电子与信息学报 2008 30 (12): 2926-2929 ISSN: 1009-5896 CN: 11-4494/TN

论文

# 基于信息增益改进贝叶斯模型的汉语词义消歧

范冬梅① 卢志茂① 张汝波① 潘树燊②

①(哈尔滨工程大学计算机学院 哈尔滨 150001)
②(哈尔滨工业大学 哈尔滨 150001)

收稿日期 2007-6-4 修回日期 2008-9-23 网络版发布日期 接受日期

摘要
词义消歧一直是自然语言处理领域的关键问题和难点之一。通常把词义消歧作为模式分类问题进行研究，其中特征选择是一个重要的环节。该文根据贝叶斯假设提出基于信息增益的特征选择方法，并以此改进贝叶斯模型。通过信息增益计算，挖掘上下文中词语的位置信息，提高贝叶斯模型知识获取的效率，从而改善词义分类效果。该文在8个歧义词上进行了实验，结果发现改进后的贝叶斯模型在消歧正确率上比改进前平均提高了3.5个百分点，改进幅度较大，效果突出，证明了该方法的有效性。

关键词 词义消歧；自然语言处理；信息增益；贝叶斯模型

分类号 TP391



# Chinese Word Sense Disambiguation Based on Bayesian Model Improved by Information Gain

Fan Dong-mei① Lu Zhi-mao① Zhang Ru-bo① Pan Shu-shen②

①(Harbin Engineering University, Harbin 150001, China)
②(Harbin Institute of Technology, Harbin 150001, China)

Abstract
Word Sense Disambiguation (WSD) is one of the key issues and difficulties in natural language processing. WSD is usually considered as an issue about pattern classification to study, which feature selection, is an important component. In this paper, according to Naïve Bayesian Model (NBM) assumption, a feature selection method based on information gain is proposed to improve NBM. Location information concealed in the context of ambiguous word is mined through information gain, to improve the knowledge acquisition efficiency of Bayesian model, thereby improving the word-sense classification. The eight ambiguous words are tested in the experiment. The experimental results show that improved Bayesian model is more correct than the NBM an average of 3.5 percentage points. The accuracy rise is bigger and the improvement effect is outstanding. These results prove also the method put forward in this paper is efficacious.

Key words Word sense disambiguation Natural language processing Information gain Naï ve Bayesian model

DOI:

通讯作者 范冬梅

作者个人主页 范冬梅① 卢志茂① 张汝波① 潘树燊②