基于对抗学习的生成式对话模型

王宝勋

三角兽(北京)科技有限公司

wangbaoxun@trio.ai

三角兽

Agenda

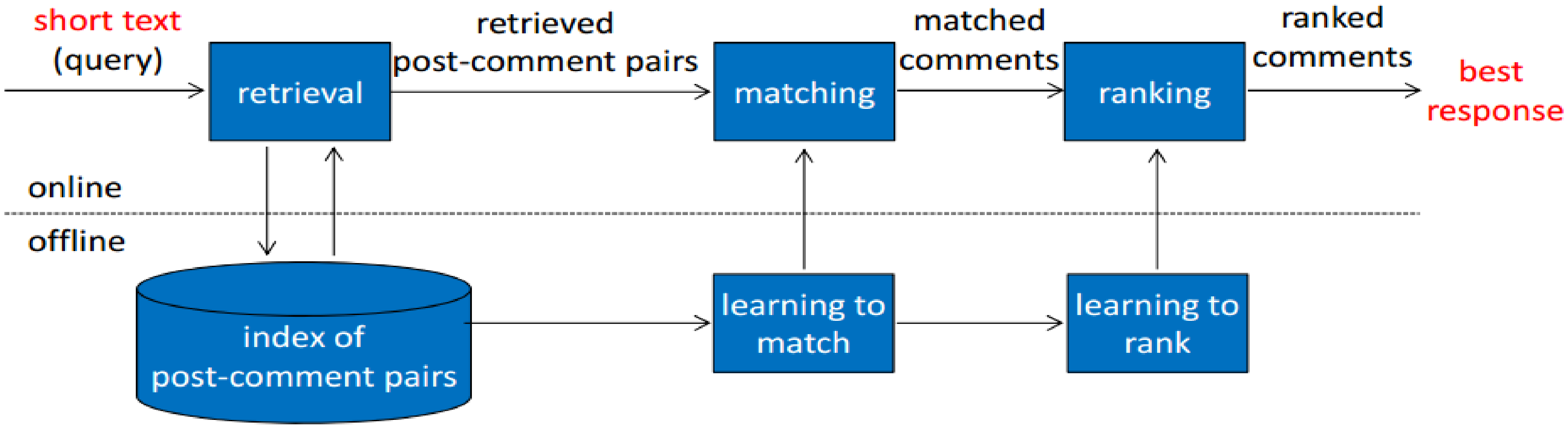
- 引言及研究背景
- 对抗学习与聊天结果多样性的直观联系
- 基于GAN的生成式聊天模型
- 实验及结果分析
- 结论

Agenda

- 引言及研究背景
  - 构建自动聊天系统的两种技术路线
  - 生成式聊天系统的技术背景及模型策略
  - 基于Seq2Seq的生成式聊天系统面临的主要问题
- 对抗学习与聊天结果多样性的直观联系
- 基于GAN的生成式聊天模型
- 实验及结果分析
- 结论

构建自动聊天系统的两种技术路线

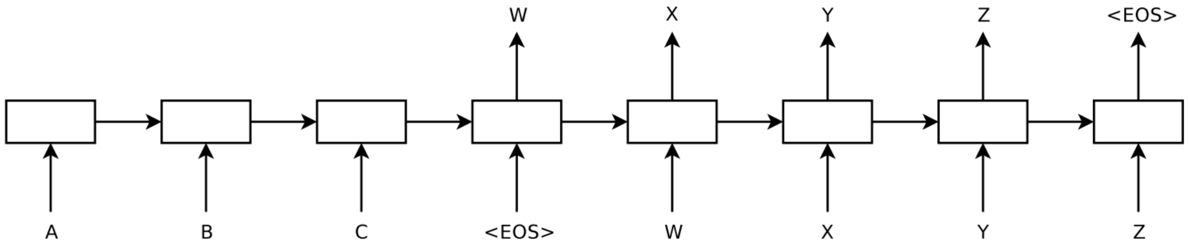
- 基于检索框架的技术路线



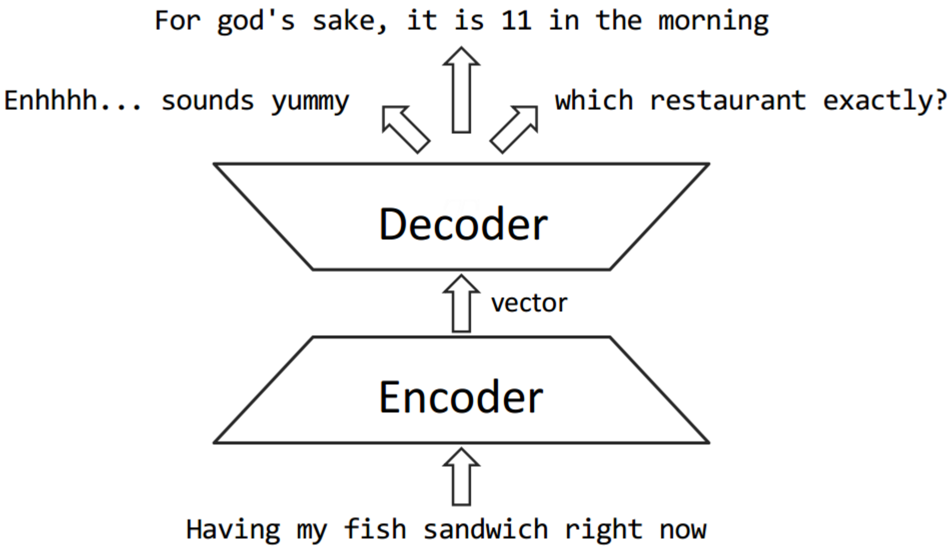
Ji, et al. 2014

构建自动聊天系统的两种技术路线

- 基于生成模型的技术路线



Ilya Sutskever et al., 2014



Lifeng Shang et al., 2015

生成式聊天系统的技术源头及模型策略

- Neural Response Generation (NRG) 溯源
  - SMT→NMT→NRG
  - 问答系统时期，SMT提供了有效的相关性feature
  - 基于Seq2Seq架构的NMT为MT提供了全新的技术范式
  - 问答/聊天可以看做是一种特殊的MT过程
  - 使用Seq2Seq框架实现聊天回复的自动生成是可能的

NRG的模型实现

- General Encoder-Decoder frameworks
  - Vinyals and Le, 2015
- Multi-view training
  - Zhou et al., 2016; Iulian et al., 2017
- Attention mechanism
  - Lifeng Shang et al., 2015; Chen et al., 2017
- Additional features
  - Li et al., 2016; Zhou et al., 2017

NRG面临的主要问题

- Safe Response
  - Boring, Boring, Boring…
  - Always breakdown the conversations.
  - NRG实用化的主要障碍

Query: You swore an oath when you put that uniform on.
Seq2Seq: I don’t know what to do.

Query: Entire town knows your son is a goon.
Seq2Seq: What do you mean?

Query: 你喜欢猫还是狗？ Do you like cats or dogs?
Seq2Seq: 喜欢养猫。I Like cats.

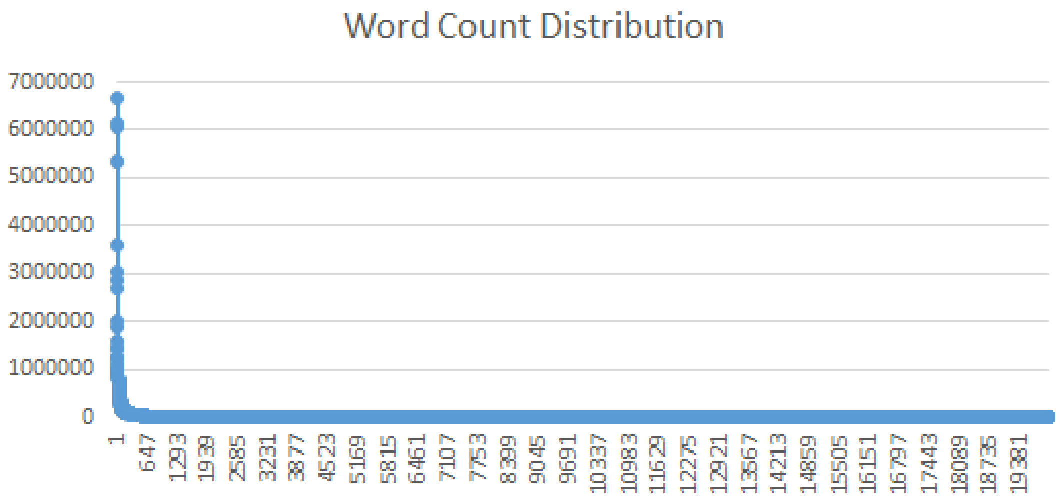
Query: 你像奥巴马的妻子。You look like Obama’s wife.
Seq2Seq: 哈哈哈哈哈哈。Haha…

Agenda

- 引言及研究背景
- 对抗学习与聊天结果多样性的直观联系
  - safe response的产生
  - 提高生成结果多样性的一个经验假设
- 基于GAN的生成式聊天模型
- 实验及结果分析
- 结论

Safe Response的产生

- 产生Safe Response的原因
  - 统计学习的天然特性
  - 词语的概率分布主导decoding过程
  - Generator陷入不合理的优化状态



Word Count Distribution of Responses

提高生成结果多样性的一个直观方案

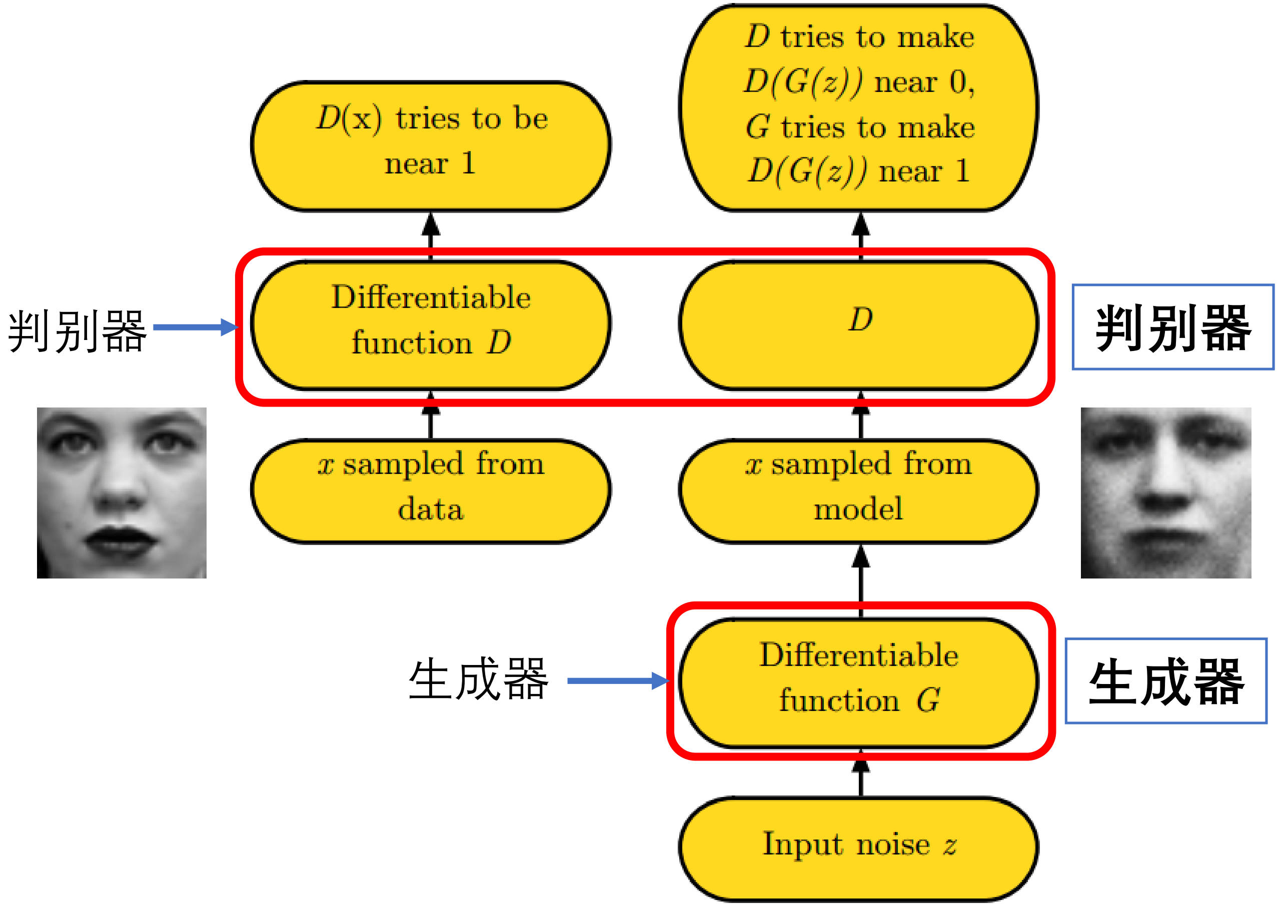
- 单纯最小化生成误差势必导致生成结果倾向于高频回复模式
- 训练一个独立的判别模型区分生成的response和真实response是可能的
- 判别模型需要影响生成模型的每一步词语选择
- 直观上，判别模型“提醒”生成模型什么样的回复是“更好”的

Agenda

- 引言及研究背景
- 对抗学习与聊天结果多样性的直观联系
- 基于GAN的生成式聊天模型
  - Generative Adversarial Nets简介
  - 在文本生成问题上应用GAN的障碍
  - GAN-AEL模型
- 实验及结果分析
- 结论

Generative Adversarial Nets简介

- Generative Adversarial Nets (Goodfellow et al., 2014)
  - 最早用于image processing领域
  - 由一个生成器G和一个判别器D组成
  - 生成器通过输入噪声生成一个尽可能迷惑判别器的样本
  - 判别器负责区分生成样本与真实样本



在文本生成问题上应用GAN的障碍

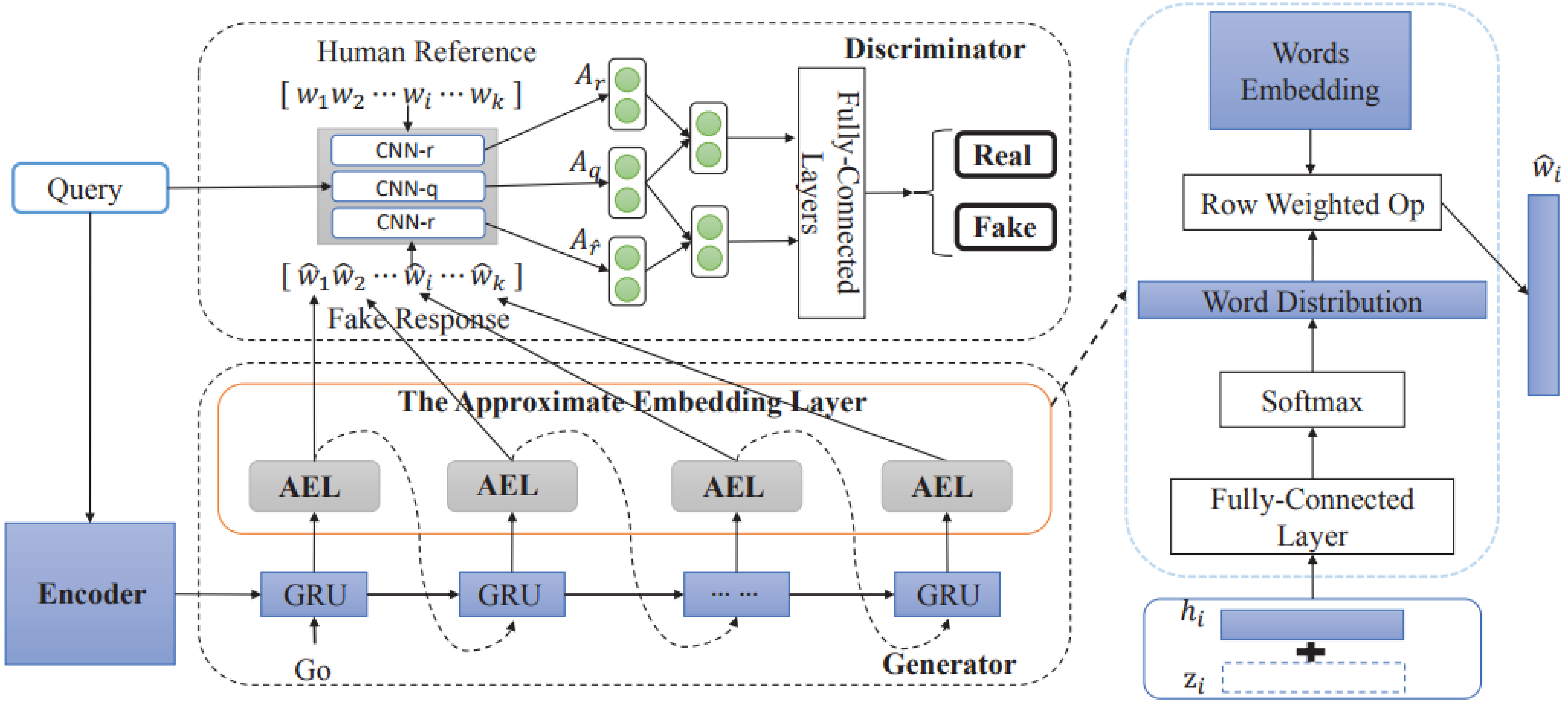
- 文本生成是通过从预测分布中采样得到离散的词序列实现的
- 在GAN中离散的词序列需要作为判别器D的输入
- 离散采样过程是不可导的，导致反向传播(Back-Propagation)中断

- 现有的方法
  - 强化学习（Reinforcement Learning）

GAN-AEL模型概述

- 动因
  - 选择合适的方法取代离散的采样过程
  - 构建连续可导的生成器输入层
  - 直接连接生成器D和判别器G
- 假设
  - 在训练充分的条件下，生成器G输出的理想词分布应该接近词的one-hot表示
- 词向量近似层 (Approximated Embedding Layer)
  - 用生成器G输出的当前step下所有词语的概率分布近似表示当前词语
  - 用近似词向量作为判别器的输入

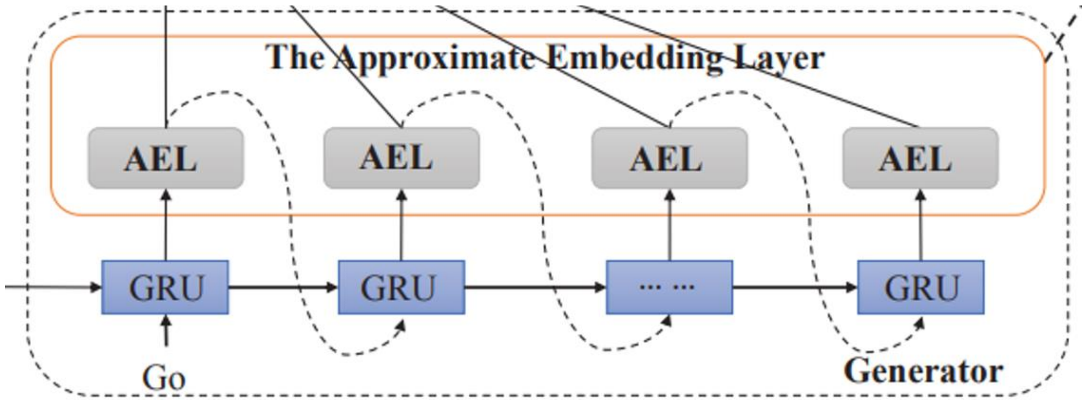
GAN-AEL模型结构

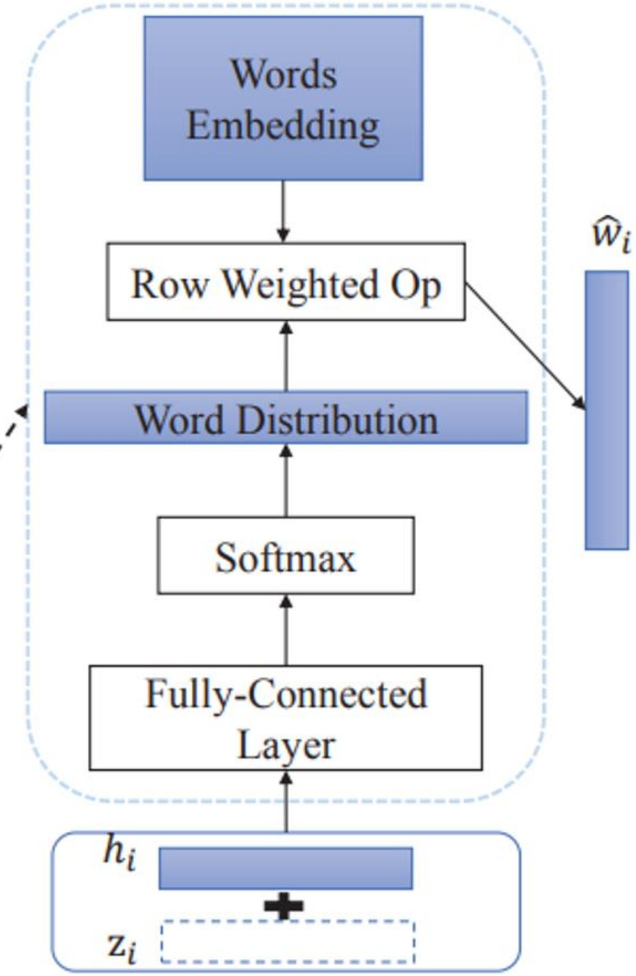

Human Reference
Discriminator
[$w_1 w_2 \cdots w_i \cdots w_k$]
A_r
A_q
$A_{\hat{r}}$
CNN-r
CNN-q
CNN-r
Query
Fully-Connected Layers
Real
Fake
[$\hat{w}_1 \hat{w}_2 \cdots \hat{w}_i \cdots \hat{w}_k$]
Fake Response
The Approximate Embedding Layer
AEL
AEL
AEL
AEL
Encoder
GRU
GRU
… …
GRU
Go
Generator
Words Embedding
Row Weighted Op
Word Distribution
Softmax
Fully-Connected Layer
h_i
z_i
$\hat{w}_i$

Approximate Embedding Layer

$$\hat{e}_{w_i} = \sum_{j=1}^{V} e_j \cdot \mathrm{Softmax}(W_p(h_i + z_i) + b_p)$$

$$h_i = g(h_{i-1}, \hat{e}_{w_{i-1}})$$





目标函数

$$D_{loss} = \log D(r|q) + \log(1 - D(\hat{r}|q))$$

$$G_{loss} = \|A_r - A_{\hat{r}}\|$$

生成器对抗训练

$$\begin{aligned}\nabla_{g_{D,G}(\theta_G)} &= \frac{\partial G_{loss}}{\partial V_{\hat{r}}} \frac{\partial V_{\hat{r}}}{\partial \theta_G} \\ &= \frac{\partial G_{loss}}{\partial V_{\hat{r}}} \frac{\partial V_{\hat{r}}}{\partial G} \frac{\partial G}{\partial \theta_G}\end{aligned}$$

Agenda

- 引言及研究背景
- 对抗学习与聊天结果多样性的直观联系
- 基于GAN的生成式聊天模型
- 实验及结果分析
  - 生成式聊天模型的评价方法
  - 数据集与Baselines
  - 实验结果分析
- 结论

生成式聊天模型的评价方法

- 目前生成式聊天模型尚无统一的评价方法
  - Human evaluation主观性较明显
- 主要的评价指标
  - BLEU
  - Perplexity
  - ROUGE
- 生成回复应注重的两个方面
  - Semantic Relevance
  - Diversity
- 本文采用的evaluation metrics
  - Relevance: word embedding based Greedy (Rus and Lintean, 2012) , Average (Mitchell and Lapata, 2008), and Extreme (Forgues et al., 2014) metrics
  - Diversity: dist-1, dist-2 (Li et al. 2016; Chen et al. 2017), and Novelty
  - Human evaluation

数据集 & Baselines

- Dataset
  - Chinese: Baidu Tieba
  - English: OpenSubtitles
  - 5,000,000 for training; 200,000 for validation; 10,000 for testing
- Baselines
  - Standard Seq2Seq
  - MMI-anti: anti-language model with MMI in the decoding phase
  - Adver-REGS: GAN model with RL proposed by Li et al. (2017)

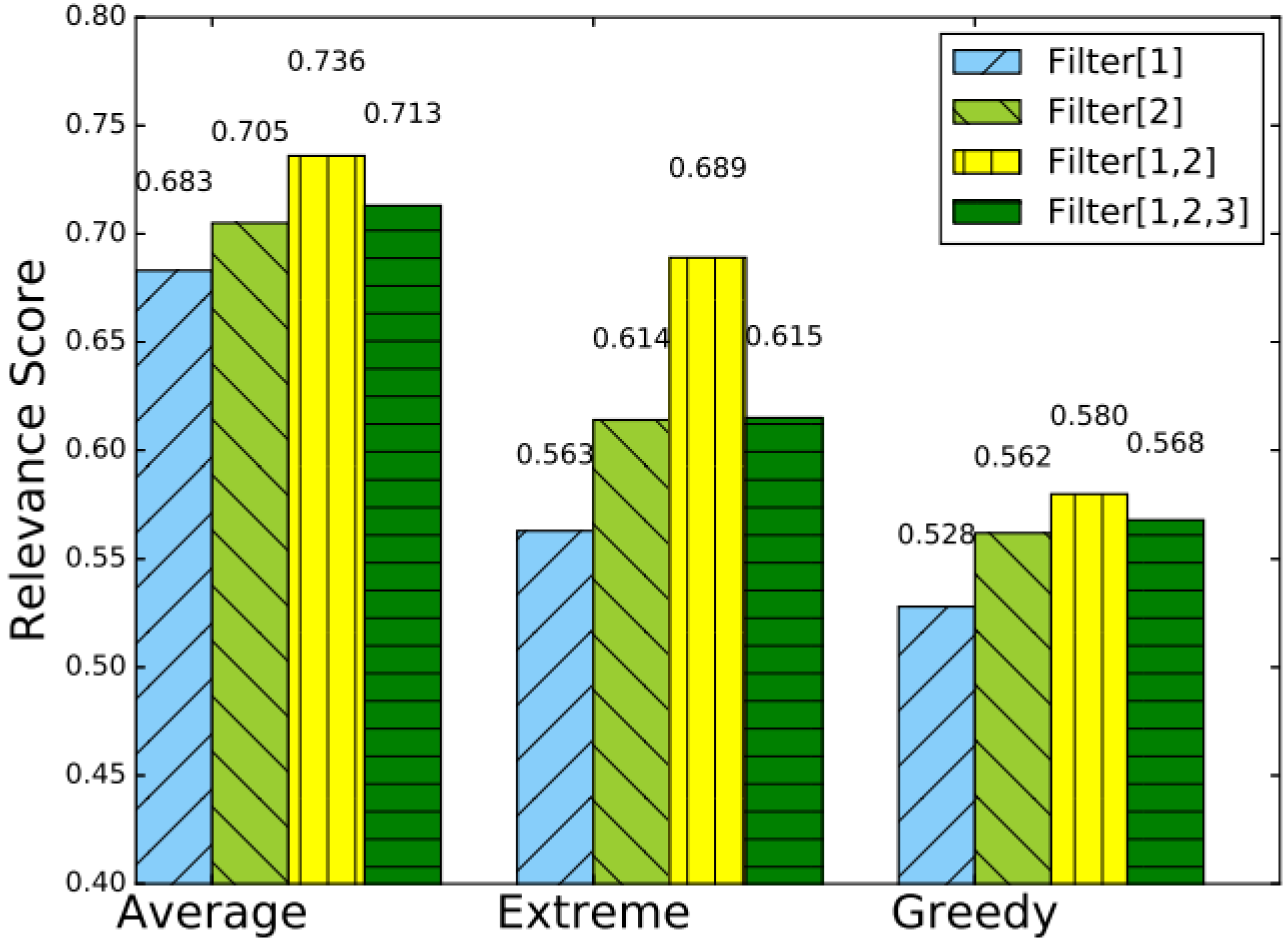
Experimental Results

Model	Relevance			Diversity		
	Average	**Greedy**	**Extreme**	**Dist-1**	**Dist-2**	**Novelty**
Seq2Seq	0.720	0.614	0.571	0.0037	0.0121	0.0102
MMI-anti	0.713	0.592	0.552	0.0127	0.0495	0.0250
Adver-REGS	0.722	0.660	0.574	0.0153	0.0658	0.0392
GAN-AEL	**0.736**	**0.689**	**0.580**	**0.0214**	**0.0963**	**0.0635**

Model	Relevance			Diversity		
	Average	**Greedy**	**Extreme**	**Dist-1**	**Dist-2**	**Novelty**
Seq2Seq	0.719	0.578	0.505	0.0054	0.0141	0.0045
MMI-anti	0.710	0.569	0.499	0.0175	0.0586	0.0097
Adver-REGS	0.726	0.590	0.507	0.0223	0.0725	0.0147
GAN-AEL	**0.734**	**0.621**	**0.514**	**0.0296**	**0.0955**	**0.0216**

判别器复杂度的影响

- 判别器的强弱对GAN模型影响明显
- 判别器能力较弱，则容易误导生成器
- 判别器能力过强，收敛速度过快，则生成器无法有效训练
- 纳什均衡



Generated Samples

Example responses generated based on the OpenSubtitle data	Example responses generated based on the Baidu Tieba data
Query: You swore an oath when you put that uniform on. **Seq2Seq:** I don’t know what to do. **GAN-AEL:** I think I had a good reason.	**Query:** 你喜欢猫还是狗？Do you like cats or dogs? **Seq2Seq:** 喜欢养猫。I Like cats. **GAN-AEL:** 我喜欢金毛。I like golden retriever.
Query: Entire town knows your son is a goon. **Seq2Seq:** What do you mean? **GAN-AEL:** His life is really a sad story.	**Query:** 你像奥巴马的妻子。You look like Obama’s wife. **Seq2Seq:** 哈哈哈哈哈哈。Haha… **GAN-AEL:** 我是大众脸。I have a common face.
Query: Turn his face away from me. **Seq2Seq:** I don’t know what to do. **GAN-AEL:** Let me explain to you about the mistake.	**Query:** 我不光搭讪，还请吃饭呢。 I wish to invite you to dinner rather than chat up only. **Seq2Seq:** 我也想吃。I want to have dinner too. **GAN-AEL:** 我是胖子，减肥。I am fat, and I am on a diet.

结论

- 有效的抑制safe response是生成式对话模型广泛应用的前提之一
- GAN用于改善Neural Response Generator的效果，思想直观，但道路曲折
- GAN-AEL能够在保证生成结果语义相关性的同时，提高生成回复的多样性
- GAN和NRG同为新的研究方向，有足够多的思路值得尝试
- 这是我们的一小步

Reference

- [Ji, et al. 2014] An Information Retrieval Approach to Short Text Conversation
- [Ilya Sutskever et al., 2014] Sequence to Sequence Learning with Neural Networks
- [Lifeng Shang et al., 2015] Neural Responding Machine for Short-Text Conversation
- [Vinyals and Le, 2015] A neural conversational model.
- [Zhou et al., 2016] Multi-view response selection for humancomputer conversation.
- [Iulian et al., 2017] A hierarchical latent variable encoder-decoder model for generating dialogues.
- [Chen et al., 2017] Topic aware neural response generation.
- [Li et al., 2016] A persona-based neural conversation model.
- [Zhou et al., 2017]Mechanism-aware neural machine for dialogue response generation
- [Rus and Lintean, 2012] A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics.
- [Mitchell and Lapata, 2008] Vector-based models of semantic composition.
- [Forgues et al., 2014] Bootstrapping dialog systems with word embeddings.
- [Li et al. 2017] Adversarial learning for neural dialogue generation.

Thanks!
Q&A