Emotional Conversation Generation

Call for Participation

In recent years, there has been a rising tendency in AI research to enhance Human-Computer Interaction by humanizing machines. However, to create a robot capable of acting and talking with a user at the human level requires the robot to understand human cognitive behaviors, while one of the most important human behaviors is expressing and understanding emotions and affects. As a vital part of human intelligence, emotional intelligence is defined as the ability to perceive, integrate, understand, and regulate emotions. Though a variety of models have been proposed for conversation generation from large-scale social data, it is still quite challenging (and yet to be addressed) to generate emotional responses.

In this challenge, participants are expected to generate Chinese responses that are not only appropriate in content but also adequate in emotion, which is quite important for building an empathic chatting machine. For instance, if user says "My cat died yesterday", the most appropriate response may be "It's so sad, so sorry to hear that"to express ***sadness***, but also could be "Bad things always happen, I hope you will be happy soon"to express ***comfort***.

Task Definition

This task is defined as follows: Given a Chinese post $X = (x_1, x_2, \cdots, x_n)$, and a user-specified emotion category of the response to be generated, the goal is to generate a response $Y = (y_1, y_2, \cdots, y_m)$ that is coherent with the emotion category. The emotion categories are {*Anger, Disgust, Happiness, Like, Sadness, Other*}, the same as defined in http://tcci.ccf.org.cn/conference/2014/dldoc/evatask1.pdf.

Each team can submit at most **TWO** runs.

Dataset

The dataset is constructed from Weibo posts and replies/comments. More than 1 million Weibo post-response pairs will be provided to participants for training their models. To ensure fair comparison, **NO** additional training data will be allowed for conversation generation, but participants can use other data to train supplementary classifiers for emotion classification, for instance. Such details should be reported in the submission of results.

The test dataset consists of about 5000 posts while 100~200 of the posts will be manually assessed, and for each post, at most 3 emotion classes will be manually specified to indicate the emotion class of a generated response. Participating systems should generate a response for each emotion class. Note that participants

should generate responses for all posts with appropriate emotion classes. Which part of the posts will be manually checked is unknown to participants for fair comparison.

The dataset will include labels of each post and response. These labels are **for reference only**, and they are obtained by a simple classifier that is based on a bidirectional LSTM model. The classifier was trained on the data from http://tcci.ccf.org.cn/conference/2014/dldoc/evatask1.pdf where the accuracy for six-way classification is about 64%. In other words, the emotion label of these data is noisy. Participants are encouraged to implement the emotion classifier by themselves and with their own data, but all details must be reported and all resources should be accessible to the community to let other researchers reproduce their results.
Note that no additional data is permitted to train the generation model.

The correspondence between the label and the emotion class can be seen in this table:

Label	0	1	2	3	4	5
Class	Other	Like	Sadness	Disgust	Anger	Happiness

Both the training dataset and the test dataset are Python List objects and are simply dumped to JSON. The training dataset looks like: [[[post,post_label],[response,response_label]],[[post,post_label],[response,response_label]],…].
There are about 1,110,000 pairs in the training data.

And the test dataset looks like: [[post,label],[post,label],[post,label],…]. These datasets are dumped to one line. We will provide about 5000 test posts to participants for emotional response generation.

Download training data here: http://www.aihuang.org/p/challenge.html

Evaluation

The following metrics will be adopted for evaluation:
Emotion Consistency: whether the emotion class of a generated response is the same as the pre-specified class.
Coherence: whether the response is appropriate in terms of both logic and content.
Fluency: whether the response is fluent in grammar and acceptable as a natural language response.

Our labeling procedure is shown by the following pseudocode:

IF (**Coherence** and **Fluency**)
    IF (**Emotion Consistency**)
        LABEL **2**
    ELSE
        LABEL **1**
ELSE
    LABEL **0**

Note that the labeling procedure is conducted on each test post for a given emotion class.

An Example is shown below:

Given	**Post**	比情人节更庸俗的事就是，对情人节的接待规格待遇提出要求。 What is more vulgar than Valentine's day is to make specific demands on the reception of the Valentine's day.	**Coherence** and **Fluency**	**Emotion Consistency**	**Label**
	Expected Emotion	Happy			
Comment1	哈哈，这也太浪漫了吧！ Aha, this is too romantic!		Yes	Yes	2
Comment2	这就是所谓的情人节！ This is the so-called Valentine's day!		Yes	No (Disgust)	1
Comment3	禽兽！放开那个女孩！ Beast, let the girl go!		Not Coherent	No (Anger)	0
Comment4	哈哈，这也太浪漫浪漫浪漫浪漫浪漫了吧！ Aha, this is too romantic romantic romantic romantic romantic!		Not Fluent	Yes	0

Contact

A/Prof. Minlie Huang
Dept. of Computer Science, Tsinghua University, Beijing 100084, China
aihuang#AT#tsinghua#DOT#edu#DOT#cn
http://aihuang.org/p