关键词抽取的研究与实现

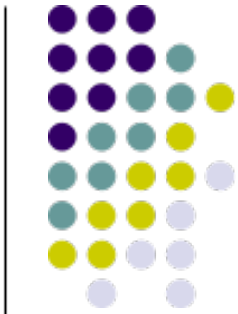
学生:罗准辰
导师:王挺教授
2008-12-06

国防科技大学计算机学院

硕士答辩

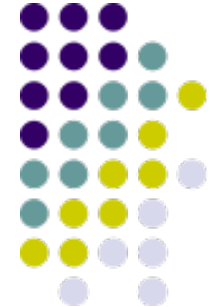
内容提要

- 研究背景和动机
- 基于分离模型的关键词提取算法
- 分离模型的特征设计
- 实验结果与分析
- 结论



一、研究动机

关键词提取问题



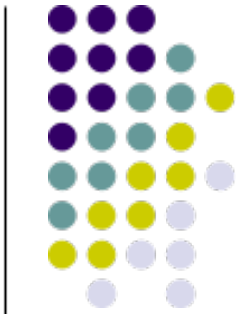
- 关键词提取：文档中候选关键词做出关键词或非关键词的分类
- 科学问题：一种二值分类问题，需要解决的难点
  - **“关键”**
    - 如何度量候选关键词的“关键”？
  - **“词”**
    - 一定是“词”吗？

课题思路

- 通常所说的关键词实际上有相当一部分是关键的短语，而这部分关键词的提取是十分困难的问题。我们提出将关键词提取分为两个问题进行处理：关键单词提取和关键词串提取，设计了一种基于分离模型的关键词提取算法。并以该算法为基础，针对关键单词提取和关键词串提取这两个问题设计了不同的特征，提高了提取的准确性。

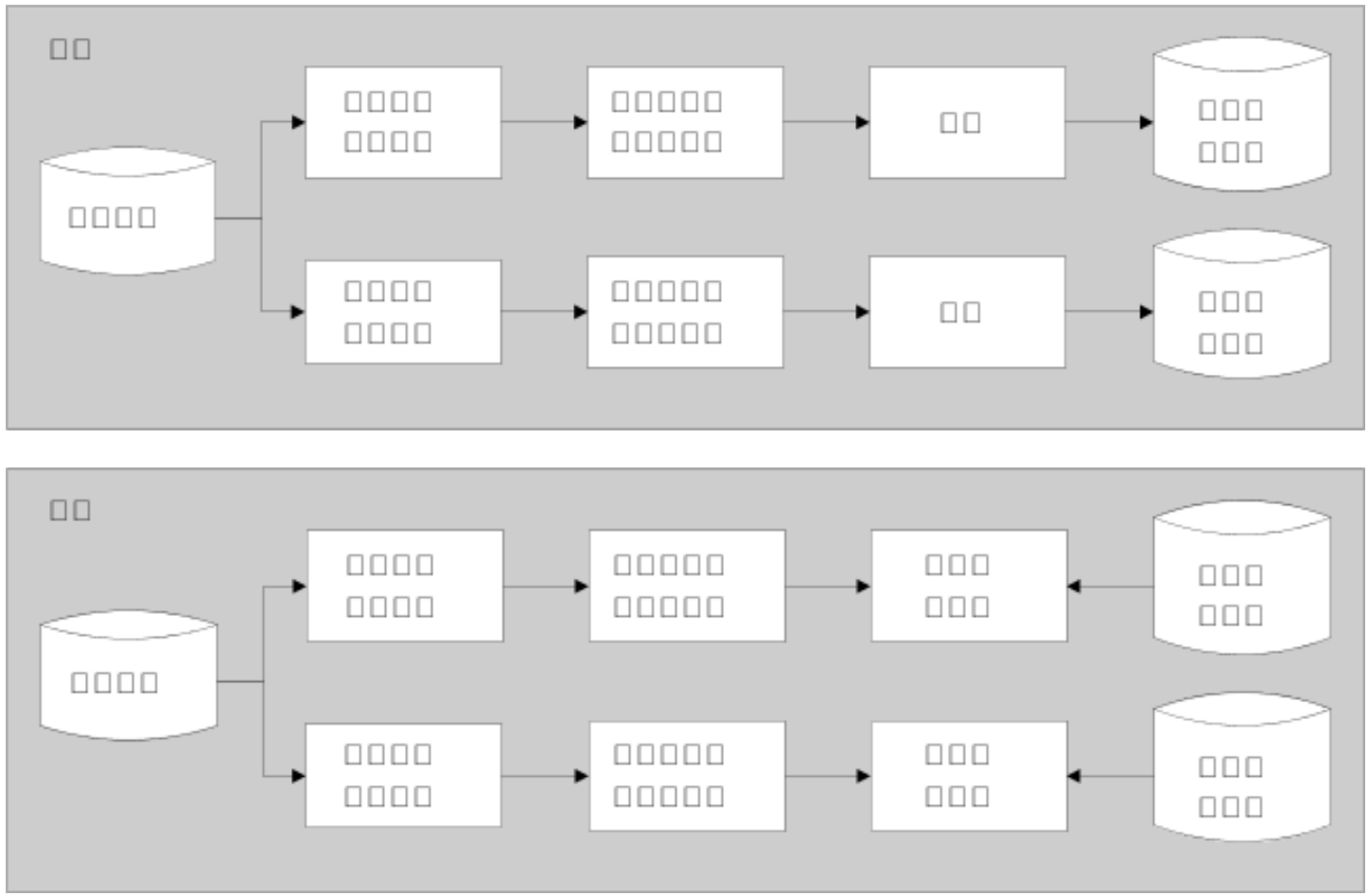
关键词分类

- 关键单词
  - **仅包含一个词的关键词**
- 关键词串
  - **含多个词的关键词**



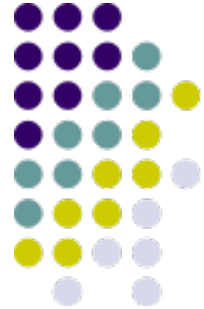
二、基于分离模型的关键词提取算法

分离模型



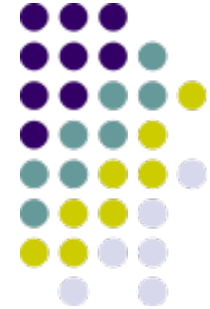
分离模型关键词提取算法

- 生成候选关键单词与候选关键词串
- 模型的训练与学习器的选择
- 提取关键词



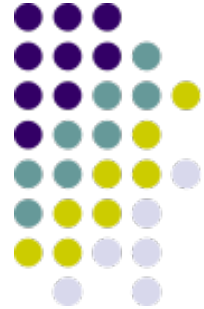
三、分离模型的特征设计

关键单词与词串公共特征设计



- TF×IDF特征
- 首次出现位置特征POS
- TF×IF特征
- 文档长度特征NWT

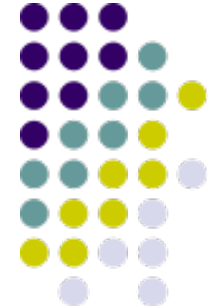
关键单词特征设计



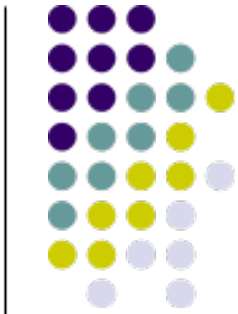
- 词性特征CKWPS

—候选关键单词是否名词

关键词串特征设计

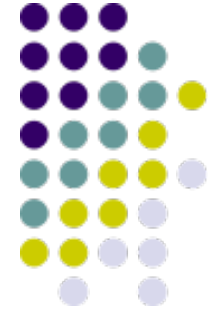


- 互信息特征
- 词串边界参数表特征
- 候选关键词串结尾词词性特征
- 候选关键词串开头词词性特征
- 候选关键词串非结尾词非形容词非名词数目
- 候选关键词串所含词数



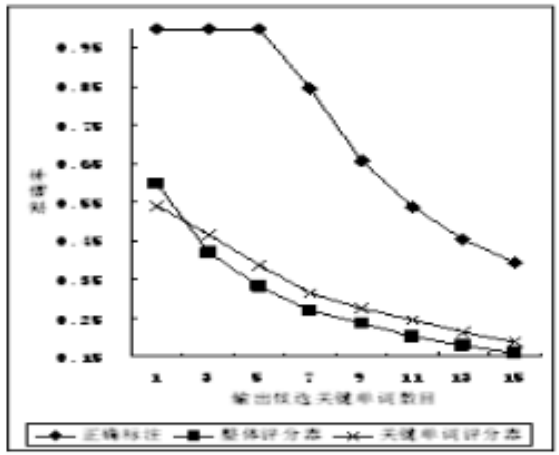
四、实验结果与分析

实验与分析

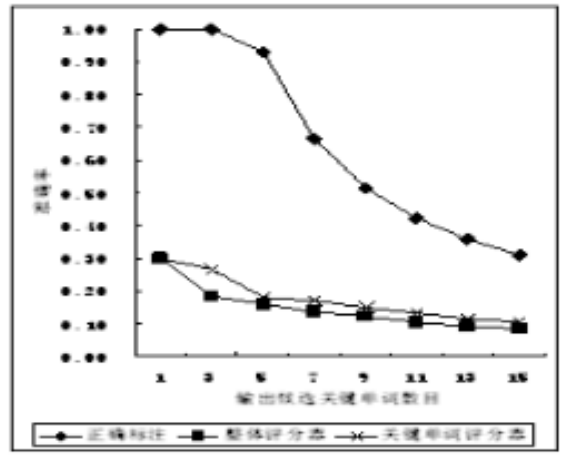


- **语料**
  - Journals（英）
  - Aliweb（英）
  - CSTR(英)
  - Blog(中)

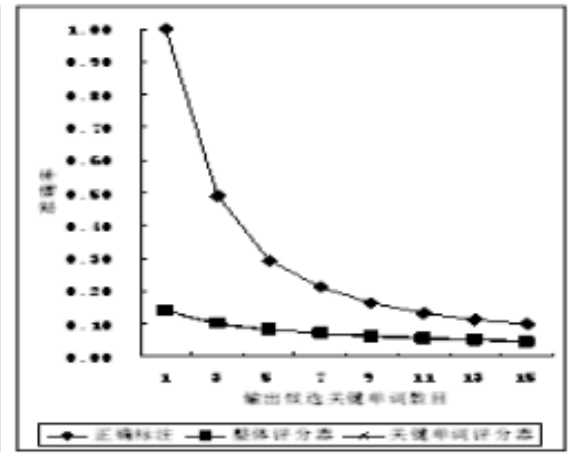
分离模型与整体模型的比较(英)



(a)· Aliweb 候选关键单词

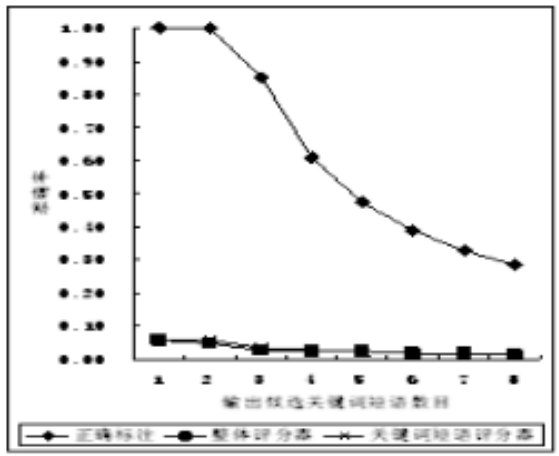


(b)· Journals 候选关键单词

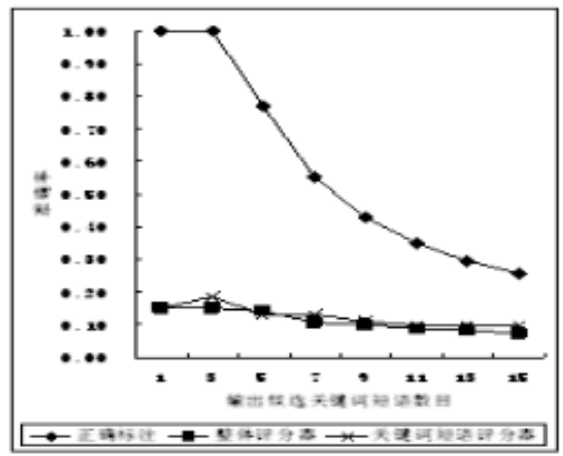


(c)· CSTR 候选关键单词

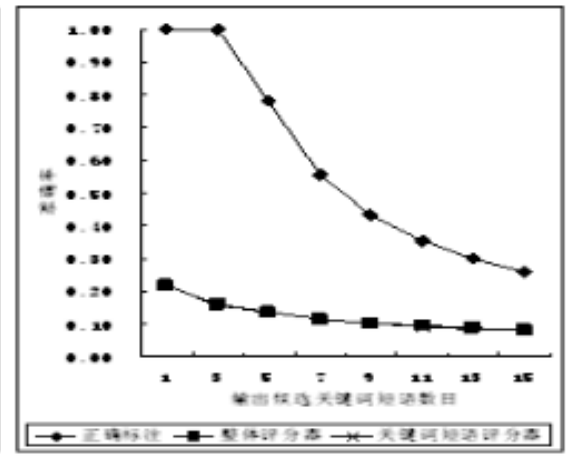
关键单词评分器与整体评分器比较



(a)· Aliweb 候选关键词短语



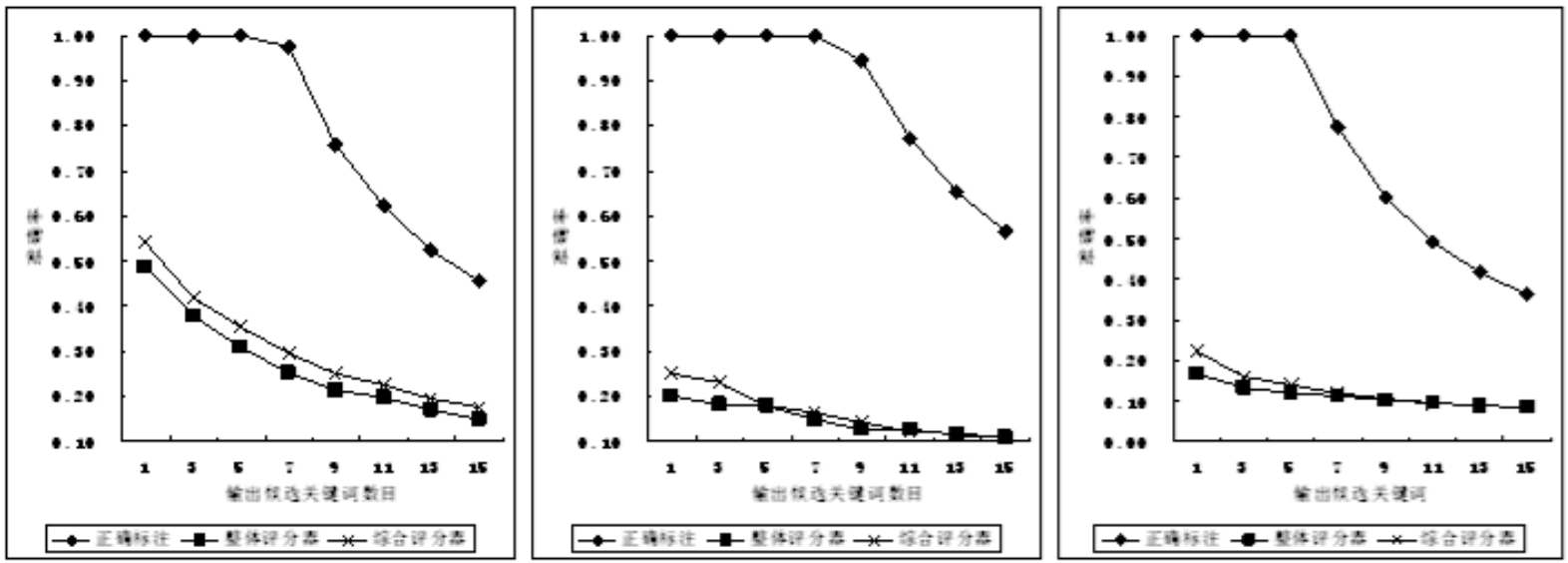
(b)· Journals 候选关键词短语



(c)· CSTR 候选关键词短语

关键词短语评分器与整体评分器比较

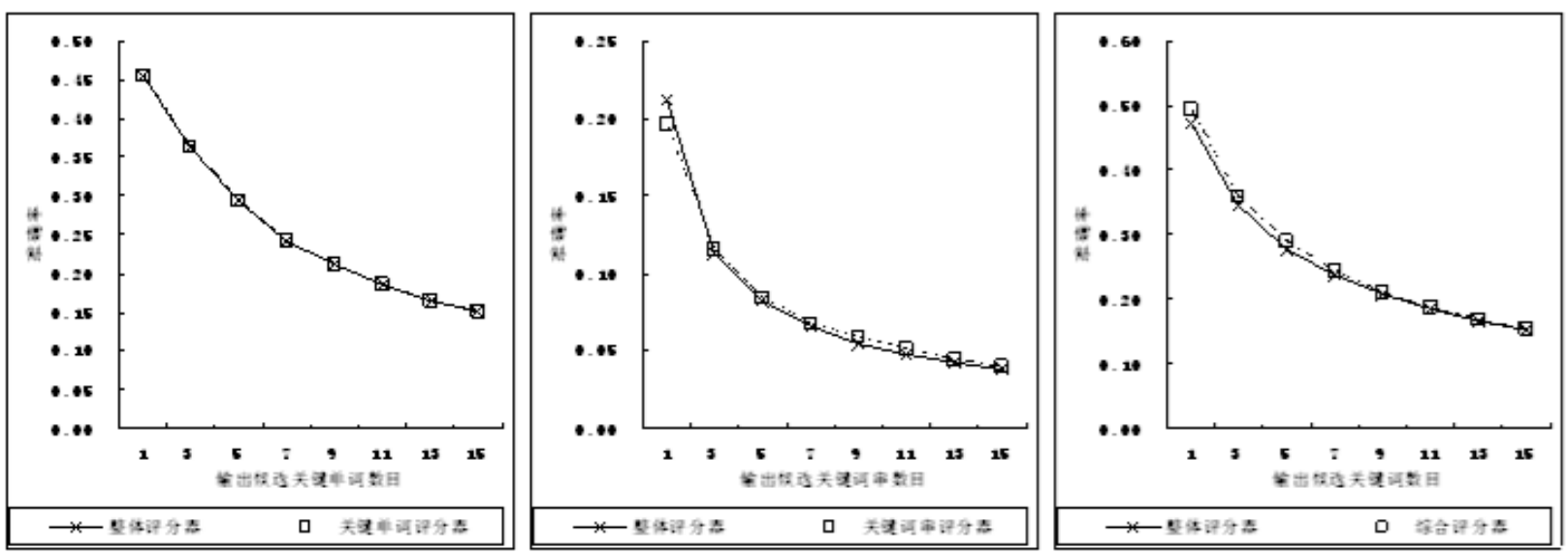
分离模型与整体模型的比较(英)



(a)·Aliweb 候选关键词　(b)·Journals 候选关键词　(c)·CSTR 候选关键词

综合评分器与整体评分器比较

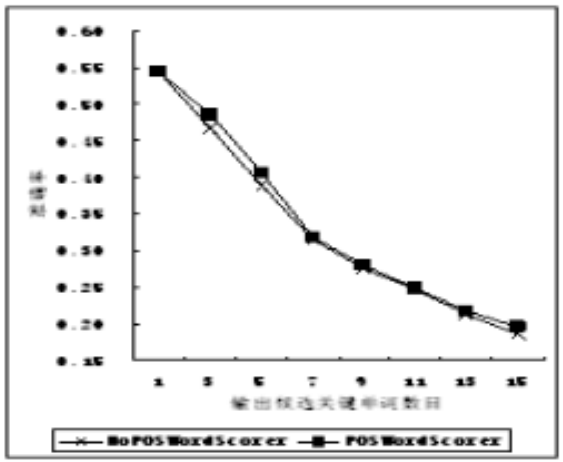
分离模型与整体模型的比较(中)



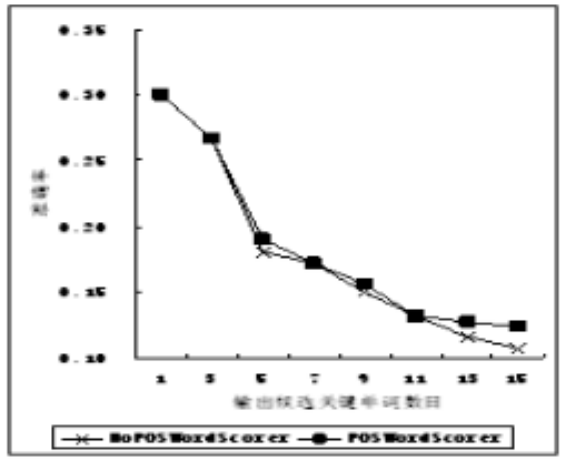
(a)· Blog 候选关键单词　　(b)· Blog 候选关键词串　　(c)· Blog 候选关键词

Blog 语料整体模型与分类模型比较

特征设计的意义(英)

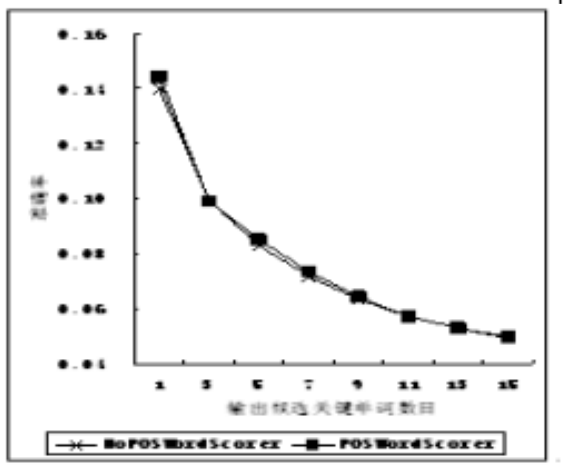


(a)· Aliweb 候选关键单词

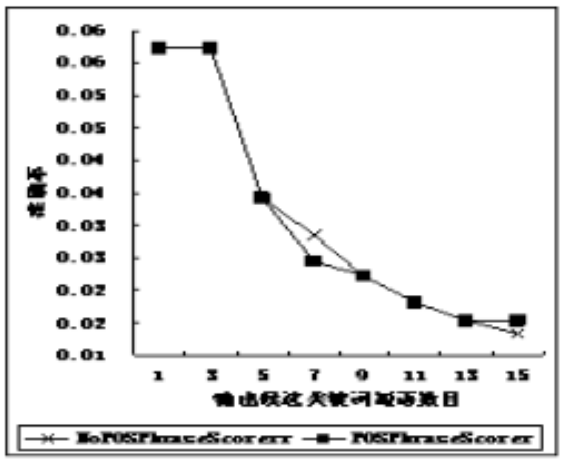


(b)· Journals 候选关键单词

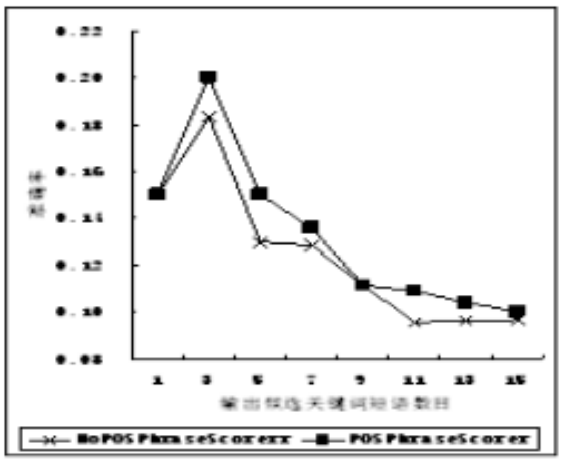
添加关键单词特征实验



(c)· CSTR 候选关键单词

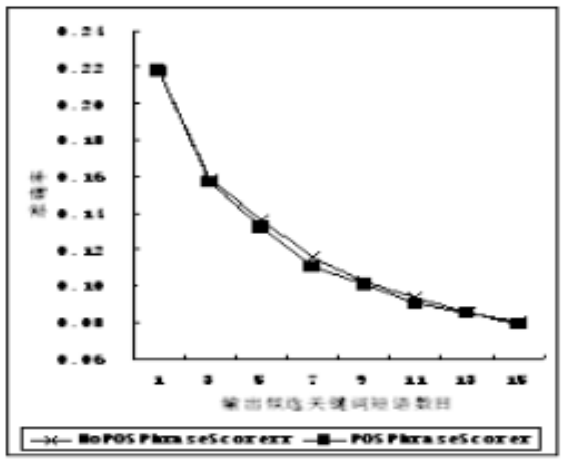


(a)· Aliweb 候选关键词短语



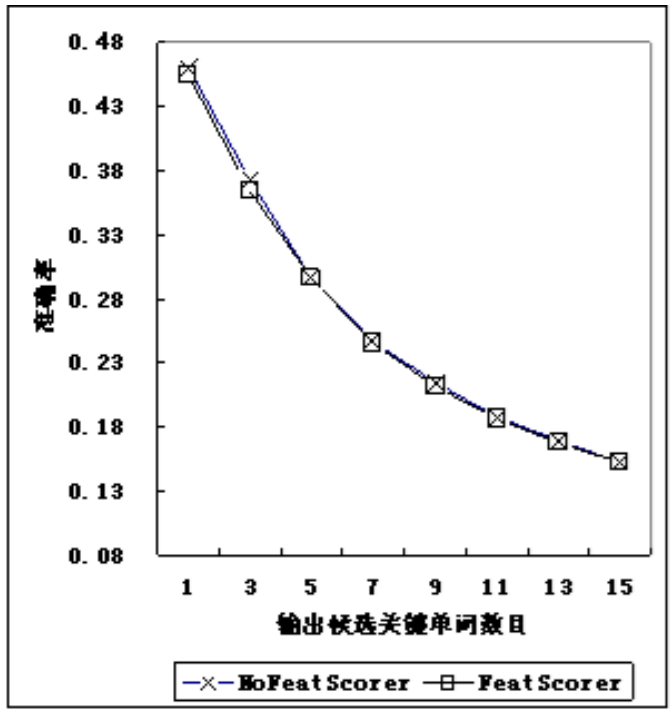
(b)· Journals 候选关键词短语

添加关键词短语特征实验

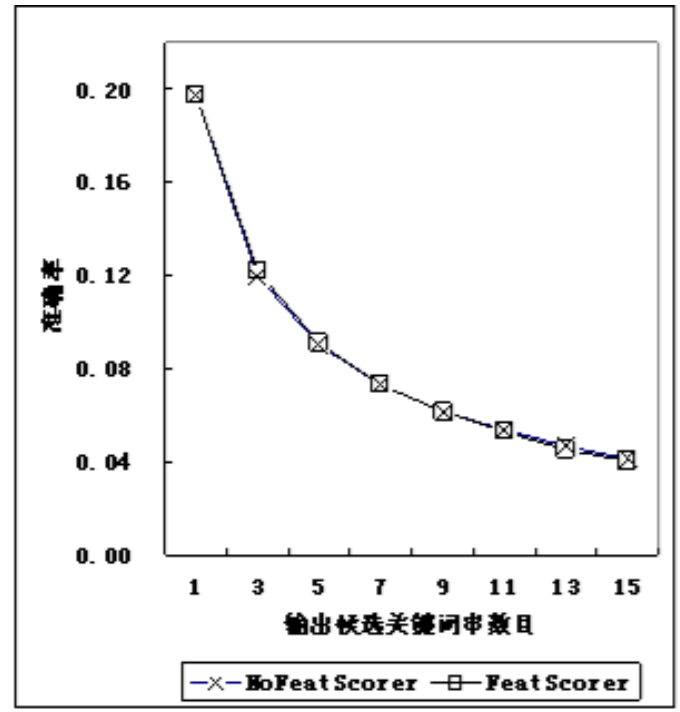


(c)· CSTR 候选关键词短语

特征设计的意义(中)



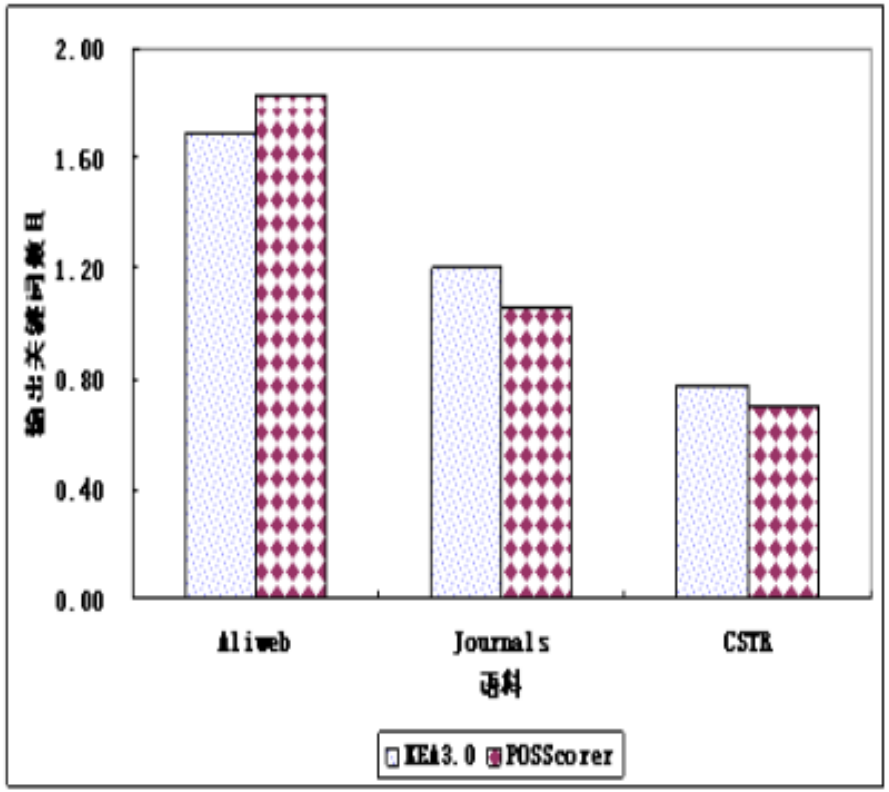
添加关键单词特征



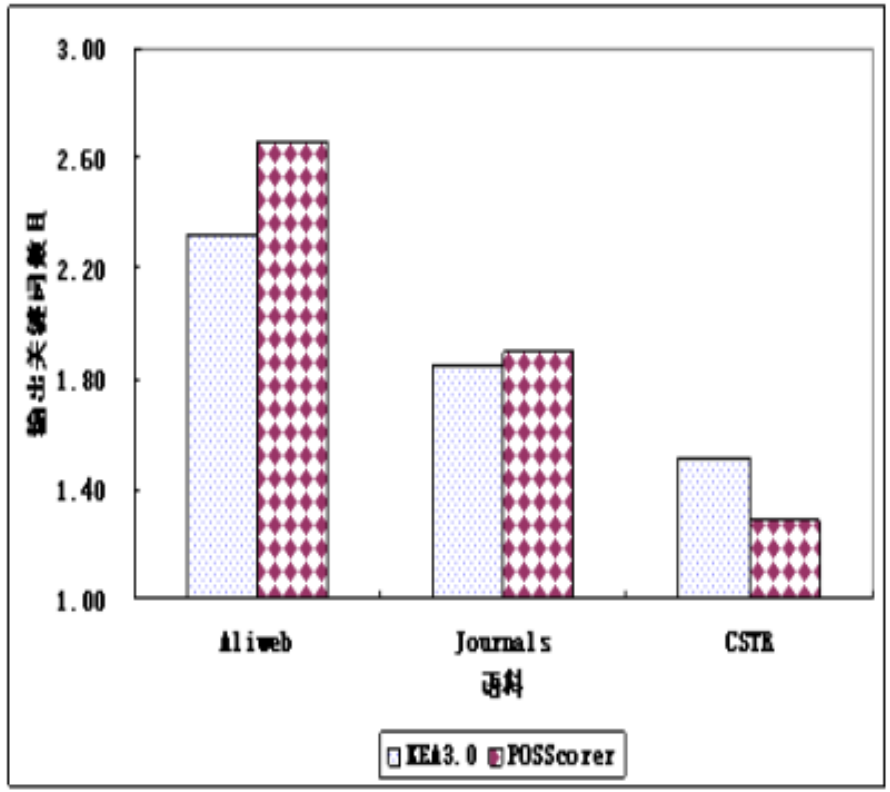
添加关键词串特征

Blog 特征实验

与KEA的比较(英)

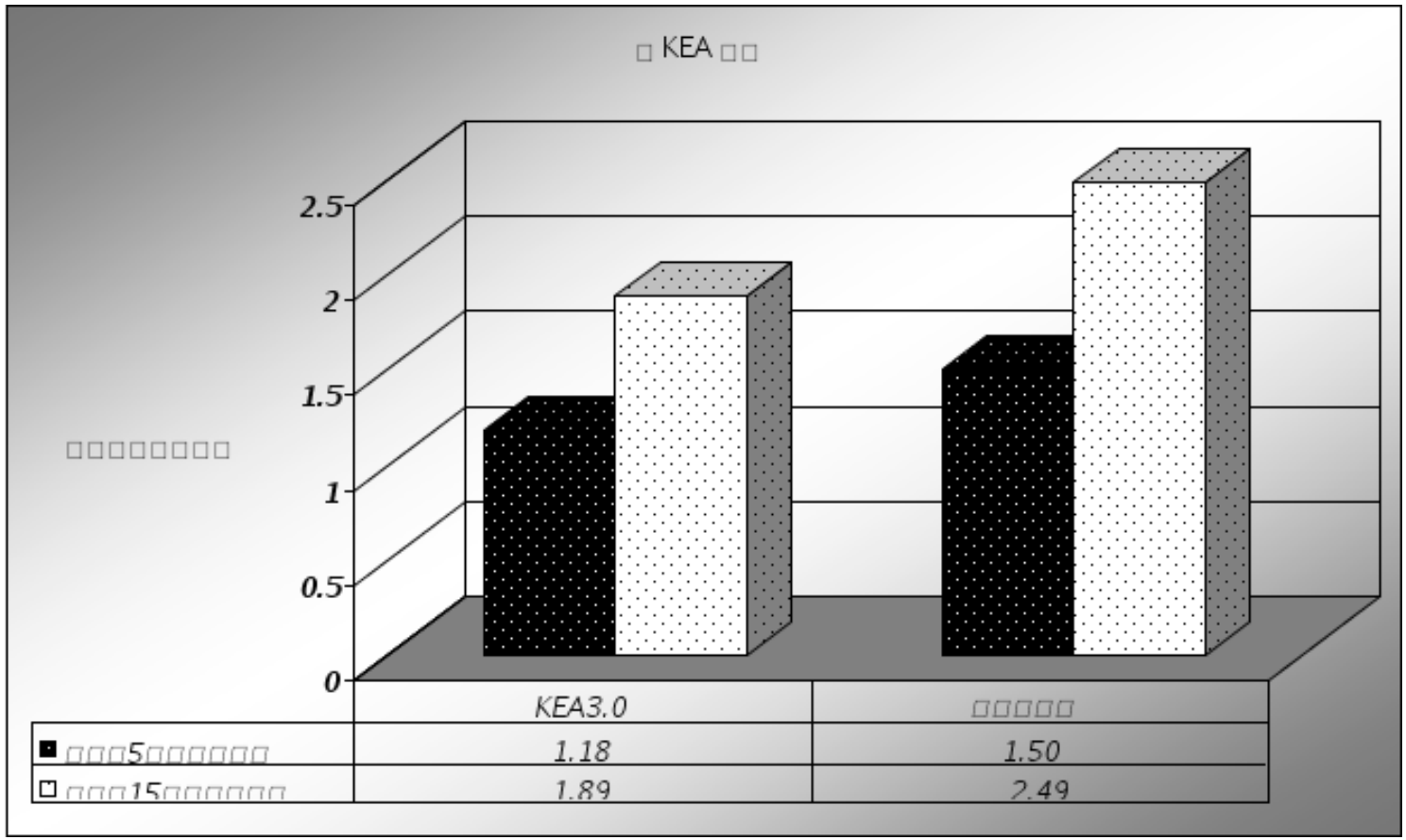


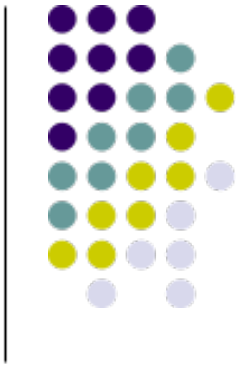
输出前5个候选关键词比较



输出前15个候选关键词比较

与KEA的比较(中)





五、结论

结束语

- 结论
  - 分离模型算法优于整体模型算法
  - 设计的特征在关键词提取中是有意义的
  - 英文中与KEA关键词提取效果相当，中文中优于KEA

下一步研究工作

- 设计更多的特征
- 关键单词与关键词串合并问题

谢谢

150544941
13548661488
zhunchenluo@nudt.edu.com

国防科技大学计算机学院NLP组

硕士答辩