# DERP: An Evaluation Metric for Dialogue Systems


| | |
|---|---|
| Barun Patra | 2013CS10773 |
| Akshay Kumar Gupta | 2013CS50275 |
| Shantanu Kumar | 2013EE10798 |
| Surag Nair | 2013EE10504 |

# Introduction

- Current automatic metrics for dialogue eval correlate poorly with human judgement (Serban et. al., How NOT to Evaluate your dialogue system)
- Obtaining human scores time consuming and costly
- Automatic metric which shows good correlation with human judgement invaluable -> learn it from data
- Hypothesis: discriminatory model easier to learn than generative dialogue system
- Need tuples of the form < context , gold_response , alternate_response , human_score > : impossible to collect a large number of these
- **IDEA**: automatically discover such tuples  and train a model on this

# Reddit Corpora

- Working with 2 data dumps of Reddit
  - 30Gbs, 53M comments from 1 month
  - 200Gbs (Compressed), all Reddit

- Pruning Rules
  - Empty/Deleted comments
  - Top level comments with no children
  - Comments with URLs
  - Comments with subreddit mentions
  - Comments with Length <= 3
  - Reply comments with Length > 40
  - Comments common across multiple contexts

# ARCtx Dataset

- Find alternate responses to the same context by matching contexts of 2 separate conversations
- Skip-thoughts sentences vectors for matching contexts

[–] **XoXFaby** Flips the script [S] 1 point 2 years ago

What is your resolution for the next year?

permalink embed save parent give gold

> [–] **UltimateSunrise** 1 point 2 years ago
>
> Balance alone time and social time better. I'm the worst about going on binges haha. I won't leave my room for two weeks then go to five events in a row :P
>
> You?

[–] **d13vs13** 4 points 2 years ago

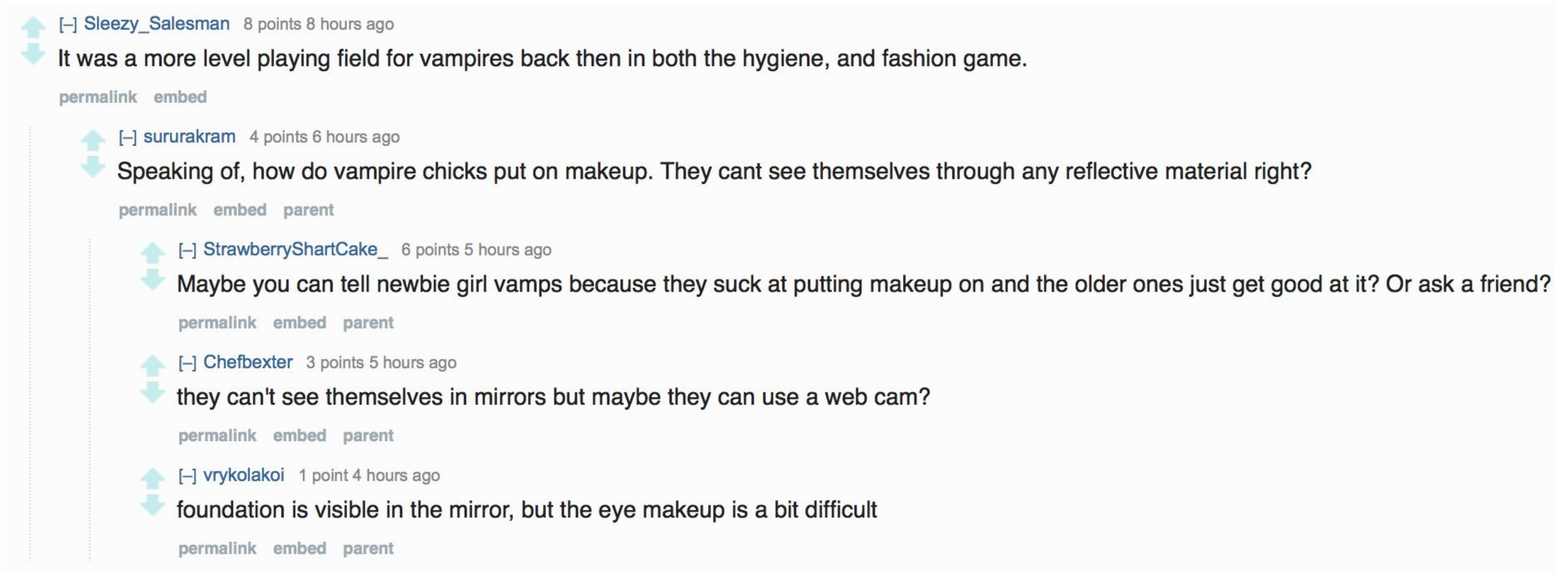
What are your new years resolutions?

permalink embed save give gold

> [–] **PittPensPats** [BrittFromPitt] 12 points 2 years ago
>
> Eat healthier, exercise more, stop being so God damn messy, get A's in my last two classes, graduate, get a job, and spend less time on my phone when around other people. I'm addicted to Twitter and reddit and it's beginning to impact my relationships with my boyfriend and my friends.

# ARCmt Dataset

- Comments on Reddits can get more replies
- We can hunt for alternate comments to any comment to build the dataset




Example:

< It was a more...right? , Maybe you can … friend? , they can’t … web cam? , 1 >

# Label Noise in ARCmt Dataset

- Manually checked 200 positive examples and 100 negative examples generated
- 90% negative examples are correct
- 80% positive examples are correct
- **TODO:** ways to deal with label noise? Better pruning? Is it even needed given the sheer amount of data

# Baseline Models

- Word-overlap based metrics
  - BLEU
  - ROUGE
- Embedding based metrics
  - Distance between skip-thought vectors of responses
  - Tfidf vectors for context and responses followed by logistic regression
  - Skip-thoughts followed by logistic regression

# Our Models

- Tfidf vectors for context and LSTM encodings of responses followed dense layer
- LSTM encodings for context and responses followed dense layer
- 2 Attention weighted representations of the context by attending using either response
- HRED model for encoding context with either response

# Results

- Data Stats
  - Training Size 6M
  - Validation Size 240K
  - Testing Size 300K

- Accuracy
  - BLEU score : 54%
  - Tfidf with Logistic Regression : 67%
  - Skip-thoughts : <Training>