P.O.Box 8718, Beijing 100080, China	Journal of Software, Mar 2006,17(3):356-363
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2006 by *Journal of Software*

面向信息检索的自适应中文分词系统

曹勇刚, 曹羽中, 金茂忠, 刘 超

曹勇刚, 曹羽中, 金茂忠, 刘 超
(北京航空航天大学 计算机学院,北京 100083)
作者简介:
曹勇刚(1977－),男,湖南长沙人,博士生,主要研究领域为知识/内容管理,文本挖掘,软件工程.曹羽中(1978－),男,硕士生,主要研究领域为文本挖掘,软件工程.金茂忠(1941－),男,教授,博士生导师,主要研究领域为编译技术,软件工程.刘超(1958－),男,教授,博士生导师,CCF高级会员,主要研究领域为软件工程.
联系人: 曹勇刚 Phn: +86-10-82324488 ext 885, E-mail: ygcao@cse.buaa.edu.cn, http://sei.buaa.edu.cn

Received 2005-08-02; Accepted 2005-10-11

Abstract
New words recognition and ambiguity resolving have vital effect on information retrieval precision. This paper presents a statistical model based algorithm for adaptive Chinese word segmentation. Then, a new word segmentation system called BUAASEISEG is designed and implemented using this algorithm. BUAASEISEG can recognize new words in various domains and do disambiguation and segment words with arbitrary length. It uses an iterative bigram method to do word segmentation. Through online statistical analysis on target article and using the offline words frequencies dictionary or the inverted index of the search engine, the candidate words selection and disambiguation are done. On the basis of the statistical methods, post-process using stopwords list, quantity suffix words list and surname list are used for further precision improvement. The comparative evaluation with the famous Chinese word segmentation system ICTCLAS, using news and papers as testing text, shows that BUAASEISEG outperforms ICTCLAS in new words recognition and disambiguation.

Cao YG, Cao YZ, Jin MZ, Liu C. Information retrieval oriented adaptive Chinese word segmentation system. *Journal of Software*, 2006,17(3):356-363.
DOI: 10.1360/jos170356
http://www.jos.org.cn/1000-9825/17/356.htm

摘要
新词的识别和歧义的消解是影响信息检索系统准确度的重要因素.提出了一种基于统计模型的、面向信息检索的自适应中文分词算法.基于此算法,设计和实现了一个全新的分词系统BUAASEISEG.它能够识别任意领域的各类新词,也能进行歧义消解和切分任意合理长度的词.它采用迭代式二元切分方法,对目标文档进行在线词频统计,使用离线词频词典或搜索引擎的倒排索引,筛选候选词并进行歧义消解.在统计模型的基础上,采用姓氏列表、量词表以及停词列表进行后处理,进一步提高了准确度.通过与著名的ICTCLAS分词系统针对新闻和论文进行对比评测,表明

BUAASEISEG在新词识别和歧义消解方面有明显的优势.

基金项目：Supported by the National High-Tech Research and Development Plan of China under Grant No.2004AA112030 (国家高技术研究发展计划(863))

References:

[1] Foo S, Li H. Chinese word segmentation accuracy and its effects on information retrieval. Information Processing and Management, 2004,40(1):161-190.

[2] Zhang HP, Yu HK, Xiong DY, Liu Q. HHMM-Based Chinese lexical analyzer ICTCLAS. In: Proc. of the 2nd SigHan Workshop. 2003. 184-187.

[3] Su KY, Chaing TH, Chang JS. An overview of corpus-based statistics-oriented (CBSO) techniques for natual language processing. Computational Linguistics and Chinese Language Processing, 1996,1(1):101-157.

[4] Zhang HP, Liu Q. Model of Chinese words rough segmentation based on N-shortest-paths method. Journal of Chinese Information Processing, 2002,16(5):1-7 (in Chinese with English abstract).

[5] Liang NY. CDWS: A word segmentation system for written Chinese texts. Journal of Chinese Information Processing, 1987,1(2): 101-106 (in Chinese with English abstract).

[6] Zhu XF, Wang H. Classification of modern Chinese quantity suffix and noun. Technical Report, 1994 (in Chinese with English abstract). http://www.icl.pku.edu.cn/icl_tr/collected_papers/chinese/collection-2/yyy23.htm

[7] Gao JF, Li M, Huang CN. Improved source-channel models for Chinese word segmentation. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. 7-12.

[8] Giles JT, Wo L, Berry MW. GTP (general text parser) Software for text mining in statistical data mining and knowledge discovery. In: Bozdogan H, ed. Boca Raton: CRC Press, 2003. 455-471.

[9] Chang JS, Lin YC, Su KY. Automatic construction of a Chinese electronic dictionary. In: Yarowsky D, Church K, eds. Proc. of the 3rd Workshop on Very Large Corpora. 1995. 107-120.

[10] Dai YB, Khoo SGT, Loh TE. A new statistical formula for Chinese word segmentation incorporating contextual information. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 1999. 82-89.

[11] Gao JF, Wu AD, Li M, Huang CN, Li HQ, Xia XS, Qin HW. Adaptive Chinese word segmentation. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. 2004. 21-26.

附中文参考文献:

[4] 张华平,刘群.基于N-最短路径方法的中文词语粗分模型.中文信息学报,2002,16(5):1-7.

[5] 梁南元.书面汉语自动分词系统——CDWS.中文信息学报,1987,1(2):101-106.

[6] 朱学锋,王惠.现代汉语量词与名词的子类划分.技术报告,1994. http://www.icl.pku.edu.cn/icl_tr/collected_papers/chinese/collection-2/yyy23.htm