Learning to Ask Questions in Open-domain Conversational Systems with **Typed Decoders**

Yansen Wang[1], Chenyi Liu[1], Minlie Huang[1], Liqiang Nie[2]

[1]Tsinghua University, China

[2]Shandong University, China

aihuang@tsinghua.edu.cn

http://coai.cs.tsinghua.edu.cn/hml

Outline

- Background
- Problem & task definition
- Motivation
- Related works
- Method
- Experiments
- Summary

Good Examples from DL Models

User: 时间啊，你这把杀猪刀。。。

Time is a killer…

Machine: 我是被杀的

I am the one who was killed.

User：最幸福的人不是拿到了最好的东西，而是学会去珍惜手上已经拥有的人。

The happiest person is not the one who gets the best thing, instead who cherishes what she/he already has.

Machine: 是的，所以我也要珍惜自己的幸福

Agree, I will cherish my happiness too.

User: 怕孤獨，又捨不下自己的小自由。

Afraid to be lonely, but wont' give up my little freedom.

Machine: 我想去流浪，但没有人陪我去。

I want to wander, but no one accompany me.

Bad Examples are More!

User: 其实，他没有那么爱你。

Ohh, he does not love you that much.

Machine: 幸福的一天开始了。

A happy day starts.

User: 回北京了！很冷很高兴！

Back to Beijing now. So cold so happy!

Machine: 什么时候来北京啊?

When will you come to Beijing?

User: 会做饭的男人是最性感的男人。

Mans who can cook are the most sexy.

Machine: 会做饭的男人是最性感的女人

Mans who can cook are the most sexy women.

Bad Examples are More!

Xiaoice by MSRA

微软小冰

语义理解的问题 **Semantic understanding**

User: 你家多大? How large is your house?

小冰: 这么快就要爆年龄了?

So soon to reveal my age?

User: 你是素食主义者还是吃肉啊?

Are you vegan or meatatarian?

小冰: 怎么减肥

How can I lose weight?

上下文理解的问题 **Context issue**

User: 早饭不好吃

小冰: 那明早一起吃

User: 行，你推荐的应该不错

小冰: 是吗?

User: 哈哈，我相信你

小冰: 呃呃

个性身份一致性问题

Inconsistency in personality

User: 你几岁了?

小冰: 我啊。。。反正是九零后!

User: 那你是几几年出生的?

小冰: 错，75年阳历一月出生。

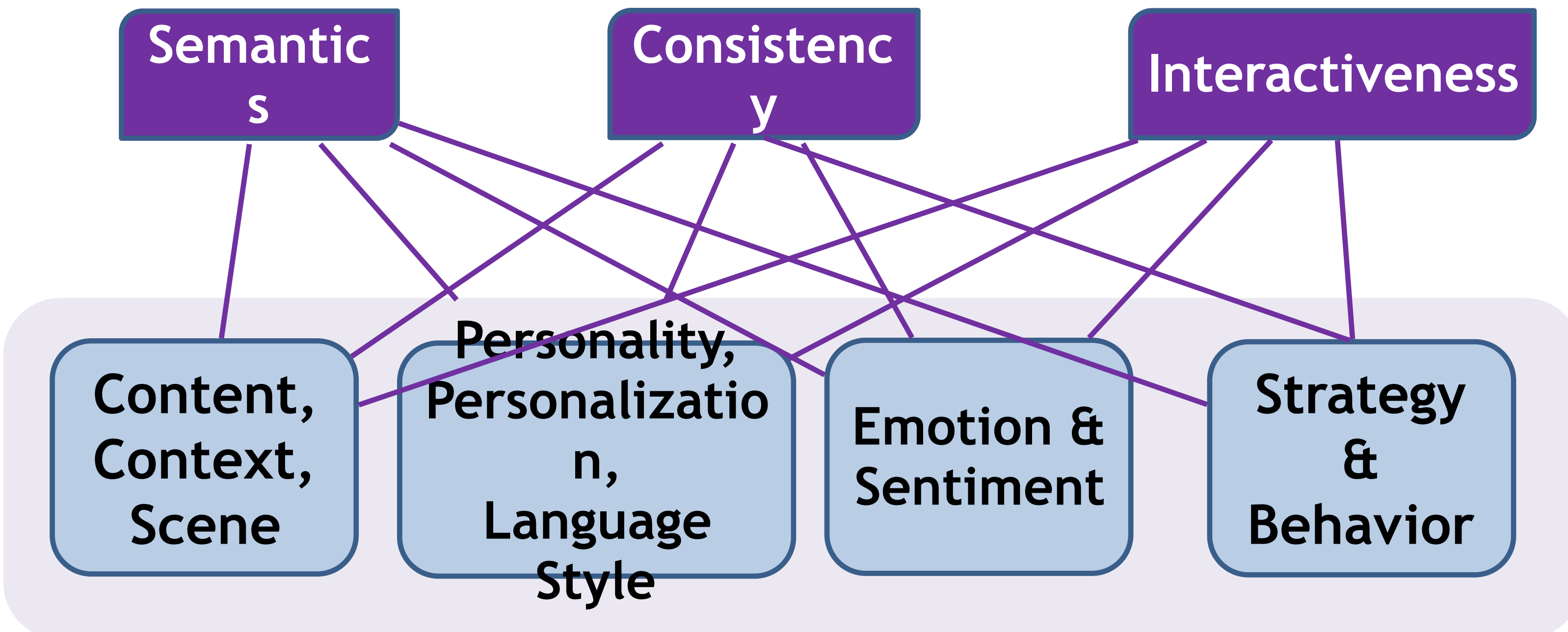
User: 你是九零后，怎么会75年出生?

小冰: 生在九零后，在深圳只能被当做八零后了。

Bad Examples (AI Ethics)

Picture from Prof. Frank Rudzicz, University of Toronto

Challenges in Chatting Machines



More Intelligent Chatting Machines

- Behaving more **interactively**:
  - **Emotional Chatting Machine** (**AAAI 2018**)
  - Proactive Behavior by **Asking Good Questions** (**ACL 2018**)
  - Controlling **sentence function** (**ACL 2018**)
- Behaving more **consistently**:
  - **Explicit Personality** Assignment (**IJCAI-ECAI 2018**)
- Behaving more **intelligently** with **semantics**:
  - Better Understanding and Generation Using **Commonsense Knowledge** (**IJCAI-ECAI 2018 Distinguished Paper**)

References:

① Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. **AAAI 2018**.

② Assigning personality/identity to a chatting machine for coherent conversation generation. **IJCAI-ECAI 2018**.

③ Commonsense Knowledge Aware Conversation Generation with Graph Attention. **IJCAI-ECAI 2018**.

④ Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders. **ACL 2018**.

⑤ Generating Informative Responses with Controlled Sentence Function. **ACL 2018**.

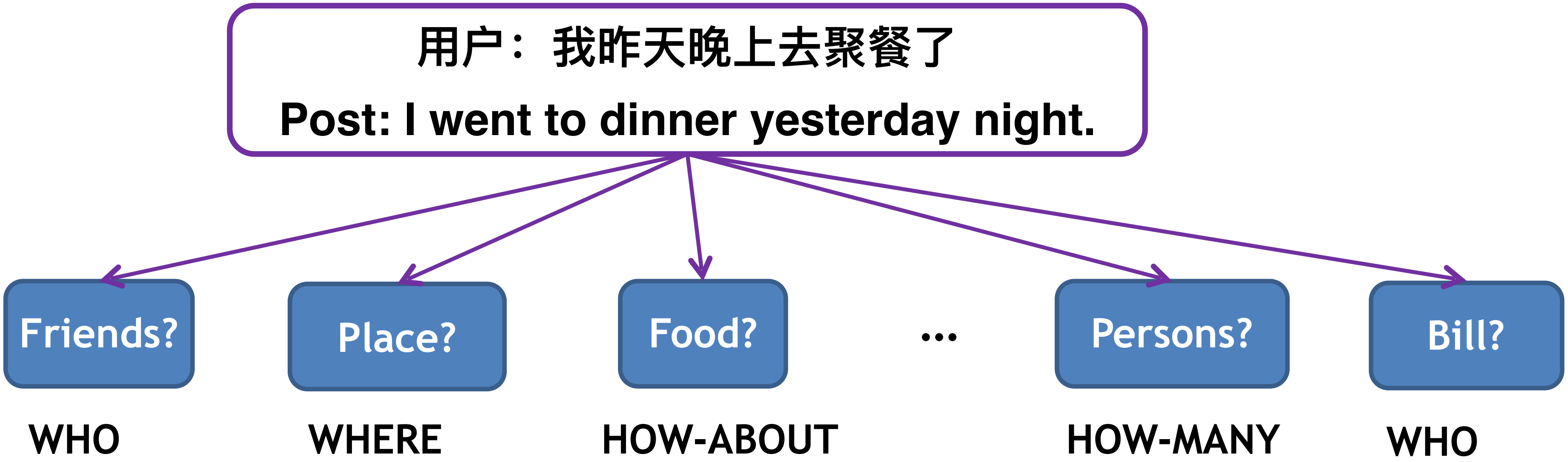
Problem & Task Definition

- How to ask **good** questions in open-domain conversational systems?

用户：我昨天晚上去聚餐了

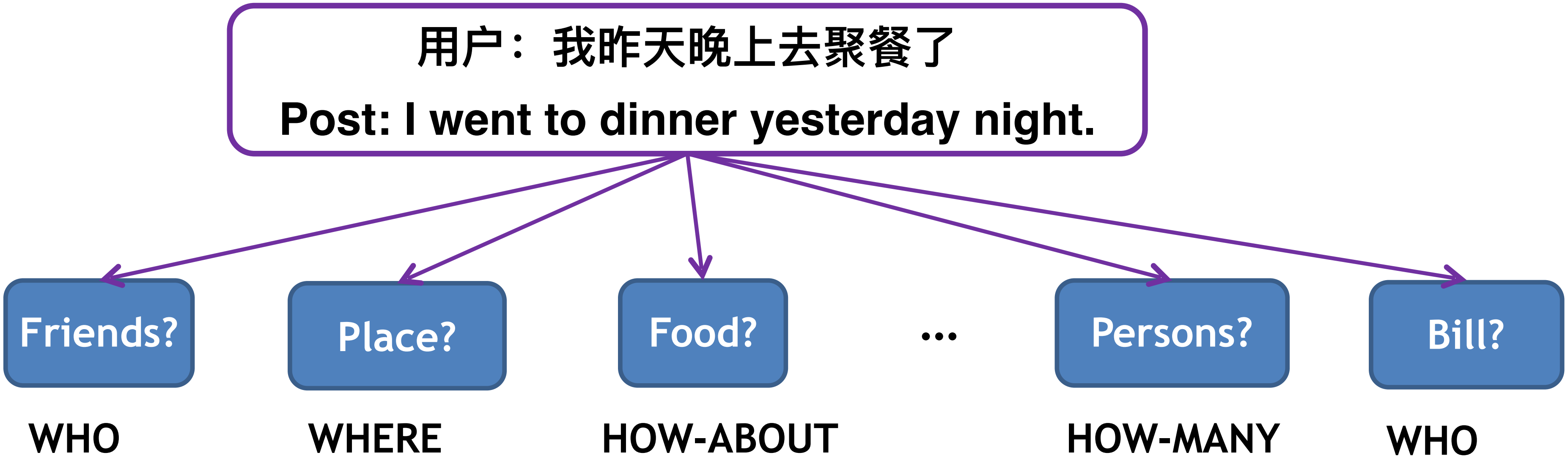
Post: I went to dinner yesterday night.

Problem & Task Definition



- **Who** were you with?
- **Where** did you have the dinner?
- **How about** the food?
- **How many** friends?
- **Who** paid the bill?
- **Is it** an Italian restaurant?

Problem & Task Definition



Scene: Dining at a restaurant

- Asking **good** questions requires **scene understandin**

Motivation

- **Responding + asking** (Li et al., 2016)
  - More interactive chatting machines
- **Key proactive** behaviors (Yu et al., 2016)
  - Less dialogue breakdowns
- Asking good questions is indication of **understanding**
  - **As in course teaching**
  - **Scene understanding** in this paper

Related Work

- Traditional question generation (Andrenucci and Sneiders, 2005; Popowich and Winne, 2013)
- **Syntactic Transformation**

- **<u>Given context</u>**: As recently as 12,500 years ago, the Earth was in the midst of a glacial age referred to as the Last Ice Age.
- **<u>Generated question</u>**: How would you describe the Last Ice Age?

Related Work

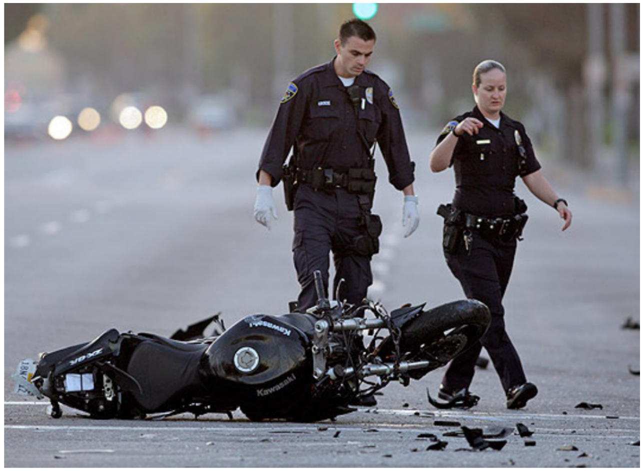
- A few neural models for question generation in **reading comprehension** (Du et al., 2017; Zhou et al., 2017; Yuan et al., 2017)

Given

- **Passage**: …Oxygen is used in cellular respiration and released by ***photosynthesis***, which uses the energy of sunlight to produce oxygen from water. …
- **Answer**: photosynthesis
- **Generated question**: What life process produces oxygen in the presence of light?

Related Work

- Visual question generation for **eliciting interactions** (Mostafazadeh, 2016): beyond image captioning
- Given image:



- Generated question: What happened?

Difference to Existing Works

- **Different goals**:
  - To enhance **interactiveness and persistence** of human-machine interactions
  - **Information seeking** in read comprehension

- **Various patterns**: YES-NO, WH-, HOW-ABOUT, etc.

- **Topic transition**: from topics in post to topics in response
  - **Dinner→food; fat → climbing; sports → soccer**

Key Observations

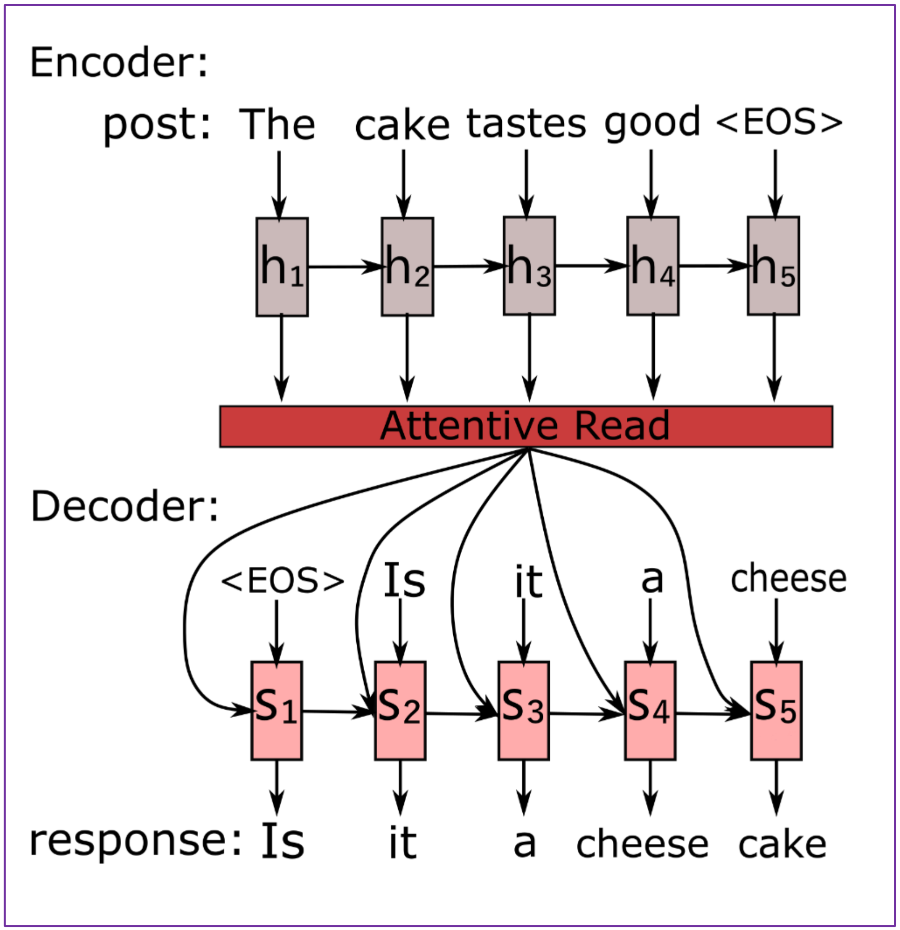
- A good question is a natural **composition** of
  - **Interrogatives** for using various questioning patterns
  - **Topic words** for addressing interesting yet novel topics
  - **Ordinary words** for playing grammar or syntactic roles

Example 1:
User: I am too fat ...
Machine: **How about** climbing this weekend?

Example 2:
User: Last night, I stayed in KTV with friends.
Machine: **Are** you happy with your singing**?**

Hard/Soft Typed Decoders (HTD/STD)

Encoder-decoder Framework



$$X = x_1 x_2 \cdots x_m$$

$$Y = y_1 y_2 \cdots y_n$$

$$Y^* = \underset{Y}{argmax}\, \mathcal{P}(Y|X).$$

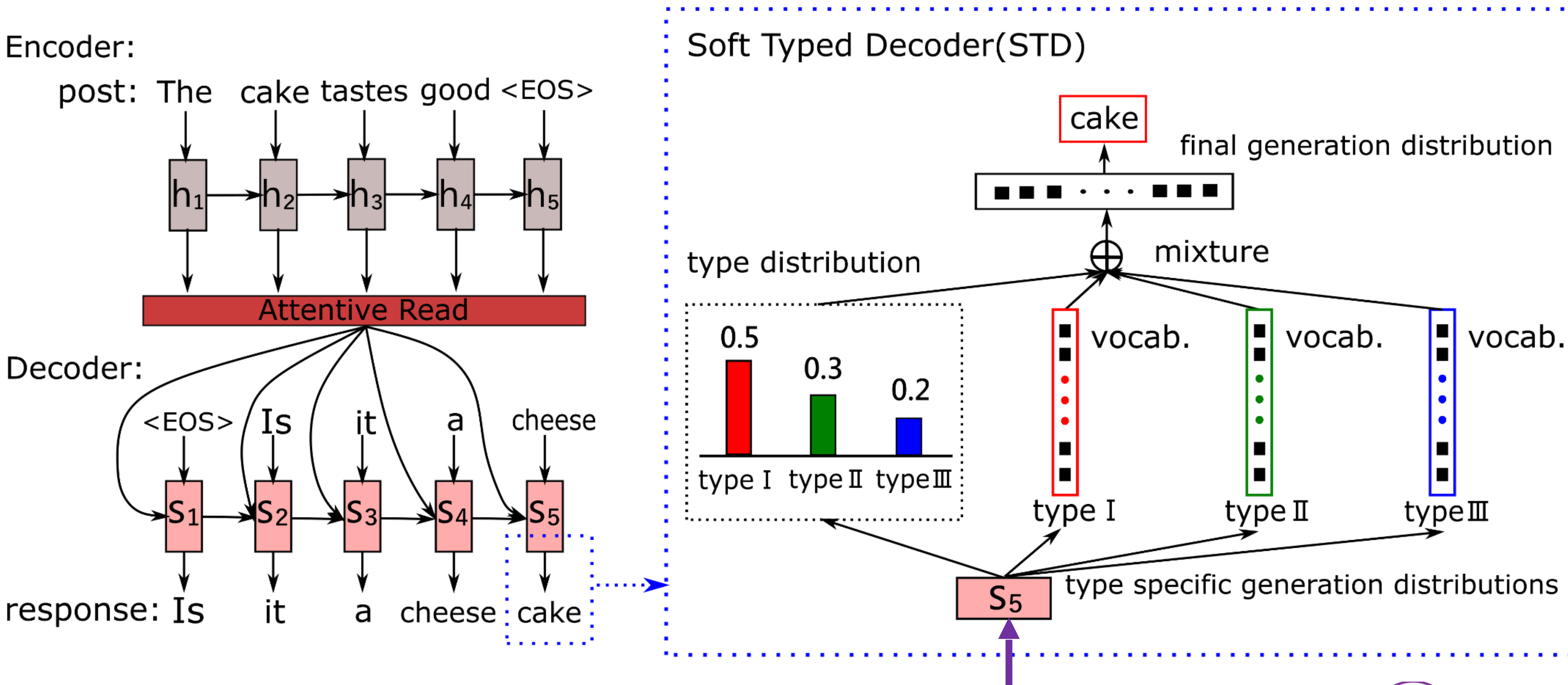
$$\mathcal{P}(y_t|y_{<t}, X) = \mathbf{MLP}(\mathbf{s}_t, \boldsymbol{e}(y_{t-1}), \mathbf{c}_t),$$

$$\mathbf{s}_t = \mathbf{GRU}(\mathbf{s}_{t-1}, \boldsymbol{e}(y_{t-1}), \mathbf{c}_t),$$

$$\mathbf{c}_t = \textstyle\sum_{i=1}^{T} \alpha_{t,i} \mathbf{h}_i$$

$$\mathbf{h}_t = \mathbf{GRU}(\mathbf{h}_{t-1}, \boldsymbol{e}(x_t)),$$

Soft Typed Decoder(STD)



Decoding state

Soft Typed Decoder(STD)

- Applying **multiple type-specific generation distributions** over the same vocabulary

- Each word has a **latent** distribution among the set

type(w)∈{***interrogative***, ***topic word***, ***ordinary word***}

$$\mathcal{P}(y_t|y_{<t},X)=\sum_{i=1}^{k}\underbrace{\mathcal{P}(y_t|ty_t=c_i,y_{<t},X)}_{\text{type-specific generation distribution}}\cdot\underbrace{\mathcal{P}(ty_t=c_i|y_{<t},X)}_{\text{word type distribution}},$$

Soft Typed Decoder(STD)

- Estimate the **type distribution** of each word:

$$\mathcal{P}(ty_t|y_{<t}, X) = softmax(\mathbf{W}_0\mathbf{s}_t + \mathbf{b}_0),$$

- Estimate the **type-specific generation distribution** of each word:

$$\mathcal{P}(y_t|\boxed{ty_t = c_i}, y_{<t}, X) = softmax(\mathbf{W}_{\boxed{c_i}}\mathbf{s}_t + \mathbf{b}_{c_i}),$$

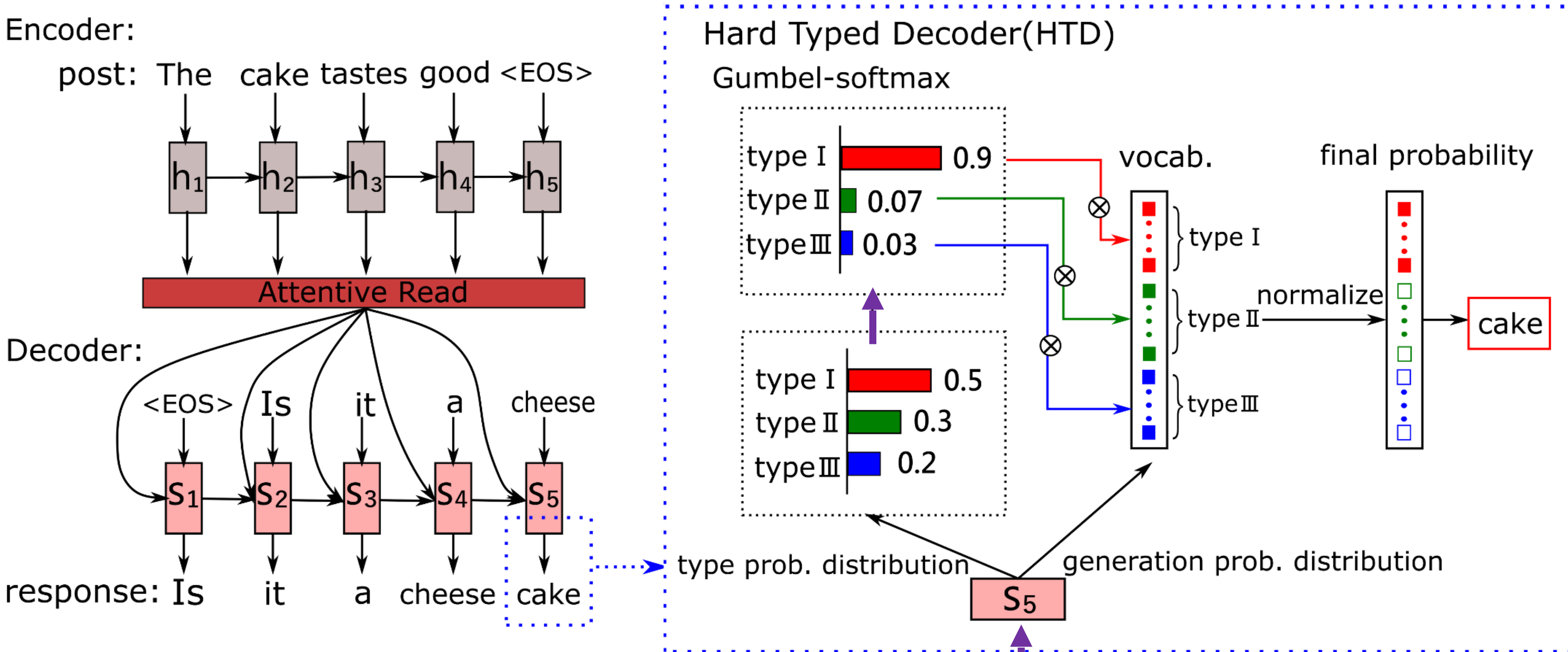
- The final generation distribution is a **mixture** of the three type-specific generation distribution.

$$\mathcal{P}(y_t|y_{<t}, X) = \sum_{i=1}^{k} \mathcal{P}(y_t|ty_t = c_i, y_{<t}, X) \cdot \mathcal{P}(ty_t = c_i|y_{<t}, X),$$

Hard Typed Decoder(HTD)

- In soft typed decoder, word types are modeled in a **latent, implicit** way
- Can we control the word type more **explicitly** in generation?
  - Stronger control

Hard Typed Decoder(HTD)



Decoding state

Hard Typed Decoder(HTD)

- Estimate the generation probability distril $\mathcal{P}(y_t|y_{<t}, X) = softmax(\mathbf{W}_0\mathbf{s}_t + \mathbf{b}_0).$
- Estimate the type probability distribution

$$\mathcal{P}(ty_t|y_{<t}, X) = softmax(\mathbf{W}_1\mathbf{s}_t + \mathbf{b}_1).$$

- Modulate words' probability by its corresponding type probability:

$$\mathcal{P}'(y_t|y_{<t}, X) = \mathcal{P}(y_t|y_{<t}, X) \boxed{\cdot \boldsymbol{m}(y_t)},$$

m(y_t) is related to the type probability of word y_t

Hard Typed Decoder(HTD)

Generation distr.			Type distr.			Modulated distr.	
what	*0.3*		$T_{interrogative}$	*0.7*		*what*	*0.8*
food	*0.2*	X	T_{topic}	*0.1*	→	*food*	*0.05*
is	*0.4*		$T_{ordinary}$	*0.2*		*is*	*0.09*

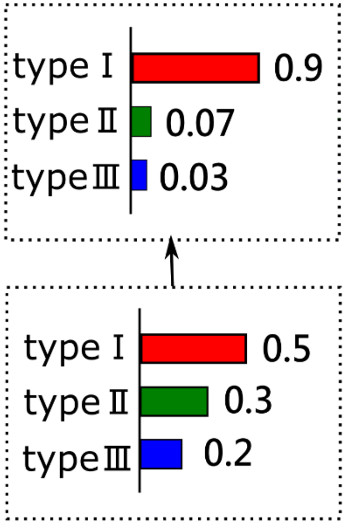
- **Argmax?** (firstly select largest type prob. then sample word from generation dist.)
  - ……Indifferentiable ……………
  - Serious grammar errors if word type is wrongly selected

Hard Typed Decoder(HTD)

- **Gumble-Softmax:**
  - A differentiable surrogate to the **argmax** function.

$$\boldsymbol{m}(y_t) = \mathbf{GS}(\mathcal{P}(ty_t = c(y_t)|y_{<t}, X)),$$

$$\mathbf{GS}(\pi_i) = \frac{e^{(log(\pi_i)+g_i)/\tau}}{\sum_{j=1}^{k} e^{(log(\pi_j)+g_j)/\tau}},$$



Hard Typed Decoder(HTD)

- In HTD, the types of words **are given in advance.**
  - *How to determine the word types?*

Hard Typed Decoder(HTD)

- **Interrogatives:**
  - A list of about 20 interrogatives are given by hand.
- **Topic words:**
  - Training: all nouns and verbs in response are topic words.
  - Test: 20 words are predicted by PMI

$$PMI(w_x, w_y) = log \frac{p(w_x, w_y)}{p_1(w_x) * p_2(w_y)},$$

$$Rel(k_i, X) = \sum_{w_x \in X} e^{PMI(w_x, k_i)},$$

- **Ordinary words:**
  - All other words, for grammar or syntactic roles

Loss Function

- Cross entropy
- Supervisions are on both final probability and the type distribution:

$$\Phi_1 = \sum_t -\log \mathcal{P}(y_t = \tilde{y}_t | y_{<t}, X),$$

$$\Phi_2 = \sum_t -\log \mathcal{P}(ty_t = \widetilde{ty}_t | y_{<t}, X),$$

$$\Phi = \Phi_1 + \lambda \Phi_2,$$

- λ is a term to balance the two kinds of losses.

Experiments

Dataset

- PMI estimation: calculated from **9 million post-response** pairs from Weibo.

- Dialogue Question Generation Dataset(DQG), about **491,000 pairs**：
  - Distilled questioning responses using about 20 hand-draft templates
  - Removed universal questions
  - Available at http://coai.cs.tsinghua.edu.cn/hml/dataset/

Baselines

- **Seq2Seq:** A simple encoder-decoder model (Luong et al., 2015)
- **Mechanism-Aware (MA):** Multiple responding mechanisms represented by real-valued vectors (Zhou et al., 2017)
- **Topic-Aware (TA):** Topic Aware Model by incorporating topic words (Xing et al., 2017)
- **Elastic Responding Machine (ERM):** Enhanced MA using reinforcement learning (Zhou et al., 2018)

Automatic Evaluation

Model	Perplexity	Distinct-1	Distinct-2	TRR
Seq2Seq	63.71	0.0573	0.0836	6.6%
MA	**54.26**	0.0576	0.0644	4.5%
TA	58.89	0.1292	0.1781	8.7%
ERM	67.62	0.0355	0.0710	4.5%
STD	56.77	0.1325	0.2509	12.1%
HTD	56.10	**0.1875**	**0.3576**	**43.6%**

Table 1: Results of automatic evaluation.

Evaluation metrics

- **Perplexity & Distinct**
- **TRR (Topical Response Ratio):**
    - 20 topic words are predicted with PMI for each post.
    - TRR is the proportion of the responses containing at least one topic word.

Manual Evaluation

- Pair-wise comparison: win, loss, tie
- Three evaluation criteria:
    - **Appropriateness:** whether a question is reasonable in logic and content, and has key info.
    - **Richness:** containing topic words or not
    - **Willingness** to respond to a generated question

Manual Evaluation(Pairwise)

Models	Appropriateness			Richness			Willingness		
	Win (%)	Lose (%)	Tie (%)	Win (%)	Lose (%)	Tie (%)	Win (%)	Lose (%)	Tie (%)
STD vs. Seq2Seq	42.0	38.6	19.4	37.2^{**}	15.2	47.6	45.4^{*}	38.6	16.0
STD vs. MA	39.6^{*}	31.2	29.2	32.6^{**}	16.8	50.6	49.4^{**}	27.0	23.6
STD vs. TA	42.2	40.0	17.8	49.0^{**}	5.4	45.6	47.6^{*}	40.2	12.2
STD vs. ERM	43.4^{*}	34.4	22.2	60.6^{**}	13.2	26.2	43.2^{*}	36.8	20.0
HTD vs. Seq2Seq	50.6^{**}	30.6	18.8	46.0^{**}	10.2	43.8	58.4^{**}	33.2	8.4
HTD vs. MA	54.8^{**}	24.4	20.8	45.0^{**}	17.0	38.0	67.0^{**}	18.0	15.0
HTD vs. TA	52.0^{**}	38.2	9.8	55.0^{**}	5.4	39.6	62.6^{**}	31.0	6.4
HTD vs. ERM	64.8^{**}	23.2	12.0	72.2^{**}	8.4	19.4	56.6^{**}	36.6	6.8
HTD vs. STD	52.0^{**}	33.0	15.0	38.0^{**}	26.2	35.8	61.8^{**}	30.6	7.6

Table 2: Annotation results. Win for "A vs. B" means A is better than B. Significance tests with Z-test were conducted. Values marked with * means *p-value* < 0.05, and ** for *p-value* < 0.01.

Score: the probability of win/lose/tie of our model vs. baseline

Examples

Post:	我们昨天去唱歌了 We went to karaoke yesterday.
Seq2Seq:	**什么时候?** **When?**
MA:	**你怎么**知道我是在问 **How** did you know I am questioning you?
TA:	**什么?** **What?**
ERM:	**什么**情况? **What** happened ?
STD:	去**哪里**唱歌? **Where** did you sing karaoke?
HTD:	你们**几个**人唱歌? **How many** people were singing with you?

Table 4: Generated question examples. Interrogative words are bolded and topic words underlined.

More Examples

- Different questioning patterns and topic

Post-3:	我在体育中心上班。 I work at the sports center.
Human:	你每天从家过去要多久? **How** long does it take you to go there from home**?**
STD:	你是**什么**部门的啊**?** **Which** department are you from**?**
HTD:	那你是体育学院的学生**吗?** **Are** you a student from the College of Physical Education**?**
Post-4:	今天的寿司真好吃。 The sushi today is so delicious!
Human:	我记得你不喜欢吃日式料理的不是**吗?** I remember you didn't like eating Japanese food, **did** you**?**
STD:	你**什么时候**来无锡，我请你。 **When** will you come to Wuxi**?** It is my treat.
HTD:	是**吗?**改天尝尝! Really**?** I will try sometime!

Work→Department
Sports→ College…

Suchi→Treat
Suchi→Try

Visualization

- Type prediction at each decoding position

Post: 我喜欢小动物(I like little animals)

Response:	你(you)	喜欢(like)	兔子(rabbit)	吗(*particle*)	?	_EOS
Interrogative	0.09	0.02	0.01	$\boxed{0.85}$	$\boxed{1.00}$	0.01
Topic word	0.26	0.35	$\boxed{0.71}$	0.14	0.00	0.02
Ordinary word	$\boxed{0.65}$	$\boxed{0.63}$	0.28	0.01	0.00	$\boxed{0.97}$
Decoding steps	1	2	3	4	5	6

Summary

- Stronger control in language generation via word semantic type
- What’s new
  - **A new task:** question generation in open-domain dialogue systems
  - **A new dataset:** Dialog Question Generation Dataset
  - **A new model with two variants:** possibly applicable to other generation tasks if word semantic types can be easily identified
- **The compatibility issue** between topic control and other word type control is NOT well solved
  - Bad grammar or not reasonable responses

Thanks for your attentions

- Dataset: http://coai.cs.tsinghua.edu.cn/hml/dataset/
- Codes: https://github.com/victorywys/Learning2Ask_TypedDecoder
- Homepage: http://coai.cs.tsinghua.edu.cn/hml
- Recruiting post-doctors!

Error Analysis

- Main error types
  - ***No topic words*** **(NoT)** in a response
  - ***Wrong topics*** **(WrT)** where topic words are irrelevant
  - ***Type generation error*** **(TGE)** where a wrong word type is predicted

Error Type	NoT	WrT	TGE	Others
STD	34%	34%	29%	3%
HTD	29%	39%	29%	3%

Table 6: Error type distribution.

Error Analysis: Examples

No topic words	**Post-1:**	今天好开心啊! I am so happy today!
	STD:	你怎么知道? **How** do you know **?**
Wrong topics	**Post-2:**	海报非常棒，期待若曦与我们男人的首度合作。 The poster is great and we look forward to our first cooperation with Ruoxi.
	HTD:	你海报怎么样啊? **How** about your poster **?**
Type generation error	**Post-3:**	又生病啦?吃点药就好了。 Got sick again? Just take some medicine and you'll be fine soon.
	STD:	我也不知道怎么回事。 I don't know **what** happened.
	HTD:	肠胃痛了，上火吗? Stomach aching, ulcer **?**