利用非对话语料来丰富对话生成模型

本文基于 ACL-2020 论文《Diversifying Dialogue Generation with Non-Conversational Text》，论文作者是腾讯微信 AI 团队。

导言

基于序列到序列（seq2seq）的神经网络模型在开放领域对话生成的任务上经常会出现低丰富度（low-diversity）的问题，即生成的回复无趣且简单。因此，作者提出利用非对话的文本语料去提高对话模型的多样性。相比于传统的对话语料，非对话的文本语料不仅容易获得而且主题包罗万象，因此作者从论坛、书籍和谚语中搜集了大量的非对话语料，结合迭代的回译（back translation）方法将非对话语料融入到对话模型的语义空间内。在豆瓣和微博的数据集上，新模型在保持相关度的同时极大提高了生成回复的多样性。

模型背景与简介

seq2seq 模型已经在很多语言生成任务上取得了很好地效果。然而，当把它应用到通用领域的闲聊生成上时，一个很大的问题就是它倾向于生成像“我不知道”、“好的”这样的通用回复。原因就在于在日常生活中，这些通用回复大量存在于我们的对话里面。Seq2seq 模型会很容易得学习到用通用回复就可以处理大部分对话场景。

目前降低 seq2seq 模型生成通用回复的方法主要有两点：（1）改变 seq2seq 的目标函数本身来让非通用回复获得更高权重。但是模型依然在有限的对话语料上训练，限制了包含广泛主题的能力。（2）用结构化信息、情感、个性等来增强训练语料。但是，这需要昂贵的人工标注，很难应用到大规模的语料。

在这篇文章里，作者提出利用非聊天语料来丰富通用的闲聊模型。与双边成对的聊天语料相比，非聊天语料往往更容易获得，同时也更多样、涵盖不同主题、不需要进一步人工标注。作者从各种数据源收集了超过一百万条非聊天语料，包括论坛评论、谚语俗语、书籍片段等等。基于此作者提出了基于迭代的回译(iterative back translation)的训练方法来利用这些非聊天语料，实验结果显示模型可以生成更多样而且不失一致性的回复。

非聊天语料收集

收集的非聊天语料每个句子长度不宜过长或者过短，可以跟日常聊天主题和风格贴近。作者考虑从以下三个来源收集：

（1） 论坛评论。论坛评论来源于知乎，在知乎上选择了所有获得超过十个喜欢，而且句子长度在 10-30 之间的评论。

（2） 谚语俗语。从多个网站抓取了谚语、俗语、名人名言、歇后语等等。这些语言大多比较精炼，可以直接用来丰富日常聊天。

（3） 书籍片段；从读书 app 上选取了 top 1000 个最受喜爱的小说或者散文。同样，只保留用户高亮过的、长度在 10-30 之间的句子。

进一步对收集的语料做过滤处理，删除了含有攻击性和歧视性语言的句子。最后语料总数超过一百万，其中有 78 万论坛评论、5 万谚语俗语和 20 万书籍片段。

模型结构

作者用{x，y}来表示聊天语料 D 中的上文和回复{context, response}对。t 代表非聊天语料 D_T 中的句子。作者首先考虑几个 baseline 系统：

（1） 检索式： 把 D_T 中的句子作为候选答复，每次要生成回复时，就从中选出最合适的回复。作者用反向 seq2seq 在 D 上训练学出的 p(x|y)来定义合适性。检索式系统最大的瓶颈就在于只能从 D_T 中选择而不能生成全新的回复。

（2） 加权平均：在 D 上训练一个普通的 seq2seq 学习 P(y|x)概率，在 D_T 上训练一个语言模型来学习 L(t)的概率。在解码回复的时候，作者用 p(y|x)和 L(t)的加权平均，这样可以考虑 D 和 D_T 两个语料中的信息。

（3） 多任务：把 D 和 D_T 混合，在混合后的语料上同时训练一个 seq2seq 模型和语言模型，解码器在两个模型之间共享参数，让模型在多任务环境下同时适应两个语料的信息。

除此以外，作者提出利用 iterative back translation 来利用非聊天语料。Iteractive back translation 在机器翻译上已经获得了广泛的使用，但是还没有被用到聊天系统中。模型首先有一个初始化阶段。初始化完成之后会不断重复反向(backward)和前向（forward）阶段。

在初始化阶段，作者在聊天预料 D 上同时训练一个 forward 模型 $p_f(y|x)$和 backward 模型 $p_b(x|y)$，训练目标如下：

$$\mathbb{E}_{X_i,Y_i\sim\mathcal{D}} - \log P_f(Y_i|X_i) - \log P_b(X_i|Y_i) \quad (2)$$

在 backward 阶段，作者用 backward 模型创建伪对(pseudo pair）来训练 forward 模型。目标函数为：

$$\begin{aligned} &\mathbb{E}_{T_i\sim\mathcal{D}_T} - \log P_f(T_i|b(T_i)) \\ &b(T_i) = \arg\max_u P_b(u|T_i) \end{aligned} \quad (3)$$

同理，在 forward 阶段，作者用 forward 模型创建伪对(pseudo pair）来训练 backward 模型。目标函数为：

$$\mathbb{E}_{X_i \sim \mathcal{D}} - \log P_b(X_i | f(X_i))$$
$$f(X_i) = \arg\max_v P_f(v | X_i) \quad (4)$$

具体的算法如下所示：

(**Inilialization**) Train by minimizing Eq. 2
  until convergence;
for *i=1 to N* **do**
  (**Backward**) Train by minimizing Eq. 3
    until convergence;
  (**Forward**) Train by minimizing Eq. 4
    until convergence;
end

Algorithm 1: Model Training Process

实验结果

作者在两个中文对话任务上进行了实验：豆瓣和微博。作者还对比了 standard seq2seq with beam search、MMI、diverse sampling、nucleus sampling 和 CVAE 模型。这些模型都只在聊天语料上进行训练，用了不同目标函数的改进来促进回复的多样化生成。

作者首先进行了自动化评论。在表 3 中，作者汇报了各个模型的 BLEU-2 分数来测量跟 ground-truth 的 overlap；dist-1、dist-2 和 ent-4 来测量生成回复的多样性；adver 来测量回复和上下文的一致性。对于 back translation (BT)模型，汇报了模型在第一个和第四个 iteration 的结果。考虑到模型引入了非聊天语料信息，生成的回复很可能跟原始聊天语料中的词频率、主题有所不同，这样在机器指标自动化评论中会有一个天然的劣势。但是，可以看到模型除了在多样性指标上获得了显著提高之外，在 BLEU-2 和 Adver 指标上也并没有下降，说明模型在学习到多样性的同时并没有丢失其它方面的性能。

Metrics	Weibo					Douban				
Model	BLEU-2	Dist-1	Dist-2	Ent-4	Adver	BLEU-2	Dist-1	Dist-2	Ent-4	Adver
STANDARD	0.0165	0.018	0.050	5.04	0.30	0.0285	0.071	0.206	7.55	0.19
MMI	0.0161	0.025	0.069	5.98	0.42	0.0263	0.143	0.363	7.60	**0.31**
DIVERSE	0.0175	0.019	0.054	6.20	0.38	0.0298	0.130	0.358	7.51	0.25
NUCLEUS	**0.0183**	**0.027**	**0.074**	**7.41**	**0.43**	**0.0312**	0.141	0.402	**7.93**	0.30
CVAE	0.0171	0.023	0.061	6.63	0.36	0.0287	**0.169**	**0.496**	7.80	0.29
RETRIEVAL	0.0142	**0.198**	**0.492**	**12.5**	0.13	**0.0276**	**0.203**	**0.510**	**13.3**	**0.17**
WEIGHTED	**0.0152**	0.091	0.316	9.26	0.22	0.0188	0.172	0.407	8.73	0.14
MULTI	0.0142	0.128	0.348	8.98	**0.27**	0.0110	0.190	0.389	8.26	0.16
BT (ITER=1)	**0.0180**	0.046	0.171	7.64	0.19	**0.0274**	0.106	0.313	8.16	0.15
BT (ITER=4)	0.0176	**0.175**	**0.487**	**11.2**	**0.35**	0.0269	**0.207**	**0.502**	**11.0**	**0.25**
HUMAN	-	0.171	0.452	9.23	0.88	-	0.209	0.514	11.3	0.85

Table 3: Automatic evaluation on Weibo and Douban datasets. Upper areas are models trained only on the conversational corpus. Middle areas are baseline models incorporating the non-conversational corpus. Bottom areas are our model with different number of iterations. Best results in every area are **bolded**.

除了自动化评论，作者也进行了人工评价，结果如表 4。作者随机从每个语料中 sample 了 500 个实例，让人工去评价每个模型生成的回复的流畅性、多样性和与上下文的一致性。实验结果跟机器指标基本一致。

Metrics	Weibo			Douban		
Model	**Rel**	**Inter**	**Flu**	**Rel**	**Inter**	**Flu**
STANDARD	0.32	0.11	0.76	0.26	0.13	0.82
NUCLEUS	**0.46**	0.19	**0.78**	0.38	0.21	**0.83**
RETRIEVAL	0.12	0.35	-	0.09	0.32	-
WEIGHTED	0.19	0.14	0.52	0.15	0.17	0.46
MULTI	0.25	0.21	0.70	0.22	0.23	0.66
BT (ITER=4)	0.43	**0.37**	0.77	**0.39**	**0.48**	0.80

Table 4: Human Evaluation Results

通过对生成回复的结果分析，发现 back translation 可以学到非聊天语料重的新词和句式，这样就可以通过不同上下文生成在原有非聊天语料中不存在的回复。

总结

在这篇文章里，作者提出了一个新的方式来丰富通用领域的闲聊模型。通过用 iterative back translation 来有效利用非聊天语料，显示模型可以从词法和语义两个层面都有效地丰富聊天回复。在跟几个基准模型对比后发现，模型显著提高回复的多样性而不降低其它方面的性能。目前的工作迈出了利用非聊天语料来丰富聊天模型的第一步，未来可以结合更加精细化的过滤、筛选来针对不同领域来自适应地选择利用的非聊天语料。