基于数据增广的渔业标准命名实体识别研究

程名 于红 杨鹤 任媛 刘巨升 付博

大连海洋大学

论文摘要

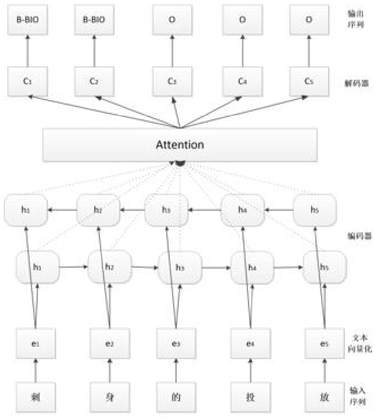
针对渔业标准命名实体识别任务中“水产品名称”类实体语料分布稀疏而导致的效果不好的问题，提出了基于数据增广的渔业标准命名实体识别方法，分别采用基于领域词典的同类词替换和基于上下文特征保护的随机删除算法进行语料库的数据增广，可以有效的提升样本的多样性。用基于领域词典的同类词替换算法进行对比实验准确率提升**13.9%**，召回率提升**36.33%**，**F**值提升**27.05%**；用基于上下文特征保护的随机删除算法进行对比实验，准确率、召回率、**F**值分别提升**8.05%**，**5.18%**，**6.31%**。经试验证明，本研究提出的基于数据增广的渔业标准命名实体识别研究可以有效解决渔业标准语料中样本稀疏问题，对渔业标准命名实体识别具有较好效果。

论文简介

针对渔业标准文本语料库中“水产品名称”类命名实体存在样本分布稀疏现象、模型无法学习较多的实体特征而导致该类实体识别效果差的问题，在传统的同义词替换和随机删除算法的基础上，提出了基于领域词典的同类词替换和基于上下文特征保护的随机删除算法，通过数据增广的方式有效的解决了渔业标准命名实体识别样本分布稀疏问题。

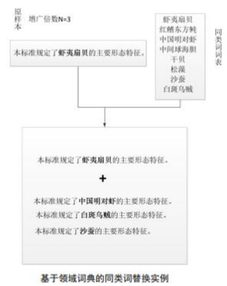
系统模型

基于数据增广的渔业标准命名实体识别采用BiLSTM+Attention+CRF模型，其中BiLSTM+Attention作为编码器，CRF作为模型的解码器，模型结构如下图所示。



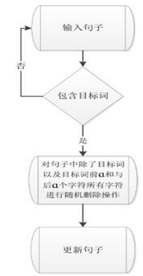
BiLSTM+Attention+CRF 模型结构

算法原理



基于领域词典的同类词替换实例

Example of replacement of similar words based on domain dictionary



基于上下文特征保护的随机删除算法

Random deletion of context feature protection

实验仿真

如图所示，基于领域词典的同类词替换，当增广系数为 **16**倍时，“水产品名称”的准确率、召回率、**F** 值分别提升 **13.9%**，**36.33%**，**27.05%**。随着增广系数的继续增大，模型的指标提升趋于平稳。

基于领域词典的同类词替换

Similar word replacement based on domain dictionary

增广参数	准确率%	召回率%	F 值%
N=1	73.77	51.72	60.81
N=2	79.79	65.22	71.77
N=4	82.99	75.25	78.93
N=8	85.98	84.97	85.47
N=16	**87.67**	**88.05**	**87.86**
N=32	86.62	88.65	87.62

如图所示，基于上下文特征保护的随机删除其中在随机删除概率为**0.02** 时，实验效果达到 最佳，准确率、召回率、**F** 值分别提升**8.05%5.18%**，**6.31%**。因此，使用上下文保护机制的随机删除可以提升样本多样性，有效的提升了模型的识别性能，增加了模型的泛化能力。

基于上下文特征保护的随机删除

Random deletion based on context feature protection

随机删除系数	准确率%	召回率%	F 值%
0	73.77	51.72	60.81
0.01	79.31	52.87	63.45
0.02	**81.82**	**56.9**	**67.12**
0.03	78.26	51.72	62.28
0.05	74.02	54.02	62.46
0.07	65.00	52.3	57.69
0.1	64.62	48.84	55.63
0.2	60.14	52.05	55.8

如图所示，使用相同的随机删除概率进行不同长 度的特征保护窗口对比试验。上下文特征保护窗 口长度为 **4** 字符单位的时候模型取得了最佳效果，因此基于上下文特征保护的随机删除算法窗口保护长度选择 **4** 字符单位可以使模型达到最 优效果。

上下文特征保护窗口长度

Context feature protection window length

窗口长度	准确率%	召回率%	F 值%
2	77.05	54.02	63.51
4	**81.82**	**56.9**	**67.12**
6	70.63	51.15	59.33
8	71.85	56.40	63.19

实验证明，本研究提出的基于数据增广的渔业标准命名实体识别研究在传统的同义词替换和随机删除算法的基础上，提出了基于领域词典的同类词替换和基于上下文特征保护的随机删除算法，通过数据增广的方式有效的解决了渔业标准命名实体识别样本分布稀疏问题。有对渔业标准命名实体识别具有较好效果。

论文结论

针对渔业标准文本语料库中“水产品名称”类命名实体存在样本分布稀疏现象、模型无法学习较多的实体特征而导致该类实体识别效果差的问题，在传统的同义词替换和随机删除算法的基础上，提出了基于领域词典的同类词替换和基于上下文特征保护的随机删除算法，通过数据增广的方式有效的解决了渔业标准命名实体识别样本分布系数问题。实验表明，该研究提出的数据增广算法可以有效丰富样本类型和样本数量，提升了模型的精度。使用深度学习模型针对其他类型的命名实体生成更接近真实文本的样本是未来研究工作的重点。

西安电子科技大学 XIDIAN UNIVERSITY