周报，2020年10月27日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. [***] EMNLP 2020会议PPT制作、视频录制。
2. [**] CCL 2020评测研讨会布置。
3. [**] 准备学校青年教师讲课比赛。
4. [**] 本科毕业设计出题。
5. [*] 实验室组会与讨论。
6. [*] 参加中国计算机大会（CNCC）和YSSNLP会议。
7. [***] 软件工程博士点申报材料。
8. [**] 其他：完善实验室在线文档、写研究生科研手册、读论文、看相关Tutorial。

- 注: *表示任务紧急程度。

2 上周计划执行情况

1. [✓×] 制作完成一个。另一个与张凯继续准备。
2. [✓×] 评测结果和日程已安排就绪并发布。研讨会细节等待大会组委会布置。评测报告待收集。
3. [✓×] PPT准备一半，比赛之前再做。
4. [✓] 毕业设计出了4道题。
5. [✓] 参加了组会与Reading Group讨论。蔺伊健讲ACL 2020最佳论文，吕高艳没做报告（会后有讨论），韩鸿飞进入实验室首次参加活动。
6. [✓] 参加了CNCC部分论坛，听了YSSNLP部分报告。
7. [✓] 完成博士点申报材料，已提交给学院。
8. [✓×] 初步建立实验室在线文档，督促学生充分利用；完成研究生科研手册初稿，已发给学生，后续有时间再完善；Stanford Transfer learning and meta learning课程看了一点。

- 注: ✓×分别表示完成与未完成。

3 本周部分重点工作详述

3.1 CNCC和YSSNLP部分报告

- CNCC文本生成论坛
  - 宋睿华介绍隐喻生成和电影剧本生成，提到一篇NLPCC隐喻质量评价的文章。
  - 万小军指出神经文本生成的三个挑战：多样性（diversity）、连贯性（coherent）和保真性（faithfulness）。其中，提高多样性的方法有：Stochastic Decoding (Top-K sampling, Top-P sampling/nucleus sampling)，Unlikelihood Training，CVAE（通过采样符合特定分布的不同向量得到不同的输出），Syntax Exemplar-based。
  - 黄民烈主要分析预训练语言模型蕴含哪些知识，以及局限性。提出应该融入外部知识帮助生成。他们有一个工作叫Knowledge-enhanced post training可以关注。
  - 史树明讲腾讯内部文本生成的应用模式，基本是传统的模板方法或基于检索的方法和神经网络方法相结合。提到一个挺有意思的任务“电竞解说”。游戏每秒钟产生的数据非常多，他们的策略是先预先概括出一些事件，根据事件去生成解说。腾讯有个模型叫SongNet，能够接受任意模式，生成符合格式的文本，包括诗词、对联、歌词等。模型显式地对格式符号、韵脚位置符号、字、句未知符号以及格式信息进行建模。应该有论文，可以关注。
  - 肖欣彦介绍了百度的一些文本生成工作。篇章图结构对长文本生成有帮助。他们有一个百度大脑智能创作平台。
  - 李磊介绍了字节跳动中一些文本生成应用，如体育新闻自动生成器等。他提到一个概念挺有意思：从个性化推荐到个性化写作，以读者为中心来写作（详细程度、深度、趣味度），高效覆盖长尾话题，针对读者关心的问题来写作，即时性。
  - 总结：大家普遍关心文本生成的一些性质：多样性、趣味性（**amusing and attractive**）和惊喜（**surprise**），这些目前都解决得不好。
- YSSNLP报告言外之意的理解。
  - 清华大学心理系张丹的报告。其中有一个观点非常有意思：From perceived speech to implied meaning，人在听别人说话的时候，实际上是边听边预测的。我一直也是这么认为的，现在有了理论依据。之前做幽默识别的时候就想到这一点，幽默、反讽、隐喻之类的语言一定是和读者的预期产生冲突才有了搞笑、讽刺、意料之外的效果。如何模拟作者的预期呢？或许预训练语言模型可以用来干这个。可惜还没有能做出来。

3.2 实验室学生情况

- 周淑慧：隐喻识别的Baseline基本完成了，有进步。错误报告发给我了，等着我给她分析呢。让她干啥，就“只”干啥。之前讨论时给了她一些方向：比如利用选择性Mask去继续训练预训练语言模型捕捉selectional preference；隐喻应该是一种关系，分类的时候可以考虑动词和名词的组合；想想能引入哪些外部知识。也不知道听没听明白。
- 屈原斌：离题数据准备过程中，要提醒他多和讯飞研究组和资源部沟通，把能想到的问题都想清楚。对已有工作有了一定了解，但是还很不系统。还是在等着别人告诉怎么做，自己缺少思考，需要改进。给自己的方法起个名字，起不出来就别着急动手，明确你做的是不是新的东西。如果是做Baseline，尽快把所有可能的Baseline都实现出来。
- 蔺伊健：写的中文论文初稿问题还很多。组会讨论达成一些共识：把客观的表达方式和趣味性等主观因素分开多个维度考虑，重新校对和标注数据，扩大标注规模。模型部分要找时间讨论细节，先确保能做对。提醒他关注大方向，不要纠结细节。
- 陈桂桦：周报论文阅读笔记写得不错，要在实验室推广。Life-long learning是一个很有意思的方向，要注意现有研究都针对哪些任务，这些任务的数据能不能拿到。Life-long learning是不是可以应用到作文评分上？比如一个人如果对多个不同题目的作文评过分，她的能力应该越来越强，但现在的机器评分模型换个题目就不行了，这有巨大的Gap。先

把Life-long learning有哪些策略了解清楚，分别是什么，有什么优缺点。而后可以在作文评分上进行一个empirical comparison，看看直接应用已有策略能达到什么水准。

- 吕高艳：做了一个问题评分的PPT，内容还可以，但还有很多问题需要细化。要完全掌握至少一个SOTA模型，能够说出模型实现的每一个细节。不能一问，就说用BERT，继续问怎么用的，就不清楚了。主观题评分有以下要素：问题、标准答案、学生作答，学生作答有的是有评分的，有的可能是没有评分的。已有模型都利用了哪些信息，怎么和BERT结合的、输入输出是什么？还没有利用的信息能不能利用，怎么加到BERT里？已有模型是怎么训练的，只利用了同问题数据还是也利用了其他问题数据？训练会不会过拟合？为什么？如果会，如何缓解？

- 韩鸿飞：刚进入实验室，请王荣荣帮助解决机位。尽快进入节奏。暂时选择PDTB篇章关系分析问题，先练练手。可以向宋子尧师兄请教。篇章分析的难度还在于训练数据规模太小，可以考虑数据增强方法。

- 宋子尧：篇章角色识别不能只看评价指标数字，要看实际效果，case by case地进行提升。人一看就能识别出来的角色，机器一定不能出错。无论用什么办法，要把效果提上去。

3.3 一些想法

- 大家关注的一些AI重要研究方向
  - Robustness/common sense：SOTA models should work in real-life. 模型应具有健壮性能够理解常识，真正有效的模型应该能解决生活中的真实问题，而不是只在特定数据集上好用。
  - few-shot learning：小样本学习。
  - continual learning：持续学习，和life-long learning一个概念
  - explainability：可解释性。我为什么能够信任这个模型？
  - efficiency：给我GPT3我也用不起啊！模型压缩未来是个大方向。
  - 评论：这几个问题都是通用问题，每个人都要结合自己的任务来思考：如何有效的评价？如何增强可解释性？能不能减少人工标注训练数据的规模也能学习得很好？

- ACL 2020最佳论文。蔺伊健介绍的这个文章挺有意思，其实和Data augmentation for NLP可以视为“矛盾”组合或对偶任务。启发是：
  - Data Augmentation是一个重要的实用手段。每个人都应该想一想怎么自动构建训练数据做数据增强。比如离题检测、作文评分能不能自动构建训练样本？需要考虑哪些问题？关于数据增强的一些参考文献：
    - EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. EMNLP 2019.
    - Does Data Augmentation Improve Generalization in NLP? Arxiv 2020.
    - ACL 2020 best paper作者的其他论文。
      这些文章主要针对底层基础任务，针对篇章级的工作还没有看到，可以尝试。
  - 论文里提及的CheckList的方法能不能应用到每个人自己的任务上。比如隐喻识别、篇章关系分析能不能设计一些像CheckList里的规则自动生成样本，验证模型的可靠性？

每个人看到或听到一些技术，都要想一想能不能迁移到自己的任务上。督促学生每周要推荐论文，更新到在线文档。

4 存在问题

- 问题1

- 问题2

5 下周计划

1. [***] CCL 2020评测研讨会组织。
2. [***] 准备学校青年教师讲课比赛。
3. [**] 考虑NAACL和ACL投稿计划。
4. [**] 组会和讨论，继续细化学生的研究任务和方向。
5. [**] 参加CCL2020线上会议。
6. [*] 学习，读ACL、EMNLP和ICLR相关最新论文、心理学相关论文。更新篇章分析论文列表。

A 附录

A.1 本周阅读论文列表

论文名称	会议/期刊	状态
EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks	EMNLP 2019	略读
Does Data Augmentation Improve Generalization in NLP?	Arxiv 2020	略读
Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking	ACL 2020	细读

Table 1: 本周论文阅读列表

A.2 阅读笔记

详见表Table **??**.

A.3 实验结果

略。

论文阅读笔记**1**	
题目	Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking
作者及单位	Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, Monica S.Lam, Stanford University
发表会议	ACL 2020
背景	自动对话代理可以减少客户支持的费用，然而训练一个特定领域的对话系统需要大量的标注数据去覆盖所有可能出现的对话情况。通常使用 Wizard-of-Oz 技术，这种技术通常是让两个外包的工作人员与对方进行交流来标注每种对话情况。但是这种数据最初的获取不仅昂贵，并且由于人为的错误和标注的延迟等问题让正确的标注成为一个巨大的挑战。并且不同领域的对话系统需要预训练大量的标注数据，这是的对话系统难以应用预料缺乏的领域，其较差的泛化能力也限制了它的应用场景。于是该文提出针对多领域对话状态追踪的零样本多领域迁移学习技术来减少对话模型过于依赖数据的问题。
贡献	1. 提出一种新的针对对话状态跟踪的零样本多领域迁移学习技术。 2. 这种方法在 MultiWOZ 2.1 数据集任务中的零样本多领域迁移学习中结果平均提高了 21%。 3. 这种方法还提高了在基于 RNN 模型的 TRADE 和基于 BERT 的 SUMBT 模型上的准确率，这也表明了该技术独立于所使用的特定模型。 4. 实验结果显示，合成后的数据能够完善 BERT 预训练过程。仅用合成数据训练时，SUMBT 模型可以取得用全部数据训练的 61% 92%的准确率。作者提出将预先训练的模型与合成的数据相结合作为一种通用的技术来引导新的对话状态跟踪器。
方法	建立一个领域独立的对话模型和一个合成算法，这个算法接受域本体以及一些与这个域相关的常识性词汇。这个算法也是基于这个模型来合成训练数据，一个新领域的数据通常是和现有其他领域的数据一起训练的。此外，调整相关领域的训练样本是通过用新领域的词典代替。从标注后的数据中提取一个模板(template)按符合语法和语义的要求来合成一个领域的对话。 文中定义了四类模板，分别是:领域主题模板 (Domain Subject Templates)、槽名模板(Slot Name Template)、槽值模板(Slot Value Templates)、信息表达模板(Information Utterance Templates)。领域主题模板描述这个领域的不同名词短语;槽名模板描述在没有值的情况下引用槽名的方法;槽值模板描述了指槽及其值的短语,它们可以是名词短语(“意大利菜餐馆”)、被动动词短语、主动动词短语(“供应意大利菜的餐馆”)、形容词短语、介词从句;信息表述模板描述提供信息的完整句子。这些模板是用于特定领域的，因为他们使用特定领域的结构。
实验	该实验是基于 MultiWOZ 数据集做的，它是一个多领域全面标注人与人谈话的语料库。分别在 TRADE 和 SUMBT 模型上评估数据合成技术。 该文还对每个特定域的对话子集计算了该域槽值的精度，还从联合精度方面对不同域的对话模型进行了评估。
评论	我认为本文还是不错，但也有瑕疵。 1. 优点： 2. 缺点：
疑问	1. 由于之前没有看过对话模型的相关文献资料，所以这篇文献中的很多术语不太理解，读起来就很吃力。 2. 实验看不懂，只知道结果，为什么从槽精度和联合精度这两面评估还是不太明白。

Table 2: 本周论文阅读笔记1