基于预训练语言模型的商品 属性抽取

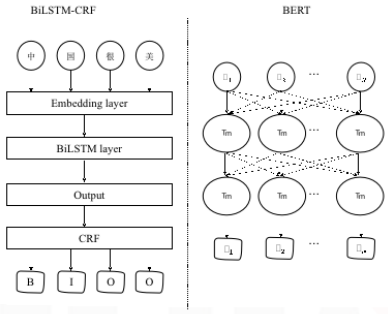
张世奇, 马进, 贾昊 ,周 夏冰, 陈文亮, 张民
苏州大学 计算机科学与技术学院, 江苏 苏州, 中国

主要贡献

■ 我们提出了一种基于扩充三元组的远程监督(EXDS)标注方法，该方法有效缓解了远程监督(DS)在标注数据产生的漏标问题。其思想是在相似类目之间进行属性和 属性值的扩充，弥补类目三元组属性缺失和 属性值覆盖度不足的缺点。

■ 我们构建了人工标注测试集，最终获得面向 电商 的多领域商品 属性抽取 标注数据集。基于新构建的数据集，本文基于多种预训练语言模型进行了领域内和 跨领域属性抽取，发 现增加少量目标领域标注数据可 以有效提高跨领域属性抽取 效果，增强了模型的领域适应性。

模型



基于扩充三元组的远程监督

原始属性值词典		
帆布鞋	颜色	黑色
帆布鞋	颜色	白色
篮球鞋	风格	复古

类目：帆布鞋

复	古	白	色	帆	布	鞋	配	黑	色	喇	叭	裤	真	好	看
O	O	B 颜	I 颜	O	O	O	O	B 颜	I 颜	O	O	O	O	O	O
		色	色					色	色						

扩充后属性值词典		
帆布鞋	风格	复古
帆布鞋	颜色	白色
帆布鞋	颜色	黑色
篮球鞋	风格	复古
篮球鞋	颜色	白色
篮球鞋	颜色	黑色

类目：帆布鞋

复	古	白	色	帆	布	鞋	配	黑	色	喇	叭	裤	真	好	看
B 风	I 风	B 颜	I 颜	O	O	O	O	B 颜	I 颜	O	O	O	O	O	O
格	格	色	色					色	色						

数据统计

属性值标注数目

数据集	DS	EXDS
微博	22491	40944
标题	22159	34020
评论	7245	8921

属性值分布

数据集	平均每句属性值种类数
微博	5.9
评论	2
标题	6.8

三类标注属性数目统计

	DS			EXDS		
数据集	颜色	风格	材质	颜色	风格	材质
微博	5179	10715	6597	8210	21085	11649
标题	2956	12360	6843	4069	19554	10397
评论	1241	2668	3336	1472	3670	3779

领域内属性抽取 实验

模 型	数据集	EXDS F	EXDS F(OOV)	DS F	DS F(OOV)
BiLSTM-CRF	微博	0.598	0.105	0.501	0.094
ELMo-BiLSTM-CRF		0.597	0.113	0.502	0.097
ALBERT		0.616	**0.156**	0.493	**0.114**
BERT		**0.623**	0.114	0.51	0.072
ELECTRA		0.618	0.076	**0.515**	0.068
RoBERTa		0.591	0.093	0.418	0.035
XLNet		0.617	0.127	0.502	0.083
BiLSTM-CRF	标题	0.746	0.097	0.639	0.084
ELMo-BiLSTM-CRF		0.75	0.122	0.619	0.09
ALBERT		0.75	0.125	0.622	0.113
BERT		0.762	0.107	0.606	0.085
ELECTRA		**0.768**	**0.169**	**0.639**	**0.132**
RoBERTa		0.737	0.089	0.576	0.035
XLNet		0.764	0.124	0.615	0.101
BiLSTM-CRF	评论	0.758	0.121	0.698	0.103
ELMo-BiLSTM-CRF		0.762	0.13	0.714	0.134
ALBERT		0.764	0.193	0.703	0.119
BERT		0.765	0.144	0.696	0.099
ELECTRA		**0.772**	0.18	**0.723**	0.13
RoBERTa		0.743	0.098	0.664	0.073
XLNet		0.77	**0.238**	0.717	**0.196**

■ 实验结果显示，各 个模型使用EXDS的效果都远好于DS。在OOV方面，绝大多数模型利用EXDS构建的标注数据训练可 大幅提升 其识别OOV的能力，该实验结果进一步证明了EXDS标注方式能有效缓解漏标问题。

跨领域属性抽取 实验

源领域：标题数据，目标领域：评论数据

模 型	领域内 F	领域内 F(OOV)	跨领域实验一 F	跨领域实验一 F(OOV)	跨领域实验二 F	跨领域实验二 F(OOV)
ALBERT	0.764	0.193	0.692	0.103	0.744	0.130
BERT	0.765	0.144	**0.749**	**0.107**	**0.760**	0.057
ELECTRA	**0.772**	0.180	0.727	0.076	**0.760**	0.145
RoBERTa	0.743	0.098	0.677	0.036	0.725	0.055
XLNet	0.770	**0.238**	0.736	0.089	0.755	**0.146**

源领域：微博 数据，目标领域：评论数据

模 型	领域内 F	领域内 F(OOV)	跨领域实验一 F	跨领域实验一 F(OOV)	跨领域实验二 F	跨领域实验二 F(OOV)
ALBERT	0.764	0.193	0.754	**0.145**	0.752	**0.154**
BERT	0.765	0.144	**0.759**	0.114	**0.756**	0.069
ELECTRA	**0.772**	0.180	0.757	0.122	0.755	0.096
RoBERTa	0.743	0.098	0.718	0.070	0.731	0.079
XLNet	0.770	**0.238**	0.746	0.117	**0.756**	0.122

实验一仅利用源领域数据
实验二在源领域数据中加入少量目标领域数据

由实验结果可 得:
■ 1）仅使用源领域数据微调模型，易使模型偏向 于建模大量属性值与类目的上下文关系，导致实验一的结果较差。
■ 2）训练集中的少量目标领域标注数据有利于缓解由两类数据间属性值分布不一致造成的模型性能下降问题。