Sentence Function Recognition Based on Active Learning

Chen Guo[1,2]　Xu Tianxiang[1]
[1](School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)
[2](Jiangsu Science and Technology Collaborative Innovation Center of Social Public Safety, Nanjing 210094, China)

Abstract: [Objective] This paper uses active learning methods, structured abstracts and a few annotations to create a classification model for sentence functions, aiming to reduce the dependence on manually labeled corpus. **[Methods]** First, we trained the SVM, CNN and Bi-LSTM classifiers with structured function sentences from abstracts. With the help of active learning techniques, we predicted the function of a large number of unlabeled common abstract sentences. Third, we automatically identified uncertain samples for manual annotation, which were used to optimize the initial classifier. Finally, we used active learning to improve the performance of classifiers. **[Results]** We examined the new method with Library and Information Science literature. The precision, recall, and F1 values were 84.65%, 84.49%, and 84.57%, which were 3.25%, 3.24%, and 3.25% higher than those of the traditional methods. **[Limitations]** We only conducted five iterations to avoid massive work of manual corpus annotation. **[Conclusions]** Active learning method could effectively discover the difference between unlabeled corpus and existing training corpus, which also reduces the manual labeling costs. The proposed method might be used in citation and full text analysis.

Keywords: Structured Abstract　Sentence Function Recognition　Active Learning　Short Text Classification

从 ACL 2019 年会看自然语言处理未来发展趋势

自然语言处理(NLP)领域的顶级盛会 ACL(计算语言学协会年会)已于 2019 年 8 月初落幕，亚马逊 Alexa AI 机器学习科学家 Mihail Eric 参加了此次会议并对本次会议进行了一次比较全面的回顾，提炼了 NLP 的关键知识点和发展趋势:

(1) NLP 应用百花齐放: NLP 领域开发的模型和工具有解决许多实际问题的潜力，包括在新闻领域、健康领域，以及生物医学领域的应用。

(2) 一种新的 NLP 范式(先预训练、再调优)已经出现: 随着强大的预训练表征的出现，一些使用语言建模目标进行训练的 NLP 技术已经可以被直接使用，例如: ELMO、OpenAI GPT 以及 BERT。它们在大规模数据上进行预训练，然后在一些较小的领域内的语料库上针对任务进行调优。实际上，这种策略已经成功地在现有的 NLP 对比基准实验中取得了目前最先进的性能。

(3) 将知识注入 NLP 架构: 虽然现有的经过预训练的语言超级模型架构十分强大，但是从原始文本语料库中所学到的东西是几乎不受限制的，本届多项研究试图将知识图谱和神经网络相结合，通过融合相关的知识资源中的信息，进行 NLP 模型构建。

(4) 开始注重模型的可解释性: 神经网络是一种黑箱模型，不具备可解释性。暂且不考虑完全解释这些模型是否必要，至少可以认为，对模型内部在某种程度上的理解可以对未来的架构设计产生深远的影响。

(5) 再次思考自然语言生成的评价与假设: 自然语言生成任务的复杂性是很棘手的，尤其是对于研究社区来说，对模型的评价仍然无法达成共识。

(6) 如何超越“预训练-调优”范式？经过不断迭代，当前的 NLP 模型近乎达到了先进的水平，但是 NLP 领域的主流观点仍然是: 还有一些问题需要改进。大部分的模型仍然是针对特定数据集做工作，而不是针对特定任务。已经建立的模型可以非常有效地收集和利用数据集特有的偏差。在这个过程中，评价指标又展示了相当具有误导性的分析结果。考虑到这些评价对比基准对于自然语言任务发展的重要意义以及模型开发的速度，对比基准也应该与时俱进，开发一套不断演化的、难度越来越大的对比基准是有必要的。

(编译自: https://www.mihaileric.com/posts/nlp-trends-acl-2019/)

(本刊讯)