Data Sellection for Supervised Dialogue Generation

Yahui Liu

Tencent AI Lab

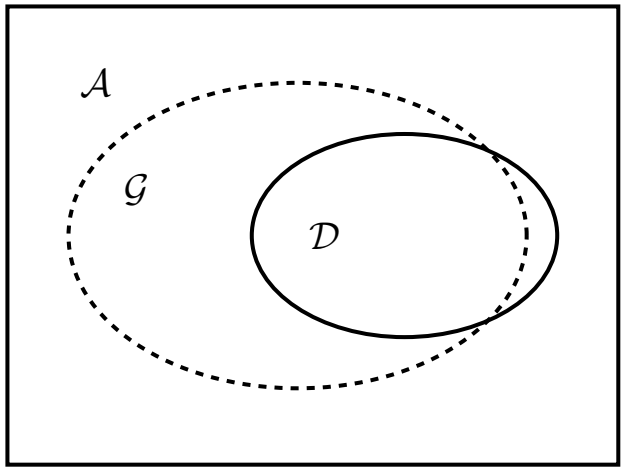
yahui.cvrs@gmail.com

July 3, 2018

Introduction

Data Theory

- The data crawled from the websites may contain only a part of ground truth as well as many noises.
- Allows high quality data to have more influences on the generation model and reduces the effect of noisy data



- Data Sellection : $\mathcal{D} \Rightarrow \mathcal{G} \bigcap \mathcal{D}$
- Data Augmentation : $\mathcal{D} \Rightarrow \mathcal{G}$?

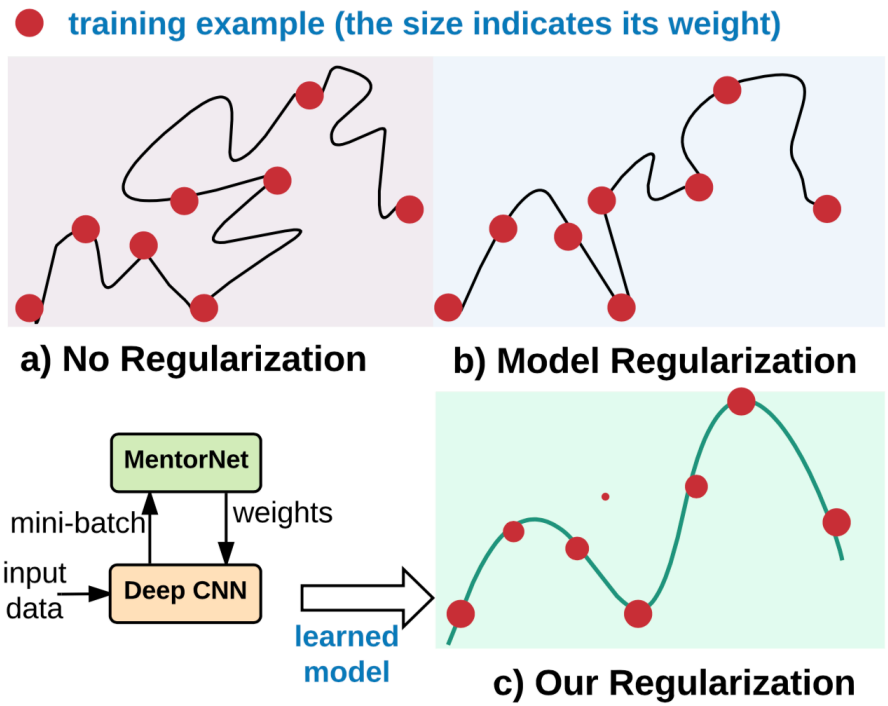
Data Sellection

Main idea

Data (subset) Sellection / Regularization : exclude or reweight the noisy data.

Given a dataset $\mathcal{D}$, we suppose that:
$f(\mathcal{M}_{\boldsymbol{\theta},\mathcal{D}'}) \geqslant f(\mathcal{M}_{\boldsymbol{\theta},\mathcal{D}})$,
s.t. $\exists\ \mathcal{D}' \subset \mathcal{D}$

where $\mathcal{M}_{\boldsymbol{\theta}}$ refers to non-convex models with parameters $\boldsymbol{\theta}$



Methods

1. Statistic weights†
2. Loss-based weights*
3. Classification weights*
4. Self-paced learning (SPL & SPCL)‡

$$\ell_w(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = w(\mathbf{y}|\mathbf{x})\ell(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$$

†Effective, domained by prior knowledge
‡Complex, consider both prior knowledge & learned information
*Unreliable

Statistic weights

Statistical features[1]:

- similarity frequency
- sentence length

Formulation

$$\mathcal{E}(\mathbf{y}) = e^{-af(\mathbf{y})}$$
$$f(\mathbf{y}) = \max\{0, Count(D(\mathbf{y}, \mathbf{y}_j) \geq \tau) - b\}, j \in |\mathbb{C}| \quad (1)$$
$$\mathcal{F}(\mathbf{y}) = e^{-c||\mathbf{y}|-|\hat{\mathbf{y}}||}$$

$$\Phi(\mathbf{y}) = \alpha\mathcal{E}(\mathbf{y}) + \beta\mathcal{F}(\mathbf{y})$$
$$w(\mathbf{y}|\mathbf{x}, \mathbb{R}, \mathbb{C}) = \frac{\Phi(\mathbf{y})}{\max_{\mathbf{r}\in\mathbb{R}}\{\Phi(\mathbf{r})\}} \quad (2)$$

[1] Liu Y. et al, Toward Less Generic Responses in Neural Conversation Models: A Statistical Re-weighting Method, submitted to EMNLP 2018

Statistic weights

你很有想法

Strict match		Similarity		
Final weight	Frequency weight	Final weight	Frequency weight	
0.7012	0.0137	0.7012	1.0000	跟你学做法棍?
0.8310	0.0000	0.2686	0.0000	我是个有内涵的人!
0.7555	1.0000	0.1931	0.0000	我也觉得我很有想法
1.0000	1.0000	1.0000	1.0000	哈哈没有男人会说这女人有想法我要了
0.8310	0.0191	0.2686	0.0000	你对我有想法???
0.6341	1.0000	0.0717	0.0000	你很有个性
0.6622	1.0000	0.6622	1.0000	年轻人想法挺多的
0.2220	0.0513	0.0717	0.0000	你很有想法
0.5655	1.0000	0.5655	1.0000	过生日不砸蛋糕乐趣就少了一半蛋糕不来个整块砸乐趣就又少了一半那生日哪来的快乐
0.6622	1.0000	0.0998	0.0000	你也很有想法
0.6622	1.0000	0.0998	0.0000	我有啥想法?
0.7555	1.0000	0.7555	1.0000	没有好想法就得变大傻
0.6140	1.0000	0.0516	0.0000	有道理哇
0.0371	0.0000	0.0371	0.0000	什么意思
0.7012	1.0000	0.7012	1.0000	一直都很有想法哈哈哈

Statistic weights

Query	Response			
	Seq2Seq	Seq2Seq-MMI	Seq2Seq-DD	Ours-RW$_{EF}$
我要生气了 I'm going to be angry.	我也生气了 So do I.	生气了 I'm angry.	我也生气了 So do I.	不要生气 Don't be angry.
最近有什么好书推荐 Would you like to recommend me some good books?	看你喜欢什么类型的 It depends on what kind of books you like?	我也想去看电影 I want to see the movie, too.	你喜欢什么类型的 What kind of books do you like?	《红楼梦》 Dream of the Red Chamber
做人不一定要风风光光 It's not necessary to be a successful and grand man.	一定要 Must be	我也不知道 I don't know, either.	是啊。。。 Yes...	做人一定要善良 Everyone should try to be a kind man.
这几天嗓子疼 I have a sore throat recently.	我也不知道 I don't know, either.	我也想去 I want to go, too	我也是。。。 So do I...	我这两天也是 So do I these days .
二号煤矿放几天假 How many days did people in No.2 coal mine take off?	假的 False	我也想放假 I want to have a holiday, too	假的，假的 False, False	我放了3天 I took three days off.
孟记普洱茶专营店一帖 An advertising of Mengji's Pu'er tea shop.	我也想吃 I want to eat, too.	我也想喝 I want to drink, too.	铁观音，茶，茶，茶 Tat-Kuan-Yin Tea, tea, tea, tea	我喜欢喝茶 I like drinking tea.

Reducing about 30% generic responses

Loss-based weights

Formulation

For a generative model, $\ell_i = \mathcal{L}(\mathbf{y}_i, G_{\boldsymbol{\theta}}(\mathbf{x}_i))$, the weight for the i-th pair example $(\mathbf{x}_i, \mathbf{y}_i)$ is:

$$w_i = \max(0, 1 - \frac{1}{\lambda}\ell_i) \tag{3}$$

or

$$w_i = (1 - e^{-\ell_i})^{\gamma} \tag{4}$$

Completely dominated by the training loss, the learning may be prone to overfitting

Classification weights

Learning to Converse with Noisy Data: Generation with Calibration[1]
RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems[2]

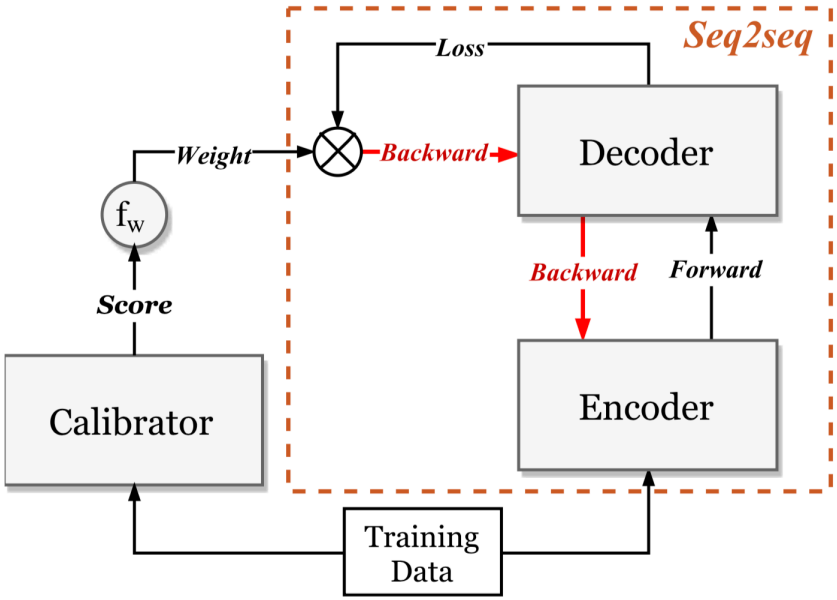
Main Idea

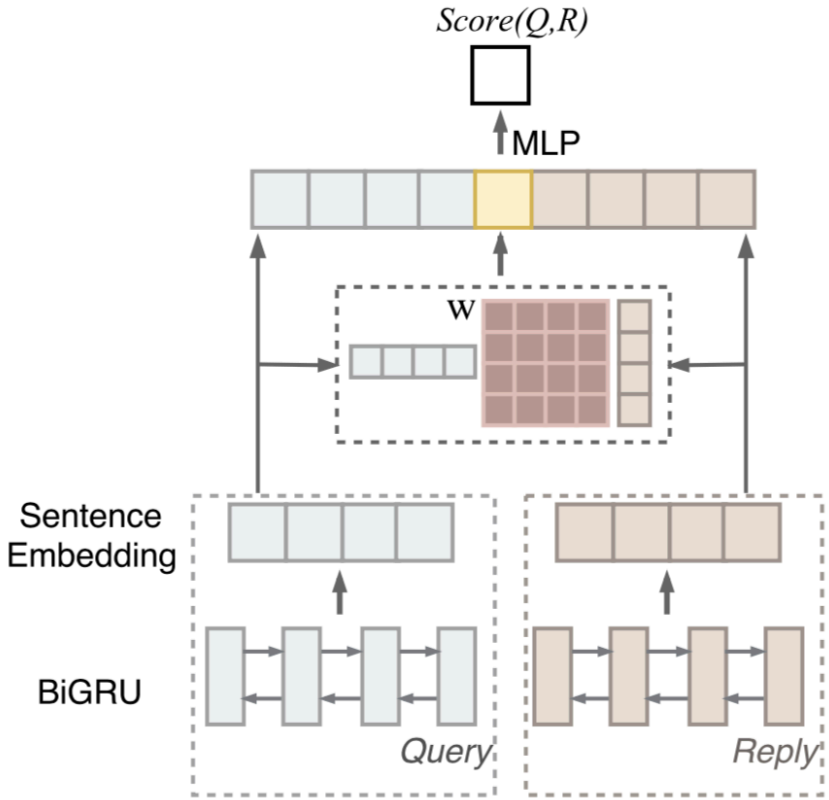
For each instance in training set, we employ a calibration network to produce a quality score for it, then the score is used for the weighted update of the generation model parameters.

[1] Shang M. et al. Learning to Converse with Noisy Data: Generation with Calibration, IJCAI 2018

[2] Tao C. et al. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems, AAAI 2018

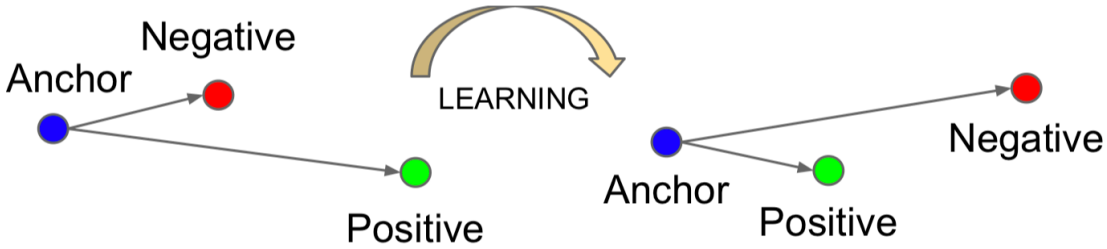
Classification weights





Classification weights

Triplet loss[1] & Negative Sampling



$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \tag{5}$$

$$\Rightarrow \quad \max\{\Delta - f_c(Q, R) + f_c(Q, R^-), 0\} \tag{6}$$

[1] Schroff F. et al. FaceNet: A Unified Embedding for Face Recognition and Clustering, CVPR 2015

Classification weights

Query	S2SA	S2SA+MMR	S2SAC+MMR
已经有个女儿了好吧, 一岁了 She has a daughter who is one year old.	也是这么想的 I think so.	好吧，那你是个孩子 Well, you are a child.	我觉得她的女儿好可爱 Her daughter is very lovely.
您好，请问主卧租出去了么？情侣租可以吗? Hello, has the master bedroom been rented out? Is it OK for a couple to rent it?	不好意思，不好意思。 Sorry, sorry!	可以，可以来看房 Yes, you can come and have a look.	房子已经租出去了 Room has been rent out.
每次看到总会想起小时候看的铁胆火车侠 It reminds me of Hikarian everytime I watch it.	我也是，我也觉得 I think you surely are.	我觉得你也是 I think you surely are.	哈哈，小时候看的时候很喜欢 LoL, I liked it very much when I was watching it as a child.

Classification weights

Data Distillation for Controlling Specificity in Dialogue Generation[1]

Main Idea

Removing training examples that are closest to the responses most commonly produced by the model trained from the last round and then retrain the model on the remaining dataset.

[1] Li J. et al. Data Distillation for Controlling Specificity in Dialogue Generation, arXiv 2016

Classification weights

Method

- Decoding a subset of the training set, **1 million** respinses in total.
- A response is considered generic if its frequency of occurrence exceeds a threshold.
- Collecting a list of most common responses, denoted by L.

Input: training data D
Output: sequence of trained models M

$M \leftarrow \emptyset$
for $i \leftarrow 1$ **to** $N = 8$ **do**
  train a SEQ2SEQ model m on D until convergence
  $M \leftarrow M + m$
  decode subset of input messages in D using model m
  collect top frequent decoded responses L
  for all instances $e \in D$ **do**
    compute relevance score $R(e)$ using Eq. 1
  end for
  $D^{\neg} \leftarrow$ top examples by $R(e)$
  distill $D^{\neg}$: $D \leftarrow D - D^{\neg}$
end for
return M

$$R(e) = \max_{e' \in L} \cos(e, e') \qquad (7)$$

Classification weights

Count	Response	Count	Response
	Iter1		**Iter2**
145575	i don 't know what you are talking about .	54227	i 'm not in the mood .
84435	i 'm not going to let you go .	29559	i 'm sorry about the way i acted .
36032	i 'm sorry i didn 't mean to offend you .	22987	you 're not in the mood .
23890	i 'm not so sure .	21392	i 'm gonna take a look at the new york times .
19405	i don 't know what to say .	20380	i 'll be there in a minute .
16888	i 'm not going to let you go !	14736	i 'm gonna take a look at this .
16048	that 's a good idea .	13753	i 'll get the money .
12782	i don 't know what to do .	13013	i 'm gonna take a shower .
11840	i 'm not going to be able to do that .	11746	i 'm in the middle of a war .
11604	i 'm sorry i can 't help you .	10130	you 're not getting any sleep .
11254	i 'm sure you 're right .	9996	i 'm gonna take a look at the other side .
9474	you don 't know what you are saying .	9644	i 'm sorry about the way you did .
9471	i 'm not going to tell you .	9169	i 've been doing a lot of things .
8905	i 'm not sure i can do it .	7837	you 're a dead man .
7905	i have no idea .	5320	i was just getting a little tired of it .
	Iter3		**Iter4**
41139	i 'm not an idiot .	30378	i 'm not from around here .
34738	i 'm not an expert on this .	26705	i 'm not from the future .
20252	i 'm sorry but i 'm not an expert on this .	9923	i was just talking to my wife .
16275	i 've got some bad news for you .	9012	i 'm not doing this .
16081	i 'll get you a new suit .	8573	you 're a goddamn liar .
13007	i 'm not an idiot !	7424	i 'll be on the way .
11254	i 'm gonna make a big deal out of this .	6919	i 'm sorry ma 'am .
6532	i 'm just an ordinary man .	5546	i 'm going back to the hotel .
5724	i 'm not an expert on the police .	4569	i 'll be on my way .
5604	i 'm not an expert on the subject .	4555	i 'm not staying here .
5168	i 'm not your enemy !	4416	you 're a goddamn genius .
4963	i 'm not an expert on the law .	4184	i 'm a little tired .
4454	i 'm gonna need some more help with this .	4183	i 'm gonna take a look at this .
4342	i was just about to get my hands on the wall .	4103	he 's a bit of a jerk .
3969	i can 't believe you 're still alive .	3819	he 's a bit of a pain in the ass .

Self-paced learning

Next Week

Self-Paced Curriculum Learning[1]

MentorNet: Regularizing Very Deep Neural Networks on Corrupted Labels[2]

$$\min_{\boldsymbol{\theta},\mathbf{w}\in[0,1]^n} \mathbb{F}(\boldsymbol{\theta},\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} w_i \mathcal{L}(\mathbf{y}_i, G_{\boldsymbol{\theta}}(\mathbf{x}_i)) \tag{8}$$

[1] Jiang L. et al. Self-Paced Curriculum Learning, AAAI 2015

[2] Jiang L. et al. MentorNet: Regularizing Very Deep Neural Networks on Corrupted Labels, arXiv 2017

Thanks!