Learning to Relate Literal and Sentimental Descriptions of Visual Properties

Mark Yatskar[1], Svitlana Volkova[2], Asli Celikyilmaz[3], Bill Dolan[3], Luke Zettlemoyer[1]

my89@cs.washington.edu svilana@jhu.edu asli@ieee.org billdol@microsoft.edu lsz@cs.washington.edu

University of Washington[1], Johns Hopkins University[2], Microsoft Research[3]

Introduction

- Literal language describes directly observable properties e.g., object color, shape, or category.
- Sentimental language describes high level, social, cultural, and emotional qualities.

Our contributions:

I. New dataset collected to describe Xbox avatars.

II. Models to learn the relationships between the avatars, their literal and sentimental descriptions.

Motivation

- Interactive Language Systems
  - Find me a romantic restaurant
- How to convey intent?
  - Sentimental Language
    - Short and concise
    - A summary of relevant criteria
    - Can be subjective
  - Literal Language
    - Cumbersome to produce
    - Might need technical expertise

A sleazy retro man from the 80s.
He is confident and wearing aviators.

Learned

A man with long orange hair.
He has normal eyebrows.
His eyes are covered by gold **aviators**.
He has a **horse shoe beard**.
His mouth is open and his jaw is cut.
He has a plaid shirt which is tucked in.
He has plaid pants and a belt.

Fixed

Data

Corpus of descriptions of avatars created by gamers. Avatar is specified by 19 attributes, (10^{20} possibilities).

i. Literal descriptions of specific facial features, clothing and accessories (50 / item):
   - blue button dress shirt with dark blue stripes
   - multi-blue striped long-sleeve button-up shirt

ii. Sentimental descriptions of avatars (5 / avatar)

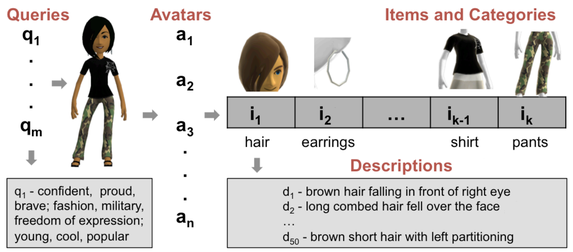
   - State of mind of the avatar (humble, satisfied)
   - Things the avatar might care about (fashion, friends, money, cars, music, education)
   - What the avatar might do for a living (teacher, singer, actor, dancer, computer engineer)
   - Overall appearance of the avatar (nerdy, smart)

iii. Multilingual literal descriptions

iv. Relative literal descriptions

v. Comprehensive full-body descriptions

Task Description



Sentimental Word Prediction

- check if a given avatar can be described with a particular sentimental word q^*.

Avatar Ranking

- rank the set of avatars according to which one best matches a sentimental description q_i.

Avatar Generation

- generate novel, unseen avatars, by selecting a set of items that best match sentimental descriptions.

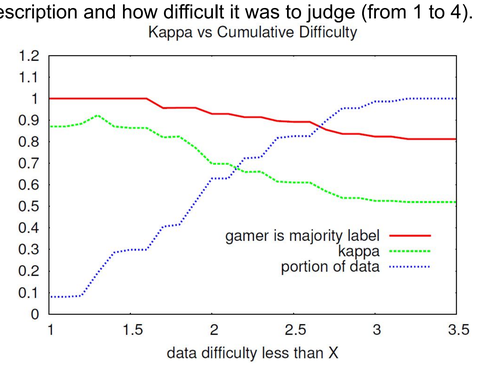
Feasibility

Sentimental language does not uniquely identify an avatar; it summarizes an overall perception.

Can we use sentimental text to select avatars?

A/B test: 100 random descriptions, 5 raters compare gamer avatars and randomly generated avatars:

- show two avatars and one sentimental description;
- ask to select which avatar is better matched by the description and how difficult it was to judge (from 1 to 4).



Sometimes random avatars are preferred over gold standard avatars; this indicates that it can be difficult to judge sentimental descriptions.

Can this avatar be specified with just literal language?

Methods

INDEPENDENT WORD MODEL (S-INDEP)

−each word independently describes the avatar
−binary classification problem for each word
−positive data: all avatars with the word $(q,\vec{a}_i,1)\ \forall i, q \in \vec{q}_i$
−negative data: $(q,\vec{a}_i,0)\ \forall i, q \notin \vec{q}_i$

$I\left(q \in \vec{q}_i, w \in \vec{d}_{a_{ij}}, j\right)$	Indicator feature: cross product of sentiment query word, a literal description word, the avatar position q = angry; w = pointy; j = eyebrows;
$I\left(q \in \vec{q}_i, a_{ij} = \varnothing, j\right)$	Bias feature to keep a position empty

Word Prediction	Avatar Ranking
Construct a separate linear model for each word in the vocab	Rank all positive instances above negative instances

Training Averaged Binary / Structured Perceptron (Collins, 02)

JOINT SENTIMENTAL MODEL (S-JOINT)

−jointly models query words to learn the relationships between literal and sentimental words with score:

$$s\left(\vec{a}\,|\vec{q},D\right) = \sum_{i=1}^{|\vec{a}|}\sum_{j=1}^{|\vec{q}|}\theta^T f\left(\vec{a}_i,\vec{q}_j,\vec{d}_{a_i}\right)$$

−every word in the query has a separate factor
−every position treated independently
−feature function is similar to the independent model

Avatar Ranking	Avatar Generation
Rank the avatar for a query above all other avatars	Score the avatar above all other valid avatars given the query

Training Averaged Structured Perceptron

Experimental Setup

SENTIMENTAL-LITERAL OVERLAP (SL-OVERLAP)

−lexical overlap in literal and sentimental descriptions

Avatar Ranking	Avatar Generation
Order avatars by the sum over every position of the # of overlapping words in q and d	For each position select the item whose literal description overlaps the most with the query

RANDOM BASELINE

Avatar Ranking	Avatar Generation
Randomly order the avatars	Select an item randomly for every position

FEATURE GENERATION

−Query vocabulary 6.1K, description vocabulary 3.5K
−400 million features that include the cross process of two vocabularies with all possible avatar positions
−Features pruned w/ stemming & freq. ≥ 10: 700k features

Results

SENTIMENT WORD PREDICTION RESULTS

Word	F-score	Precision	Recall	N
happi	0.84	0.89	0.78	149
student	0.78	0.82	0.74	129
friend	0.76	0.84	0.70	153
music	0.74	0.89	0.63	148
confid	0.74	0.82	0.76	157
sport	0.69	0.62	0.76	76
casual	0.63	0.60	0.67	84
youth	0.60	0.57	0.64	88

- Prediction is possible for a subset of words
- 33.2% zero f-score ex. unusual, bland, sarcastic, prepared

AVATAR RANKING RESULTS

	S-Joint	S-Indep	SL-Overlap	Random
Percentile	77.3	73.5	60.4	48.8

- 56.2 % of avatars marked as relevant @ 5 by humans
- many reasonable avatars for a description

S-Joint Ranking Results for the sentimental description: pensive, confrontational; music, socializing; musician, bar tending, club owner; smart, cool.

AVATAR GENERATION RESULTS

	S-Joint	SL-Overlap	Random
Item Overlap w/ Gold	0.126	0.049	0.041

	Kappa	Majority, %	Rand, %	Sys, %
SL-Overlap	0.20	0.35	0.34	0.32
S-Joint	0.52	**0.90**	0.07	0.81
Gamer	0.52	0.81	0.08	0.77

- Generated avatars as easy to distinguish from random as Gamer avatars

Ambition; business, fashion, success; salesman; smooth, professional

Stressed, bored, discontent; emo music; record store; goth, dark, drab

Happy, confident, home, career, family, secretary, student, classy, clean, casual

S-Joint generation results for the sentimental descriptions.

Conclusion

- Explored how both **literal** and **sentimental** visual language maps to the overall physical appearance
  - Some sentimental words highly ambiguous
  - Many avatars can match a description
- Allow for natural-driven dialog scenarios by formulating a high-level description of user intent