using an URL link rendered on a browser, with the use of a microphone and a camera.

The conversation flow is controlled via the dialogue management system that keeps track of the various states. It also decides between two different types of conversation, one is where Zara asks the question, or machine-initiative, and the other is user-initiative questions or challenges to Zara.

2.2 Facial and Speech Recognition

At the initial stage, when the system is started, a snapshot of the user's face is taken, and the facial recognition algorithm tries to identify the user's gender and ethnicity, along with a confidence score.

For speech recognition, we collected acoustic data from different public domain and LDC corpora, which makes a total of 1385 hours of speech. The acoustic models are trained by the Kaldi speech recognition toolkit (Povey et al., 2011), using the raw audio together with encode-decode parallel audio to train Deep Neural Network - Hidden Markov Models (DNN-HMMs). We apply sequence discriminative training using state Minimum Bayes Risk (sMBR) criterion, and layer wise training of restricted Boltzmann machines (RBMs), along with frame cross-entropy training via mini-batch stochastic gradient descent (SGD). We use text data, that includes Cantab filtering sentences on Google 1 billion word LM benchmark (Chelba et al., 2013), acoustic training transcriptions, and other web crawled news, and music and weather domain queries, making a total of around 90M sentences. Our decoder supports streaming of raw audio or CELP encoded data via TCP/IP or HTTP protocol, and performs decoding in real time. The ASR system achieves 7.6% word error rate on our clean speech test data[1].

2.3 Real-Time Emotion Recognition from Raw Audio

Most of the benchmark systems on classification of Emotional speech (Mairesse et al., 2007) or music genres or moods (Schermerhorn and Scheutz, 2011), involves feature extraction and classifier learning, which is both time-consuming and requires a lot of hand tuning. Therefore, we have developed a Convolutional Neural Network model that can recognise emotions directly from time-domain audio signal, bypassing the feature engineering. This is suitable for use in applications like interactive dialogue systems, which have real-time requirements.

We built a dataset from the TED-LIUM corpus release 2 (Rousseau et al., 2014), that includes 207 hours of speech extracted from 1495 TED talks. After initially annotating the data using a commercially available API, we hand-corrected the annotations. Six categories of emotions are used: criticism, anxiety, anger, loneliness, happiness, and sadness, and the audio data is divided into 13 second segments for annotations.

Using 8kHz as the sampling rate, and a single filter in the CNN, we set the convolutional window size to be 200, which is 25 ms, and an overlapping step size of 50, equivalent to 6 ms. The convolutional layer uses the differences between neighbouring and overlapping frames, and also performs its own feature extraction from the raw audio. Max pooling is done later that gives an output of a segment-based vector, which is then fed to a fully connected layer that acts like a Deep Neural Network (DNN), thereby mapping the output to a probabilistic distribution over the emotion categories via a final softmax layer.

For baseline, we use Support Vector Machine (SVM) classifier with a linear kernel using the INTERSPEECH 2009 emotion feature set (Schuller et al., 2009). The results are shown in Table 1. By using a single filter CNN architecture, we achieve real-time decoding, around 1.62 ms on average for each segment of longer than 13s, and also we achieve a notable 4.5% improvement on average when compared to the baseline SVM method.

[1] https://catalog.ldc.upenn.edu/LDC94S13A

Emotion class	SVM (%)	CNN (%)
Criticism/Cynicism	55.0	61.2
Defensiveness/Anxiety	56.3	62.0
Hostility/Anger	72.8	72.9
Loneliness/Unfulfillment	61.1	66.6
Love/Happiness	50.9	60.1
Sadness/Sorrow	71.1	71.4
Average	61.2	**65.7**

Table 1: Accuracies obtained in the Convolutional Neural Network model for emotion classification from raw audio samples.

2.4 Sentiment Recognition from Text

Previous research by Kim (2014) has shown that Convolutional Neural Networks (CNNs) can perform impressively in the sentiment classification task. We use word embedding vectors (Word2Vec) trained on the Google News corpus (Mikolov et al., 2013) of size 300, to train a CNN with one layer of convolution and max pooling (Collobert et al., 2011). Using convolutional sliding window of sizes 3, 4 and 5 to represent different features, we apply a max-pooling operation on the output vectors from the convolutional layer. Two different CNN channels are used, one that keeps the word vectors static throughout, and the other fine tunes the vectors via back-propagation (Kim, 2014). The two sentence encoding vectors from the two channels are fed to the final softmax layer, that gives as output the probability distribution over the binary sentiment classification of the transcribed speech text. To improve the performance accuracy, we have used a larger Twitter sentiment 140[2] dataset, and have compared to the original Movie Review dataset used in Kim (2014). Results are shown in Table 2.

Model	Accuracy	Precision	Recall	F-score
Movie Review	67.8%	91.2%	63.5%	74.8
Twitter	72.17%	78.64%	86.69%	82.47

Table 2: Sentiment analysis results tested on human-machine conversations when trained from Twitter and Movie Review datasets

2.5 Personality Analysis

Our task is to identify the user personality from sixteen different MBTI personality types[3], and we designed six different domain specific personal questions for the classification. A group of training users were asked to fill up the original MBTI personality test questionnaire, that contains about 70 questions, and this was used as the gold standard label for training. The user responses to Zara's questions were used to calculate scores in four personality dimensions (Introversion - Extroversion, Intuitive - Sensing, Thinking - Feeling, Judging - Perceiving). Based on previous research done by Mairesse et al. (2007), we use the scores from the emotion and sentiment recognition as speech and linguistic cues to calculate the personality dimension scores.

3 Handling Challenges

Sometimes users can respond in a way that does not answer the question directly, and therefore impose a challenge on Zara. From a preliminary study on the recorded responses, it was found that 12.5% of users asked irrelevant questions to Zara, 24.62% challenged Zara in some other way, and 37.5% tried to avoid the topic. According to Wheeless and Grotz (1977), such cases are also common in human-human conversations.

[2] www.sentiment140.com

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

Most common challenges were avoidance of topic, followed by usage of abusive language. Although Zara is made empathetic in nature, it is also given some witty traits, for example, if multiple swearing or use of inappropriate language is detected, then Zara stops conversing with the user unless they apologise. A general question to Zara (like "What is the capital of China?") will be answered from a general knowledge database using a search engine API.

4 Conclusion

We have described our prototype system, Zara the Supergirl, that uses real-time emotion and sentiment recognition to converse with a user by attempting to give emotionally intelligent responses. Such systems will help future robots to have a better and more advanced empathy module in them, thereby enabling them to build an emotional connection with humans. Also, we have shown that current research on deep learning can help come up with better and faster models to recognise different aspects of human behaviour like personality, in real-time conversations. This advancement can help us build robots that will help humans in the future, and instead of bringing mischief, they can be our companions and caregivers.

References

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Wan Yan, and Ricky Chan Ho Yin. 2015. Zara the supergirl: An empathetic personality recognition system.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tim Polzehl, Sebastian Möller, and Florian Metze. 2010. Automatically assessing personality from speech. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 134–140. IEEE.

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.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2014. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*, pages 3935–3939.

Paul Schermerhorn and Matthias Scheutz. 2011. Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In *Proceedings from the International Conference on Advances in Computer-Human Interactions*, pages 236–241.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315. Citeseer.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, volume 2010, pages 2795–2798.

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