电子与信息学报 2009 31 (12): 2969-2974 ISSN: 1009-5896 CN: 11-4494/TN

论文

# 基于Web的无指导译文消歧词模型与*N*-gram模型及对比研究

刘鹏远[1], 赵铁军[2]

[1]北京大学计算语言学研究所 北京 100871; [2]哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001

收稿日期 2008-12-5 修回日期 2009-5-7 网络版发布日期 2009-12-4 接受日期

摘要
该文提出了基于Web的无指导译文消歧的词模型及*N*-gram模型方法，并在尽可能相同的条件下进行了比较。两种方法均利用搜索引擎统计不同搜索片段在Web上的Page Count作为主要消歧信息。词模型定义了汉语词汇与英语词汇之间的双语词汇Web相关度，根据汉语上下文词汇与英语译文之间的相关度进行消歧；*N*-gram模型首先假设不同语义下的多义词*N*-gram序列行为模式不同，从而可对多义词不同语义类下词汇在实例中的N-gram序列进行统计与分析以进行消歧。两个模型的性能均超过了在国际语义评测SemEval2007的task#5上可比较的最好无指导系统。对这两个模型进行试验对比可发现*N*-gram模型性能优于词模型，也表明组合两类模型的结果有进一步提升消歧性能的潜力。

关键词 计算语言学 无指导译文消歧 词模型 *N*-gram模型 Page Count 双语词汇Web相关度

分类号 TP391

# Comparison of Web-Based Unsupervised Translation Disambiguation Word Model and *N*-gram Model

Liu Peng-yuan[1], Zhao Tie-jun[2]

[1]Institute of Computational Linguistics, Peking University. Beijing 100871, China;
[2]Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract
This paper describes and compares web-based unsupervised translation disambiguation word model and *N*-gram model. For acquiring knowledge of disambiguation, both two models put differents queries to search engine and statistic page counts which it returned. Word model defines Web Bilingual Relatedness(WBR) between Chinese words and English words and disambiguates word sense by maxmizing Web Bilingual Relatedness between contexts and the translations of target word. Based on the hypothesis that the pattern of a polysemant is different while different sense of it is being used, *N*-gram model makes disambiguation by statisticing and analyzing *N*-grams of words in different semantic class of that polysemant. Both of the two models are evaluated on the SemEval2007 task#5, achieving the top performance against the state-of-the-art comparable unsupervised systems. Furthmore, *N*-gram model outperforms word model and the performence has potential for promotion when combine the results of that two class model.

Key words Computational linguistics Unsupervised translation disambiguation Word model *N*-gram model Page Count Web Bilingual Relatedness(WBR)

DOI:

通讯作者

作者个人主页 刘鹏远[1]; 赵铁军[2]

扩展功能
本文信息
- Supporting info
- PDF(223KB)
- [HTML全文](0KB)
- 参考文献[PDF]
- 参考文献

服务与反馈
- 把本文推荐给朋友
- 加入我的书架
- 加入引用管理器
- 复制索引
- Email Alert
- 文章反馈
- 浏览反馈信息

相关信息
- 本刊中 包含“计算语言学”的 相关文章
- 本文作者相关文章

· 刘鹏远
· 赵铁军