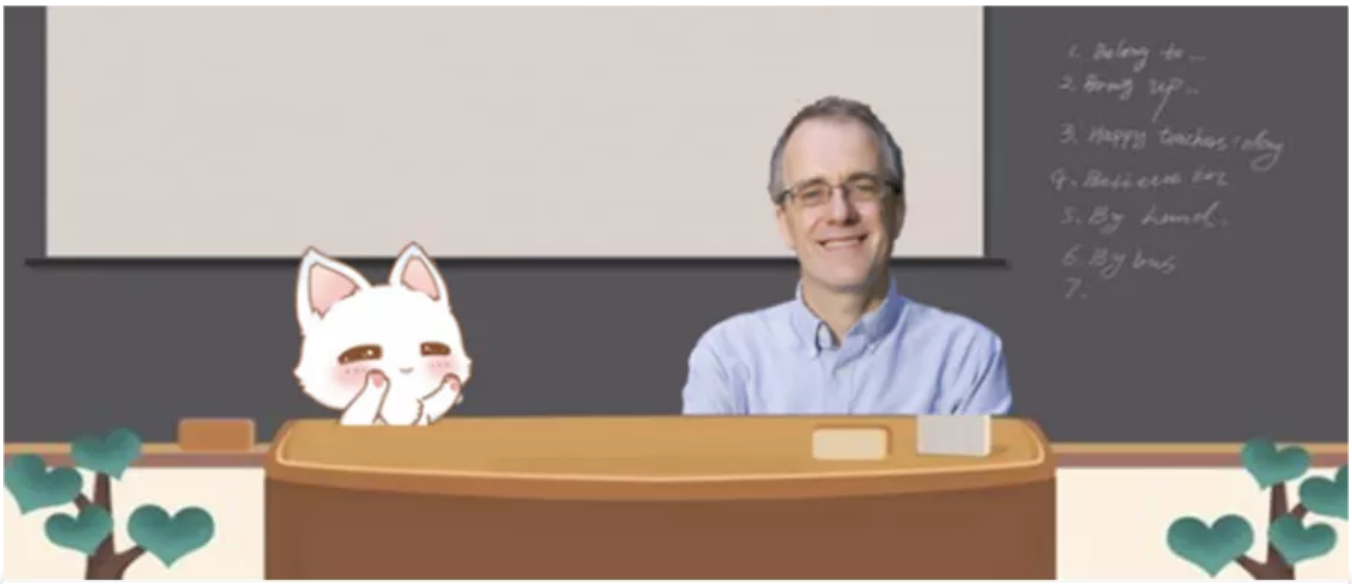
在斯坦福，做 Manning 的 phd 要有多强?

原创 付瑶 夕小瑶的卖萌屋 2021-09-20 12:05



文 | 付瑶
编 | 小轶

博士的毕业论文是我们博士学位教育重要的一环，不仅仅是获得学位的最后一个难关，也是读博期间工作的总结展现。那么**一个优秀的博士在读博期间会做出多少成果？ta 的博士论文又长什么样**？今天，让我们打开一篇最新的斯坦福博士的毕业论文，来看看都讲了些什么。

作者是刚刚8月份毕业于斯坦福的女博士Abigail See。**Abigail 的研究方向是开放式的文本生成，导师是大名鼎鼎的 Chris Manning。目前在谷歌学术上已经拥有 2139 的引用量**。同时，她也是斯坦福 AI Salon，AI woman 两个组织的主要负责人，还连续担任过是斯坦福 cs224n（NLP导论）的助教组长。

Abigail 在读博期间共计发表了 6 篇一作文章。她在博士毕业论文中对自己读博 6 年间的科研成果进行了总结。单论数量而言，平均每年一篇的产量，可能即使放之国内普通高校也不能算十分突出。难得的是篇篇高质量，其中不乏**引用量 1700+ 的超高影响力论文**，以及获得**最佳论文提名的高认可度工作**。

Understanding and predicting user dissatisfaction in a neural generative chatbot
Abigail See, Christopher D. Manning
⋆ Nominated for Best Paper Award ⋆
Special Interest Group on Discourse and Dialogue (SIGDIAL), 2021 (Video presentation)

Neural Generation Meets Real People: Towards Emotionally Engaging Mixed-Initiative Conversations
Ashwin Paranjape*, Abigail See*, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, Christopher D. Manning
3rd Proceedings of Alexa Prize, 2020

Do Massively Pretrained Language Models Make Better Storytellers?
Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, Christopher D. Manning
CoNLL 2019 (Poster presentation)

What makes a good conversation? How controllable attributes affect human judgments
Abigail See, Stephen Roller, Douwe Kiela, Jason Weston
NAACL 2019 (Oral presentation)

Get To The Point: Summarization with Pointer-Generator Networks
Abigail See, Peter J. Liu, Christopher D. Manning
ACL 2017 (Oral presentation)

Compression of Neural Machine Translation Models via Pruning
Abigail See*, Minh-Thang Luong*, Christopher D. Manning
CoNLL 2016 (Poster presentation)

▲Abigail See 读博期间的一作论文

博士论文标题：

NEURAL GENERATIONOF OPEN-ENDED TEXT AND DIALOGUE

论文链接：

https://purl.stanford.edu/hw190jq4736

作者主页：

https://cs.stanford.edu/people/abisee/

工作概述

Abigail 博士期间的研究方向在**开放式文本生成**，但具体应用的下游任务并不集中，主要涉及**摘要**、**对话**、**故事生成** 三类。在这三个子领域上，作者对自己的 contribution 总结如下：

- **摘要**：提出指针生成器模型（pointer-generator network）来提高复制的准确性，以及一个覆盖机制来减少生成摘要的重复。
- **对话**：通过收集大规模用户评价，揭示了机器人行为(如重复、特异性、话题停留和提问)和用户质量判断之间的关系，改善用户体验
- **故事生成**：描述了大规模预训练和解码算法对生成文本的句法、语义、结构和文体方面的影响。作为成果，作者部署研究了一个生成式聊天模型，能够通过分析机器人与用户的交互，确定了机器人的主要错误类型、与用户不满的关系，从而改善对话系统。

文章架构

作者在毕业论文中分为了5大部分来主要叙述自己的研究工作分别是：

- 引言
- 研究背景
- 指针生成网络
- 控制聊天对话的属性
- 预训练对故事生成的影响
- 用户聊天对话中的不满

引言和背景介绍部分我们就略去不表了，主要关注后面四个部分。

指针生成网络概述

本章节中主要叙述了作者构建的指针生成网络 Pointer-Generator的相关工作。该文发表于ACL'17，目前引用量已达1700+。对 NLG 有过了解的同学想必都听说过。

相关论文：

Get to the point: Summarization with pointer-generator networks

论文链接：

https://arxiv.org/pdf/1704.04368.pdf

Pointer-Generator 构建了一个融合网络以及指针网络的混合模型，既允许通过指针复制单词，也允许从固定词汇表中生成新的单词。把sequence-to-sequence模型应用于摘要生成时存在两个主要的问题：（1）难以准确复述原文的事实细节、无法处理原文中的未登录词(OOV)；（2）生成的摘要中存在重复的片段。针对这两个问题，本文提出的融合了seq2seq模型和pointer network的pointer-generator network以及覆盖率机制(coverage mechanism)，在CNN/Daily Mail数据集上，相比于state-of-art，ROUGE分数提升了两个点。

	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	**39.53**	**17.28**	**36.38**	17.32	18.72
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

Table 1: ROUGE F_1 and METEOR scores on the test set. Models and baselines in the top half are abstractive, while those in the bottom half are extractive. Those marked with * were trained and evaluated on the anonymized dataset, and so are not strictly comparable to our results on the original text. All our ROUGE scores have a 95% confidence interval of at most ±0.25 as reported by the official ROUGE script. The METEOR improvement from the 50k baseline to the pointer-generator model, and from the pointer-generator to the pointer-generator+coverage model, were both found to be statistically significant using an approximate randomization test with $p < 0.01$.

控制聊天对话的属性

相关论文：

What makes a good conversation? How controllable attributes affect human judgments

论文链接：

https://arxiv.org/pdf/1902.08654.pdf

作者提出：一个好的对话需要有以下特性：简洁与细节 持续主题与更换主题 问问题和回答问题，对应四种属性：重复性、独特性、回复相关性和问与答。在这部分内容中作者旨在设计通用且易于调整的的控制方法，研究了两种控制方法条件训练（conditional Traning）和加权解码(weighted decoding)。使用条件训练和加权解码来控制四个属性：repetition重复性、secificity特异性、response-relatedness反映相关性和question-asking提问。在测试该任务改进的效果子作者对28种模型配置进行了大规模的人工评估，并进行了人机对话以进行比较。

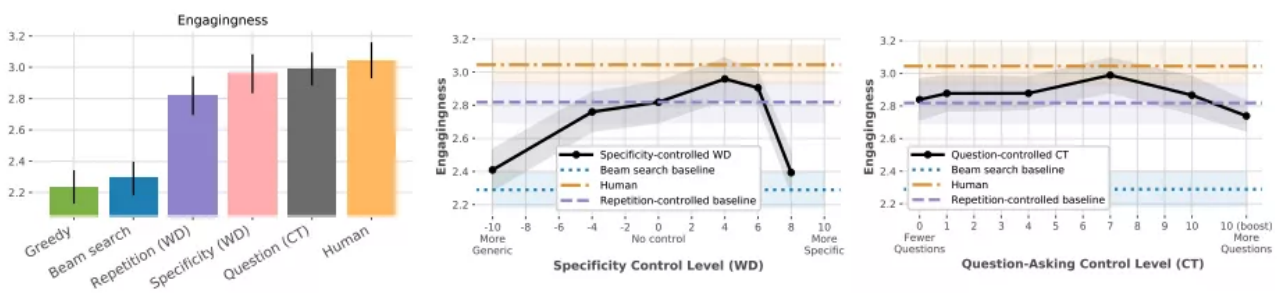
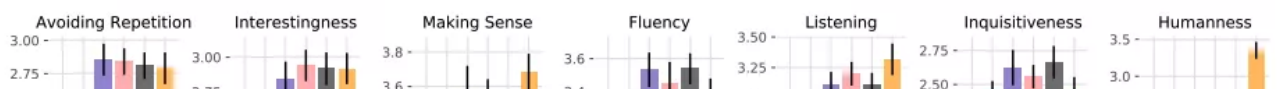


Figure 3: Calibrated human judgments of engagingness for the baselines and best controlled models (left); for different specificity control settings (middle); and for different question-asking control settings (right).



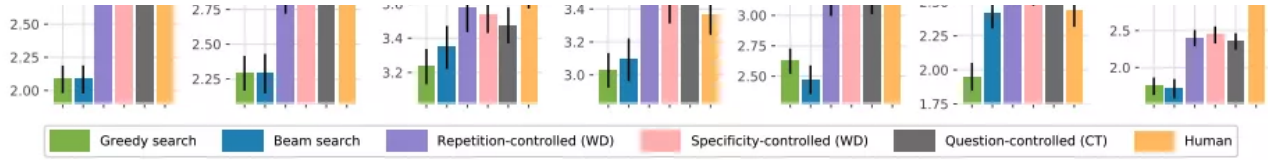


Figure 4: Calibrated human judgments of conversational aspects for the baselines and best controlled models. Note: In Figure 3 and here, the Specificity and Question controlled models both include Repetition control, but Question control doesn't include Specificity control, or vice versa. See Table 8 for exact numbers.

Model	Win%	Top 3 reasons for preferring model
Specificity WD (weight = 6)	84.1%	*More information; Better flow; More descriptive*
Specificity WD (weight = 4)	75.5%	*More information; They describe their life in more detail; Funny*
Specificity CT ($z = 7$)	56.2%	*More information; Better flow; Seems more interested*

Table 3: A/B tests comparing various specificity-controlled models to the repetition-controlled baseline on interestingness. We find all comparisons are significant ($p < .05$; binomial test).

预训练对故事生成的影响

相关论文：

Do Massively Pretrained Language Models Make Better Storytellers?

论文链接：

https://arxiv.org/pdf/1909.10705.pdf

在大规模语料中训练得到的预训练语言模型在很多NLP任务中都取得了较好的表现，但是在开放文本生成中的能力仍未被明确。一些实验结果虽然展现了其潜在的能力，但是并没有关于预训练模型在文本生成的能力的具体研究。作者通过在WritingPrompts-1024上评估，对比了GPT2-117与Fusion model等模型在故事生成的表现。通过多种指标评估生成文本后，研究人员发现了一些可以很好生成故事的模型，以及一些表现不太好的模型。虽然 GPT2-117 在语境上更好，对事件的顺序更敏感，而且使用了更多不常用的词汇，但是它在使用最大似然解码算法时只能生成重复的、没有多样性的文本。

用户聊天对话中的不满

相关论文：

Understanding and predicting user dissatisfactionin a neural generative chatbot

论文链接：

https://sigdial.org/sites/default/files/workshops/conference22/Proceedings/pdf/2021.sigdial-1.1.pdf

🏆 **Nominated for Best Paper Award**

神经生成对话代理已经显示出越来越多的能力进行简短的闲谈对话，神经生成可以实现更强大的社交聊天机器人，能够比以前基于规则或基于检索的对话系统灵活地讨论更广泛的主题。然而，它们在实际部署中的表现-尤其是在嘈杂的环境中与内在动机的用户对话，却没有得到很好的研究。

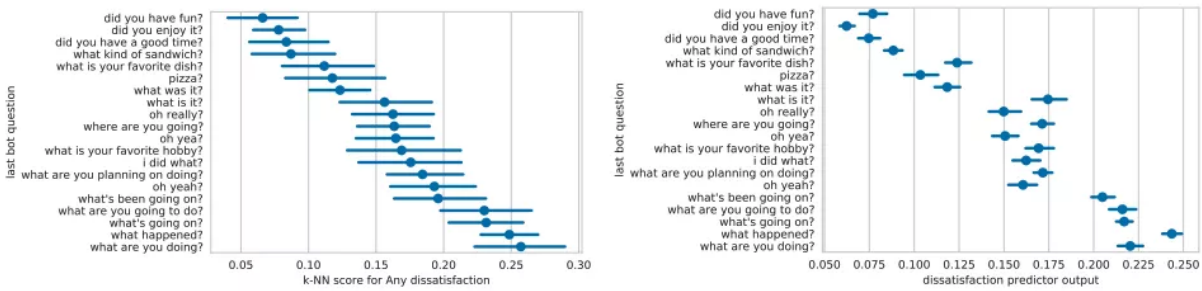


Figure 5: For each of the 20 most common bot questions, mean scores and 95% CIs for Any dissatisfaction given by the k-NN classifier (left) and the predictor (right).

作者对一个神经生成模型进行了详细的案例研究，该模型部署在Chirpy Cardinal (Alexa Prize socialbot)上，在一系列的实验中，发现了不够明确的话语是生成错误的主要来源，如忽略、幻觉、不清楚和重复。除此之外,作者证明了不满意的用户话语可以作为半监督学习信号来改进对话系统，训练了一个predictor用于改进下一轮来减少不满，并通过人类评价表明，作为一个排名函数，它选择了更高质量的神经生成的话语。

Problem	Definition	% in ctrl set	% when no user prob.
User already dissatisfied	The user has already expressed dissatisfaction in c.	12.0%	0.0%
User unclear	The main gist of the user's latest utterance in c is unclear or obscured.	22.0%	0.0%
Bot repetitive	The primary content of b was already said/asked by the bot earlier in c.	6.0%	4.3%
Bot redundant question	b is asking for information that the user has already provided earlier in c.	12.0%	15.9%
Bot unclear	It's hard to find an interpretation of b that makes sense.	12.0%	7.2%
Bot hallucination	b refers to something that hasn't been mentioned, acts like the user said something they didn't, confuses self with user, or seems to be responding to own utterance.	17.0%	10.1%
Bot ignore	b ignores or fails to acknowledge the user's latest utterance, doesn't answer a question, doesn't adequately respond to a request, or switches to an unrelated topic.	20.0%	14.5%
	b is generally on-topic, but makes an assumption or		

Bot logical error	association that's incorrect, unfounded or strange.	15.0%	17.4%
Bot insulting	*b* says or implies something insulting about the user, or about others in a way that might offend the user.	1.0%	1.4%
Any bot error	True iff any of the above *bot* errors are true.	53.0%	46.4%

Table 3: Definitions of problems that may be present in a NeuralChatTurns example (c = context, b = bot utterance); prevalence in the control set ($n = 100$); prevalence in control set examples with no user problems ($n = 69$).

小结

当我们打开论文来看作者在读博期间的研究工作，虽然她在读博期间的论文数量并不算多，但是每一篇章的质量都很高，不仅获得过最佳论文的提名，而且有引用量高达1700的文章，即使有的论文没有太高的引用量，也是对在该领域有深刻影响，是立足所研究课题长远发展的角度进行科研工作。比起快速切换热点来迎合顶会的青睐，她选择了坚定沿着自己的思路，来创立自己的学术宇宙。对一个普通研究生来说，能有一两篇顶会论文已实属不易。但如果志存高远，以领域内的贡献要求自己，你将会看到不一样的峰顶。Chris Manning 和他的 phd 给我们树立了一个很好的榜样。

> 博士毕业文是各位攻读博士学位的同学获取学位必须经历的一道难关，除了学术态度之外，写作的技巧也非常重要。通过这次的拜读经历，小编总结了几条tips分享给大家：
> （1）**梳理脉络**：博士毕业论文篇幅较大，如果作者脉络梳理的不够清晰，不仅会显得研究工作、学术思路杂乱无章，而且会导致读者一头雾水，读不透论文的内容。
> （2）**内容组织**：毕业论文是在读博期间几年围绕课题开展的研究的集合，框架的设定、章节的展开都应与你的研究历程关联，层层剥茧，互为支撑。
> （3）**凸出重点**：支撑大论文写作的研究内容和数据的数量会非常庞大，将与论文相关性较弱的数据剔除，删掉旁支末节，以此来突出自己的主要研究重点及关键实验结果。
> （4）**撰写细节**：在大篇幅写作中，搭配不当、语义重估、语序颠倒等错误的出现不是罕事，这些会对你的论文将会非常的减分。因此，一定要多检查几遍细节。除此之外，论文中的图片也是一种重要的成果展展示，控制所有图片的颜色、尺寸、图中文字的字体、字号，使得你的论文看起来整洁统一。

喜欢此内容的人还喜欢

清华轮椅博士荣获CCF优秀博士学位论文奖！陈文光、谢涛等入选CCF新任会士

新智元

只需2040张图片，训练视觉Transformer：南大吴建鑫团队提出IDMM

机器之心

Science：神经再生电子器件助推人工智能硬件集成

知社学术圈