Sounding Board: A User-Centric and Content-Driven Social Chatbot

Yahui Liu

Tencent AI Lab

yahui.cvrs@gmail.com

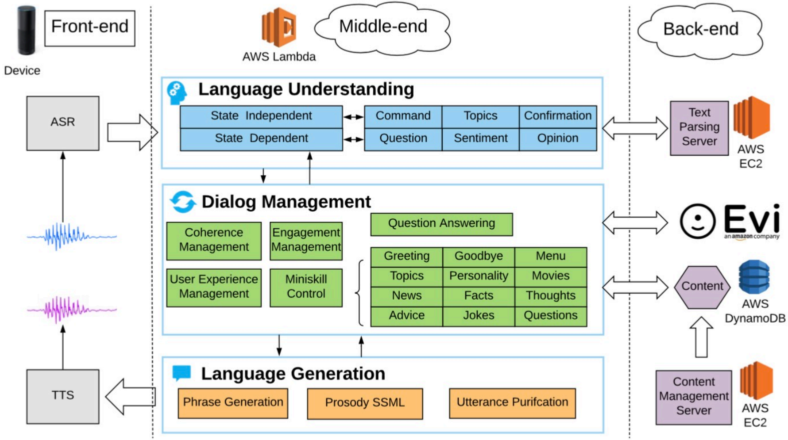
August 1, 2018

Modules

- Spoken Language Understanding (SLU): extracts the speaker's intent or goals, topic or subtopic, stance or sentiment [1].
- Dialog Management (DM): executes the dialog policy and decides the next dialog state.
- Natural Language Generation (NLG): generates and combines the response components according to the speech acts and content provided by the DM.

[1] Kumar A. et al. Just ASK: building an architecture for extensible self-service spoken language understanding, NIPS 2017.

Framework



Front-end
Device
Middle-end
AWS Lambda
Back-end
Language Understanding
State Independent
State Dependent
Command
Topics
Confirmation
Question
Sentiment
Opinion
Text Parsing Server
AWS EC2
ASR
Dialog Management
Coherence Management
Engagement Management
User Experience Management
Miniskill Control
Question Answering
Greeting
Goodbye
Menu
Topics
Personality
Movies
News
Facts
Thoughts
Advice
Jokes
Questions
Evi
an amazon company
Content
AWS DynamoDB
Content Management Server
AWS EC2
TTS
Language Generation
Phrase Generation
Prosody SSML
Utterance Purifcation

Example & DM

Turn	User	Sounding Board
1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
2	I'm five.	Glad to hear it! We could talk about robots, Batman, or Superman. Say "next", to chat about other things.
3	Superman.	It looks like you wanna chat about Superman. I was high up in the cloud when I realized: If people don't recognize Clark Kent as Superman because of his glasses, does his eye doctor know his true identity?
4	I guess so.	Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing "World of Warcraft"
5	Really, I didn't know that.	Weird, right? Speaking of Superman, Did you see: Superman? It was released in 1997. It's a comedy.
6	Yes, it was hilarious.	I'm glad you feel this is hilarious. Which part do you like best about this movie?
7	The part when he met Lewis Leah.	Interesting. Meccartin and Raffi co-directed this film. The movie has a 6.3 out of 10 on IMDB. which seems pretty good! do you like the movie's director?

Engagement, Coherence,
User experience

DM use a hierarchically-structured, state-based:

- a master manages the overall conversation
- a collection of miniskills handle different types of conversation segments (thoughts, facts, movies)

NLG

Method

- Phrase Generation: speech acts: grounding, inform, request, and instruction
- Prosody: convey information more clearly
- Utterance Purification: replaces profanity with a non-offensive word

Content Management

Content is stored in a knowledge graph (KG) at the back-end.

The KG is organized based on miniskills.

The DM drives the conversation forward and generates respones by either traversing links between content nodes associated with the same topic or through relation edges to content nodes on a relevant new topic.

Analysis

Personality

- Five Factor model [1]
- mini-IPIP questionnaire [2]

We find users who are more extraverted, agreeable, or open to experience tend to rate our socialbot higher

Content

- The length distribution has a long tail.
- Longer conversation tended to get higher rating.

[1] McCrae R. et al. An introduction to the five-factor model and its applications, 1992.

[2] Donnellan M. et al. The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality, 2006.

Token-level and sequence-level loss smoothing for RNN language models

Yahui Liu

Tencent AI Lab

yahui.cvrs@gmail.com

August 1, 2018

Motivation

Two limitations of MLE

- It treats all sentences that do not match the ground truth as equally poor, ignoring the structure of the output space.
    - possible outputs is practically unbounded
    - evaluation measures don't take into account structural similarity.
- Exposure bias

Method

- token-level smoothing: using word-embedding, to achieve smoothing among semantically similar terms, and instroduce a procedure to promote rare tokens.
- sequence-level smoothing: using restricted token replacement vocabularies.

MLE

Objective function:

$$\begin{aligned}\ell_{MLE}(y^*, x) &= -\ln p_\theta(y^*|x) \\ &= -\sum_{t=1}^{T} \ln p_\theta(y_t^*|h_t^*)\end{aligned} \tag{1}$$

$\Rightarrow$

$$\begin{aligned}\ell_{MLE}(y^*, x) &= D_{KL}(\delta_{y^*} \| p_\theta(y|x)) \\ &= \sum_{t=1}^{T} D_{KL}(\delta_{y_t^*} \| p_\theta(y_t|h_t^*))\end{aligned} \tag{2}$$

Sequence-level loss smoothing

Replacing the sequence-level Dirac δ_{y^*} in Eq. 2 with a distribution:

$$r(y|y^*) \propto \exp r(y, y^*)/\tau \tag{3}$$

$\Rightarrow$

$$\begin{aligned} \ell_{seq}(y^*, x) &= D_{KL}(r(y|y^*) \| p_\theta(y|x)) \\ &= H(r(y|y^*)) - \mathbb{E}_r[\ln p_\theta(y|x)] \end{aligned} \tag{4}$$

- Entropy $H(r(y|y^*))$ does not depend on the model parameters θ,
- Replacing the expectation $\mathbb{E}_r[\cdot]$ with Monte-Carlo approximation:

$$\mathbb{E}_r[-\ln p_\theta(y|x)] \approx -\sum_{l=1}^{L} \ln p_\theta(y^l|x) \tag{5}$$

Sequence-level loss smoothing

Sampling

- Stratified sampling[1]: using Hamming or edit distance
- Importance sampling: previous introduction

$$\mathbb{E}_r[-\ln p_\theta(y|x)] \approx -\sum_{l=1}^{L} w_l \ln p_\theta(y^l|x) \quad (6)$$

- Restricted vocabulary sampling:
  - $\mathcal{V}$: the full vocabulary
  - $\mathcal{V}_{refs}$: the set of tokens appears in the ground-truth sentence(s)
  - $\mathcal{V}_{batch}$: the tokens appear in the ground-truth sentences in a given training mini-batch

[1] Norouzi M. et al. Reward aug- mented maximum likelihood for neural structured prediction, NIPS 2016.

Token-level loss smoothing

Replacing the token-level Dirac $\delta_{y_t^*}$ in Eq. 2 with a distribution:

$$r(y_t|y_t^*) \propto \exp r(y_t, y_t^*)/\tau \tag{7}$$

Using the cosine similarity between y_t and y_t^* in a semantic word-embedding space (GloVe[1]).

Promoting rare tokens

Encourages frequent tokens into considering the rare ones:

$$\begin{aligned} r^{freq}(y_t, y_t^*) &= r(y_t, y_t^*) \\ &\quad - \beta \min\left(\frac{freq(y_t)}{freq(y_t^*)}, \frac{freq(y_t^*)}{freq(y_t)}\right) \end{aligned} \tag{8}$$

[1] Pennington J. et al. GloVe: Global vectors for word representation, EMNLP 2014.

Combining loss

Sequence-level:

$$\begin{aligned} \ell_{seq}^{\alpha}(y^*, x) &= \alpha \ell_{seq}(y^*, x) + \bar{\alpha}\ell_{MLE}(y^*, x) \\ &= \alpha \mathbb{E}_r[\ell_{MLE}(y, x)] + \bar{\ell}_{MLE}(y^*, x) \end{aligned} \tag{9}$$

Token-level:

$$\ell_{Tok}^{\alpha}(y^*, x) = \alpha \ell_{Tok}(y^*, x) + \bar{\alpha}\ell_{MLE}(y^*, x) \tag{10}$$

and the combination of Eq. 9 and Eq. 10.

Lazy sequence smoothing

To speed up training in Eq. 5, replacing the MLE loss:

$$\ell_{MLE}(y^l, x) = -\sum_{t=1}^{T} \ln p_\theta(y_t^l | h_t^l) \tag{11}$$

into:

$$\ell_{lazy}(y^l, x) = -\sum_{t=1}^{T} \ln p_\theta(y_t^l | h_t^*) \tag{12}$$

Experiments

Loss	Reward	$\mathcal{V}_{sub}$	Without attention BLEU-1	Without attention BLEU-4	Without attention CIDER	With attention BLEU-1	With attention BLEU-4	With attention CIDER
MLE			70.63	30.14	93.59	73.40	33.11	101.63
MLE + γH			70.79	30.29	93.61	72.68	32.15	99.77
Tok	Glove sim		71.94	31.27	95.79	73.49	32.93	102.33
Tok	Glove sim r^{freq}		72.39	31.76	97.47	74.01	33.25	102.81
Seq	Hamming	$\mathcal{V}$	71.76	31.16	96.37	73.12	32.71	101.25
Seq	Hamming	$\mathcal{V}_{batch}$	71.46	31.15	**96.53**	73.26	**32.73**	101.90
Seq	Hamming	$\mathcal{V}_{refs}$	**71.80**	**31.63**	96.22	**73.53**	32.59	**102.33**
Seq, lazy	Hamming	$\mathcal{V}$	70.81	30.43	94.26	73.29	32.81	101.58
Seq, lazy	Hamming	$\mathcal{V}_{batch}$	71.85	31.13	**96.65**	73.43	32.95	**102.03**
Seq, lazy	Hamming	$\mathcal{V}_{refs}$	**71.96**	**31.23**	95.34	**73.53**	**33.09**	101.89
Seq	CIDER	$\mathcal{V}$	71.05	30.46	94.40	73.08	32.51	101.84
Seq	CIDER	$\mathcal{V}_{batch}$	71.51	31.17	95.78	**73.50**	**33.04**	**102.98**
Seq	CIDER	$\mathcal{V}_{refs}$	**71.93**	**31.41**	**96.81**	73.42	32.91	102.23
Seq, lazy	CIDER	$\mathcal{V}$	71.43	**31.18**	**96.32**	73.55	**33.19**	**102.94**
Seq, lazy	CIDER	$\mathcal{V}_{batch}$	71.47	31.00	95.56	73.18	32.60	101.30
Seq, lazy	CIDER	$\mathcal{V}_{refs}$	**71.82**	31.06	95.66	**73.92**	33.10	102.64
Tok-Seq	Hamming	$\mathcal{V}$	70.79	30.43	96.34	73.68	32.87	101.11
Tok-Seq	Hamming	$\mathcal{V}_{batch}$	72.28	31.65	96.73	73.86	33.32	102.90
Tok-Seq	Hamming	$\mathcal{V}_{refs}$	72.69	32.30	98.01	73.56	33.00	101.72
Tok-Seq	CIDER	$\mathcal{V}$	70.80	30.55	96.89	73.31	32.40	100.33
Tok-Seq	CIDER	$\mathcal{V}_{batch}$	72.13	31.71	96.92	73.61	32.67	101.41
Tok-Seq	CIDER	$\mathcal{V}_{refs}$	**73.08**	**32.82**	**99.92**	**74.28**	**33.34**	**103.81**

Experiments

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr		SPICE	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC^{+} (Vinyals et al., 2015)	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6	18.2	63.6
Hard-Attention (Xu et al., 2015)	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3	17.2	59.8
ATT-FCN^{+} (You et al., 2016)	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8	18.2	63.1
Review Net^{+} (Yang et al., 2016)	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9	18.5	64.9
Adaptive^{+} (Lu et al., 2017)	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9	19.7	67.3
SCST:Att2all$^{+\dagger}$ (Rennie et al., 2017)	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7	-	-
LSTM-A3$^{+\dagger\circ}$ (Yao et al., 2017)	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116	118	-	-
Up-Down$^{+\dagger\circ}$ (Anderson et al., 2017)	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5	-	-
Ours: Tok-Seq CIDEr	72.6	89.7	55.7	80.9	41.2	69.8	30.2	58.3	25.5	34.0	53.5	68.0	96.4	99.4	-	-
Ours: Tok-Seq CIDEr $^{+}$	74.9	92.4	58.5	84.9	44.8	75.1	34.3	64.7	26.5	36.1	55.2	71.1	103.9	104.2	-	-

Table 2: MS-COCO 's server evaluation . ($^{+}$) for ensemble submissions, (†) for submissions with CIDEr optimization and ($^{\circ}$) for models using additional data.

Thanks!