

n-gram统计特征在中文 分词领域自适应中的应用

郭振 张玉洁 苏晨 徐金安

北京交通大学
计算机与信息技术学院

发表内容

- 中文分词领域自适应问题
- n-gram统计特征
- 实例研究——中文分词科技文献领域的自适应
- 实验评测与结论

中文分词领域自适应问题

- 基于统计的分词方法对标注语料具有很强的依赖性。
- 现实中的文本分属不同的领域，具有不同的构词规律。
- 大规模、高质量的标注语料较难获得。
- 由于训练语料领域的限制，基于统计的分词方法领域适应能力较差。

从新闻领域到科技文献领域

分词系统性能严重下降

- CTB测试语料
- 348句
- OOV: 3.47%

训练语料	测试语料	综合性能 F_1
新闻领域	新闻领域	97.62%
	科技文献领域	83.89%



CTB训练语料

- CTB分词标准人工标注
- 300句
- OOV: 22.4%

相关研究

- 利用网络获取的大规模生语料生成特征，经CRF训练得到的分词模型在CTB测试语料上获得目前的最好结果(Yiou Wang et al, 2011)。
- 利用领域词典生成特征，经CRF训练得到的分词模型能有效提高模型的领域适应能力(张梅山 et al., 2012)。

n-gram统计特征

- n-gram: 文本中连续出现的n个字
- 出现频率越高的n元字串，越有可能是一个词。
- 某一领域文本中n元字串的出现频率反应了此领域文本中由字构词的规律。
- 很容易将大规模语料中n元字串的频率信息转化为CRF学习的特征。
- 特定领域的大规模生语料比标注语料、领域词典更易获取。

n-gram频度值和AV值

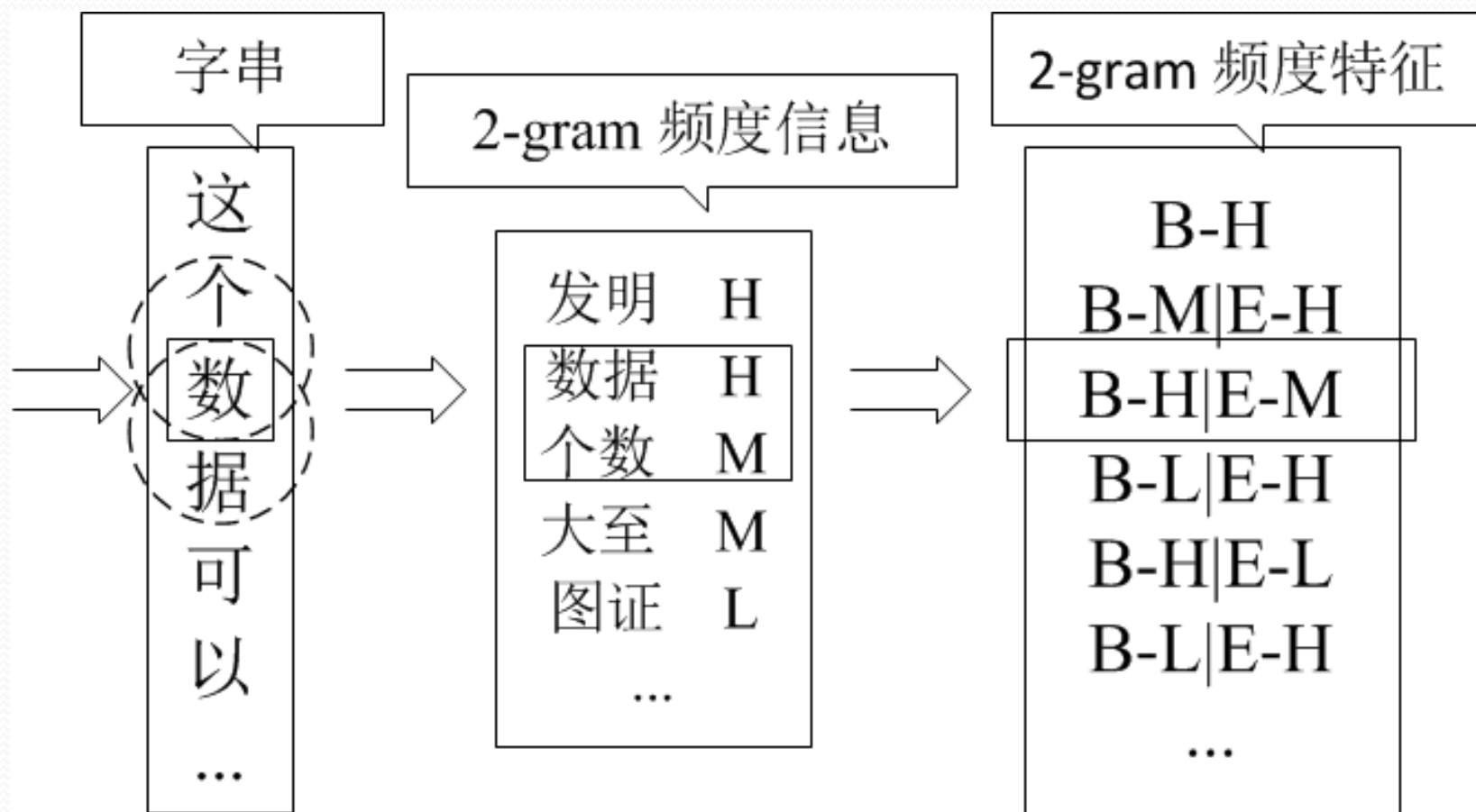
- n-gram频度值：字串在语料中出现的次数
- n-gram AV值：字串在语料中的上下文环境数

H: 高
M: 中
L: 低

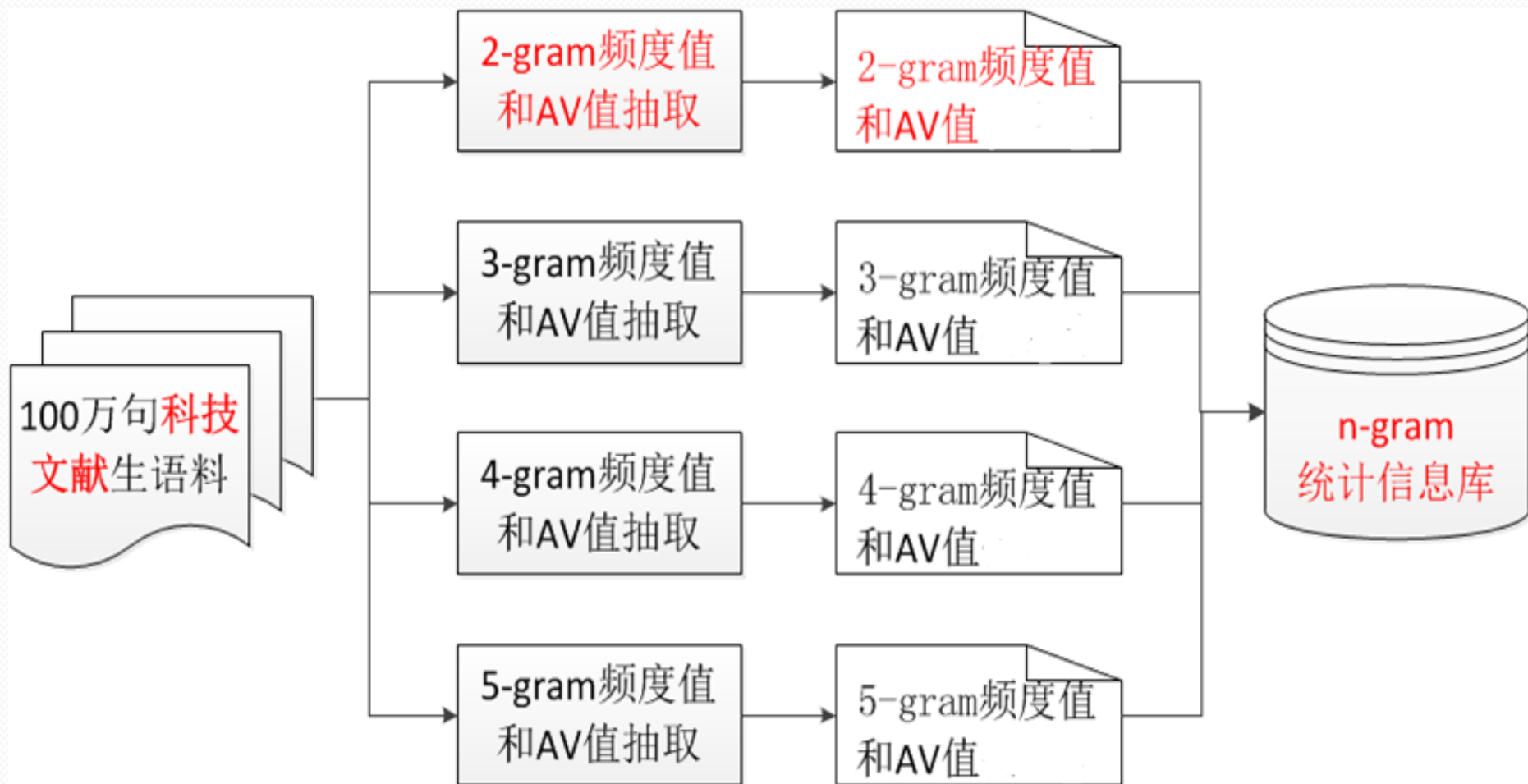
他将门把手弄坏了
小明把门把手修好了
这个门把手很漂亮
这个门把手坏了

字串	频度值	AV值
门把手	4	3

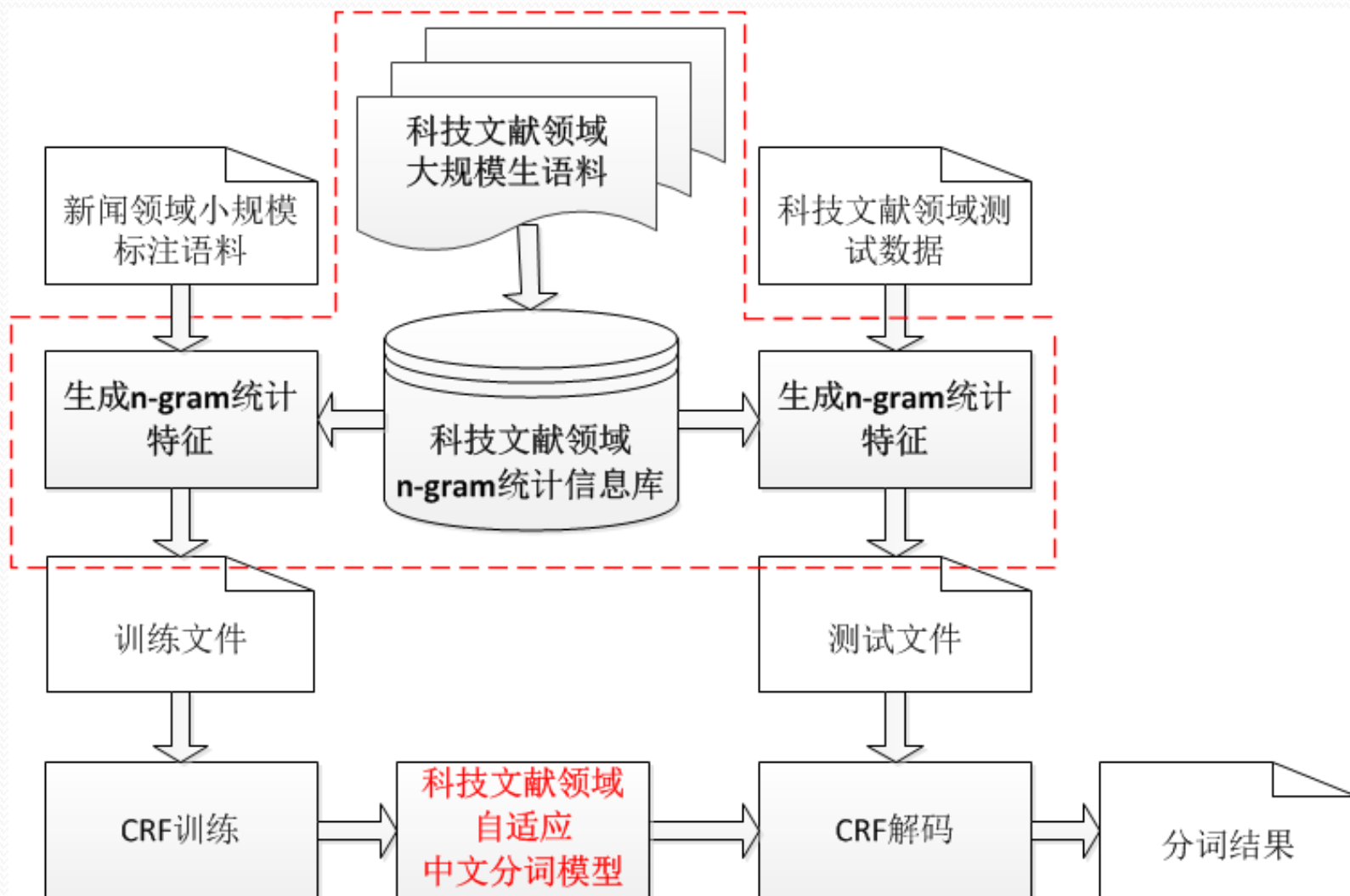
n-gram特征的生成



实例研究——中文分词科技文献领域的自适应



科技文献领域自适应的总体框架



实验评测

■ 实验数据

■ 训练数据：

- 标注语料：CTB训练集

- 生语料：NTCIR-9提供的科技文献领域的100万中英句对中的中文句子

- 测试数据：人工标注的300句，共10636单词的科技文献语料，未登录词占 22.4%。

评测指标

- **准确率P**: 系统输出结果中正确结果的比例
- **召回率R**: 标准答案中被系统正确识别的比例
- **综合性能 F_1** : P与R的调和值
- **未登录词召回率 R_{OOV}** : 标准答案中的未登录词被系统正确识别的比例
- **已知词召回率 R_{IV}** : 标准答案中的已知词被系统正确识别的比例

评测结果

F_1 ↑ 2.53%

分词系统	召回率 R	准确率 P	综合性 能 F_1	未登录词召 回率 R_{OOV}	已知词召 回率 R_{IV}
未加n-gram特 征	86.05%	81.83%	83.89%	63.70%	92.51%
+(a)n-gram频度 值特征	87.95%	84.32%	86.09%	69.15%	93.37%
+(b)n-gram AV 值特征	88.29%	84.28%	86.24%	69.15%	93.82%
+(a)+(b)	88.31%	84.61%	86.42%	69.91%	93.63%

实验结果分析

- n-gram统计特征，包括频度值特征和AV值特征，使分词系统在新领域上的召回率和准确率都有提高，两种特征结合使用效果更好，增强了系统的领域适应能力。
- 从综合指标 F_1 值看，AV值特征比频度值特征的提高幅度更大；召回率方面 AV值特征比频度值特征更有效；准确率方面频度值特征比 AV值特征更有效。
- n-gram统计特征使分词系统对已知词的召回率提高了1.12%，对未登录词的召回率提高了6.21%。

实验结果的具体实例分析

- 实例一：“它/可用于/驱动/电光层/到/显示/数据...”
 - 未加入n-gram统计特征：“.../电光/层到/...”
 - 加入n-gram统计特征后得到正确结果
- 实例二：“而/不/使用/压电/元件/或/楔形/换能器”
 - 未加入n-gram统计特征：“.../换/能器”
 - 加入n-gram统计特征后得到正确结果

结论

- **n-gram**统计特征可以从大规模生语料中获取其领域单词的构词规律，特别是可有效提高未登录词的召回率和准确率，从而提高系统的分词精度。
- 相比利用标注语料和领域词典的方法，本研究提出的中文分词领域自适应方法有效利用了大规模生语料，更加简洁、行之有效。

谢谢！