智源学者成果展示——自然语言处理

作者　穗志方（北京大学）

2020年6月

项目负责人：穗志方

单位 / 团队：自然语言处理

成果名称：自然语言推断中唯假设偏置的消除方法

成果简介：

传统的自然语言推断任务模型仅凭借假设句做出正确判断而完全忽略了前提句。此项研究旨在针对这种唯假设偏置提出对抗样本并探索消除偏置的方法。我们首先从训练数据集的假设句子中导出不同粒度的短语（人造特征）并展示他们与特定标签有着强关联，接着基于这些短语提出了对抗样本集。在这些对抗样本上测试了预训练模型和非预训练模型的表现。基于所挖掘的短语，本研究提出了下采样和对抗训练两种方式来消除唯假设偏置并取得良好效果。

Tianyu Liu, Xin Zheng, Baobao Chang and **Zhifang Sui.**(2020) HypoNLI: Exploring Artificial Patterns of Hypothesis-only Bias in Natural Language Inference. The 12th Language Resources and Evaluation Conference, LREC 2020

SNLI 和 MultiNLI 数据集中人造特征与特定标签的共现概率 (%)

	Multi-word Patterns						Unigram Patterns					
	Entailment		Neutral		Contradiction		Entailment		Neutral		Contradiction	
SNLI	in this picture	96.4	tall human	99.7	Nobody # # .	99.8	outdoors	78.8	vacation	91.0	Nobody	99.7
	A human	96.4	A sad	95.6	dog # sleeping	97.5	sport	75.1	winning	89.9	No	95.8
	A # # outdoors .	95.9	A # human	94.1	There # no	96.2	instrument	74.4	favorite	88.7	cats	93.4
	A # # outside .	89.8	the first	88.6	in # bed	94.2	animal	68.5	date	87.4	naked	88.7
	is near # # .	87.6	on # way	87.0	at home	93.5	moving	67.8	brothers	85.6	tv	88.4
MultiNLI	It # possible	71.7	, said the	93.6	There are no	92.4	Several	54.7	addition	69.6	None	85.4
	There # a # # the	70.8	They wanted to	81.4	does not # any	91.9	Yes	54.4	also	68.6	refused	80.5
	There is an	68.8	the most popular	78.7	no # on	91.5	various	53.7	locals	65.7	never	79.0
	are two	67.0	addition to	78.4	are # any	90.1	· · ·	53.1	battle	63.3	perfectly	77.3
	There # some	65.9	because he was	77.8	are never	89.9	According	53.1	dangerous	63.2	Nobody	77.1

Beijing Academy of Artificial Intelligence

微信关注

北京智源人工智能研究院