ALOHA: Artificial Learning of Human Attributes

Aaron W. Li,[1] Veronica Jiang*,[1] **Steven Y. Feng***,[1] Julia Sprague,[1] Wei Zhou,[2] Jesse Hoey[1]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Huawei Technologies Co., Ltd.

1. Summary

- **Major Challenge** - Giving chatbots and virtual assistants human emotions/personality
- **Human-Level Attributes (HLAs)** - Propose to use HLAs to train dialogue agents based on tropes: characteristics of fictional characters representative of their profile/identity
- **HLA-Chat**: Dataset of hundreds of characters along with their HLA and dialogue data
- **Artificial Learning of Human Attributes** (**ALOHA**): Design and implement a three-stage system to retrieve character (or personality) specific language models
- **Two Variations of ALOHA**: ALOHA-BERT and ALOHA-Poly

2. Overall Task

- **Task:** 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

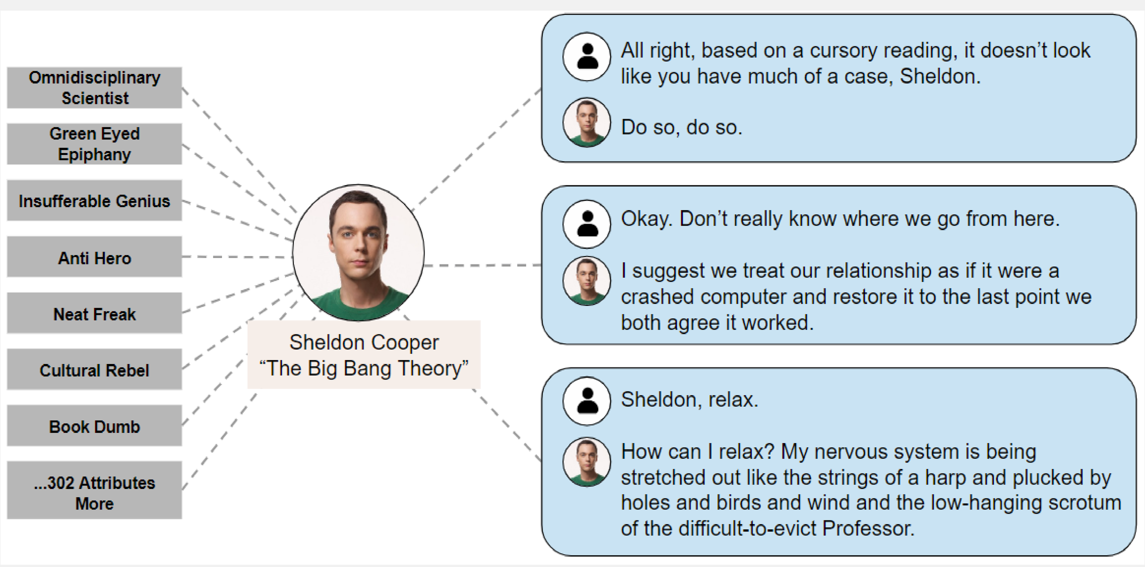

Figure 1: Example character with sample HLAs and dialogue

Input: ***What do you think of my AAAI paper?***

Sheldon (Reviewer 1): *Very good. Author of one of the greatest horror stories ever written: Turn of the Screw. And I'm looking for one.*
Picard (Reviewer 2): *Class M in all respects. Quite Earth-like.*
Marge (Reviewer 3): *I think it's awesome. I love it.*

Figure 2: ALOHA interaction example

3. Datasets and HLA-Chat

- **Tropes Dataset:** Collect tropes (HLAs) for thousands of characters from TV Tropes
- **Dialogue Dataset:** Collect dialogues from 327 major characters from 38 TV shows
- **HLA-Chat:** For each character in the dialogue dataset, we include their HLA data

4. Artificial Learning of Human Attributes (ALOHA)

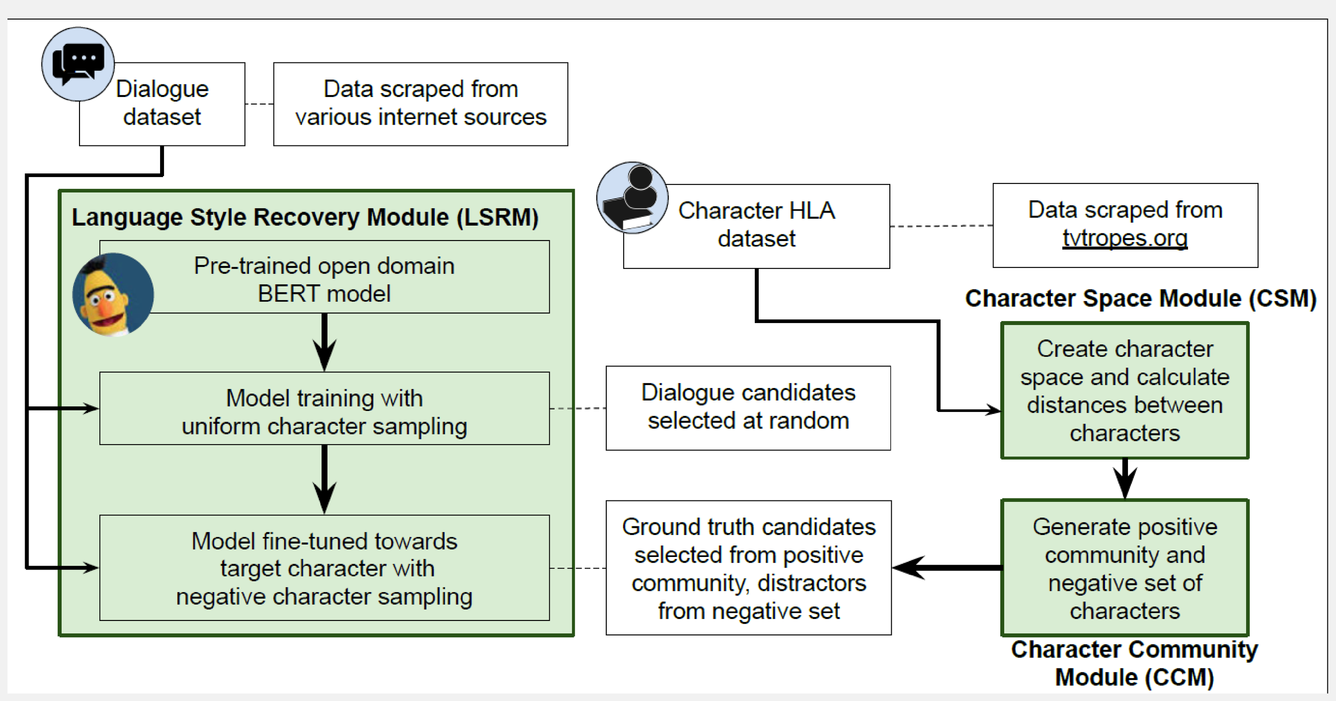

Figure 3: Diagram of ALOHA's overall architecture (ALOHA-BERT variation)

5. Character Space Module (CSM)

- **CSM** learns to rank characters based on the similarity between their HLAs
- **Collaborative filtering** procedure is used [1]

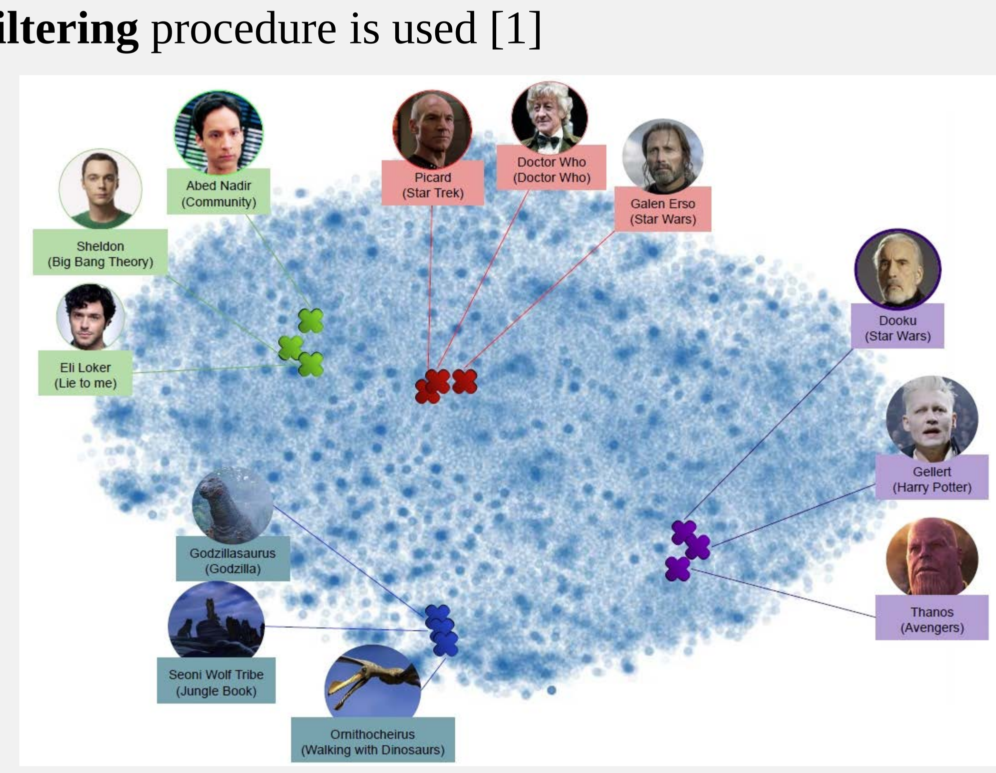

Figure 4: t-SNE visualization of character space generated by the CSM

6. Character Community Module (CCM)

- Given a target character c_t, **CCM** learns to divide other characters into a **positive community** and **negative set** using a two-level connection representation

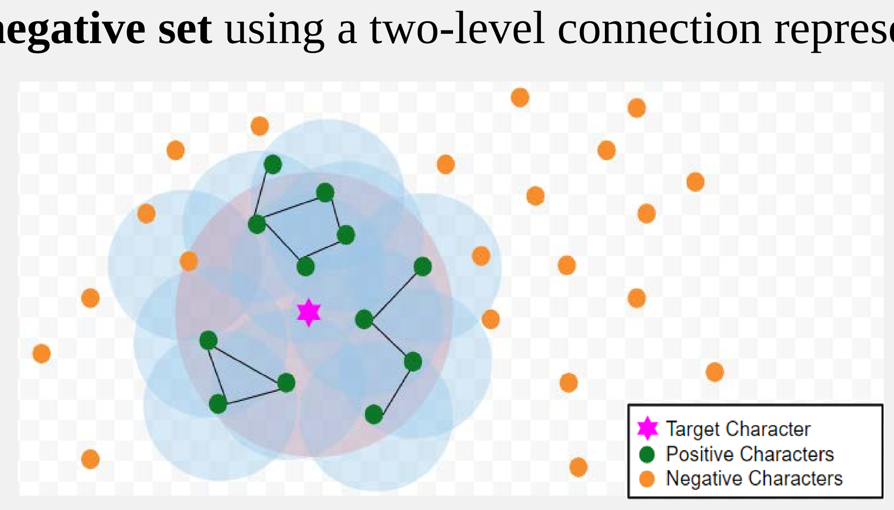

Figure 5: Illustration of two-level connection representation procedure

7. Language Style Recovery Module (LSRM)

- **LSRM**: Uses the positive community and negative set determined in the CCM along with the dialogue dataset to recover the language style of a target character c_t
- Uses the **BERT bi-ranker** [2] and **Poly-encoder** [3] to rank responses (two variations)
  - Task: choose the best response out of 20 candidate responses (by ranking them)
- **Uniform Model**: train the BERT bi-ranker on the dialogue dataset using uniform character sampling: randomly sample 19 distractor responses
- **LSRM-BERT**: fine-tuned Uniform Model on HLA-Chat with negative character sampling: randomly sample 19 distractor responses from only the negative character set
- **LSRM-Poly:** train the Poly-encoder on HLA-Chat using negative character sampling

8. Automatic Evaluation

- **Five-Fold Cross Validation**: Dialogue data for 80% of TV shows as training, and remaining 20% of TV shows for validation/testing
- **Five Evaluation Characters**: Five distinct well-known characters from different genres of TV shows, one from each test set, are chosen
  - Sheldon Cooper – The Big Bang Theory
  - Jean-Luc Picard – Star Trek
  - Monica Geller – Friends
  - Gil Grissom – CSI
  - Marge Simpson – The Simpsons
- **Baselines:**
  - BERT bi-ranker
  - Poly-encoder
  - Kvmemnn
  - Feed Yourself
- **Evaluation Metrics:**
  - Hits@n/N (1/20, 5/20, 10/20)
  - Mean Rank
  - F_1-score
  - BLEU
  - Mean Reciprocal Rank (MRR)

9. Human Evaluation

- **12 participants** from the University of Waterloo
- Each participant evaluates one or two characters (from the five evaluation characters)
- Each **questionnaire** made up of ten samples, where each sample includes an initial line of dialogue with 20 candidate responses (one of which is the ground truth)
- Each candidate **prescreened** to ensure they have sufficient knowledge of the character
- Task is to choose the **ground truth response** for the given character

10. Evaluation Results

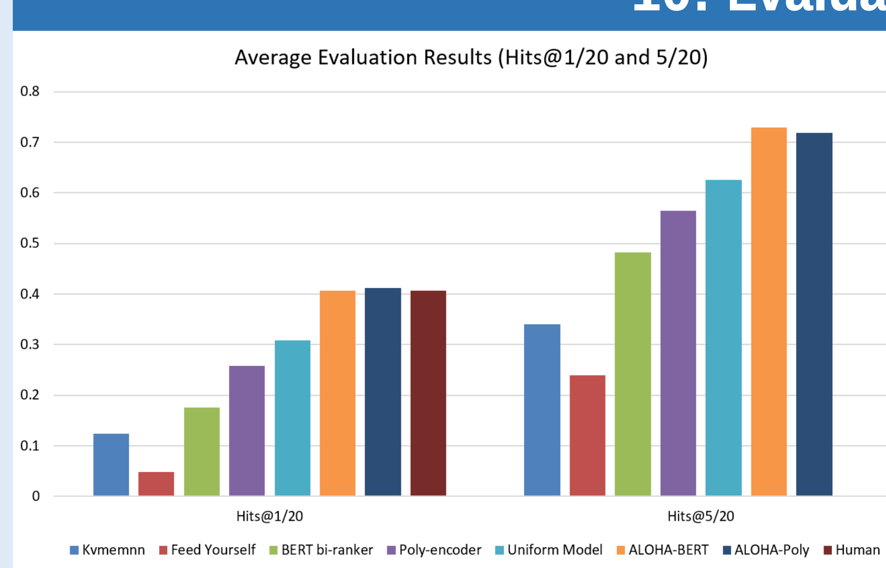

Figure 6: Average Hits@1/20, 5/20 evaluation results

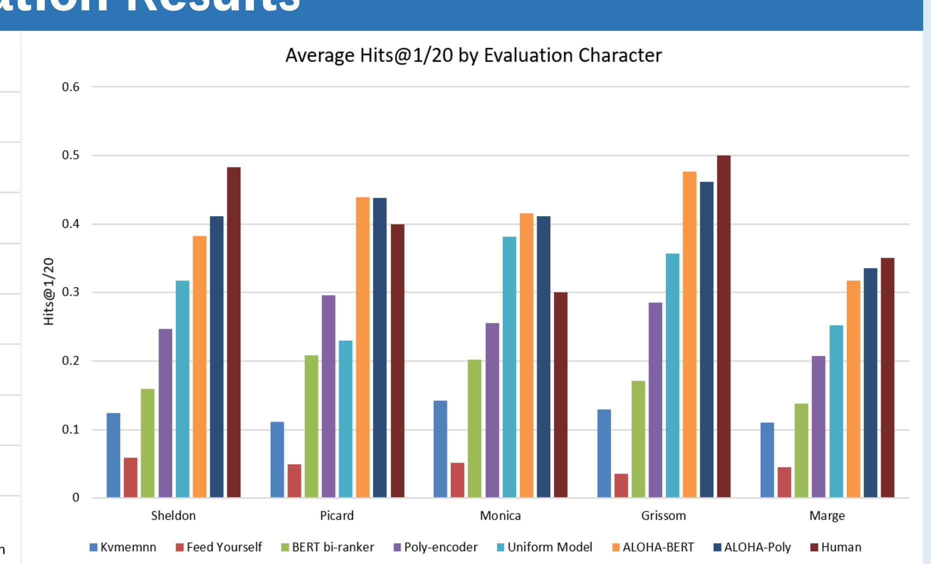

Figure 7: Average Hits@1/20 by character

11. Analysis and Discussion

- ALOHA has **closest performance** to humans
- Humans do not perform very well on this task either
- Human scores higher for **specific characters** due to their distinct personalities
- ALOHA **correlates** much better with human scores than Uniform Model, demonstrating that **HLAs accurately model human impressions** of characters
- ALOHA **outperforms all baselines** on every metric
- Lack of HLAs limits ability to recover language styles of specific characters

12. Conclusion and Future Work

- HLA-based character dialogue retrieval **improves personality learning** for chatbots
- ALOHA is **robust and stable** across all characters from a variety of TV shows
- **Future directions** for exploration:
  - Training ALOHA with a **multi-turn response** approach
  - Modeling of the **dialogue counterpart**
  - Perform **semantic text exchange** on the chosen response (e.g. using SMERTI) [4]
  - HLA-aligned **generative** models
  - Reverse direction: determining **HLAs from text**
  - Larger and more diverse **participant pool** for human evaluation

13. Acknowledgements

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

14. References

[1] Hu, Yifan, et al. "Collaborative filtering for implicit feedback datasets." IEEE International Conference on Data Mining (2008).
[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." NAACL 2019: Human Language Technologies.
[3] Humeau, Samuel, et al. "Real-time Inference in Multi-sentence Tasks with Deep Pretrained Transformers." arXiv preprint arXiv:1905.01969 (2019).
[4] Feng, Steven Y., Aaron W. Li, and Jesse Hoey. "Keep Calm and Switch On! Preserving Sentiment and Fluency in Semantic Text Exchange." EMNLP 2019.

Email: stevenyfeng@gmail.com