ALOHA: Artificial Learning of Human Attributes for Dialogue Agents

Aaron W. Li,
Veronica Jiang,*
Steven Y. Feng,*
Julia Sprague,
Wei Zhou,
Jesse Hoey
1David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
2Huawei Technologies Co., Ltd.
{w89li, r4jiang, sy2feng, jsprague, jhoey}@uwaterloo.ca, wei.zhou1@huawei.com

Abstract

For conversational AI and virtual assistants to communicate with humans in a realistic way, they must exhibit human characteristics such as expression of emotion and personality. Current attempts toward constructing human-like dialogue agents have presented significant difficulties. We propose Human Level Attributes (HLAs) based on tropes as the basis of a method for learning dialogue agents that can imitate the personalities of fictional characters. Tropes are characteristics of fictional personalities that are observed recurrently and determined by viewers’ impressions. By combining detailed HLA data with dialogue data for specific characters, we present a dataset, HLA-Chat, that models character profiles and gives dialogue agents the ability to learn characters’ language styles through their HLAs. We then introduce a three-component system, ALOHA (which stands for Artificial Learning of Human Attributes), that combines character space mapping, character community detection, and language style retrieval to build a character (or personality) specific language model. Our preliminary experiments demonstrate that two variations of ALOHA, combined with our proposed dataset, can outperform baseline models at identifying the language styles of specific characters. We test the performance are most similar in profile and identity, and recovers language styles based on their HLAs, determines which character space responses of chosen target characters, and are stable regardless of the character’s identity, the genre of the show, and the context of the dialogue.

1 Introduction

Attempts toward constructing human-like dialogue agents have met significant difficulties, such as maintaining conversation consistency (Zhang et al. 2018). This is largely due to inabilities of dialogue agents to engage the user emotionally because of an inconsistent personality (Rashkin et al. 2019). Many agents use personality models that attempt to map personality attributes into lower dimensional spaces (e.g. the Big Five (John and Srivastava 1999)). However, these represent human personality at a very high-level and lack depth. They prohibit the ability to link specific and detailed personality traits to characters, and to construct large datasets where dialogue is traceable back to these traits.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
* Authors contributed equally

Figure 1: Example of a character and its associated HLAs (tropes) on the left and dialogue lines on the right.

For this reason, we propose Human Level Attributes (HLAs), which we define as characteristics of fictional characters representative of their profile and identity. We base HLAs on tropes collected from TV Tropes (ttvttropes.org 2004), which are determined by viewers’ impressions of the characters. See Figure 1 for an example. Based on the hypothesis that profile and identity contribute effectively to language style (Pennebaker and King 1999), we propose that modeling conversation with HLAs is a means for constructing a dialogue agent with stable human-like characteristics. By collecting dialogues from a variety of characters along with this HLA information, we present a dataset, HLA-Chat, with novel labelling of this dialogue data traceable back to both its context and associated human-like qualities.

We also propose a system called ALOHA (Artificial Learning of Human Attributes) as a novel method of incorporating HLAs into dialogue agents. ALOHA maps characters to a latent space based on their HLAs, determines which are most similar in profile and identity, and recovers language styles of specific characters. We test the performance of ALOHA in character language style recovery against four baselines, demonstrating outperformance and system stability. We also run a human evaluation supporting our results.

Our major contributions are: (1) We propose HLAs as personality aspects of fictional characters from the audience’s perspective based on tropes; (2) We provide a large dialogue dataset, HLA-Chat, traceable back to both its context and associated human-like attributes; (3) We propose a system...
called ALOHA that is able to recommend responses linked to specific characters. We demonstrate that ALOHA, combined with the proposed dataset, HLA-Chat, outperforms baselines. ALOHA also shows stable performance regardless of the character’s identity, genre of the show, and context of the dialogue. We release all of ALOHA’s data and code along with additional information for reproduction.1

2 Related Work

Task completion chatbots (TCC), or task-oriented chatbots, are dialogue agents used to fulfill specific purposes, such as helping customers book airline tickets, or a government inquiry system. Examples include the AIML based chatbot (Satu, Parvez, and others 2015) and DIVA Framework (Xuetao, Bouchet, and Sansonnet 2009). While TCC are low cost, easily configurable, and readily available, they are restricted to working well for particular domains and tasks.

Open-domain chatbots are more generic dialogue systems. An example is the Poly-encoder from Humenau et al. (2019). It outperforms the Bi-encoder (Mazaré et al. 2018; Dinan et al. 2018) and matches the performance of the Cross-encoder (Wolf et al. 2019) while maintaining reasonable computation time. It performs strongly on downstream language understanding tasks involving pairwise comparisons, and demonstrates state-of-the-art results on the ConvAI2 challenge (Dinan et al. 2019). Feed Yourself (Hancock et al. 2019) is an open-domain dialogue agent with a self-feeding model. When the conversation goes well, the dialogue becomes part of the training data, and when the conversation does not, the agent asks for feedback. Lastly, Kvmemnn (Eric and Manning 2017) is a key-value memory network with a knowledge base that uses a key-value retrieval mechanism to train over multiple domains simultaneously. We use all three of these models as baselines for comparison. While these can handle a greater variety of tasks, they do not respond with text that aligns with particular human-like characteristics.

Li et al. (2016) defines persona (composite of elements of identity) as a possible solution at the word level, using backpropagation to align responses via word embeddings. Burtl and Spanakis (2017) uses sentence embeddings and a retrieval model to achieve higher accuracy on dialogue context. Liu et al. (2019) applies emotion states of sentences as encodings to select appropriate responses. Pichl et al. (2018) uses knowledge aggregation and hierarchy of sub-dialogues for high user engagement. Mairesse and Walker (2007)’s PERSONAGE focuses on generating language using the extraversion personality trait of the Big Five. However, these agents all represent personality at a high-level and lack detailed human qualities. We model language styles through HLAs which are much more detailed and specific. Hence, the language styles we are recovering may likely capture additional information.

3 Methodology

3.1 Human Level Attributes (HLA)

We collect HLA data from TV Tropes (tv.tropes.org 2004), a knowledge-based website dedicated to pop culture, containing information on characters from a variety of sources. Similar to Wikipedia, its content is provided and edited collaboratively by a massive user-base. These attributes are determined by human viewers and their impressions of the characters, and are correlated with human-like characteristics. Furthermore, many tropes include context information (e.g. jealous girlfriend) versus high-level personality models such as the Big Five. We believe that TV Tropes is better for our purpose of fictional character modeling than data sources used in works such as Shuster et al. (2019) because TV Tropes’ content providers are rewarded for correctly providing content through community acknowledgement.

TV Tropes defines tropes as attributes of storytelling that the audience recognizes and understands. We use tropes as HLAs to calculate correlations with specific target characters. We collect data from numerous characters from a variety of TV shows, movies, and anime. We filter and keep characters with at least five HLA, as those with fewer are not complex enough to be correctly modeled due to reasons such as lack of data. We end up eliminating 5.86% of total characters, and end up with 45,821 characters and 12,815 unique HLA, resulting in 945,519 total character-HLA pairs. Each collected character has 20.64 HLAs on average. See Figure 1 for an example character and their HLAs.

3.2 Overall Task

Our task is the following, where $t$ denotes “target”:

*Given a target character $c_t$ with HLA set $H_t$, recover the language style of $c_t$ without any dialogue of $c_t$ provided.*

For example, if Sheldon Cooper from The Big Bang Theory is $c_t$, then $H_t$ is the set of HLA on the left side of Figure 1.

We define the language style of a character as its diction, tone, and speech patterns. It is a character specific language model refined from a general language model. We must learn to recover the language style of $c_t$ without its dialogue as our objective is to imitate human-like qualities, and hence the model must understand the language styles of characters based on their traits. If we feed $c_t$’s dialogue during training, the model will likely not effectively learn to imitate language styles based on HLAs, but based on the correlation between text in the training and testing dialogues (Joshi et al. 2019).

We define character space as the character representations within the HLA latent space (see Figure 2), and the set $C = \{c_1, c_2, ..., c_n\}$ as the set of all characters. We define Observation (OBS) as the input that is fed into any dialogue model. This can be a single or multiple lines of dialogue along with any additional information. The goal of the dialogue model is to find the best response to this OBS. We show an example of our model’s responses imitating five distinct characters in Table 1.

3.3 ALOHA

We propose a three-component system called ALOHA to solve the task (see Figure 3). The first component, Charac-
ter Space Module (CSM), generates the character space and calculates confidence levels using singular value decomposition (Sarwar et al. 2000) between characters $c_j$ (for $j = 1$ to $n$ where $j \neq t$) and $c_t$ in the HLA-oriented neighborhood.

The second component, Character Community Module (CCM), ranks the similarity between our target character $c_t$ with any other character $c_j$ by the relative distance between them in the character space.

The third component, Language Style Recovery Module (LSRM), recovers the language style of our character space using the CSM. Our results demonstrate higher accuracy at retrieving $c_t$’s ground truth response. Our system is able to pick responses which are correct both in context as well as character space.

Hence, the overall process for ALOHA works as follows. First, given a set of characters, determine the character space using the CSM. Next, given a specific target character, determine the positive community and negative set of associated characters using the CCM. Lastly, using the positive community and negative set determined above along with a dialogue dataset, recover the language style of the target character.

3.4 Character Space Module (CSM)

CSM learns how to rank characters. We can measure the interdependencies between the HLA variables (Hu, Koren, and Volinsky 2008) and rank the similarity between the TV show characters. We use implicit feedback instead of neighborhood models (e.g. cosine similarity) because it can compute latent factors to transform both characters and HLAs into the same latent space, making them directly comparable.

We define a matrix $P$ that contains binary values, with $P_{u,i} = 1$ if character $u$ has HLA $i$ in our dataset, and $P_{u,i} = 0$ otherwise. We define a constant $\alpha$ that measures our confidence in observing various character-HLA pairs as positive. $\alpha$ controls how much the model penalizes the error if the ground truth is $P_{u,i} = 1$. If $P_{u,i} = 1$ and the model guesses incorrectly, we penalize by $\alpha$ times the loss. But if $P_{u,i} = 0$ and the model guesses a value greater than 0, we do not penalize as $\alpha$ has no impact. This is because $P_{u,i} = 0$ can either represent a true negative or be due to a lack of data, and hence is less reliable for penalization. See Equation 1. We find that using $\alpha = 20$ provides decent results.

We further define two dense vectors $X_u$ and $Y_i$. We call $X_u$ the “latent factors for character $u$”, and $Y_i$ the “latent factors for HLA $i$”. The dot product of these two vectors produces a value ($X_u^T Y_i$) that approximates $P_{u,i}$ (see Figure 4). This is analogous to factoring the matrix $P$ into two separate matrices, where one contains the latent factors for characters, and the other contains the latent factors for HLAs. We find that $X_u$ and $Y_i$ being 36-dimensional produces decent results. To bring $X_u^T Y_i$ as close as possible to $P_{u,i}$, we minimize the following loss function using the Conjugate Gradient Method (Takács, Pilászy, and Tikk 2011):

$$loss = \sum_{u} \sum_{i} (\alpha P_{u,i} - X_u^T Y_i)^2 + \lambda(||X_u||^2 + ||Y_i||^2)$$

(1)

The first term penalizes differences between the model’s prediction ($X_u^T Y_i$) and the actual value ($P_{u,i}$). The second term is an L2 regularizer to reduce overfitting. We find $\lambda = 100$ provides decent results for 500 iterations (see Section 5.1).

3.5 Character Community Module (CCM)

CCM aims to divide characters (other than $c_t$) into a positive community and a negative set. We define this positive
community as characters that are densely connected internally to $c_t$ within the character space, and the negative set as the remaining characters. We can then sample dialogue from characters in the negative set to act as the distractors (essentially negative samples) during LSRM training.

As community finding is an ill-defined problem (Fortunato and Hric 2016), we choose to treat CCM as a simple undirected, unweighted graph. We use the values learned in the CSM for $X_u$ and $Y_t$ for various values of $u$ and $t$, which approximate the matrix $P$. Similar to Hu, Koren, and Volinsky (2008), we can calculate the correlation between two rows (and hence two characters).

We then employ a two-level connection representation by ranking all characters against each other in terms of their correlation with $c_t$. For the first level, the set $S_i^{FL}$ is the top 10% most highly correlated characters with $c_t$ out of the 45,820 total other characters we have HLA data for. For the second level, for each character $s_i$ in $S_i^{FL}$, we determine the 30 most heavily correlated characters with $s_i$ as set $S_i^{SL}$. The positive set $S_i^{pos}$ are the characters present in at least 10 $S_i^{SL}$ sets. We call this value 10 the minimum frequency. All other characters in our dialogue dataset make up the negative set $S_i^{neg}$. These act as our positive community and negative set, respectively. See Figure 5 for an example.

### 3.6 Language Style Recovery Module (LSRM)

LSRM creates a dialogue agent that aligns with observed characteristics of human characters by using the positive character community and negative set determined in the CCM, along with a dialogue dataset, to recover the language style of $c_t$ without its dialogue. We use the BERT bi-ranker model from the Facebook ParlAI framework (Miller et al. 2017) and the Poly-encoder (Humeau et al. 2019), where the models have the ability to retrieve the best response out of 20 candidate responses. These are trained to produce LSRM-BERT and LSRM-Poly, respectively. (Dinan et al. 2019; Urbanek et al. 2019; Zhang et al. 2018) choose 20 candidate responses, and for comparison purposes, we do the same.

**BERT** (Devlin et al. 2019) is first trained on massive amounts of unlabeled text data. It jointly conditions on text on both the left and right, which provides a deep bidirectional representation of sentence inference. BERT is proven to perform well on a wide range of tasks by simply fine-tuning on one additional layer. We are interested in its ability to predict the next sentence, called Next Sentence Prediction. We perform further fine-tuning on BERT for our target character language style retrieval task to produce our LSRM-BERT model by optimizing both the encoding layers and the additional layer. We use BERT to create vector representations for the OBS and for each candidate response. By passing the first output of BERT’s 12 layers through an additional linear layer, these representations can be obtained as 768-dimensional sentence-level embeddings. It uses the dot product between these embeddings to score candidate responses and is trained using the ranking loss.

**Poly-encoder** (Humeau et al. 2019) is a transformer architecture that learns global rather than local level token features to perform attention on. The model has state-of-the-art accuracy on response retrieval on the Persona-Chat dataset. As in the Bi-encoder, a given candidate response is first encoded into a vector. Then, softmax attention against multiple context vectors encoded from the input observation is performed to compute the final score.

**Candidate response selection** is similar to the procedure from previous work done on grounded dialogue agents (Zhang et al. 2018; Urbanek et al. 2019). Along with the ground truth response, we randomly sample 19 distractor responses from other characters from a uniform distribution of characters, and call this process uniform character sampling. Based on our observations, this random sampling provides multiple context correct responses. Hence, the BERT bi-ranker model is trained by learning to choose context correct responses, and the model learns to recover a domain-general language model that includes training on every character. This results in a Uniform Model that can select context correct responses, but not responses corresponding to a target character with specific HLAs.

We then fine-tune on the above model to produce our LSRM-BERT model with a modification: we randomly sam-

---

Figure 4: Illustration of our Collaborative Filtering procedure. Green check-marks indicate a character having an HLA, and ‘X’ indicates otherwise. We randomly mask 30% of this data for validation, as marked by the ‘?’.

Figure 5: Illustration of the two-level connection representation procedure, using a minimum frequency of two. The transparent red circle indicates the first level set ($S_i^{FL}$), while the blue ones indicate the sets $S_i^{SL}$. The lines indicate connections between the characters within the community structure of $c_t$.
ple the 19 distractor responses from only the negative character set instead. We choose the responses that have similar grammatical structures and semantics to the ground truth response, and call this process negative character sampling. This guides the model away from the language style of these negative characters to improve performance at retrieving responses for target characters with specific HLAs.

We train a second version of our LSRM, LSRM-Poly, by training Poly-encoder directly on HLA-Chat using negative character sampling following the same procedure as above with 19 distractor responses. Our results demonstrate higher accuracy from both ALOHA-BERT and ALOHA-Poly variations of our system at retrieving the correct response from character $c_i$, which is the ground truth.

## 4 Experiment

### 4.1 Dialogue Dataset and HLA-Chat

To train the Uniform Model and LSRM, we collect dialogues from 327 major characters (a subset of the 45,821 characters we have HLA data for) in 38 TV shows from various existing sources of clean data on the internet, resulting in a total of 1,042,647 dialogue lines. We use a setup similar to the Persona-Chat dataset (Zhang et al. 2018) and Cornell Movie-Dialogs Corpus (Danesu-Niculescu-Mizil and Lee 2011), as our collected dialogues are also paired in terms of valid conversations. See the right side of Figure 1 for an example of these dialogue lines. We combine these dialogue lines with our collected HLA (tropes) data for these characters to form our proposed dataset, HLA-Chat.

### 4.2 HLA Observation Guidance (HLA-OG)

We define HLA Observation Guidance (HLA-OG) as explicitly passing a small subset of the most important HLAs of a given character as part of the OBS rather than just an initial line of dialogue. This is adapted from the process used in Zhang et al. (2018) and Wolf et al. (2019) which we call Persona Profiling. Specifically, we pass eight HLAs that are randomly drawn from the top 40 most important HLAs of the character. We train the Uniform Model using No HLA-OG by explicitly passing eight HLAs of ‘none’ along with the initial line of dialogue as the OBS. We use HLA-OG during training of the LSRM (both BERT and Poly-encoder variations) and testing of all models. This is because the baselines (see Section 5.3) already follow a similar process (Persona Profiling) for training. For testing, HLA-OG is necessary as it provides information about which HLAs the models should attempt to imitate in their response selection. Just passing an initial line of dialogue without HLAs replicates a typical dialogue response task based only on context correctness. See Table 2.

### 4.3 Training Details

**BERT bi-ranker** is a baseline model trained by us on the Persona-Chat dataset. Similar to Zhang et al. (2018), we cap the length of the OBS at 360 tokens and the length of each candidate response at 72 tokens. We use a batch size of 80, learning rate of 5e-5, and perform warm-up updates for 1000 iterations. The learning rate scheduler uses SGD optimizer with Nesterov’s accelerated gradient descent (Sutskever et al. 2013) and is set to have a decay of 0.4 and to reduce on plateau. We initialize using pretrained fastText embeddings.

**LSRM-BERT** is produced by finetuning on the Uniform Model discussed above using HLA-Chat and negative character sampling. We use the same hyperparameters as the BERT bi-ranker along with half-precision operations (i.e. float16 operations) to increase batch size as recommended (Humeau et al. 2019). We initialize using pretrained fastText embeddings.

**LSRM-Poly** is produced by training the Poly-encoder directly on HLA-Chat using negative character sampling. Humeau et al. (2019) introduced an effective pretraining procedure on Reddit data which we fine-tune on. Other than using a smaller batch size of 80, we adapt all parameters used in Humeau et al. (2019): Adam optimizer with learning rate of 2e-4, $\beta_1 = 0.9, \beta_2 = 0.98$, no L2 weight decay, linear learning rate warmup, and inverse square root decay of the learning rate. We refer to our Poly-encoder model as ALOHA-Poly.

## 5 Evaluation

### 5.1 CSM Evaluation

We begin by evaluating the ability of the CSM component of our system to correctly generate the character space. To do so, during training, 30% of the character-HLA pairs (which are either 0 or 1) are masked, and this is used as a validation.

Table 2: Example for Persona Profiling, HLA-OG, and No HLA-OG. All lines under OBS are fed together as input to the language style retrieval model.

<table>
<thead>
<tr>
<th>Persona Profiling</th>
<th>HLA-OG</th>
<th>No HLA-OG</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBS:</td>
<td>OBS:</td>
<td>OBS:</td>
</tr>
<tr>
<td>Ground Truth Response:</td>
<td>All right, based on a cursory reading, it doesn’t look like you have much of a case, Sheldon.</td>
<td>All right, based on a cursory reading, it doesn’t look like you have much of a case, Sheldon.</td>
</tr>
</tbody>
</table>

Ground Truth Response:

Ground Truth Response:

2Our dataset has much more dialogue per character compared to Persona-Chat and Cornell Movie-Dialogs Corpus as we need sufficient data to learn each character’s language style.

3Tokens here refer to the WordPiece tokens used by BERT.

4We are able to recover up to 78% Hits@1 accuracy on Persona-Chat (see Section 5.4).
set (see Figure 4). For each character \( c \), the model generates a list of the 12,815 unique HLA ranks similarly to Hu, Koren, and Volinsky (2008) for \( c \). We look at the recall of our CSM model, which measures the percentage of total ground truth HLA ranks (over all characters \( c \)) present within the top \( N \) ranked HLAs for all \( c \) by our model. That is:

\[
\text{recall} = \frac{\sum_c |H L A_c^g \cap H L A_c^{tN}|}{\sum_c |H L A_c^g|} \tag{2}
\]

where \( H L A_c^g \) are the ground truth HLAs for \( c \), and \( H L A_c^{tN} \) are the top \( N \) ranked HLAs by the model for \( c \). We use \( N = 100 \), and our model achieves 25.08% recall.

To inspect the CSM performance, we use the T-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008) to reduce each high-dimensionality data point to two-dimensions via Kullback-Leibler Divergence (Kullback and Leibler 1951). This allows us to map our character space into two-dimensions, where similar characters from our embedding space have higher probability of being mapped close by. We sampled characters from four different genres or groups. As seen in Figure 2, our learned character space effectively groups these characters, as similar characters are adjacent to one another in four regions.

### 5.2 Automatic Evaluation Setup

#### Five-Fold Cross Validation

is used for the training and testing of the Uniform Model and two LSRM variations. The folds are divided randomly by the TV shows in our dialogue data. We use the dialogue data for 80% of these shows as the four-folds for training, and the dialogue data for the remaining 20% as the fifth-fold for testing. The dialogue data used is discussed in Section 4.1. This ensures no matter how our data is distributed, each part of it is tested, allowing our evaluation to be more robust to different characters. See Table 3 for detailed statistics.

#### Five Evaluation Characters

are chosen, one from each of the five testing sets described above. Each is a well-known character from a separate TV show, and acts as a target character \( c_t \) for evaluation of every model. We choose Sheldon Cooper from The Big Bang Theory, Jean-Luc Picard from Star Trek, Monica Geller from Friends, Gil Grissom from CSI, and Marge Simpson from The Simpsons. We choose characters of significantly different identities and profiles (intelligent scientist, ship captain, outgoing friend, police leader, and responsible mother, respectively) from shows of a variety of genres to ensure that we can successfully recover the language styles of various types of characters. We choose well-known characters because humans require knowledge on the characters they are evaluating (see Section 5.5).

For each of these five evaluation characters, all the dialogue lines from the character act as the ground truth responses. The initial dialogue lines are the corresponding dialogue lines to which these ground truth responses are responding. For each initial dialogue line, we randomly sample 19 other candidate responses from the associated testing set using uniform character sampling. Note that this is for evaluation, and hence we use the same uniform character sampling method for all models including ALOHA. The use of negative character sampling is only in ALOHA’s training.

### 5.3 Baselines

We compare against four dialogue system baselines: Kvmemnn, Feed Yourself, Poly-encoder, and a BERT bi-ranker baseline trained on the Persona-Chat dataset using the same training hyperparameters (including learning rate scheduler and length capping settings) described in Section 4.3.\(^5\)

For the first three models, we use the provided pretrained (on Persona-Chat) models. For the Poly-encoder baseline, we use the official model trained on ConvAI2. We evaluate all four baselines on our five evaluation characters discussed in Section 5.2.

### 5.4 Key Evaluation Metrics

#### Hits@n/N

is the accuracy of the correct ground truth response being within the top \( n \) ranked candidate responses out of \( N \) total candidates. We measure Hits@1/20, Hits@5/20, and Hits@10/20.

#### Mean Rank

is the average rank that a model assigns the ground truth response among the 20 total candidates.

#### Mean Reciprocal Rank (MRR)

looks at the mean of the multiplicative inverses of the rank of each correct answer out of a sample of queries \( Q \):

\[
\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{3}
\]

where \( \text{rank}_i \) refers to the rank position of the correct response for the \( i \)-th query, and \(|Q|\) refers to the total number of queries in \( Q \).

\(^5\)See Section 2 for more details about the first three models.

---

<table>
<thead>
<tr>
<th>TV SHOW TRAINING/TESTING SPLITS</th>
<th>Fold 1</th>
<th>Fold 2</th>
<th>Fold 3</th>
<th>Fold 4</th>
<th>Fold 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total shows (Train Fold)</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Total shows (Test Fold)</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>DIALOGUE COUNT INFO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total dialogue (Train Fold)</td>
<td>764168</td>
<td>836414</td>
<td>852802</td>
<td>929025</td>
<td>788179</td>
</tr>
<tr>
<td>Total dialogue (Test Fold)</td>
<td>278479</td>
<td>206233</td>
<td>189845</td>
<td>113622</td>
<td>254468</td>
</tr>
<tr>
<td>Total target dialogue (Testing)</td>
<td>9133</td>
<td>11036</td>
<td>7448</td>
<td>6100</td>
<td>10244</td>
</tr>
<tr>
<td>CHARACTER COUNT INFO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total characters (Train Fold)</td>
<td>256</td>
<td>255</td>
<td>276</td>
<td>275</td>
<td>246</td>
</tr>
<tr>
<td>Total characters (Test Fold)</td>
<td>71</td>
<td>72</td>
<td>51</td>
<td>52</td>
<td>81</td>
</tr>
<tr>
<td>Total characters (Testing)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CHARACTER HLA COUNTS AND POSITIVE/NEGATIVE SPLITS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target character HLA count</td>
<td>217</td>
<td>125</td>
<td>97</td>
<td>43</td>
<td>110</td>
</tr>
<tr>
<td>Positive community (to target)</td>
<td>125</td>
<td>101</td>
<td>122</td>
<td>106</td>
<td>86</td>
</tr>
<tr>
<td>Negative set (to target)</td>
<td>131</td>
<td>154</td>
<td>154</td>
<td>169</td>
<td>160</td>
</tr>
</tbody>
</table>

Table 3: Detailed five-fold cross validation statistics.
BLEU scores as the average scores of 1 to 4-grams. Two pieces of text are in content and structure, with higher BLEU in the ground truth response contained in the chosen one. Precision is the fraction of words in the chosen response contained in the candidate responses to prevent human participants from responding with for each of the ten initial dialogue lines. See Figure 6 for an example. We mask any character names within the candidate responses to prevent human participants from knowing which show the response is from. We conduct a human evaluation to get an upper bound on the other candidate responses. For example, dialogue correctness. Both ALOHA variations perform similarly and so a random guess over these context-correct responses indicates human accuracy is around 40%, demonstrating that our target metric, which is the ability to retrieve the correct response of a target character by its HLAs. However, from manual observation, we noticed that in the 20 chosen candidate responses, there are typically numerous context correct responses, but only one ground truth for the target character (for an example, see Figure 6).

To investigate this, we randomly chose 50 sets of input and candidate responses (a total of 1000 candidate responses: 10 sets per target character and 20 responses per set), and manually labelled the number of context correct responses for each set. We found a total of 333 context correct responses (79, 71, 68, 53, 62 for characters 1 to 5 respectively) which means an average of 6.66 (out of 20) per input, and so a random guess over these context-correct responses would give an accuracy of 15%. Our empirical results indicate human accuracy is around 40%, demonstrating that humans make a choice relying on much more than just context correctness. Both ALOHA variations perform similarly (around 41%) and show that human performance is seemingly achievable by our system.

2. Retrieving responses for the target character depends on the other candidate responses. For example, dialogue

Table 4: Average automatic evaluation on HLA-OG and human evaluation results. Bold indicates the best performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Hits@1/20</th>
<th>Hits@5/20</th>
<th>Hits@10/20</th>
<th>Mean Rank</th>
<th>MRR</th>
<th>F1-score</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kvmemnn</td>
<td>0.1236</td>
<td>0.1117</td>
<td>0.1419</td>
<td>0.1289</td>
<td>0.1097</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feed yourself</td>
<td>0.0590</td>
<td>0.0492</td>
<td>0.0514</td>
<td>0.0359</td>
<td>0.0455</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bert Bi-ranker</td>
<td>0.2468</td>
<td>0.2959</td>
<td>0.2551</td>
<td>0.2848</td>
<td>0.2071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poly-encoder</td>
<td>0.3176</td>
<td>0.3201</td>
<td>0.3816</td>
<td>0.3564</td>
<td>0.2527</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALOHA-BERT</td>
<td>0.3525</td>
<td>0.4357</td>
<td>0.4160</td>
<td>0.4790</td>
<td>0.3717</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALOHA-Poly</td>
<td>0.4114</td>
<td>0.4585</td>
<td>0.4110</td>
<td>0.4618</td>
<td>0.3356</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>0.4833</td>
<td>0.4000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.3500</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Average Hits@1/20 scores by evaluation character on HLA-OG data. Bold indicates the best performance (excluding humans).

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Model</th>
<th>Sheldon</th>
<th>Picard</th>
<th>Monica</th>
<th>Grissom</th>
<th>Marge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kvmemnn</td>
<td>0.1236</td>
<td>0.1117</td>
<td>0.1419</td>
<td>0.1289</td>
<td>0.1097</td>
<td></td>
</tr>
<tr>
<td>Feed yourself</td>
<td>0.0590</td>
<td>0.0492</td>
<td>0.0514</td>
<td>0.0359</td>
<td>0.0455</td>
<td></td>
</tr>
<tr>
<td>Bert Bi-ranker</td>
<td>0.2468</td>
<td>0.2959</td>
<td>0.2551</td>
<td>0.2848</td>
<td>0.2071</td>
<td></td>
</tr>
<tr>
<td>Poly-encoder</td>
<td>0.3176</td>
<td>0.3201</td>
<td>0.3816</td>
<td>0.3564</td>
<td>0.2527</td>
<td></td>
</tr>
<tr>
<td>ALOHA-BERT</td>
<td>0.3525</td>
<td>0.4357</td>
<td>0.4160</td>
<td>0.4790</td>
<td>0.3717</td>
<td></td>
</tr>
<tr>
<td>ALOHA-Poly</td>
<td>0.4114</td>
<td>0.4585</td>
<td>0.4110</td>
<td>0.4618</td>
<td>0.3356</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>0.4833</td>
<td>0.4000</td>
<td>0.5000</td>
<td>0.5000</td>
<td>0.3500</td>
<td></td>
</tr>
</tbody>
</table>

6 Results and Analysis

6.1 Evaluation Results

Table 4 shows average results of our automatic and human evaluations. Table 5 shows average Hits@1/20 scores by evaluation character. See Table 1 for a demo interaction example between a human and ALOHA-Poly for all five evaluation characters.

6.2 Evaluation Challenges

The evaluation of our task (retrieving the language style of a specific character) is challenging and hence the five-fold cross validation is necessary for the following reasons:

1. The ability to choose a context correct response without attributes of specific characters may be hard to separate from our target metric, which is the ability to retrieve the correct response of a target character by its HLAs. However, from manual observation, we noticed that in the 20 chosen candidate responses, there are typically numerous context correct responses, but only one ground truth for the target character (for an example, see Figure 6).

To investigate this, we randomly chose 50 sets of input and candidate responses (a total of 1000 candidate responses: 10 sets per target character and 20 responses per set), and manually labelled the number of context correct responses for each set. We found a total of 333 context correct responses (79, 71, 68, 53, 62 for characters 1 to 5 respectively) which means an average of 6.66 (out of 20) per input, and so a random guess over these context-correct responses would give an accuracy of 15%. Our empirical results indicate human accuracy is around 40%, demonstrating that humans make a choice relying on much more than just context correctness. Both ALOHA variations perform similarly (around 41%) and show that human performance is seemingly achievable by our system.

2. Retrieving responses for the target character depends on the other candidate responses. For example, dialogue
retrieval performance for Grissom from CSI, which is a crime/police context, is higher than other evaluation characters (see Table 5), potentially due to other candidate responses not falling within the same crime/police context.

6.3 Performance: ALOHA vs. Humans
As observed from Tables 4 and 5, ALOHA (both variations) has a performance relatively close to humans. Human Hits@1/20 scores have a mean of 40.67% and a median over characters of 40%. The limited human evaluation scale limits what can be inferred, but it indicates the problem is solved to the extent that ALOHA is able to slightly outperform humans on two folds and perform closely on another two folds. Even humans do not perform extremely well, demonstrating this task is more difficult than typical dialogue retrieval tasks (Urbanek et al. 2019; Dinan et al. 2019).

Looking more closely at each character from Table 5, we can see that human evaluation scores are higher for Sheldon and Grissom. This may be due to these characters having more distinct personalities, making them more memorable. ALOHA performs worse on Sheldon compared to humans. This is possibly due to the large number of Sheldon’s HLAs (217) compared to the other four evaluation characters (average of 93.75), along with the limited amount of HLAs we are using for guidance due to the models’ limited memory.

We also look at Pearson correlation values of the Hits@1/20 scores across the five evaluation characters. For human versus Uniform Model, this is 0.047 (heavily uncorrelated), demonstrating that the Uniform Model, without knowledge of HLAs, fails to imitate human impressions. For human versus ALOHA-BERT and ALOHA-Poly, these are 0.4149 and 0.5468, respectively, demonstrating that ALOHA is able to retrieve character responses somewhat similarly to human impressions. The difference between the ALOHA variations and the Uniform Model is based on the additional knowledge of the HLAs (e.g. by using HLA-OG and negative instead of uniform character sampling). This demonstrates that HLAs are indeed an accurate method of modeling human impressions of character attributes and that ALOHA is able to effectively use them to improve upon dialogue retrieval performance.

6.4 Performance: ALOHA vs. Baselines
ALOHA, combined with HLA-Chat, achieves a significant improvement on the target character language style retrieval task compared to the baseline open-domain chatbot models across all five folds. As observed from Tables 4 and 5, ALOHA achieves a significant boost in Hits@n/N accuracy and other metrics for retrieving the correct responses from five diverse characters (see Section 5.2). Paired t-tests between the Hits@1/20 scores of ALOHA-BERT against BERT Bi-ranker and ALOHA-Poly against Poly-encoder across all five evaluation folds are statistically significant with p-values of 0.0004 and less than 0.0001, respectively, showing that the consistent improvement is meaningful.

6.5 Performance: ALOHA vs. Uniform Model
We observe a noticeable improvement in performance between ALOHA and the Uniform Model in recovering the language styles of specific characters that is consistent across all five folds (see Tables 4 and 5). Paired t-tests between the Hits@1/20 scores of ALOHA-BERT and ALOHA-Poly against the uniform model across all five evaluation folds are statistically significant with p-values of 0.0329 and 0.0234, respectively, showing that the consistent improvement is meaningful. This indicates that lack of knowledge of HLAs limits the ability of the model to successfully recover the language style of specific characters.

We claim that, to the best of our knowledge, we have made the first step in using HLA-based character dialogue clustering to improve upon personality learning for chatbots. ALOHA demonstrates an accuracy boost for all five evaluation characters, showing that the system is robust and stable and has the ability to recover the dialogue styles of fictional characters regardless of the character’s profile and identity, genre of the show, and context of the dialogue.

7 Conclusion and Future Work
We proposed Human Level Attributes (HLAs) as a novel approach to model human-like attributes of characters, and collected a large volume of dialogue data for various characters with complete HLA profiles which we release in a dataset, HLA-Chat. We also proposed and evaluated a system, ALOHA, that uses HLAs to recommend tailored responses by specific characters. We demonstrated both ALOHA-BERT and ALOHA-Poly’s outperformance of the baselines, and their ability to effectively recover language styles of various characters, showing promise for learning character or personality styles. Further, we demonstrated ALOHA’s slight outperformance of humans for two out of five evaluation characters, with close performance on two others. ALOHA was shown to be stable regardless of the character’s identity, genre of show, and context of dialogue, and ALOHA-Poly was shown to be particularly robust.

Potential directions for future work include training ALOHA with a multi-turn response approach (Zhang et al. 2018) that tracks dialogue over multiple responses. Another potential is the modeling of the dialogue counterpart (e.g. the dialogue of other characters speaking to the target character). Further, performing semantic text exchange on the chosen response with a model such as SMERTI (Feng, Li, and Hoey 2019) may improve the ability of ALOHA to converse with humans. This is because the response may be context and HLA correct, but incorrect semantically (e.g. it may say the weather is sunny when it is actually rainy). HLA-aligned generative models is another area of exploration. Typically, generative models produce text that is less fluent, but further work in this area may lead to better results. Lastly, a more diverse and larger participant pool is required due to the limited size of our human evaluation. We can also investigate other factors affecting human performance on specific characters such as their familiarity with TV series (e.g. they may be more familiar with more recent shows).

Acknowledgments
We thank our anonymous reviewers, study participants, and Huawei Technologies Co., Ltd. for financial support.
References


Satu, M. S.; Parvez, M. H.; et al. 2015. Review of integrated applications with aiml based chatbot. In 2015 International Conference on Computer and Information Engineering (ICCIE), 87–90. IEEE.


