Conference Report

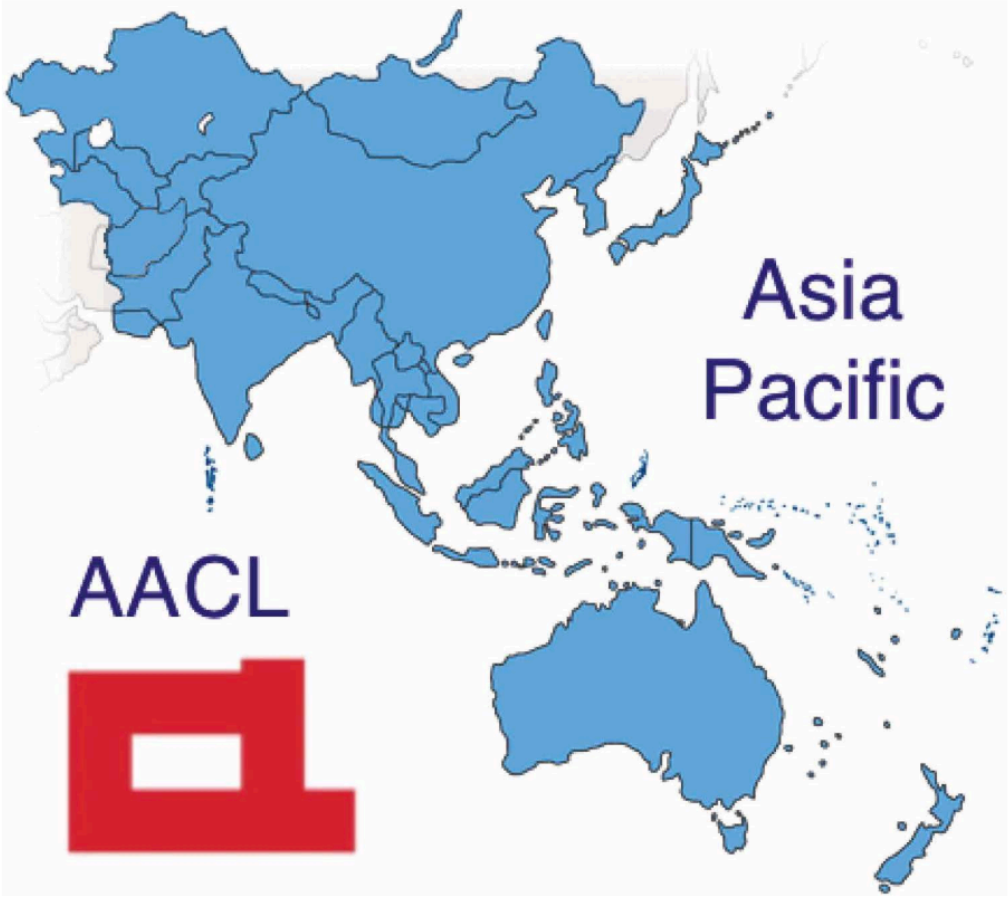
AI Lab / NLP center

Yang Zhao

Basic Statistics

- Long paper submitted/accepted/rate: 1018/256/25.1%
- Short paper submitted/accepted/rate: 526/125/23.8%
- 1610 reviewers involved
- ACL is growing!

亚太地区ACL 横空出世



Officers of AACL executive board

Position	Name	Affiliation, Country / territory	Term
Chair	Haifeng Wang	Baidu, China	2018-2020
Chair-elect	Keh-Yih Su	Institute of Information Science, Taiwan	2018-2020
Secretary	Yang Liu	Tsinghua University, China	2018-2020
Treasurer	Seung-won Hwang	Yonsei University, Korea	2018-2020
At large	Yusuke Miyao	National Institute of Informatics, Japan	2018-2020
At large	Jian Su	Institute for Infocomm Research, Singapore	2018-2020
At large	Mark Dras	Macquarie University, Australia	2018-2020

© Asia-Pacific Chapter of the Association for Computational Linguistics

The Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) provides a regional focus for members of the Association for Computational Linguistics (ACL) in Asia-Pacific, organizes annual conferences, promotes cooperation and information exchange among related scientific and professional societies, encourages and facilitates ACL membership by people and institutions in the Asia-Pacific, and provides a source of information on regional activities for the ACL Executive Committee.

My Focus - Text Generation and Summarization

1. Thoughts on text generation by Yejin Choi
2. Sentiment Adaptive End-to-End Dialog Systems
3. MOJITALK: Generating Emotional Responses at Scale
4. Personalizing Dialogue Agents: I have a dog, do you have pets too?
5. Hierarchical Neural Story Generation
6. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

Outline

1. Thoughts on text generation by Yejin Choi
2. Sentiment Adaptive End-to-End Dialog Systems
3. MOJITALK: Generating Emotional Responses at Scale
4. Personalizing Dialogue Agents: I have a dog, do you have pets too?
5. Hierarchical Neural Story Generation
6. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

Thoughts on Text Generation

Yejin Choi

- Seq2seq **good** for **strong alignment** (between source and target) task, e.g., neural machine translation
- Seq2seq **not good** for **weak-alignment** task which requires abstraction, knowledge or commonsense, e.g. chit-chatbot where response could be anything engaging
- Language models are passive learners and surface learners
  - Even RNNs need to "practice" writing
  - We also need *world* models
- Learning objective isn't quite right
  - people don't write to maximize the probability of the next token

Learning to Write with Cooperative Discriminators

Ari Holtzman[†] **Jan Buys**[†] **Maxwell Forbes**[†]
Antoine Bosselut[†] **David Golub**[†] **Yejin Choi**[†‡]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
{ahai,jbuys,mbforbes,antoineb,golubd,yejin}@cs.washington.edu

- 量的准则（The maxim of quantity)
  - 所说的话应包含当前交谈目的所需要地信息
  - 所说地话不应包含多于需要地信息
- 质的准则（The maxim of quality)
  - 不要说自知是虚假的话
  - 不要说缺乏足够证据的话
- 关联准则（The maxim of relevance):
  - 所说的话与话题要相关联
- 方式准则（The maxim of manner) 清楚明白地表达出要说的话，尤其是：
  - 避免晦涩, 避免歧义, 简练, 有条理

- **Motivation**: Long-form text generated from RNNs is often generic, repetitive, and even self-contradictory.
- **Proposed method:** a unified learning framework with four discriminators.
  - discriminators inspired by Grice's maxims (Grice+, 1975) of quantity, quality, relation, and manner.

Details

- Conditional language generation of a sequence **y** given a context **x**:

Decoding objective for generation

$$f_\lambda(\mathbf{x}, \mathbf{y}) = \log(P_{\text{lm}}(\mathbf{y}|\mathbf{x})) + \sum_k \lambda_k s_k(\mathbf{x}, \mathbf{y})$$

Language Models

Cooperative Communication Models

- When the scores s_k are log probabilities, this corresponds to a Product of Experts (PoE) model (Hinton, 2002)

- Each model is trained to discriminate between good and bad generations, produce a binary probability

Details

$$f_\lambda(\mathbf{x}, \mathbf{y}) = \log(P_{\text{lm}}(\mathbf{y}|\mathbf{x})) + \sum_k \lambda_k s_k(\mathbf{x}, \mathbf{y})$$

Language Models

Cooperative Communication Models

- Repetition Modular (maxim of Quantity) cosine similarity between word embeddings within a fixed window of the previous k words
- Entailment Model (maxim of Quality) we would like to guide the generator to neither contradict its own past generation.
- Relevance Model (maxim of Relation): predicting whether the content of a candidate continuation is relevant to the given context.
- Lexical Style Model (maxim of Manner): based on observed lexical distributions which captures writing style as expressed through word choice.

	BookCorpus					TripAdvisor				
Model	BLEU	Meteor	Length	Vocab	Trigrams	BLEU	Meteor	Length	Vocab %	Trigrams
L2W	**0.52**	**6.8**	**43.6**	**73.8**	98.9	1.7	11.0	83.8	**64.1**	**96.2**
ADAPTIVELM	**0.52**	6.3	43.5	59.0	92.7	**1.94**	**11.2**	**94.1**	52.6	92.5
CACHELM	0.33	4.6	37.9	31.0	44.9	1.36	7.2	52.1	39.2	57.0
SEQ2SEQ	0.32	4.0	36.7	23.0	33.7	1.84	8.0	59.2	33.9	57.0
SEQGAN	0.18	5.0	28.4	73.4	**99.3**	0.73	6.7	47.0	57.6	93.4
REFERENCE	100.0	100.0	65.9	73.3	99.7	100.0	100.0	92.8	69.4	99.4

Table 1: Results for automatic evaluation metrics for all systems and domains, using the original continuation as the reference. The metrics are: Length - Average total length per example; Trigrams - % unique trigrams per example; Vocab - % unique words per example.

Human evaluation shows that the quality of the text produced by our model exceeds that of competitive baselines by a large margin.

BookCorpus	**Specific Criteria**				**Overall Quality**		
L2W vs.	Repetition	Contradiction	Relevance	Clarity	Better	Equal	Worse
ADAPTIVELM	+0.48	+0.18	+0.12	+0.11	**47%**	20%	32%
CACHELM	+1.61	+0.37	+1.23	+1.21	**86%**	6%	8%
SEQ2SEQ	+1.01	+0.54	+0.83	+0.83	**72%**	7%	21%
SEQGAN	+0.20	+0.32	+0.61	+0.62	**63%**	20%	17%
LM VS. REFERENCE	-0.10	-0.07	-0.18	-0.10	41%	7 %	**52%**
L2W VS. REFERENCE	+0.49	+0.37	+0.46	+0.55	**53%**	18%	29%

TripAdvisor	**Specific Criteria**				**Overall Quality**		
L2W vs.	Repetition	Contradiction	Relevance	Clarity	Better	Equal	Worse
ADAPTIVELM	+0.23	-0.02	+0.19	-0.03	**47%**	19%	34%
CACHELM	+1.25	+0.12	+0.94	+0.69	**77%**	9%	14%
SEQ2SEQ	+0.64	+0.04	+0.50	+0.41	**58%**	12%	30%
SEQGAN	+0.53	+0.01	+0.49	+0.06	**55%**	22%	22%
LM VS. REFERENCE	-0.10	-0.04	-0.15	-0.06	38%	10%	**52%**
L2W VS. REFERENCE	-0.49	-0.36	-0.47	-0.50	25%	18%	**57%**

Table 2: Results of crowd-sourced evaluation on different aspects of the generation quality as well as overall quality judgments. For each sub-criteria we report the average of comparative scores on a scale from -2 to 2. For the overall quality evaluation decisions are aggregated over 3 annotators per example.

Outline

1. Thoughts on text generation by Yejin Choi
2. Sentiment Adaptive End-to-End Dialog Systems
3. MOJITALK: Generating Emotional Responses at Scale
4. Personalizing Dialogue Agents: I have a dog, do you have pets too?
5. Hierarchical Neural Story Generation
6. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

Sentiment Adaptive End-to-End Dialog Systems

Weiyan Shi
[24]7.ai
weiyan.shi@247.ai

Zhou Yu
University of California, Davis
joyu@ucdavis.edu

Motivation

- User's frustrating experience and even expressed anger towards automated customer service systems

- Traditional sentiment response modular in chatbot are strictly-written rules based and hard to train, to update with new data and to debug errors.

- No previous work tried to incorporate sentiment information in the end-to-end trainable systems.

Solution: they manually annotated 50 dialogs consisting of 517 conversation turns for user sentiment (negative, neutral and positive).

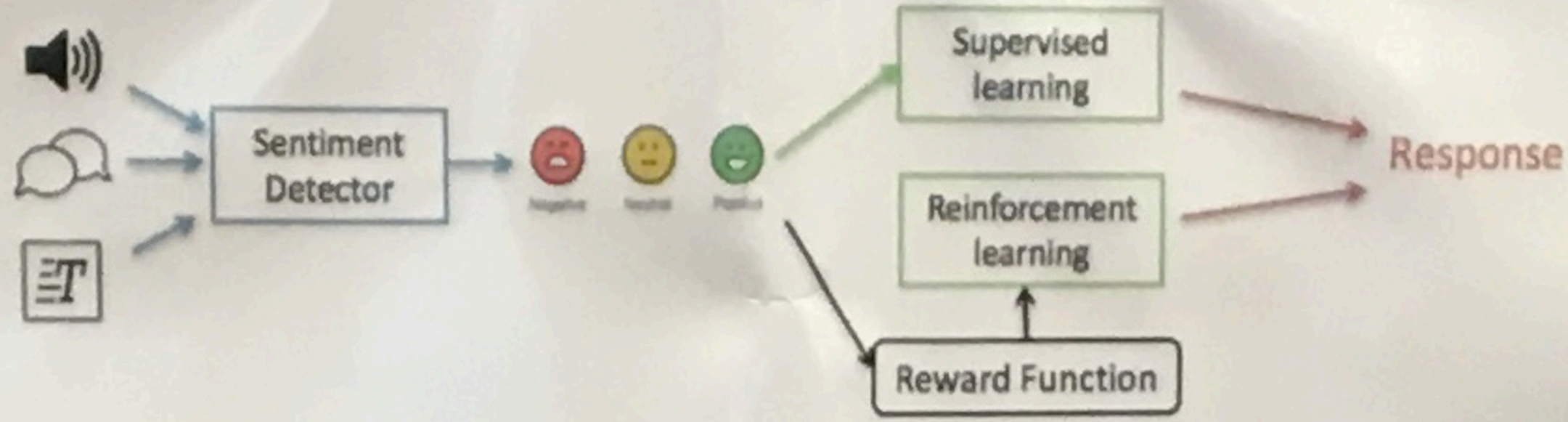


Figure 1. Proposed sentiment adaptive end-to-end dialog framework

- A sentiment detector is built on an annotated subset and is used to predict sentiment labels and sentiment scores for the supervised and reinforcement learning.
- Supervised learning uses the predicted sentiment labels from the sentiment detector as additional context features for the training.
- Reinforcement learning simulates the dialogs and uses the predicted sentiment scores from the sentiment detector as immediate rewards to guide the training.
- The whole model is end-to-end trainable and user-adaptive.

System with sentiment V.S. System without sentiment

Sentiment Adaptive System	**Baseline System without Sentiment**
SYS: The <route>. Where would you like to leave from?	**SYS**: The <route>. Where would you like to leave from?
USR: *Yeah* ***[negative sentiment]***	**USR**: *Yeah*
SYS: Where are you leaving from? For example, you can say, <place>.	**SYS**: Right. Where would you like to leave from?

Table 7: An example dialog by different systems in the supervised learning setting. The sentiment-adaptive system gives a more detailed error-handling strategy than the baseline system.

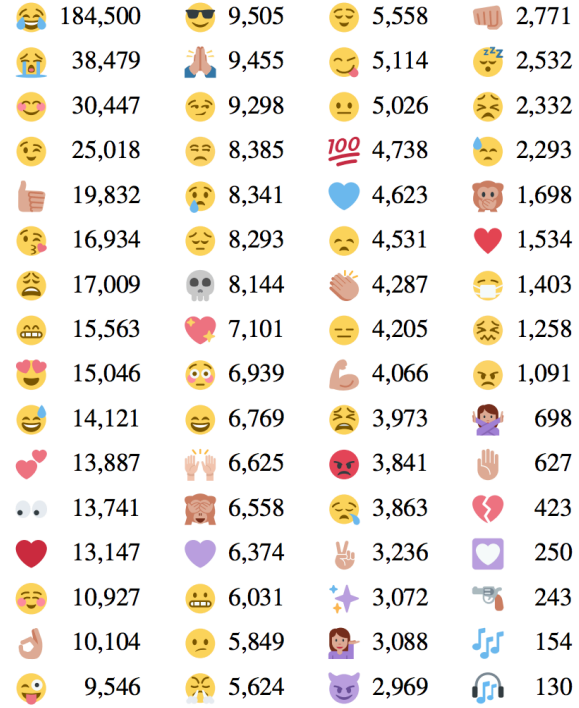
Outline

1. Thoughts on text generation by Yejin Choi
2. Sentiment Adaptive End-to-End Dialog Systems
3. MOJITALK: Generating Emotional Responses at Scale
4. Personalizing Dialogue Agents: I have a dog, do you have pets too?
5. Hierarchical Neural Story Generation
6. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

MOJITALK: Generating Emotional Responses at Scale

Xianda Zhou
Dept. of Computer Science and Technology
Tsinghua University
Beijing, 100084 China
zhou-xd13@mails.tsinghua.edu.cn

William Yang Wang
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106 USA
william@cs.ucsb.edu



Challenge: building empathetic NLP agents need large-scale labeled training data and human annotation is limited in size

Solution: leveraging Twitter data that are naturally labeled with emojis
- they assume the emojis convey the underlying emotions of the sentence

Approach: apply generative models to train an emotional response generation system

Details

- Crawled conversation pairs consisting of an original post and a response including at least one of the 64 emoji labels on Twitter
- Yielding training set/val/test 596,959/32,600/32,600
- Generative models
  - Attention-Based seq2seq
  - Conditional Variational Autoencoder
  - Reinforced CVAE

Model	Perplexity	Emoji Accuracy	
		Top1	Top5
	Development		
Base	127.0	34.2%	57.6%
CVAE	37.1	40.7%	75.3%
Reinforced CVAE	38.1	42.2%	76.9%
	Test		
Base	130.6	33.9%	58.1%
CVAE	36.9	41.4%	75.1%
Reinforced CVAE	38.3	42.1%	77.3%

Table 2: Generation perplexity and emoji accuracy of the three models.

Details

Setting	Model v. Base	Win	Lose	Tie
reply	CVAE	42.4%	43.0%	14.6%
reply	Reinforced CVAE	40.6%	39.6%	19.8%
emoji	CVAE	48.4%	26.2%	25.4%
emoji	Reinforced CVAE	50.0%	19.6%	30.4%

Table 4: Results of human evaluation. Tests are conducted pairwise between CVAE models and the base model.

- Their model is capable of generating high-quality emotional responses, without the need of laborious human annotations.

Outline

1. Thoughts on text generation by Yejin Choi
2. Sentiment Adaptive End-to-End Dialog Systems
3. MOJITALK: Generating Emotional Responses at Scale
4. Personalizing Dialogue Agents: I have a dog, do you have pets too?
5. Hierarchical Neural Story Generation
6. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

Personalizing Dialogue Agents: I have a dog, do you have pets too?

Saizheng Zhang[†,1], **Emily Dinan**[‡], **Jack Urbanek**[‡], **Arthur Szlam**[‡], **Douwe Kiela**[‡], **Jason Weston**[‡]
[†] Montreal Institute for Learning Algorithms, MILA
[‡] Facebook AI Research
saizheng.zhang@umontreal.ca, {edinan,jju,aszlam,dkiela,jase}@fb.com

- Chatbot Problem: (1) no personality (2) can't remember what itself said or what other speakers said (3) generic response like "I don't know"

- Solution: add profile information to chatbot (persona 个性化角色档案)
  - profile stored in a memory-augmented neural network

Data collection

(i) Personas: we crowd-source a set of 1155 possible personas, each consisting of at least 5 profile sentences, setting aside 100 never seen before personas for validation, and 100 for test.

(ii) Revised personas: to avoid modeling that takes advantage of trivial word overlap, we crowd-source additional rewritten sets of the same 1155 personas, with related sentences that are rephrases, generalizations or specializations, rendering the task much more challenging.

(iii) Persona chat: we pair two Turkers and assign them each a random (original) persona from the pool, and ask them to chat. This resulted in a dataset of 164,356 utterances over 10,981 dialogs, 15,705 utterances (968 dialogs) of which are set aside for validation, and 15,119 utterances (1000 dialogs) for test.

Conclusion

- we show that models trained on PERSONA-CHAT (with or without personas) are more engaging than models trained on dialogue from other resources (movies, Twitter).

Outline

1. Thoughts on text generation by Yejin Choi
2. Sentiment Adaptive End-to-End Dialog Systems
3. MOJITALK: Generating Emotional Responses at Scale
4. Personalizing Dialogue Agents: I have a dog, do you have pets too?
5. Hierarchical Neural Story Generation
6. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

Hierarchical Neural Story Generation

Angela Fan **Mike Lewis** **Yann Dauphin**

Facebook AI Research, Menlo Park
{angelafan, mikelewis, ynd}@fb.com

Prompt: The Mage, the Warrior, and the Priest

Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

Figure 1: Example prompt and beginning of a story from our dataset. We train a hierarchical model that first generates a prompt, and then conditions on the prompt when generating a story.

- Task: explore story generation: creative systems that can build coherent and fluent passages of text about a topic.
- Contribution:
  - Dataset 303,358 human generated stories paired with writing prompts from an online forum
  - Evaluating free form text is challenging, they introduce new evaluation metrics (fluency & adhere to the prompt) Stories are decoded under 10 different prompts—9 randomly sampled prompts and 1 true corresponding prompt—and the likelihood of the story given the various prompts is recorded.

Outline

1. Thoughts on text generation by Yejin Choi
2. Sentiment Adaptive End-to-End Dialog Systems
3. MOJITALK: Generating Emotional Responses at Scale
4. Personalizing Dialogue Agents: I have a dog, do you have pets too?
5. Hierarchical Neural Story Generation
6. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction

Onno Kampman, Elham J. Barezi, Dario Bertero, Pascale Fung
Center for AI Research (CAiRE)
Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
ejs,dbertero@connect.ust.hk, pascale@ece.ust.hk

- Task: predicting personality from speech, language and video frames (facial features), （15s youtube）
- Novelty: proposed a multimodal fusion approach for personality prediction
- Application: emotional intelligence of virtual agents

Personality: set of identifiers that can be used with reasonable consistency to predict behavior

大五人格测试的五个维度

开放性（**O**penness）：具有想象、审美、情感丰富、求异、创造、智能等特质。

责任心（**C**onscientiousness）：显示胜任、公正、条理、尽职、成就、自律、谨慎、克制等特点。

外倾性（**E**xtraversion）：表现出热情、社交、果断、活跃、冒险、乐观等特质。

宜人性（**A**greeableness）：具有信任、利他、直率、依从、谦虚、移情等特质。

神经质或情绪稳定性（**N**euroticism）：具有平衡焦虑、敌对、压抑、自我意识、冲动、脆弱等情绪的特质，即具有保持情绪稳定的能力。

Mean Acc		Big Five Personality Traits				
Model	Mean	E	A	C	N	O
Audio	.8941	.8920	.9047	.8840	.8923	.8976
Text	.8868	.8823	.9023	.8794	.8833	.8865
Video	.8965	.8960	.9040	.8913	.8936	.8976
DLF	.9033	.9030	.9107	.8951	.9021	.9053
NNLB	.9034	.9030	.9104	.8962	.9027	.9049
NNFB	.9062	.9042	.9093	.9078	.9036	.9062
DCC	.9121	.9104	.9154	.9130	.9097	.9119
evolgen	.9133	.9145	.9157	.9135	.9098	.9130
Train labels avg	.8835	.8806	.8991	.8739	.8791	.8847

Contribution of different modalities to personality detection task

	Big Five Personality Traits				
Model	E	A	C	N	O
Audio	0.44	0.32	0.27	0.45	0.54
Text	-0.03	0.22	0.13	0.03	-0.06
Video	0.59	0.46	0.60	0.52	0.52

Table 1: Optimal weights learned for combining the three modalities for each trait. E, A, C, N, and O stand for Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness, respectively.

“Takeaways”

- Emotional response is important because user is emotional
  - *“evidence”: less than 5% of posts on Twitter are questions, whereas around 80% are about personal emotional state, thoughts or activities.*
- Better learning objective for text generation is highly needed
  - *MLE biases high frequency words*
- Dataset construction is hard but important
  - *王威廉, et al.*

END