分类号_____ 密级_____
UDC_____ 编号_____

中国科学院研究生院

博士学位论文

汉语问答系统关键技术研究

吴 友 政

指导教师 徐波 研究员 中国科学院自动化研究所
赵军 副研究员 中国科学院自动化研究所

申请学位级别 工学博士 学科专业名称 模式识别与智能系统

论文提交日期_____ 论文答辩日期 _____

培养单位 中国科学院自动化研究所

学位授予单位 中国科学院研究生院

答辩委员会主席_____

Research on the Key Technologies for Chinese Question Answering

Dissertation Submitted to
Institute of Automation, Chinese Academy of Sciences
in partial fulfillment of the requirements
for the degree of
Doctor of Engineering

By
Youzheng Wu
(Pattern Recognition and Intelligent System**)**

Dissertation Supervisor: Professor Bo Xu & Jun Zhao

独创性声明

本人声明所成交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确地说明并表示了谢意。

签名：_____导师签名：_____ 日　期：_____

关于论文使用授权的说明

本人完全了解中国科学院自动化研究所有关保留、使用学位论文的规定，即：中国科学院自动化研究所有权保留送交论文的复印件，允许论文被查阅和借阅；可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：_____导师签名：_____ 日　期：_____

本博士论文的研究工作得到以下项目资助

[1] 国家自然科学基金（项目编号60372016）

[2] 北京市自然科学基金（项目编号4052027）

[3] 富士通研究开发中心有限公司项目

摘 要

互联网的迅猛发展和广泛普及，使人们可以方便地从网络上获得信息。但是网络信息的爆炸性增长，又把人们淹没在信息的海洋里，准确、快速地获得有价值信息的难度大大地增加了。问答系统的出现旨在提供更有力的信息获取工具，以应对信息爆炸带来的严重挑战。

相对于英文问答技术研究的迅速发展以及实用英文问答系统的推出，从事中文问答技术研究的科研机构还不多，而且基本没有成型的中文问答系统问世。本文就是在这样的情况下针对汉语问答技术展开深入研究，主要工作包括：

[1] 建立了一个具有一定规模并可扩充的汉语问答技术评测平台

论文在吸收英文、日文和多语言问答系统评测的成功经验基础上，研发了面向汉语问答系统的评测平台。平台的语料规模约为 1.8GB；测试集现包括 7050 个汉语提问句；打分标准主要是借鉴 TREC 的评分标准。

[2] 提出了汉语问答系统的提问分类体系及基于多特征的提问分类算法

论文从新的角度提出了一种提问分类体系，即提问的技术分类和提问的语义分类，并在此基础上实现了基于多特征的支持向量机提问分类算法。与英文层级分类体系相比，论文提出的汉语平行分类体系的特点是，既能为提问选择最合适的技术方案，也能确定提问答案的语义类型。实验数据表明，论文实现的分类算法能够获得较高性能的分类效果。

[3] 设计了基于多特征的汉语命名实体识别算法

论文提出的基于多特征的汉语命名实体识别算法具有以下特点：① 强调大颗粒度特征(词性特征)和小颗粒度特征(词形特征)的结合；② 强调统计模型和专家知识的结合；③ 为准确刻画不同实体的内部特征，设计了多个细分类的实体模型以识别不同国家的人名、单字地名与多字地名、简称机构名和全称机构名。在不同测试语料上的实验结果表明：基于多特征的汉语命名实体识别模型要优于使用单一特征的命名实体识别模型。

[4] 提出了基于主题语言模型的汉语问答系统句子检索算法

论文提出了基于主题语言模型的汉语问答系统句子检索算法，该算法利用问答系统中特有的提问分类信息(即提问的答案语义信息)对句子初检结果进行主题聚类，通过 Aspect Model 将句子所属的主题信息引入到语言模型中，从而获得对句子语言模型更精确的描述。对初检结果的主题聚类，本文提出“一个句子

多个主题”和“一个句子一个主题”两种聚类算法。实验结果表明论文提出的方法可以有效地改善汉语问答系统的句子检索性能。

[5] 提出了基于无监督学习的问答模式抽取技术

论文研究利用模式匹配技术处理由于自然语言的灵活性和多变性给问答技术带来的挑战，并提出了一种基于无监督学习算法的问答模式抽取技术，从互联网上抽取应用于汉语问答系统的答案模式。该算法可以避免有监督学习算法的不足，它无需用户提供<提问，答案>对作为训练集，只需用户提供每种提问类型两个或以上的提问实例，算法即可通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类等步骤完成该类型提问的答案模式的学习。实验结果表明，论文提出的无监督问答模式学习方法是有效的，基于模式匹配的答案抽取技术能够较大幅度地提高汉语问答系统的性能。

关键词：汉语问答系统，问答评测，命名实体识别，句子检索，问答模式抽取

Abstract

With the rapid development and broad popularization of Internet, it's very convenient for people to get information from web. Meanwhile, the volatile increase of web information brings users into the ocean of information and it becomes very difficult to obtain the exact and correct information quickly. Question Answering (QA) aims at providing the more powerful information access tools to help users to overcome the problems of information overloading.

Compared with the development of the technologies of English QA and the application of English QA System, researches on Chinese QA are still at its early stage. This thesis focuses on the research on the key technologies of Chinese Question Answering. The main contributions and novelties are summarized as follows.

1. Build an Evaluation Platform for Chinese Question Answering

Based on the experiences from TREC, NICIR and CLEF, we built an evaluation platform for Chinese QA which has the extendable framework. The evaluation platform is composed of the corpus as the primary source of answers (about 1.8GB from Internet), the testing questions set (7050 testing questions), and the evaluation metrics in terms of Mean Reciprocal Rank.

2. Present a Chinese Question Taxonomy and SVM Classifiers Based on Multiple Features

In this thesis, we present a new question taxonomy from the views of semantic typology and methodological typology, and the SVM classification algorithm based on multiple features. Compared with English hierachical taxonomy, the Chinese parallel question taxonomy can not only choose a most suitable approach for the answer-finding of the question, but also determine the semantic category of the expected answer. The experimental results show that the performances of the proposed SVM classifiers are satisfying.

3. Propose a Chinese Named Entity Recognition Model Based on Multiple Features

The thesis proposes a Chinese named entity recognition (NER) model based on multiple features. It differs from the most of the previous approaches mainly as follows. First, the proposed hybrid model integrates coarse particle feature (POS

model) with fine particle feature (word model), so that it can make up the disadvantages of each other. Second, in order to reduce the searching space and improve the efficiency, we introduce the heuristic human knowledge into statistical model, which could increase the performance of NER significantly. Third, we use several sub-models to respectively describe three kinds of transliterated person name, single character and multi words location name, abbreviative and full organization name. From the experimental results on different testing data, we can conclude that the hybrid model is better than the models which only use one kind of the features.

4. Present Topic-based Language Model for Sentence Retrieval in Chinese Question Answering

For sentence retrieval in Chinese QA, we present a novel topic-based language model. The main idea is to make use of the peculiar characteristc in question answering scenario, that is, the semantic category of the expected answer, to conduct topic segmentation, then incorporate the information of the sentence topic into the standard language model. For the topic segmentation, we propose two approaches that are 'One-Sentence-One-Topic' and 'One-Sentence-Multi-Topics' respectively. The experimental results show that the performance of sentence retrieval based on the proposed topic-based language model is improved significantly.

5. Present Unsupervised Answer Pattern Learning for Answer Extraction in Chinese Question Answering

The thesis adopts the pattern matching technology to tackle with the flexibility and variability of natural language in Chinese question answering. For answer pattern learning, we present an unsupervised learning algorithm. Given two or more questions of one question type, the algorithm can learn answer patterns from internet via web search, topic segmentation, pattern extraction, vertical clustering and horizontal clustering, etc. The experimental results show that the performance of pattern-based answer extraction of Chinese QA is improved significantly.

Keywords: Chinese Question Answering, Evaluation of Question Answering, Chinese Named Entity Recognition, Sentence Retrieval, Answer Pattern Extraction

第一章 绪论 1

1.1 研究的动机 1
1.2 问答系统的分类 3
1.3 问答系统的国内外研究动态 4
1.3.1 国际动态 4
1.3.2 国内动态 5
1.4 论文的研究内容 6
1.5 论文的结构组织 10

第二章 问答系统评测及问答技术研究综述 11

2.1 问答系统评测 11
2.1.1 英语问答评测平台-TREC 11
2.1.2 日语问答评测平台-NTCIR 14
2.1.3 多语言问答系统评测平台-CLEF 15
2.2 问答技术分析 16
2.2.1 基于检索的问答技术 16
2.2.2 基于模式匹配的问答技术 16
2.2.3 基于浅层自然语言处理的问答技术 17
2.2.4 基于统计翻译模型的问答技术 18
2.2.5 四类问答技术的比较分析 19
2.3 应用于问答系统的自然语言处理技术 20
2.3.1 命名实体识别技术 20
2.3.2 短语结构分析或依存结构分析技术 21
2.3.3 逻辑形式转换技术 21
2.3.4 词汇链技术 23
2.3.5 复述技术 23
2.4 本章小结 24
2.5 本章研究成果 24

第三章 构建汉语问答评测平台 25

3.1 引言 25
3.2 汉语问答系统的发展阶段 25
3.3 构建语料库 28
3.4 建立测试集 29
3.4.1 测试集的建立原则和步骤 29
3.4.2 测试集类型 31
3.4.3 测试集答案 32
3.5 制定打分标准 32
3.5.1 事实问题 33
3.5.2 列表问题 33
3.5.3 定义问题 34
3.6 本章小结 35
3.7 本章研究成果 35

第四章 汉语问答系统中的提问分类技术 36

4.1 引言 36
4.2 相关工作 37
4.3 汉语提问分类体系 38
4.4 提问分类的特征向量构造 41
4.4.1 特征类型 41
4.4.2 特征的选择和加权 42
4.4.3 分类器的选择 43
4.4.3.1 支持向量机简介 44
4.5 提问分类实验 45
4.5.1 最佳的特征向量权值算法及其参数 46
4.5.2 不同特征对分类器的贡献 46
4.5.3 BI-GRAM 特征与依存句法特征的性能对比研究 47
4.5.4 验证多类别输出的系统性能 48
4.6 错误分析 49
4.7 本章小结 50
4.8 本章研究成果 51

第五章 基于多特征的汉语命名实体识别 52

5.1 引言 52
5.2 相关工作 53
5.2.1 系统评测和系统性能 53
5.2.2 代表方法 54
5.2.3 汉语命名实体识别的代表系统 54
5.3 基于多特征的汉语命名实体识别模型 56
5.3.1 基本思想 56
5.3.2 基于多特征的汉语命名实体识别模型 58
5.3.3 词形和词性上下文模型 59
5.3.4 实体模型 59
5.3.4.1 人名实体模型 60
5.3.4.2 地名和机构名实体模型 61
5.3.4.3 单字地名实体模型 62
5.3.4.4 简称机构名实体模型 62
5.4 专家知识 64
5.5 模型训练 65
5.6 模型评测 66
5.6.1 平衡因子 B 对系统性能的影响 66
5.6.2 模型的一致性检验 69
5.6.3 专家知识对统计模型的贡献 69
5.6.4 863 评测 70
5.7 本章小结 71
5.8 本章研究成果 72

第六章 基于主题语言模型的句子检索技术 73

6.1 引言 73
6.2 基于语言模型的信息检索 76

6.2.1 文档模型(DOCUMENT MODEL) 76
6.2.2 查询模型(QUERY MODEL) 77
6.2.3 距离模型(DIVERGENCE MODEL) 78
6.2.4 翻译模型(TRANSLATION MODEL) 78
6.2.5 语言模型工具包-LEMUR 79
6.2.6 语言模型的检索方法小结 79
6.3 基于主题语言模型的问答系统句子检索 79
6.3.1 一个句子多个主题 80
6.3.2 一个句子一个主题 81
6.3.3 和 PLSI 的比较 83
6.4 实验部分 83
6.4.1 BASELINE 系统 84
6.4.2 主题语言模型与 BASELINE 系统的对比试验 85
6.4.2.1 基于“一个句子多个主题”聚类的主题语言模型 85
6.4.2.2 基于“一个句子一个主题”聚类的主题语言模型 86
6.4.2.3 不同评测标准下系统的性能 86
6.4.3 聚类效果的分析 88
6.4.4 基于 PLSI 模型的句子检索 89
6.4.5 基于常规伪相关反馈的句子检索 91
6.5 相关工作 92
6.6 本章小结 93
6.7 本章研究成果 93

第七章 基于无监督学习的问答模式抽取技术 94

7.1 引言 94
7.2 基于无监督的问答模式学习算法 96
7.2.1 模式抽取 98
7.2.2 垂直聚类 99
7.2.3 水平聚类 100
7.3 实验结果与分析 102
7.3.1 评测主题划分 103
7.3.2 评测基于模式匹配的答案抽取系统 104
7.3.2.1 基于检索的答案抽取系统 104
7.3.2.2 基于模式匹配的答案抽取系统 105
7.3.2.3 对比分析字符表层模式和句法模式 106
7.4 本章小结 107
7.5 本章研究成果 108

第八章 结论与展望 109

8.1 结论 109
8.2 展望 111

参 考 文 献 116

附录 A：命名实体识别结果样例 126

附录 B：北京大学汉语文本词性标注标记集 128

个人简历 **130**

在学期间发表的学术论文 **130**

在学期间参加的科研项目 **131**

致 谢 **132**

第一章 绪论

1.1 研究的动机

互联网的迅猛发展和广泛普及，使人们可以方便地从网络上获得信息，但是网络信息的爆炸性增长，又使人们准确、快速地获得有价值信息的难度大大增加。2001年世界各大调查机构和搜索引擎公司分别发布了人们使用搜索引擎的现状。其中，PARC 的调查报告显示：网上搜索信息的人很少考虑如何找到他们所需要的信息，因此，他们搜索信息时象动物猎食般盲目。MORI 的民意调查结果表明：只有 18%的用户表示总能在网上搜索到需要的信息，68%的用户说他们对搜索引擎很失望，28%表示还可以，其余 5%为不知道；Roper Starch 的调查指出：36%的互联网用户一个星期要花超过 2 个小时时间在网上搜索；71%的用户在使用搜索引擎的时候遇到过麻烦；搜索受挫中 46%都是因为链接错误；平均每个搜索者在 12 分钟的徒劳搜索后就感到恼火和受挫；绝大部分(86%)的互联网用户感到应当出现更有效的、准确的信息搜索技术。Keen 的调查报告显示：人们平均每天有四个问题需要从外界获取答案；其中 31%的人使用搜索引擎寻找答案；平均每周花费 8.75 个小时找寻答案；53.3%时间花在从旁人那里获得答案，其中 29%的时间花在亲戚朋友身上，24.3%的时间花在销售商那里；网上查找答案的，半数以上都不成功；他们每周将花费 14.5 美元以上，以获取正确的信息[1]。

从这些调查数据中不难看出，尽管一些优秀的搜索服务提供商(Google、Yahoo、百度等)在研发搜索技术方面已经花费了大量的时间和精力，但传统的搜索引擎仍然存在不少的局限性，比如信息丢失、返回信息太多、信息无关等等。这使得网络用户对于现有的搜索技术仍然不满，期盼更完美的搜索技术的出现。造成这种情况的根本原因是传统检索系统具有先天性不足，表现在以下几个方面：

- 以几个关键词的简单组合无法表达用户复杂的检索需求：用户的检索需求往往是非常复杂而特殊的，通过简单的几个关键词组合无法清楚表达用户的检索意图，搜索引擎自然也就无法找出令用户满意的答案了。

- 以关键词匹配为基础的检索算法无法获得令人满意的检索效果：以关键词为基础的索引、匹配算法尽管简单易行，但毕竟停留在语言的表层，而没有触及语义，因此，检索效果很难进一步提高。

[1] http://www.itlearner.com/article/2003/14.shtml

■ 检索结果的冗余信息太多：传统的搜索引擎返回的相关文章太多，用户很难快速准确地定位到所需的信息。在 Google 上输入几个关键字，它有可能返回成千上万个网页，用户将浪费很多时间在这些网页中查找自己所需要的信息。

为了克服传统搜索引擎的弊端，研究人员正尝试探索一种更高效、更人性化的搜索引擎技术-问答系统(Question Answering System)。

[**问答系统定义**] 问答系统是指系统接受用户以自然语言形式描述的提问，并从大量的异构数据中查找出能回答该提问的准确、简洁答案的信息检索系统。

下面是问答系统的几个实例：

Q1：世界上最大的宫殿是什么宫殿？
A1：紫禁城

Q2：NASDAQ 的英文全称是什么？
A2：National Association of Securities Dealers Automated Quotations

Q3：骨质疏松症是一种什么病？
A3：骨质疏松症，英文名字是 osteo-porosis，是一种以低骨量和骨组织微结构破坏为特征，导致骨骼脆性增加和易发生骨折的全身性疾病。

Q4：列举世界四大洋
A4：太平洋、大西洋、印度洋和北冰洋

因此，问答系统和根据关键词检索并返回相关文档集合的传统搜索引擎有着根本的区别。可以说，问答系统能够提供用户真正有用、精确的信息，它是集知识表示、信息检索、自然语言处理于一身的新一代搜索引擎。问答系统目前已成为国际上新兴的一个研究热点。

表 1-1 归纳了问答系统和传统检索系统的主要差别。

表 1-1 问答系统与传统检索系统的区别

	问答系统	**传统检索系统**
系统的输入	**自然语言提问**	关键词组合
系统的输出	**准确的答案**	相关文档的列表
所属的领域	**涉及 NLP 和 IR 两个领域**	纯 IR 领域
信息确定性	用提问表示的用户信息需求相对明确	用关键词组合表示的用户信息需求很模糊

1.2 问答系统的分类

问答系统根据用户自然语言的提问，从大量异构数据中查找提问的答案。根据不同的分类标准，作者把问答系统分为如图 1-1 所示的体系。

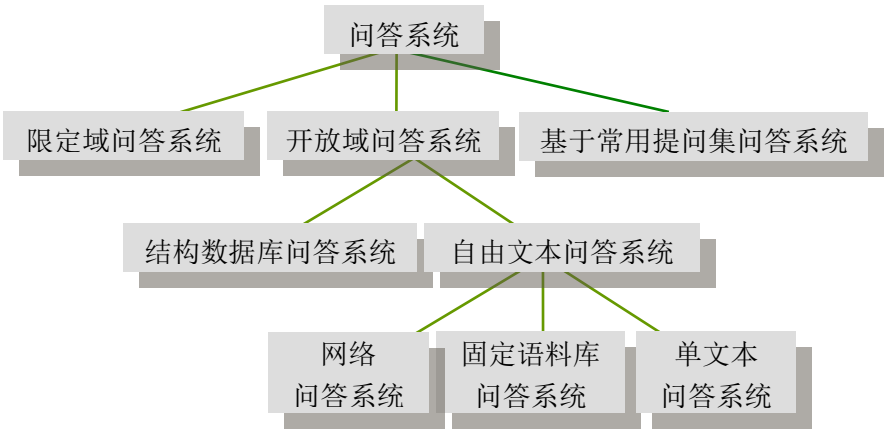

图1-1 问答系统分类体系

在不同的应用环境，需要设计不同类型的问答系统。它们的特点分析如下：

■ 固定语料库问答系统：答案是从预先建立的大规模真实文本语料库中进行查找，如 TREC QA Track。所以，语料库中无法涵盖用户所有类型提问的答案，但能够提供一个优良的算法评测平台，适合对不同问答技术的比较研究。

■ 网络问答系统：从互联网(Internet)中查找提问的答案。所以，可以认为它基本涵盖所有问题的答案。虽然它是在真实环境下研发的问答技术，但由于网络是一个变化的“语料库”，不适合评价各种问答技术的优劣。

■ 单文本问答系统：也可以称之为阅读理解式的问答系统，它是从一篇给定的文章中查找答案。系统在“阅读”完一篇文章后，根据对文章的“理解”给出用户提问的答案。这种系统非常类似于我们在学习英语时做的阅读理解。由于仅在一篇文章中查找提问的答案，数据冗余性不高，所以要求的技术也相对复杂。

■ 结构数据库问答系统：基于结构数据库的问答系统是从一个预先建立的结构化的数据库中查找提问的答案，可以设计出具有较强推理能力的问答技术。但建立大规模的结构知识库是一个非常困难的问题。所以，基于知识库的问答系统只能限定在特定领域。

■ 自由文本问答系统：由于是从自由文本中提取答案，所以它的技术难度相对于结构数据问答系统要大得多。因此，自由文本问答系统是国际上的研

究热点。本论文就是针对该类型汉语问答系统展开研究。

■ 基于常用提问集问答系统：该类问答系统是在已有的“提问—答案”对集合中找到与用户提问相匹配的提问，并将其对应的答案返回给用户。“百度知道[2]”和“新浪爱问[3]”从技术上可以看作是这一类的问答系统。

■ 限定领问答系统：限定领域问答系统的用户提问只能限定在某一特定领域。所以，设计领域知识，实现相对复杂的推理算法是可行的。但当领域转换时，系统需要重新进行设计，领域知识需要重新构建。

■ 开放域问答系统：用户问题不受任何领域限制的问答系统。由于领域的开放，对系统算法的性能要求更高。从领域的开放性看，“百度知道”，“新浪爱问”属于这一类的问答系统。

1.3 问答系统的国内外研究动态

1.3.1 国际动态

从第一个英文问答系统 STUDENT 系统[T. Winograd, *et al*. 1977]，到早期著名的 LUNAR 系统[W. A. Woods. 1977]， MURAX 系统[J. Kupiec. 1993]，DARPA 支持的 HPKB 工程[P. Cohen, *et al*. 1998]和现今由美国 NIST 组织的 TREC QA Track [E. M. Voorhees, *et al*. 1999; E. M. Voorhees. 2000; E. M. Voorhees. 2001; E. M. Voorhees. 2002; E. M. Voorhees. 2003; E. M. Voorhees. 2004]，英文问答技术已经获得了长足的发展，研究领域也从初期的限定领域(Moon Rock, Crisis Management)拓展到如今的开放领域；研究对象从当初的固定语料库拓展到互联网。

目前，比较成功的英文问答式检索系统有 Ask Jeeves[4]，AnswerBus[5]和 START[6]等等。其中，Ask Jeeves 接受自然语言提问，但返回的结果还是和用户提问相关的“**文档**”。AnswerBus 是一个句子级的多语种的问答系统，对于用户提出的法语、西班牙语、德语、意大利语或葡萄牙语问题，系统返回可能包含答案的 8 个“**句子**”。因此，从严格意义上讲，Ask Jeeves 和 AnswerBus 都不是真正的问答系统。而 START 才是真正的问答系统，它直接向用户的自然语言问题并提供用户简洁答案。例如输入问题：How many people in China? 系统的返回则是：1,286,975,468

[2] http://zhidao.baidu.com/
[3] http://iask.com/
[4] http://www.ask.com
[5] http://www.answerbus.com/about/index.shtml
[6] http://www.ai.mit.edu/projects/infolab

(July 2003 est.)。

实际上，现如今的问答系统离实用还有相当大的距离。所以，学术界这几年召开了很多关于问答技术的研讨会。表 1-2 列出了其中几个主要研讨会的情况。

表 1-2 问答系统研讨会情况列表

问答系统研讨会名	**主会名**
Workshop on Open-domain Question Answering	ACL-2001
Workshop on Multilingual Summarization and Question Answering	COLING-2002
Workshop on Multilingual Summarization and Question Answering	ACL-2003
Question Answering in Restricted Domains	ACL 2004
Workshop on Pragmatics of Question Answering	HLT/NAACL2004
IR4QA: Information Retrieval for Question Answering	SIGIR-2004
Workshop on Question Answering in Restricted Domains	AAAI-05
Workshop on Inference for Textual Question Answering	AAAI-05
Workshop on Multilingual Question Answering	EACL-2006
Workshop on Task-Focused Summarization and Question Answering	ACL2006

1.3.2 国内动态

近年来，国内从事问答系统的研究机构也在不断地增加。在往届的 TREC QA Track 评测中，复旦大学[L.D. Wu, *et al*. 2001; L.D. Wu, *et al*. 2002]、中科院计算所[H. B. Xu, *et al*. 2002]都获得了良好的成绩。此外，中科院计算所[7]、哈尔滨工业大学[张刚等, 2001]、复旦大学[Y. Zhang, *et al*. 2003]等[S. J. Li, *et al*. 2002; X. Y. Li, *et al*. 2001; 崔恒等, 2004; 郑实福等, 2002]在汉语问答技术的研究中也作了有益的探索。2003 年，哈尔滨工业大学，北京大学和复旦大学联合承担了一个关于中文问答系统的国家自然科学基金重点项目。但是，和国际研究相比，国内从

[7] http://www.nki.net. cn/

事问答技术尤其是中文自动问答技术研究的科研机构还不多，而且基本没有成型的汉语自动问答系统问世。在这样的情况下，本论文针对汉语问答系统的相关技术进行系统研究和探讨是非常有意义的。

1.4 论文的研究内容

典型的问答系统通常由提问处理模块，检索模块和答案抽取模块三部分组成，如图 1-2 所示。

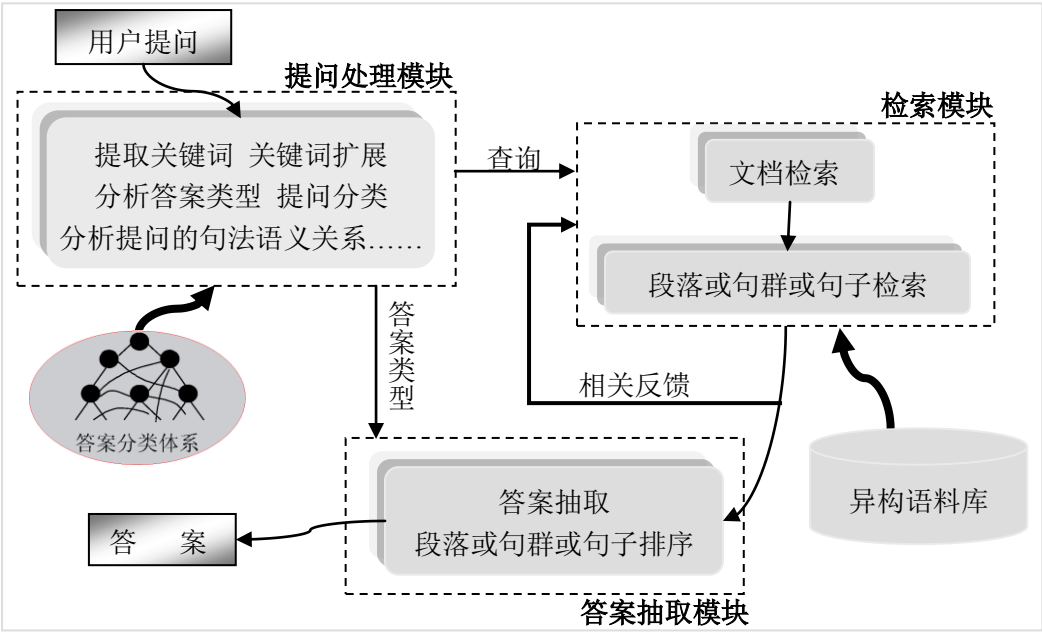

图 1-2 问答系统原理图

[**提问处理模块**] 提问处理模块负责对用户的提问进行处理，生成查询关键词(提问关键词，扩展关键词，...)，确定提问答案类型(PER, LOC, ORG, TIM, NUM, ...)以及提问的句法、语义表示等等。

[**检索模块**] 检索模块根据提问处理模块生成的查询关键词，使用某种检索方式，检索和提问相关的信息。返回的信息可以是段落、也可以是句群或者句子。

[**答案抽取模块**] 答案抽取模块则从检索模块检索出的相关段落、句群或句子中抽取出和提问答案类型一致的实体，根据某种原则对候选答案进行打分，把概率最大的候选答案返回给用户。

从问答系统的原理框图可以看出，问答系统面临解决的关键技术问题可以概括为以下四大类技术：

[1] 提问分类技术

提问分类的任务是为每个用户提问指定一个或多个预先定义的类别，它是问答系统重要的基础环节。Moldovan[D. Moldovan, *et al*. 2003]的研究表明：大约36.4% 的开放域问答系统的错误是由于提问分类的错误造成的。虽然提问分类类似于文本分类，但相对于一整篇文章，一个自然语言的提问中包含的信息量更少，分类难度更大。因此，必须提出有效的方法来解决提问分类中的难点。

[2] 语义标注技术

因为问答系统的答案抽取是基于特定的语义类别，所以应该有这类语义标注工具。该工具可以自动识别文本的某些类语义实体，如人名、地名、机构名、时间词和数量词等等。现阶段的语义标注主要指的是命名实体的识别。

[3] 句子检索技术

问答系统的答案抽取基本上都是在句子检索的基础上进行的。因而，句子检索是问答系统的另一个核心模块，其检索质量不仅在一定程度上影响问答系统的准确率，而且对系统的运行速度也有一定的影响。但由于数据稀疏问题，传统的检索方法在应用于问答系统句子检索时不可避免地会存在不少问题。

[4] 答案匹配技术

对同一语义的形式不同的表述是普遍存在于自然语言中的一种现象，而一个成熟的问答系统应该能够处理由于语言本身的这种灵活性和多变性给问答技术带来的挑战。但在现阶段，由于自然语言处理的各种底层技术仍然不完善和不成熟，对文本进行深层分析，从语义层面来处理语言的灵活性和多变性仍然是一件十分艰难的任务。所以，必须提出一种有效的方法来解决这一问题。

本论文紧紧围绕汉语问答评测环境的建立以及汉语问答系统中四个重要的关键技术问题展开了深入的研究，主要包括：

① 汉语问答系统评测环境的建立

在问答系统的研发过程中，系统评估对于系统的研发和应用有显著的影响。近几年来，“通过系统化、大规模的定量评测推动研发向前发展”的研究方法和技术路线受到越来越多的研发人员的重视。例如，国际上著名的 TREC(Text Retrieval Conference)[8]、NTCIR(NII-NACSIS Test Collection for IR Systems)[9]和CLEF(Cross Language Evaluation Forum)[10]都设立了问答技术评测专项。但到目前为止，还没有一个公开、公认的汉语问答技术评测平台，这是制约汉语问答技术

[8] http://trec.nist.gov/
[9] http://research.nii.ac.jp/ntcir/index-en.html
[10] http://www.clef-campaign.org/

发展的主要障碍。

本论文在吸收 TREC、NTCIR 和 CLEF 等问答系统评测成功经验的基础上，建立了面向汉语问答系统的评测环境 EPCQA(Evaluation Platform for Chinese Question Answering)，并规划了汉语问答系统评测的几个阶段。除此之外，本论文还对 EPCQA 语料库、测试集和打分标准等构建过程进行了详细的介绍。

② 汉语问答系统中的提问分类技术

本论文从新的角度提出了一种新的提问分类体系：提问的技术分类和提问的语义分类。

不同的提问，应该采用不同的技术路线。没有一个统一的技术方案可以解决所有类型的用户提问。因此，问答系统的第一步是要进行提问分类，给当前用户提问指定最合适的解决方案。这就是提问的技术分类。例如提问：毛泽东是哪一年出生的？模式匹配技术可能最合适。

此外，用户提问的答案具有明确语义类别，问答系统的答案抽取应该是基于提问答案的语义类别，而不是考查所有可能的词组或者短语。因此，提问分类还要确定当前提问答案的语义类别。这是提问的语义分类。例如提问：世界上最大的宫殿是什么宫殿？其答案语义类别是地名。

在此基础上，本论文还提出了基于多特征的支持向量机提问分类算法。实验结果表明：本文提出的分类算法可以获得较好的分类效果，其中提问技术分类器和提问语义分类器的分类准确率分别达到了 96.20%和 94.37%。

③ 汉语命名实体识别技术

命名实体识别是信息提取、问答系统、句法分析、机器翻译、面向 Semantic Web 的元数据标注等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位。在问答系统中，命名实体主要用于段落或句子排序以及支撑候选答案的抽取。一般来说，命名实体识别的任务就是识别出待处理文本中三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)命名实体。汉语命名实体，尤其是人名、地名和机构名的识别是命名实体识别的重点和难点。

本论文在对已有算法进行分析的基础上，提出大颗粒度特征(词性特征)和小颗粒度特征(词类特征)相结合、统计模型和专家知识相结合的多特征融合的汉语命名实体识别模型。

用 1998 年 1 月份人民日报生语料进行测试，本文提出的模型对人名、地名

和机构名的识别性能(精确率，召回率)分别达到了(94.06%，95.21%)、(93.98%%，93.48%)和(84.69%，86.86%)。

④ 基于主题语言模型的汉语问答系统句子检索算法

问答系统的检索算法基本上是从传统的文档信息检索中直接借鉴过来的，没有针对问答系统的特殊性挖掘更多的有用信息。而通过观察发现，问答系统中的某些特殊信息可以用来提高句子检索的质量。

利用问答系统的特殊性，本文提出了基于主题语言模型的汉语问答系统句子检索算法。该算法可以概括为：① 依据从初检结果中抽取的候选答案对初检结果进行聚类，即对初检结果进行主题划分；② 统计词语在主题上的概率分布以及句子关于主题的概率分布；③ 通过 Aspect Model 将句子所属的主题引入句子语言模型中，从而获得对句子语言模型更精确的逼近。对于初检结果的主题聚类，本文提出了“一个句子多个主题”和“一个句子一个主题”两种方法。

对比实验的结果表明，本文提出的基于主题语言模型汉语问答系统的句子检索算法在多个评测指标下，相对于标准语言模型的句子检索算法均有明显的提高。其中，基于“一个句子多个主题”聚类思想的主题语言模型对句子检索性能的最大相对提高幅度约为 7.7%。

⑤ 基于无监督学习的问答模式抽取技术及其在汉语问答系统中的应用

对于语言的灵活性和多样性问题，作者希望通过基于字符的表层文本分析技术来解决。模式匹配技术即是这种方法的代表。已有英文问答系统已经采用了这种技术，并在 TREC 评测中获得了很好的成绩[M.M. Soubbotin, *et al*. 2002]。

对此，本论文提出了无监督的学习算法从网络文本中提取应用于汉语问答系统的问答模式。该方法和有监督学习算法的不同在于：无监督学习算法无需用户提供<提问，答案>对作为训练集，只需用户提供每种提问类型两个或以上的提问实例，算法即可通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类等步骤完成该类型提问的答案模式的学习。

实验结果表明：本文提出的无监督问答模式学习的方法是有效的，能够较大幅度地提高汉语问答系统的性能。相对于基于检索的答案抽取系统，基于字符表层模式的答案抽取系统性能提高幅度约 9.0%；基于句法模式的答案抽取系统性能提高幅度约 14.0%。

1.5 论文的结构组织

本论文的主要结构分为八章，具体安排如下：

第一章是绪论部分，旨在介绍论文的选题意义、问答系统的定义与分类、国内外的研究现状，以及论文所要解决的主要问题和结构安排。

第二章综述了三种典型问答系统国际评测项目(即 TREC、NICIR 和 CLEF)以及四类主流的问答技术，即基于信息检索和信息抽取的问答技术、基于模式匹配的问答技术、基于深层自然语言处理的问答技术和基于统计翻译模型的问答技术。

第三章建立了汉语问答系统评测环境 EPCQA，主要介绍了评测环境的建立原则和建立过程。

第四章提出了汉语问答系统的提问分类体系，以及基于支持向量机的提问分类算法，并对基本特征(词和词性)、结构特征(Bi-gram 和依存关系)和词汇语义特征(同义词集合和命名实体)对提问分类器的影响进行了详细的对比实验。

第五章介绍了支撑汉语问答系统的命名实体识别技术，以及本文提出的基于多特征(大颗粒度特征与小颗粒度特征的融合，统计模型和专家知识的融合等)的汉语命名实体识别模型。章节最后给出了系统在 2004 年国家 863 评测命名实体识别专项评测中的情况。

第六章介绍了本论文提出的基于主题语言模型的汉语问答系统句子检索算法和“一个句子多个主题”和“一个句子一个主题”两种主题划分方法。

第七章提出了基于无监督学习的问答模式抽取技术及其在汉语问答系统中的应用，并对模式抽取、垂直聚类和水平聚类等核心模块进行了重点介绍。

第八章为结论与展望，旨在对论文工作进行了总结，并指出了下一步研究的可能方向。

第二章 问答系统评测及问答技术研究综述

2.1 问答系统评测

所有的研究者在设计、研发问答技术的时候都会遇到同样一个问题：如何比较不同问答技术的优劣？问答系统评测即是完成这个任务，它对问答技术的发展有着很大的推动作用。目前，对问答系统进行评测的国际会议有：英语问答评测平台 TREC QA Track (Text REtrieval Conference Question Answering Track)[11]、日语问答评测平台 NTCIR(NII-NACSIS Test Collection for IR Systems)[12]和多语言问答评测平台 CLEF(Cross-Language Evaluation Forum)[13]。

2.1.1 英语问答评测平台-TREC

关于问答系统评测，首先要提到由美国 NIST(National Institute of Standards and Technology)资助的 TREC QA Track。从 1999 年的 TREC-8 到 2005 年的 TREC-13[E. M. Voorhees, *et al*. 1999; E. M. Voorhees. 2000; E. M. Voorhees. 2001; E. M. Voorhees. 2002; E. M. Voorhees. 2003; E. M. Voorhees. 2004]，QA Track 已经成功举办了 7 届。表 2-1 给出 TREC-8 ~ TREC-13 QA Track 发展情况。

从表 2-1 可以看出，TREC QA Track 每年的评测任务和评测指标都在不断地变化，其评测任务主要有：

- Factoid 任务

测试系统对基于事实、有简短答案的提问的处理能力。例如，Where is Belize located？Who is the president of China？而那些需要总结、概括的提问不在测试之列。例如，如何办理出国手续？如何赚钱？等。

- List 任务

要求系统列出满足条件的几个答案。在 TREC2003 之前，任务要求被测试系统给出不少于给定数目的实例。如：Name 22 cities that have a subway system。TREC2003 之后要求系统给出满足条件的尽可能多的实例，如：List the names of chewing gums。

[11] http://trec.nist.gov/
[12] http://research.nii.ac.jp/ntcir/index-en.html
[13] http://www.clef-campaign.org/

表 2-1 TREC-8 ~ TREC-13 QA Track 比较

<table>
<tr><td></td><td>TREC-8
(1999)</td><td>TREC-9
(2000)</td><td>TREC-10
(2001)</td><td>TREC-11
(2002)</td><td>TREC-12
(2003)</td><td>TREC-13
(2004)</td></tr>
<tr><td>任务类型</td><td>Main Task</td><td>Main Task</td><td>Main/List/
Context Task</td><td>Main/List
Task</td><td>Main/
Passage Task</td><td>Series Task</td></tr>
<tr><td>测试提问数</td><td>200</td><td>693</td><td>500/50/10
Series</td><td>500/25</td><td>500</td><td>65 Series
(351 问题)</td></tr>
<tr><td>测试集来源</td><td>FAQ Finder 日志
QA 参赛者
TREC Team</td><td>Encarta/
Excited 日志</td><td colspan="2">MSNSearch/
Ask Jeeves 日志</td><td>---</td><td rowspan="2">Series 是从搜索日志中选取的，提问是人工编写的</td></tr>
<tr><td>测试集真实度</td><td>部分提问采用逆构法产生，故缺乏一定的真实度</td><td colspan="4">测试集都是从自然语言检索系统的日志中提取出来的，只进行了适当的修正</td></tr>
<tr><td>答案个数</td><td colspan="3">每个提问给出按概率大小排列的 5 个答案</td><td colspan="3">Factoid 类型提问只给出一个答案，
List 类型提问给出满足要求的全部答案</td></tr>
<tr><td rowspan="2">答案长度</td><td colspan="3">≤50 字节</td><td colspan="3" rowspan="2">准确答案</td></tr>
<tr><td colspan="3">≤250 字节</td></tr>
<tr><td>评测指标</td><td colspan="3">Main Task: MRR
List/Context Task: Accuracy</td><td colspan="2">CWS</td><td>准确率</td></tr>
<tr><td>最佳性能[14]</td><td>66.0%</td><td>58% (50 字节)
76%(250 字节)</td><td>77%</td><td>85.6%</td><td>55.9%</td><td>60.1%</td></tr>
</table>

[14]事实上，各届 TREC 的系统性能的打分标准不一致，所以并不具备可比性。

■ Definition 任务

要求系统给出某个概念，术语或现象的定义、解释。例如：What is Iqra? 等。

■ Context 任务

测试系统对相关联的系列提问的处理能力，即对提问 i 的回答还依赖对提问 j(i>j)的理解。例如：1、佛罗伦萨的哪家博物馆在 1993 年遭到炸弹的摧毁？2、这次爆炸发生在哪一天？3、有多少人在这次爆炸中受伤？

■ Passage 任务

这是 TREC2003 提出的新任务，Passage 任务对答案的要求偏低，不需要系统给出精确答案，只要给出包含答案的一个字符序列(a small chunk of text that contains an answer)。

■ Other 任务

这是 TREC2004 才定义的任务。TREC2004 的测试集包括 65 个目标(Target)，每个 Target 由数个 Factoid 问题，0~2 个 List 问题和一个 Other 问题组成。例如 TREC2004 测试集中的第四个 Series：

```
- <target id="4" text="James Dean">
  - <qa>
    <q id="4.1" type="FACTOID">When was James Dean born?</q>
  </qa>
  - <qa>
    <q id="4.2" type="FACTOID">When did James Dean die?</q>
  </qa>
  - <qa>
    <q id="4.3" type="FACTOID">How did he die?</q>
  </qa>
  - <qa>
    <q id="4.4" type="LIST">What movies did he appear in?</q>
  </qa>
  - <qa>
    <q id="4.5" type="FACTOID">Which was the first movie that he was in?</q>
  </qa>
  - <qa>
     <q id="4.6" type="OTHER">Other</q>
  </qa>
</target>
```

其中，Other 问题的返回答案应该是一个非空的、无序的、不限定内容的关于 Target 的描述，但不能包括 Factoid、List 问题已经回答的内容。

TREC QA Track 的评测指标主要有平均排序倒数(Mean Reciprocal Rank，简称 MRR)、准确率(Accuracy)、CWS(Confidence Weighted Score)等，计算公式分别如(2-1)和(2-2)。

$$MRR = \frac{\sum_{i=1}^{N} 1/\text{标准答案在系统给出的排序结果中的位置}}{N} \tag{2-1}$$

如果标准答案存在于系统给出的排序结果中的多个位置，以排序最高的位置计算；如果标准答案不在系统给出的排序结果中，本题得 0 分。

$$CWS = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{前}i\text{个提问中被正确回答的提问数}}{i} \tag{2-2}$$

CWS 指标希望系统把最确定的答案排在前面。公式(2-1) ~(2-2)中的 N 表示测试集中总的提问个数。

2.1.2 日语问答评测平台-NTCIR

日语问答评测平台是从 NTCIR-3(2001~2002)/QAC1 年开始的，每两年举办一届，到 2005 年，已经成功举办了三届。NTCIR-3，NTCIR-4 定义了三个子任务[J. Fukumoto, *et al*. 2003]。NTCIR-3/QAC1 和 NTCIR-4/QAC2 日文问答系统的评测情况基本如表 2-2 所示。

表 2-2 各届日文问答评测情况

		任务 1	任务 2	任务 3	语料库
QAC1 2001/2002	测试提问数	200	200	40	Mainichi Newspaper (1998~1999)
	最佳系统性能	0.61	0.36	0.17	
QAC2 2003/2004	测试提问数	200	200	200	Mainichi Newspaper Yomiuri Newspaper (1998~1999)
	最佳系统性能	0.607	0.321	0.21	
QAC3 2004/2005		--	--	360 (50 series)	Mainichi Newspaper Yomiuri Newspaper (2000~2001)
评测指标		MRR，MF，MMF			

[**任务** 1] 每个提问，系统给出五个按概率大小排列的答案列表；采用 MRR 打分标准；系统必须给出支持每个答案的文档。

[**任务** 2] 每个提问，系统只能给出一个答案；如果某个提问在语料中有几个答案，系统须给出所有答案，且必须给出支持每个答案的文档。

[**任务** 3] 这个任务评测系统对关联提问的处理能力；关联提问是指提问之间可能有互指关系、省略等，类似 TREC 中的 Context Task；系统必须给出支持每个答案的文档。

NTCIR5/QAC3(2004~2005)设置了跨语言问答系统的评测[Y. Sasaki, *et al*. 2005]，包括 JE，EJ，CE，CC，EC 等五个子任务。其中 C，E，J 分别表示汉语，英语和日语，XY 表示给定用 X 语言表示的提问，要求系统必须从 Y 语言的文档集合中提取答案，最后翻译成源语言 X 返回。跨语言问答评测系统仅定义了一个任务：对于答案类型是命名实体的提问，要求系统返回一个答案或者空答案，即相当于日文问答评测中的任务 2。严格评测[15]结果具体如表 2-3 所示。

表 2-3 NTCIR-5/CLQA 的评测情况

任务	CC	EC	CE	JE	EJ
最好结果	33%	11.5%	6%	30%	12.5%

2.1.3 多语言问答系统评测平台-CLEF

由 IST Programme of the European Union 资助的 CLEF 在 2003 年设立第一届多语言问答系统评测(Multilingual Question Answering)专项，并计划每年举办一次。CLEF QA Track 定义了单语和多语两个任务，具体情况如表 2-4。

表 2-4 CLEF QA Track 每届基本情况

		CLEF2003	CLEF2004
单语言任务	语种	Dutch Italian Spanish	Dutch French German Italian Spanish
	最佳系统性能	0.422	0.455
多语言任务	语言	Dutch, French, German, Italian, Spanish	Dutch, French, German, Italian, Spanish, others
	最佳系统性能	0.393(bilingual Italian)	0.35
评测指标		MRR	CWS，Accuracy

[15] 对于系统返回的[Answer，DocNo]对，不仅答案 Answer 正确，且文档 DocNo 支持该答案

[**单语言任务定义**] 输入提问是某种语言，输出的答案就是某种语言文字。

[**多语言任务定义**] 输入提问可以是任何一种语言，但是系统给出的答案必须是英文。

2.2 问答技术分析

根据问答系统的技术特色，作者把现有的问答技术分为基于检索的问答技术(IR-based)、基于模式匹配的问答技术(PM-based)、基于深层自然语言处理的问答技术(NLP-based)和基于统计翻译模型的问答技术(SMT-based) 四大类。

2.2.1 基于检索的问答技术

候选答案的排序是这类技术的核心，排序的依据通常是提问处理模块生成的查询关键词。由于不同类别的关键词对排序的贡献不同，算法把查询关键词分为几类，①普通关键词(O)：即从提问中直接抽取的关键词；② 扩展关键词(E)：从 WordNet 或者 Web 中扩展的关键词；③ 基本名词短语(B)；④ 引用词(Q)：通常是引号中的词；⑤ 其他关键词(T)等等。

公式(2-3)给出常用关键词的一种加权方法。

$$Score = wo \times O + we \times E + wb \times B + wq \times Q + wt \times T + ... \quad (2\text{-}3)$$

式(2-3)中的 *wo*，*we*，*wb*，*wq*，*wt* 分别是普通关键词、扩展关键词、基本名词短语、引用词和其他关键词的加权因子，它们体现各种关键词的重要程度。通常 *wo* > *we*，*wq* > *wb* > *wt*。式(2-3)中的 *O*、*E*、*B*、*Q*、*T* 是关键词本身的得分，系统[H. Yang, *et al*. 2002]使用答案关键词和提问关键词的覆盖度来表示；系统[A. Ittycheriah, *et al*. 2002]使用 ISF(Inverse Sentence Frequency)表示。基于检索的问答技术代表系统参见新加坡国立大学 Hui Yang 等人研发的系统[H. Yang, *et al*. 2002]。

2.2.2 基于模式匹配的问答技术

如何自动获取某些类型提问(某人的出生日期、某人的原名、某物的别称等)的尽可能多的答案表述模式是基于模式匹配问答系统的关键技术。也就是说，如果能够获得某类提问答案所有可能的答案表达方式(模式)，问答系统的设计将会变得相对简单。

基于模式匹配的方法往往先离线地获得各类提问答案的模式[D. K. Lin, *et al* 2001; D. Zhang, *et al*. 2002; D. Ravichandran, *et al*. 2002; E. Brill, *et al*. 2001]，在运行阶段，系统首先判断当前提问属于哪一类，然后使用这类提问的所有模式来对抽取的候选答案进行验证。例如，英文中“某人生日年月日”类提问的部分答案模式如下：

1.0 <NAME>(<ANSWER> -)
0.85 <NAME> was born on <ANSWER>,
0.6 <NAME> was born in <ANSWER>
0.59 <NAME> was born <ANSWER>
0.53 <ANSWER> <NAME> was born
0.50 – <NAME>(<ANSWER>
0.36 <NAME>(<ANSWER> -

值得一提的是，在信息抽取领域，人们已经开始把注意力从原来的基于深层文本分析方法转移到基于字符的表层文本分析技术上[M. M. Soubbotin, *et al*. 2001] [M. M. Soubbotin, *et al*. 2002]。基于模式匹配的问答技术代表系统参见俄罗斯 InsightSoft-M 公司 Martin Soubbotin 等人研发的系统[M. M. Soubbotin, *et al*. 2001] [M. M. Soubbotin, *et al*. 2002]。

2.2.3 基于浅层自然语言处理的问答技术

虽然前两种方法相对简单、有效，在 TREC2001、TREC2002 中获得了良好的成绩。但是，人们普遍认为：前两种方法有它们本身的缺陷，要想改进或者说更大程度地提高问答系统的性能，必须引入自然语言处理的技术[E. Nyberg, *et al*. 2003]。现阶段，自然语言处理的技术还不成熟，对句子的深层句法、语义分析还不能达到实用的效果。因此，大多数系统都是基于对句子进行浅层分析，获得句子的浅层句法、语义表示，作为对前两种方法的补充和改进。这方面代表性工作有[D. Moldovan, *et al*. 2001; D. K. Lin, *et al* 2001; E. Hovy, *et al*. 2001; M. Pasca. 2001; D. Moldovan, *et al*. 2002; S. Harabagiu, *et al*. 2000]。

基于自然语言处理的问答技术典型系统参见美国 Language Computer Corporation 公司 Sanda Harabagiu 等人研发的 LCC 系统[S. Harabagiu, *et al*. 2000]，系统的原理图如 2-1 所示。该系统一个最大的特点是使用了词汇链和逻辑形式转换技术，即把提问句和答案句转化成统一的逻辑形式(Logic Form)，通过词汇链，实现答案的推理验证，这在 2.3.3 节有详细的描述。

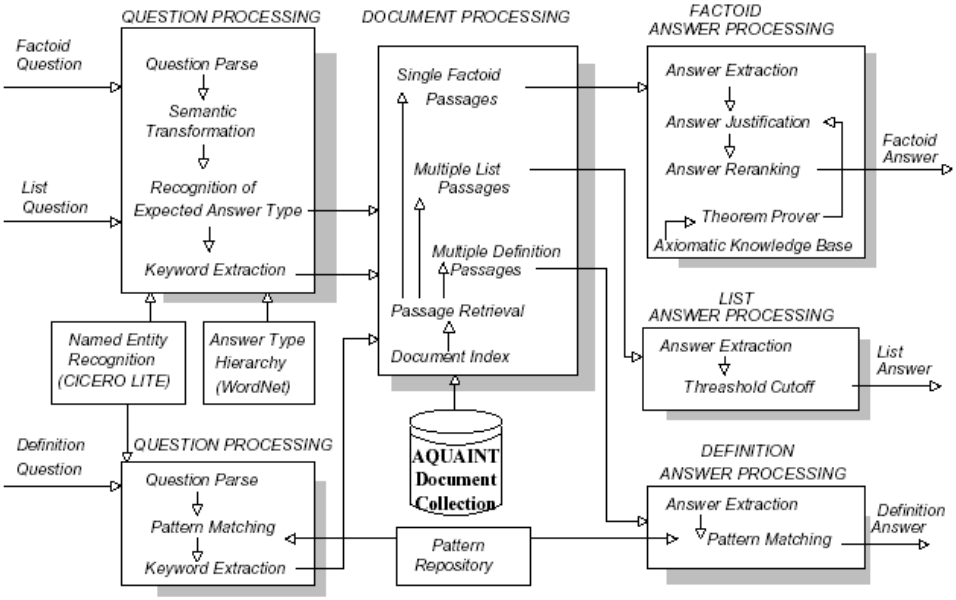

图 2-1　LCC 系统原理图

LCC 系统在 TREC QA Track 2001 ~ 2004 连续三年的评测中均获得第一名的成绩，且具有较大的领先优势。

2.2.4 基于统计翻译模型的问答技术

基于统计翻译模型的问答技术把提问句看作是答案句在同一语言内的一种翻译形式，答案句中和提问句的疑问词对应的词即是该问句的答案。其代表系统是 A. Berger[A. Berger, *et al* .2000]，A. Echihabi[A. Echihabi, *et al*. 2003]和 V. Murdock[V. Murdock, *et al*. 2004]等人的工作。这里以 A. Echihabi 工作为代表介绍这类方法的基本思想。

对于句子检索结果，A. Echihabi 方法通过以下几个步骤实现基于统计翻译模型的答案抽取：

- 对检索句进行句法分析
- 保留句中的提问词
- 候选答案用其句法或者语义类表示，并加上前缀”A_”
- 所有不包含提问词和候选答案词的非叶子节点用其句法或语义类表示
- 保留所有剩下的叶子节点
- 使用翻译模型(主要是对齐技术)抽取提问的答案

图 2-2 给出了该方法的原理示意图。

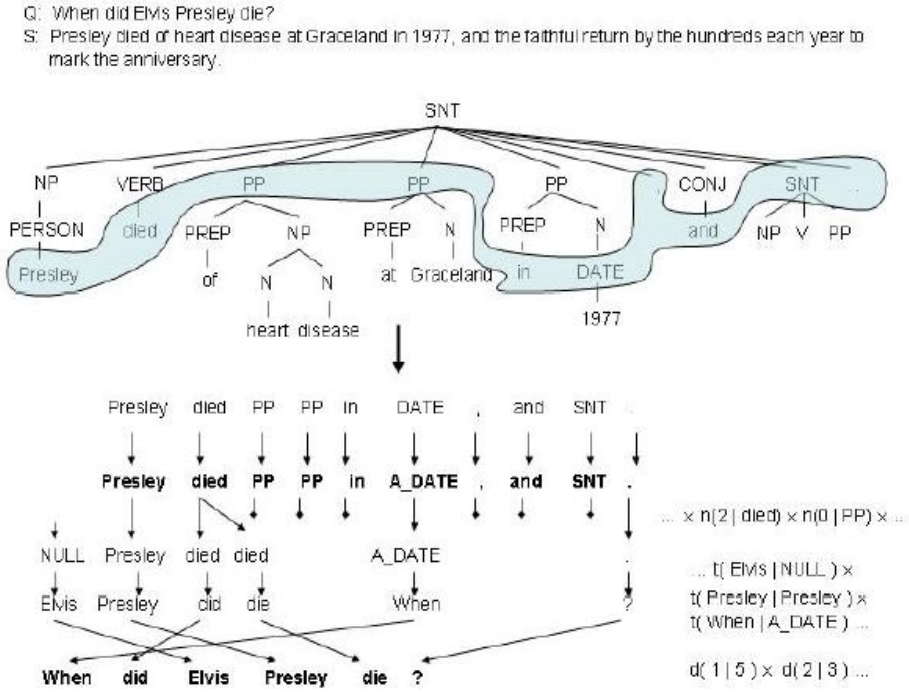

图 2-2 A. Echihabi 方法的原理图

2.2.5 四类问答技术的比较分析

基于检索的问答技术相对简单，容易实现。但它以基于关键词的检索技术(也可被称为词袋检索技术)为重点，只考虑离散的词，不考虑词之间的关系。因此，无法从句法关系和语义关系的角度解释系统给出的答案，也无法回答需要推理的提问。

基于模式匹配的问答技术虽然对于某些类型提问(如定义，出生日期提问等)有良好的性能，但模板不能涵盖所有提问的答案模式，也不能表达长距离和复杂关系的模式，同样也无法实现推理。

基于自然语言处理的问答技术可以对提问和答案文本进行一定程度的句法和语义分析，从而实现推理。但目前自然语言处理技术还不成熟，除一些浅层的技术(汉语分词、词性标注、命名实体识别等)外，其他技术还没有达到实用的程度。所以，这种技术的作用目前还很有限，只能作为对前两种方法有效的补充。

基于统计翻译模型的问答技术在很大程度上依赖训练语料的规模和质量，而对于开放域问答系统，这种大规模训练语料的获取是非常困难的。

针对这四类技术的总结分析，作者认为：基于字符表层的文本分析技术(例如模板技术)必须和快速、浅层自然语言处理技术有效结合，才能获得性能优良的问答系统。

表 2-5 给出了前三类方法的典型系统以及它们在各届 TREC Factoid 子任务中取得的名次[16]。

表 2-5 三类问答技术代表系统在 TREC 评测中的成绩比较

		IR-based	PM-based	NLP-based
代表系统		代表系统[H. Yang, *et al*. 2002]	代表系统[M. M. Soubbotin, *et al*. 2001; M. M. Soubbotin, *et al*. 2002]	代表系统[D. Moldovan, *et al*. 2001; D. Moldovan, *et al*. 2002]
代表方法在各届 TREC 中的名次	2000	-	-	1
	2001	-	1	2
	2002	3	2	1
	2003	2	-	1

2.3 应用于问答系统的自然语言处理技术

通过前面章节的分析和总结，可以得出这样的结论：问答系统整体性能的优劣在很大程度上依赖于对 NLP 技术和资源的有效利用。本节将详细介绍现阶段有哪些自然语言处理技术可以被应用于问答系统以及它们是如何应用于问答系统的。这些技术主要包括：命名实体识别技术、短语结构分析或依存结构分析技术、逻辑形式转换技术、词汇链技术和复述(Paraphrase)技术等等。

2.3.1 命名实体识别技术

相对于其他技术而言，命名实体识别技术是使用最广泛的一项自然语言处理技术，对问答系统的性能有着至关重要的影响。命名实体识别技术主要被用于问答系统的段落或句子排序和答案抽取两个阶段。

■ 段落或句子排列

问答系统首先根据查询关键词进行检索，然后对于检索出来的段落或句子重新进行排序：当某个句子包含所期望的实体时，则给句子适当的加分。例如，系统[A. Ittycheriah, *et al*. 2002]采用如式(2-4)的加分策略。

$$Sne,i = Si + (E + (3 - dne) \times Dne) \tag{2-4}$$

式(2-4)中，Sne,i 表示句中第 i 个候选答案的得分；Si 表示句子检索算法的得

[16] 基于统计翻译模型问答技术的系统没有参加 TREC 评测

分；*E* 表示当该句中出现期望实体时，给该句的一个加分；*dne* 表示期望实体和提问关键词之间的距离；*Dne* 是对距离的惩罚因子。

- 答案抽取

大多数的问答系统都是在答案抽取阶段使用命名实体的技术[D. Moldovan, *et al*. 2001; E. Brill, *et al*. 2001; H. Yang, *et al*. 2002; D. Moldovan, *et al*. 2002]，答案抽取模块只抽取和期望答案类型一致的实体作为答案，而命名实体不参与句子或段落的排序。

2.3.2 短语结构分析或依存结构分析技术

短语结构分析或依存结构分析的结果是得到句子的短语结构句法树或依存结构句法树。在句子排序或答案抽取阶段，使用更合理的句法信息。举个例子：

> 提问：Who killed Lee Harvey Oswald?
> 文本：Belli's clients have included Jack Ruby, who killed John F. Kennedy assassin Lee Harvey Oswald and Jim and Tammy Bakker.

对于候选答案 Jack Ruby 和 John F. Kennedy，如果采用基于词袋的检索问答技术，系统很有可能返回 John F. Kennedy，因为 John F. Kennedy 和查询关键词 killed、Lee Harvey Oswald 的距离更近。但是，如果引入句法信息，系统只会返回答案 Jack Ruby。因为 Jack Ruby 在文本中是 killed 的逻辑主语，Lee Harvey Oswald 是 killed 的逻辑宾语，这和问句的句法结构完全相似。

算法[D. Mollá 2003; D. K. Lin, *et al* 2001; E. Hovy, *et al*. 2001; K. C. Litkowski. 1999; T. Tetsuro, *et al*. 2002; U. Hermjakob, *et al*. 2001; S. Harabagiu, *et al*. 2000]在使用句法树(短语句法树或依存句法树)的细节上有所不同，但他们的目的都是比较提问句法树和文本句法树的相似性，使系统给出的答案有句法上的解释。

2.3.3 逻辑形式转换技术

通过比较提问和文本的句法树来抽取答案虽然提高了系统的性能，但这种基于句法树分析的方法还是非常浅层的。因为对句法树的分析基本上就是合一(Unification)运算，比较两棵句法树的相似性，无法回答那些需要推理才能回答的提问。举个例子：

提问：Who is the first Russian astronaut to walk in space?

文本：The broad-shouldered but paunchy Leonov, who in 1965 became the first man to walk in space, signed autographs.

提问依存树和文本依存树分别如图 2-3 和 2-4 所示。

图 2-3 提问依存关系数

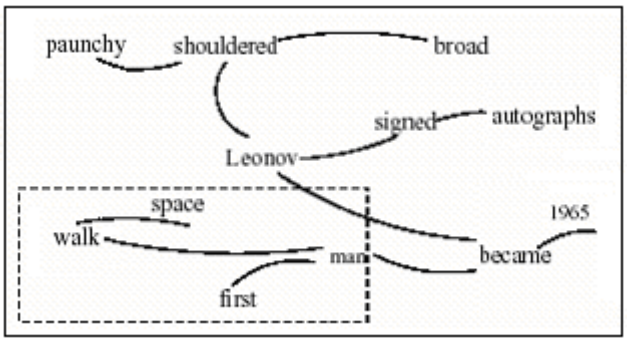

图 2-4 文本依存关系树

图 2-4 中的虚线框是问句依存树和文本依存树合一的结果。可以看到，合一的结果并不能给出提问的答案 Leonov，因为 Leonov 和 man 没有依存关系。这个时候，必须使用语义信息(Leonov 和 man 在这里是指同一个实体)才能给出正确答案。

[D. Moldovan, *et al*. 2001; D. Moldovan, *et al*. 2002]提出逻辑形式转换(Logic Form Transformation)来解决这个问题，即把问句和文本同时转化成统一的逻辑形式(QLF 和 ALF)，通过对 QLF 和 ALF 的运算实现答案的抽取。逻辑形式最大的特色是它通过词汇链表达语义推理知识，这也是 LCC 系统成绩优异的主要原因。举个例子来说明：

Q：When did Lucelly Garcia, former ambassador of Colombia to Honduras, die?

A：Several gunmen on a highway leading to the Colombian city of Ibague murdered Colombian Ambassador to Honduras Lucelly Garcia in 1994.

QLF：Lucelly_Garcia(x1) & former(x1) & ambassador(x1) & of(x1, x2) & Colombia(x2) & to(x1, x3) & Honduras(x3) & die(e1, x1) & TIME_STAMP(e1)

ALF：gunman(x2') & murder(e1', x2', x1') & Colombian(x1') & ambassador(x1') & to(x1', x3') & Honduras(x3') & Lucelly_Garcia(x1') & TIME_STAMP(e1')

从 WordNet 中获取的推理知识库：

Rule1: Colombian(x1) --->of(x1, x2) & Colombia(x2)
Rule2: murder(e1, x2, x1) ---> kill(e1, x2, x1) & intentionally(e1)
Rule3: kill(e, x1, x2) ---> cause(e1, x1, e2) & die(e2, x2)

根据 Rule1，Rule2 和 Rule3，由 ALF 可以推理出 ALF1。

ALF1：gunman(x2') & cause (e2', x2', e3') & die (e3', x1') & intentionally (e1') & of(x1', x7') & Colombia(x7') & ambassador(x1') & to(x1', x3') & Honduras(x3') & Lucelly_Garcia(x1') & TIME_STAMP(e2') & TIME_STAMP(e3')

通过合一运算 QLF 和 ALF1，系统给出正确答案"in 1994"。在整个推理过程中，词汇链起到了非常重要的作用。

2.3.4 词汇链技术

很多情况下，提问关键词和文本关键词是不一致，但它们却表达相同的意思。词汇链对于解决这类提问非常的重要。2.3.3 节利用 WordNet 构建词汇链，连接提问关键词和答案关键词，实现推理[D. Moldovan, *et al*. 2001]。

例如，WordNet 对 kill 的一个解释：cause to die，这样就可以把 kill 和 die 连接起来，即 kill 分解为 cause 和 die 两个动作，而且 kill 的宾语是 die 的主语，其 LF 为：Kill(e, x1, x2) ---> cause(e1, x1, e2) & kill(e2, x2)。

2.3.5 复述技术

复述(Paraphrase)是指用不同的词汇-句法结构表达同样的意思。2.3.4 节描述的词汇链就是一种特殊的 Paraphrase：词汇 Paraphrase。Paraphrase 技术可以解决因提问和答案的表述不同给问答系统的设计带来的麻烦。举个例子：

When did Colorado become a state?
(1a) Colorado became a state in 1876.
(1b) Colorado was admitted to the Union in 1876.
Who killed Abraham Lincoln?
(2a) John Wilkes Booth killed Abraham Lincoln.
(2b) John Wilkes Booth ended Abraham Lincoln's life with a bullet.

如果上述两个提问的答案都是以(1a)(2a)的形式来表述的，问答系统使用非常简单的技术(命名实体识别技术)就可以找出答案。但是如果答案以(1b)(2b)的形式

出现，问答系统要找到答案将是非常困难的。但是通过 Paraphrase 技术获得如下的 Paraphrase 规则：

X became a state in Y ⟺ X was admitted to the Union in Y

X killed Y ⟺ X ended Y's

问答系统就能容易地找出提问的答案。将 Paraphrase 技术应用于问答系统的代表工作有[A. Ibrahim, *et al*. 2003; F. Duclaye, *et al*. 2003; D. K. Lin, *et al* 2001; F. Rinaldi, *et al*. 2003]等。

2.4 本章小结

基于自然语言提问的问答系统经过这几年的发展，已经成为自然语言处理研究领域的一个重要分支和新兴的研究热点，其"通过系统化、大规模地定量评测推动研究向前发展"的发展轨迹，以及某些成功启示，都推动了自然语言处理研究的发展，促进了 NLP 研究与应用的紧密结合。

但是，目前的问答技术也不成熟，问答系统能够处理的提问非常有限，其性能离实用的目标还很远。作者认为，在问答系统(尤其是汉语问答系统)的发展过程中，应该注意以下一些问题。

- 处理好问答技术研究和系统实用性之间的关系。目前的问答系统基本上都是针对具有简短答案的事实问题研发的，但这样的系统在实际应用中到底能够解决用户真正关心问题的百分之多少，或者说应该研究哪种类型问答系统。这点非常值得去研究。
- 重视大规模的公开评测技术，以评测推动问答技术的发展。现阶段对于汉语问答技术的研究，迫切需要一个公开、公认、合理的问答评测平台。
- 从问答技术的研究角度看，需要重视基于字符表层文本分析技术和基于自然语言处理技术的有效结合，扬长避短。

2.5 本章研究成果

吴友政, 赵军, 段湘煜, 徐波. 问答式检索技术及其评测研究综述. 中文信息学报. 2005 年第 3 期, pp1~13.

第三章 构建汉语问答评测平台

3.1 引言

在问答系统的研发进程中，系统评估对于系统的研发和应用有显著的影响。近几年来，“通过系统化、大规模的定量评测推动研发向前发展”的研究方法和技术路线受到越来越多的研发人员的重视。例如，国际上著名的 TREC，MUC(Message Understanding Conference)，DUC(Document Understanding Conference)，国内的 863、973 评测等等。这种以评测推动研究发展的思路意义在于：

- 以系统化、大规模测试为基础，推动研究的向前发展；
- 通过开放式的论坛，使与会者能交流研究的成果与心得，增进学术界和产业界的交流互通；
- 通过对真实环境的模拟，加速将实验室研究成果转化为产品；
- 发展适当且具应用性的评估技术，供各界遵循采用。

然而，和英语、日语、欧洲跨语言的问答系统相比，缺乏大规模的汉语问答评测环境已经成为制约汉语问答技术发展的主要障碍。

本文旨在吸收 NTCIR，TREC 和 CLEF 成功经验的基础上，推出汉语问答系统评测环境 EPCQA (Evaluation Platform for Chinese Question Answering)，希望能与国内外问答检索领域的团队合作，在各个研究小组的共同参与下，互相验证彼此的研究成果，完善汉语问答系统测试集，一起推动汉语问答技术的发展。

3.2 汉语问答系统的发展阶段

2000 年，美国国防部高级研究规划局 TIDES(The Trans-lingual Information Detection, Extraction and Summarization)曾对问答技术未来的发展进行了规划[J. Burger, *et al*. 2001]。但此规划的可操作性不强[E. M. Voorhees. 2000]。所以，本论文在规划汉语问答系统评测的阶段性任务时，遵循黑箱子原则和可操作性原则，从用户提问的答案角度将之划分为四个阶段，如图 3-1 所示。

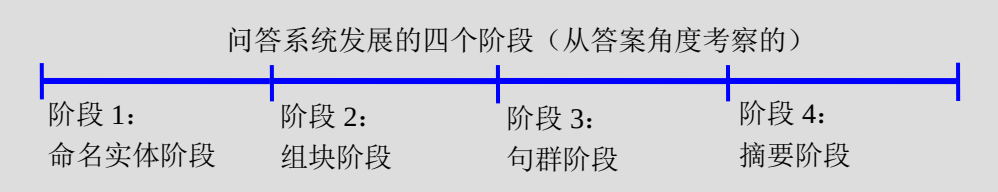

图 3-1 汉语问答系统发展阶段

所谓黑箱子原则，是指评测系统只从答案的角度考察问答系统，而不考虑系统使用的技术以及系统给出答案的依据文档。例如：

用户提问：世界上平均海拔最高的洲是哪个洲？
文档 258991：……亚洲地形的总特点是地势高、地表起伏大，中间高、周围低，平均海拔约 950 米，是除南极洲外世界上地势最高的一洲……
文档 258191：……世界海拔最高的洲——南极洲，平均高度海拔 2 350m……

很显然，从文档 258991 中给出答案的难度要大于从文档 258191 中抽出答案的难度，但评测平台平等对待文档 258991 和文档 258191 中的答案，没有区别打分。

采用黑箱子原则的主要原因包括两个方面：

I. 便于评测的需要。因为对用户提取答案的源句进行打分，标准很难把握。

II. 对于基于大规模语料库的问答系统，选择从包含答案的简单句中提取答案本身就是一项艰难的技术。

■ 命名实体阶段

该阶段评测系统的主要任务评测问答系统处理答案类型是命名实体(人名、地名、机构名、时间、数量等等)的用户提问的能力，而且测试集中的每个问题都可以从语料库中查找到答案。例如提问：

Q1：二战期间的美国总统是谁？
A1：罗斯福

Q2：国际奥委会成立于哪一年？
A2：1894 年

■ 组块阶段

问答系统在该阶段必须能够处理答案类型是组块(包括词，命名实体，短语等)的用户提问，其候选答案选择模块可能还需要一个知识本体作支撑。例如：

用户提问：美国邮递员的制服是什么颜色的？
文档 107110：……美国大城市的邮差多是步行，在投递上分片包干。他们穿蓝灰色的制服，推着个小型轻便推车，走到各家门前就停下来放上一把已经分检好的邮件……

对于上述问题，如果问答系统在知识本体的支撑下，知道蓝灰色是一种颜色，那么，系统极有可能从文档 107110 中找出提问的正确答案：蓝灰色。

此外，第二阶段的提问类型还包括列表型提问和定义型提问等。例如提问：

Q3：请问初唐四杰是哪四位？
A3：王勃、杨炯、卢照邻、骆宾王

Q4：什么是 H 股？
A4：注册地在内地、上市地在香港的外资股

■ 句群阶段

这一阶段的问答技术不一定会难于第一、二阶段，但它主要目的是评测问答技术的实用化程度，把问答系统推向实用。所以，在句群阶段，问答系统处理的用户提问范围更广，除了包括前两个阶段全部的用户提问类型，更主要的是处理答案类型是句子，或句群的用户提问类型。例如提问：

Q5：人在死海游泳不会沉到水底，是什么原因？

Q6：如何办理出国手续？天空为什么是蓝色的？

■ 摘要阶段

该阶段的评测系统将评测问答系统是否已成为真正的“专业信息分析师”。此阶段的问答系统应该能够基本上满足用户所有的要求，能够处理任何用户提出的问题，并从大量的异构语料(结构化、半结构化、自由文本、多种语言、多个媒介)中提取、判断、概括、总结出答案。例如提问：

Q7：美国在伊拉克战场上的战况如何？

Q8：2004 年我国的财政状况如何？

摘要阶段和句群阶段的最大区别是：句群阶段的提问答案只是机械地从文档中抽取出句子，返回给用户，而摘要阶段用户提问的答案可能还需要在理解之后进行生成。

本文主要针对评测的命名实体阶段和组块阶段建立面向汉语的问答评测平台。第一、二阶段的区别主要在于用户提问的答案类型不同，其他方面基本上没

有差别。所以，剩下章节的介绍不再分第一阶段还是第二阶段，而是将重点介绍汉语问答评测平台(EPCQA)的三个重要环节：构建语料库、测试集和打分标准。

3.3 构建语料库

虽然基于 Web 的问答系统更能满足用户的需求，且不需要收集大规模的语料库，但 Web 是一个动态变化的海量“语料库”，语料每天都在发生变化，这一点不利于对不同算法以及同一算法在不同阶段的评测。因此，论文主要对基于固定语料库的问答技术进行评测。

为了能够在更大程度上模拟系统实际使用的情况，论文收集的语料均来自互联网网页。目前，EPCQA 语料库的规模已达 1.8GB，主要分布于国内、国际、娱乐、体育、社会和财经等领域。

此外，为了评测需要，论文对 EPCQA 语料库进行一定程度的加工。表 3-1 给出了 EPCQA 语料标注的标记集。

表 3-1 EPCQA 语料标注的标记集

标记形式	*说明*
<DOC>…</DOC>	文章开始和结束
<DOCNO> … </DOCNO>	文章编号
<SOURCE> …</SOURCE>	文章来源
<PUBLISHER> …</PUBLISHER>	文章出版者
<BODY>…</BODY>	文章主题
<HEADLINE> … </HEADLINE>	文章标题
<TEXT>…</TEXT>	文章内容
<P>…</P>	**段落标记**

下面是 EPCQA 对一篇文档的标注实例。

```
<DOC>
<DOCNO> INTERNET.0001 </DOCNO>
<SOURCE> INTERNET </SOURCE>
<PUBLISHER> ChineseLDC </PUBLISHER>
<BODY>
<HEADLINE>
1987 诺贝尔文学奖
</HEADLINE>
```

<TEXT>
<P>
约瑟夫·布罗茨基(Joseph Brodsky， 1940～1996)苏裔美籍诗人。生于列宁格勒一个犹太家庭，父亲是摄影师，布罗茨基自小酷爱自由，因不满学校的刻板教育，15 岁便退学进入社会。他先后当过火车司炉工、板金工、医院陈尸房工人、地质勘探队的杂务工等。业余时间坚持写诗，译诗。
</P>
……
<P>
1987 年，由于他的作品“超越时空限制，无论在文学上及敏感问题方面，都充分显示出他广阔的思想和浓郁的诗意”，获得诺贝尔文学奖。
</P>
</TEXT>
</BODY>
</DOC>

3.4 建立测试集

EPCQA 已从多个不同的渠道，例如，自然语言搜索网站日志、百科知识问答题库、实验室人员对热点问题的提问和对英语提问句(主要来源于 TREC–QA 测试和训练集)的意译[吴友政等. 2004]等，收集了 4250 个基于事实的汉语提问句。

此外，哈尔滨工业大学信息检索实验室向本文提供了 2800 个汉语提问句，因此，评测平台现有 7050 个汉语提问句[17]。

3.4.1 测试集的建立原则和步骤

EPCQA 测试集的建立遵循全面性、真实性和无歧义性三个原则。

全面性指测试集中的提问要尽量涵盖多个主题，避免千篇一律地全是提问人物或者地点等。

真实性指测试集中的提问应尽量反映用户使用疑问句的习惯，避免千篇一律的疑问句法。TREC-8 在测试集的真实性方面做得不够，因为测试集中的一部分

[17] 本文也向哈尔滨工业大学信息检索实验室提供了 4250 个汉语提问句。

提问是通过逆构法[18]产生的。这样的问句通常会包括较多的提示信息，比较容易回答[E. M. Voorhees. 1999]。从 TREC-9 开始，测试集都是从自然语言检索系统的搜索日志(例如 MSNSearch 和 AskJeeves 的搜索日志)中提取出来的。

无歧义性指测试集中的每个提问都不能有歧义。

按照上述三个原则，论文分三步完成了汉语问答系统测试集的建立：

第一步 自动过滤

过滤原则是问句中应该包括一个疑问词(谁、哪、什么时候等)；或者以情态词或动词开始；或者以问号结束。

第二步 人工过滤

过滤掉的问题包括非事实问题、程序问题、某物在网络中位置问题，模糊性的问题等。

第三步 人工修正

对测试集进行的人工修正的工作主要包括拼写检查，标点符号检查和语法规则的检查等。

最终，从自然语言搜索网站的日志中共提取 5400 多个提问。但是，其中很多提问还不是现阶段问答系统研究的重点，例如：非基于事实的提问、省略了疑问词的提问、表达模糊的提问、要求回答的是完成某件事的程序而非简短答案的提问等等。论文对这些提问进行人工剔除，例如：“如何网上赚钱？女朋友过生日送什么礼物？如何申请免费空间？成龙的近况如何？”等。还有一些符合要求但表达不当的提问，论文对它们进行了一定的修改。

百科知识问答题库中的提问相对比较书面化，不能够真实反映用户使用问句的方式。对此论文进行了一些口语化的处理。例如提问：香港电影《花样年华》最近在第 53 届戛纳国际电影节上获最佳男主角奖，在该片中饰演男主角的哪一位演员？中国第一次派运动员参加的奥运会和中国夺得第一枚金牌的奥运会是在同一城市举行，它是什么城市？论文分别把它们修改成：谁在香港电影《花样年华》中饰演男主角？中国夺得第一枚金牌的奥运会是在哪个城市举办的？作者认为这样更能反映系统在使用中的实际情况。

实验室工作人员可以提出任何他们感兴趣的问题，只是要求对提问的表达要尽可能的多样化，不要总是用是同一种方式进行提问。

对英语提问句的翻译是论文获取汉语问答系统测试集的另一个重要的途径。其中，英语提问句的来源主要是往届的 TREC 评测的测试集。这里的“翻译”不全

[18] 问题设计者先找一个自己感兴趣的 topic，然后根据检索到的文本把陈述句改为疑问句。

是对英语提问句的直接翻译，而是对于部分可能在文中找不出答案的提问在不改变提问类型的情况下，进行了适当的修改，例如：

英语提问：Who wrote "East is east, west is west and never the twain shall meet"?
汉语提问：名著《红楼梦》是谁的作品？

英语提问：What is the name of CEO of Apricot Computer?
汉语提问：联想公司的 CEO 叫什么名字？

目前，作者通过上述四个途径已建立了一个有 4250 个提问的汉语问答系统测试集。这个测试规模还很小，作者希望能够在以后的工作中逐步扩大和完善。

3.4.2 测试集类型

EPCQA 的 7050 个测试集问题可以分为三大类：事实问题、列表问题和描述问题。关于它们的例子可以参看表 3-2。

表 3-2 汉语问答系统测试集的部分实例

提问类型	例子	答案
事实问题	氧气占空气体积的百分之多少？	20.95%/21%/五分之一
	目前国际奥委会总部在哪里？	瑞士洛桑
	谁获得 1987 年的诺贝尔文学奖？	约瑟夫·布罗茨基
	被称为我国“瓷都”的是指哪一城市？	景德镇
	中国最大的商业银行是什么银行？	中国工商银行
	布什是谁？	美国总统
	林肯是怎么死的？	暗杀
	UPS 全称是什么？	不间断电源系统/Uninterruptible Power System
列表问题	请问初唐四杰是哪四位？	王勃、杨炯、卢照邻、骆宾王
	列举出联合国安理会常任理事国？	中国 美国 俄罗斯 英国 法国
定义问题	UFO 的是什么？	Unidentified Flying Objects/不明飞行物
	什么是 H 股？	注册地在内地、上市地在香港的外资股

[**事实问题**] 事实问题是指用户的提问是客观事实，不是个人的主观想法或者意见，其答案通常都是一个组块(包括词和短语)。

[**列表问题**] 列表问题实际上是事实问题的一个子类，不同的是系统返回的答案是不少于提问指定数目的实例。

[**描述问题**] 描述问题则是要求系统给出对一个人、一件事物或组织的简短描述。

3.4.3 测试集答案

在确定了测试集的提问之后，接下来要做的就是从语料库中找出这些提问的简洁答案[19]。如果某个提问在语料库中没有答案，问答系统应返回 NIL；否则系统返回的答案应该是如下形式的三元组：[问题编号 答案 支持答案的文档编号]。

对于某些问题，语料库中的不同文档给出的答案可能不相同，有的甚至是错误的答案。但只要文档能够支持这个答案，评测时就将之作为正确答案对待。例如：

> 用户提问 1：18K 金含金量是多少？
> 文档 5891：…… 24K 为足金,含量为 99.9%,18K 含金量为 75%.……
> 文档 5892：……每 K 金含金量为 4.15%，含金量为 99.6%以上的为 24K，含金量 91.3%为 22K，含金量 74.4%为 18K，其余以此类推……

此时文档 5891 支持的答案(75%)和 5892 中的答案(74.4%)都被作为正确答案对待。

3.5 制定打分标准

汉语问答系统的评分标准采用国际上通用的 *MRR*、准确率(*Precision*)、召回率(*Recall*)和 *F-Measure* 等指标。且针对不同的问题类型，对答案的具体要求和打分标准也有所差异。

[19] 简洁答案是指问答系统给出的答案不能包括除答案之外的字符串。例如提问:《哈利 波特》一书的作者是谁？答案 a)、b)、c)都不能作为正确答案，只有 d)正确。a)37 岁的罗琳；b)《哈利-波特》作者罗琳成英国第一女富豪；c)《哈利 波特》小说的作者罗琳；d) 罗琳。

3.5.1 事实问题

事实提问采用 *MRR* 打分标准。即，每个事实问题，问答系统可以给出按照概率大小排列的五组[问题编号 答案 支持答案的文档编号]对。如果第一个答案是对的，那么这个问题就得 1 分，如果第二个答案是对的，那么这个问题得 1/2 分，如果第三个答案是对的，那么这个问题得 1/3 分，依此类推。如果所有给出的答案都是错误的，那么就得 0 分。把每个问题所得的分加起来再除以问题的总数就可以得到整个事实问题测试集的 *MRR*。*MRR* 越高，说明该系统的性能越高。具体参见公式 3-1。

$$MRR = \frac{\sum_{i=1}^{N} 1/\text{标准答案在系统给出的排序结果中的位置}}{N} \tag{3-1}$$

说明：如果标准答案存在于系统给出的排序结果中的多个位置，以排序最高的位置计算；如果标准答案不在系统给出的排序结果中，本题得 0 分。

3.5.2 列表问题

对于列表问题，问答系统给出的答案是一个是非空、无序、无重复、不超过指定大小的列表。EPCQA 可以保证语料库中至少包含提问中指定数量的实例，但不能保证每个列表问题的所有指定大小的实例都能在语料库中的某一篇文章中找到，有时实例可能分散在多个文章中。在这种情况下，EPCQA 要求问答系统能够从这多篇文章中概括出列表问题的实例。例如：

用户提问 2：东北三宝是哪三宝？
文档 5893：闻名于世的东北三宝之一的貂皮，可称得上是裘皮之冠，……
文档 5894：人参是“东北三宝”第一宝。山参的生长在深山老林之中……鹿茸是“东北三宝”之一，是雄鹿额骨上生长的尚未骨化的幼角……

问答系统需要从文档 5893 和 5894 中概括出如下的答案列表：

用户提问 2	文档 5893	貂皮
用户提问 2	文档 5894	人参
用户提问 2	文档 5894	鹿茸

每一个列表问题的答案评分采用事例召回率(*IR*)、事例准确率(*IP*)和事例 *F-Measure*(*IF*)，具体计算方法参见公式(3-2)~(3-4)。

$$IR = \frac{\text{正确的、无重复的事例数}}{\text{列表问题要求给出的事例数}} \tag{3-2}$$

$$IP = \frac{\text{正确的、无重复的事例数}}{\text{系统返回的事例数}} \tag{3-3}$$

$$IF = \frac{2 \times IR \times IP}{(IR + IP)} \tag{3-4}$$

所有列表问题的 *IR*、*IP*、*IF* 值是各个列表问题 *IR*、*IP*、*IF* 值的算术平均值。

3.5.3 定义问题

对每一个定义问题，评测员会列出一个基本信息和可接受信息的表单。基本信息是指这一问题的答案中不可缺少的描述部分。可接受信息是指可以构成一个正确的答案的，但还不是必需的信息。超出基本信息和可接受信息的部分将在评分体系中给予扣分。EPCQA 用片断召回率(*NR*)、片断准确率(*NP*)和片段 F 值(*NF*)来评测一个描述提问的得分。具体参见公式(3-5)~(3-7)。

$$NR = \frac{\text{系统返回的基本信息个数}}{\text{全部基本信息个数}} \tag{3-5}$$

用允许长度(Allowance)和实际长度(Length)来定义 NP 如下：

$$NP = \begin{cases} 1 & if\ Length < Allowance \\ 1 - \dfrac{Length - Allowance}{Length} & else \end{cases} \tag{3-6}$$

Allowance ＝ 100 ×(返回的基本信息个数 + 返回的可接受信息个数)

Length ＝ 返回答案的全部长度

NF 是 *NR* 和 *NP* 的平均，公式如下：

$$NF = \frac{(1 + \beta^2) \times NP \times NR}{\beta^2 \times NP + NR} \tag{3-7}$$

同样，所有描述问题的 *NR*，*NP* 和 N*F* 是单个描述问题 *NR*，*NP* 和 *NF* 的算术平均值。

3.6 本章小结

构建汉语问答系统评测平台的出发点是想通过对真实环境的模拟，以系统化、大规模的评测为基础，推动问答技术研究向前发展，加速将实验室研究成果转化为产品，并发展适当且具应用性的评估技术。

目前的 EPCQA 还不成熟，无论是语料库的规模、测试集的规模、测试集的合理性与否，还是打分标准都有待在实践中逐步的改进和完善。下一步工作重点主要包括以下几个方面：

- 扩展现有提问类型

目前，事实提问，列表提问和定义提问是组成评测平台的三大类提问。下一步应该扩展其他类型的问题，例如交互式问答(Interactive Question Answering)，多视角问答(Multi-Perspective Question Answering，简称 MPQA)等。

- 构建更为合理的测试集

测试集的合理性主要体现在几个方面：

a. 测试集是否具有开放性
b. 测试集的提问方式能否反应用户实际使用时的情况
c. 测试指标能否有效、合理地比较各个问答系统的性能

- 构建更为合理的打分标准

目前的评分标准只是从问答系统返回的答案的角度进行打分。此外，如果还考虑问答系统返回答案的文档，打分会更合理。而对于其他类型的问题，如程序型提问、解释型提问、摘要型提问、比较型提问等等，应该有一个更客观的打分标准。

- 逐步扩大用户提问的广度和深度

3.7 本章研究成果

吴友政, 赵军, 段湘煜, 徐波. 构建汉语问答系统评测平台. 第一届全国信息检索与内容安全学术会议论文集, 2004.11, 上海, 中国, pp315~323.

第四章 汉语问答系统中的提问分类技术

4.1 引言

提问处理模块中的提问分类是整个问答系统的基石，它的任务是为每个用户提问指定一个或多个预先定义的类别。按照作者的理解，其作用主要体现在以下两个方面：

1. 确定用户提问答案的语义类型

确定提问答案的语义类型是为了限制候选答案的产生。例如提问 1：哪位医生第一次成功移植肝脏？提问分类确认该提问希望的答案类型是一位医生(PERSON)，这样候选答案提取模块产生的候选答案将大大减少，同时也提高了答案选择模块的效率和准确性。

2. 确定对用户提问应采取的问答技术

不同的提问类型，应该对症下药，采取不同的技术策略。没有一个统一的方法可以解决所有类型的提问。例如，提问书籍作者，某人出生年月日等类型的问题，简单有效的模板技术是最合适的。例如提问 2:《红楼梦》是谁写的？提问分类确认该提问希望的答案是书籍的作者(AUTHOR)，这样答案抽取就可以使用准确率较高的模式匹配技术。而对某人或者某物的描述类型提问，模板技术却表现较差，检索技术更有效。对于其他类型提问，需求助于基于自然语言处理的理解技术。

因此，提问分类模块是非常重要，它是整个问答系统重要的基础环节。提问分类的错误将不可避免导致对用户提问回答的失败。Moldovan [D. Moldovan, *et al*. 2003]的研究已经表明：大约 36.4% 的开放域问答系统的错误都是由于提问分类的错误造成的。

基于对提问分类模块作用的理解和现有提问句语料的分析，本文提出一种新的提问分类体系：提问的技术分类和提问的语义分类，以及在此基础上的基于多特征的支持向量机(SVM)提问分类算法。本论文通过实验分析，比较了不同类型特征以及权重计算方法对 SVM 分类器的性能影响；Bi-Gram 结构特征与依存句法结构特征对分类器的贡献的对比分析，以及多类别输出的系统性能。

4.2 相关工作

大部分问答系统[U. Hermjakob, *et al*. 2001; E. Hovy, *et al*. 2002; H. Yang, *et al*. 2002; D. R. Radev, *et al*. 2002]均使用人工编写规则的方法实现对提问的分类，但这种方法费时费力，且不利于系统的移植和扩充。近年来，研究人员开始利用统计学习方法实现提问的分类，这些学习算法大致包括支持向量机算法[D. Zhang, *et al*. 2003; J. Brown. 2003; K. Hacioglu, *et al*. 2003; J. Suzuki, *et al*. 2003]，SnoW 算法[X.Li, *et al*. 2002; X.Li, *et al*. 2003]，统计语言模型[W. Li. 2002; J. Brown. 2003]，最大熵模型[K. Kocik. 2003]和改进贝叶斯模型[张宇等.2004]等等[T. Solorio,*et al*. 2004; 李鑫等. 2004]。下面介绍这些模型的代表性工作。

[X.Li, *et al*. 2002; X.Li, *et al*. 2003]提出了层级(二级)提问分类体系及在此基础上的 SnoW 分类算法。论文设计了包括 ABBREV.，ENTITY，DESCRIPTION，HUMAN，LOCATION，NUMERIC 等在内的 6 个大类，50 个小类别。SnoW 分类器使用词汇(词、词性、Chunk、head Chunk)和语义(命名实体、WordNet 词义和半自动建立的特定类同义词)两类特征，在 21,5000 个训练提问句和 1000 个测试提问句上，算法的大类最优划分性能为 92%，小类最优划分性能为 89.3%。

[D. Zhang, *et al*. 2003]使用同[X.Li, *et al*. 2002]相同的分类标准对提问分类算法进行研究，针对提问分类的特殊情况设计了基于句法树的 SVM 核函数-Tree Kernel。算法在 5500 个训练句和 1000 个测试句上获得的最优大类划分性能为 90%。Jun Suzuki[J. Suzuki, *et al*. 2003]在 68 个日语提问类别的基础上设计了基于 HDAG (Hierarchical Directed Acyclic Graph)核函数的 SVM 算法。算法在 5011 个日语训练、测试实例上的获得的最优性能为 88.2%。

[W. Li. 2002]则借用统计语言模型和正则表达式相结合的策略实现对用户提问的划分，算法对 PERSON，LOCATION，DATE，ORGANIZATION，NUMBER，OBJECT，REFERENCE 等 7 种提问类型的最优划分结果达到 85.4%。

[T. Solorio, *et al*. 2003]提出了一种和语言无关的提问分类思路：基于 Internet 特征的提问分类算法。目前，大部分提问分类特征的提取均使用了自然语言处理的工具，比如词性标注、Chunk 和命名实体识别等。但对于某些类语言，这些工具还没有开发出来，在这种情况下，如何提高提问分类的正确率。[T. Solorio, *et al*. 2003]提出了一种解决思路。本论文通过一个例子介绍这种算法的实现过程。例如提问 4：Who is the President of the France Republic? 算法首先组合提取的限定数目关键词和可能的提问类别，如提问 4 的查询关键词是{President France Republic}，可能的提问类别是 s_i，$s_i \in${ORGANIZATION/机构名，PERSON/人名，PLACE/地名，DATE/日期，MEASURE/度量}；再将查询关键词和可能的提问类

别组合后，如(President France Republic is a s_i)，提交给 Internet 搜索引擎；接下来计算搜索引擎返回的结果数并进行归一化计算；最后，归一化得分同提问关键词一道作为 SVM 分类器的特征。令人遗憾的是，引入 Internet 特征可以提高提问分类正确率约 3%左右，但系统的整体性能仍然不很理想。

[崔桓等. 2001]对汉语提问句的分类进行了较全面的研究，并把汉语提问句分为两类：一类是疑问词和答案类型直接映射；另一类是疑问词和答案类型不能直接映射。第一种情况容易确定提问类型。对于第二类情况，算法选择最靠近疑问词的名词或者名词短语作为提问焦点词，并依据 Hownet 对焦点词的解释决定提问的类型。该算法的缺点是：对于某些类型的提问，很难获取焦点词，而且获取的焦点词往往也是多义词，不能确定其语义属性。

[张宇等. 2004]提出改进贝叶斯算法划分提问的语义类型。

4.3 汉语提问分类体系

设计问答系统提问分类模块的第一步是确立提问分类体系。受到问答技术和相关资源工具的影响，不同的问答系统定义的提问类型差别较大。例如，[E. Hovy, *et al*. 2002]定义了 141 个提问类型；[X.Li, *et al*. 2002; X.Li, *et al*. 2003]定义了 6 大类，50 个小类的层级提问分类体系；[H. Yang, *et al*. 2002]定义的提问分类体系是 6 个大类，54 个小类。随着问答技术及相应资源工具的不断完善，这些分类体系也在不断的改进和细化。在建立汉语问答系统提问分类体系时，论文遵守下述 3 个原则：

1. 提问类型用来确定后续候选答案提取模块和答案选择模块应该采用的技术类型，这是提问的技术分类

因为没有一个统一的方法可以解决所有类型的提问，所以，必须对不同类型的提问采取不同的技术策略。例如，对于出生日期、简称、定义等类型的问题，模板技术具有简单有效的特质[D. Ravichandran, *et al*. 2002; M. M. Soubbotin, *et al*. 2001; M. M. Soubbotin, *et al*. 2002]；而对某人或者某物的描述类型提问，检索技术更有效；对于其他类型提问，引入语义计算技术是获取高性能提问分类器的保证。

2. 提问类型用来确定提问答案的语义类型，限制候选答案的产生，这是提问的语义分类

提问答案的语义类型可以限制候选答案的产生，同时由于噪音的减少提高了答案选择模块的效率和准确性，因为从所有词或者词组中选择正确答案要比从某

一类语义实体中选择正确答案的难度更大。

3. 设计的提问分类体系在现有的资源和工具上是可实现的，且具有一定的普遍性和覆盖度

基于上述提问分类体系的建立原则和论文收集的 7050 个汉语提问句，本文定义如表 4-1，4-2 所示的汉语问答系统提问分类体系[20]。表 4-1 和 4-2 中的#表示某类型提问在语料中出现的次数。

表 4-1 技术类分类体系

提问类型	#	提问类型	#
ABBR/简称	*42*	*OLD-NAME/原名*	*41*
CH-ABBR/中文简称	*11*	*CAPITAL-PLACE/首都*	*48*
EN-ABBR/英文简称	*39*	*POPULATION/人口*	*52*
ABBR-EXPANSION/全称	*92*	*WHY-FAMOUS/身份*	*79*
CH-ABBR-EXPANSION/中文全称	*6*	*DEFINITION/定义*	*257*
EN-ABBR-EXPANSION/英文全称	*7*	*LIST/列表*	*234*
YES-NO-QUESTION/是非问题	*10*	*REASON/原因*	*129*
TRANS/翻译	*25*	*MANNER/方式*	*90*
TRANS-TO-ENGLISH/中到英	*19*	*CONTRAST/对比*	*41*
TRANS-TO-CHINESE/英到中	*22*	*FUNCTION/功能*	*39*
SYNONYM/昵称	*190*	*DESCRIPTION/描述*	*128*
BIRTHDAY/生日	*52*	*COMPONENTS/部件*	*51*
BIRTH-PLACE/出生地	*25*	*CAUSE-OF-DEATH/死因*	*41*
AUTHOR/作者	*71*	*OTHER-APPROACH/其他*	

表4-2 语义类分类体系

提问类型	#	提问类型	#
OTHER-TEMP/日期时间	*399*	*OTHER-ORG/机构名*	*148*
DURATION/段时间	*123*	*PARTY/党派*	*20*
SEASON/季节	*28*	*SPORTS-TEAM/运动队*	*35*
YEAR/年份	*206*	*UNIVERSITY/大学*	*37*
MONTH/月份	*37*	*MAGNEWS/报纸杂志*	*18*
DATE/日期	*90*	*BANK/银行*	*26*
TIME/时间	*16*	*OTHER-ENTITY/其他实体*	*641*
AGE/年龄	*48*	*DYNASTY/朝代*	*48*
PERSON/人名	*769*	*LANGUAGE/语言*	*56*

[20] http://www.nlpr.ia.ac.cn/english/cip/Yzwu/QcData/questiontypology.html

OTHER-PLACE/地名	*606*	*ANIMAL/动物*	*129*
CONTINENT/洲名	*73*	*PLANT/植物*	*58*
COUNTRY/国家名	*334*	*PRODUCT/产品*	*11*
PROVINCE/省份	*99*	*OCCUPATION/职业*	*19*
CITY/城市	*189*	*HUMAN-FOOD/食物*	*50*
BODY-OF-WATER/水域名	*142*	*BODY-PART/肢体*	*32*
ISLAND/岛屿	*35*	*DISEASE/疾病*	*26*
MOUNTAIN/山脉	*45*	*SPORT/体育项目*	*36*
SPHERE/星体	*72*	*COLOR/颜色*	*51*
NUMBER/数量	*426*	*UNIT/单位*	*27*
MONEY/货币	*67*	*NATIONALITY/民族*	*31*
SPATIAL-NUMBER/空间距离	*209*	*MONETARY-UNIT/货币单位*	*37*
SPEED/速度	*41*	*BOOK-NAME/书名*	*48*
WEIGHT/重量	*44*	*MOVIE-NAME/电影名*	*62*
ACCELERATION/加速度	*25*	*MUSIC-INSTRUMENT/乐器*	*35*
ORDINAL/序数	*34*	*PHONE-NUMBER/电话号码*	*37*
PERCENTAGE/百分数	*68*	*ZIP-CODE/邮编*	*33*
TEMPERATURE/温度	*56*	*EMAIL/电子邮箱*	*4*
RANGE- NUMBER/数量范围	*19*	*URL/超链接*	*13*

本文之所以确定这样一个分类体系，是由论文的问答技术路线决定的：基于字符表层的文本分析技术(模板技术)和快速、浅层自然语言处理技术必须紧密结合。例如提问 6：莫扎特是哪年出生的？该提问在技术类中属于 BIRTHDAY 类别，在语义类别中属于 YEAR 类别。因此，技术类别确定该提问可以使用高性能的模板技术，而语义类别又可以限制模板技术的机械性匹配。

和其他提问分类体系相比，本论文提出的汉语提问分类体系是由两个平行的类别组成，这不同于英文的层级分类体系，分别如图 4-1 和 4-2 所示。

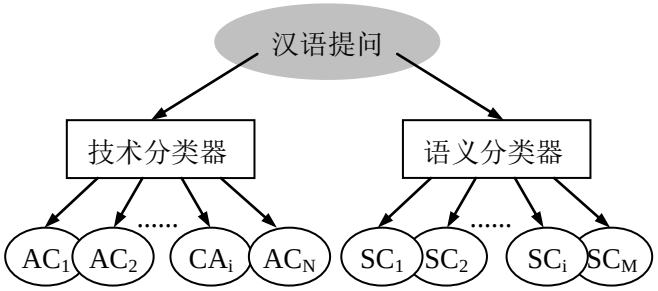

图4-1 汉语平行分类体系

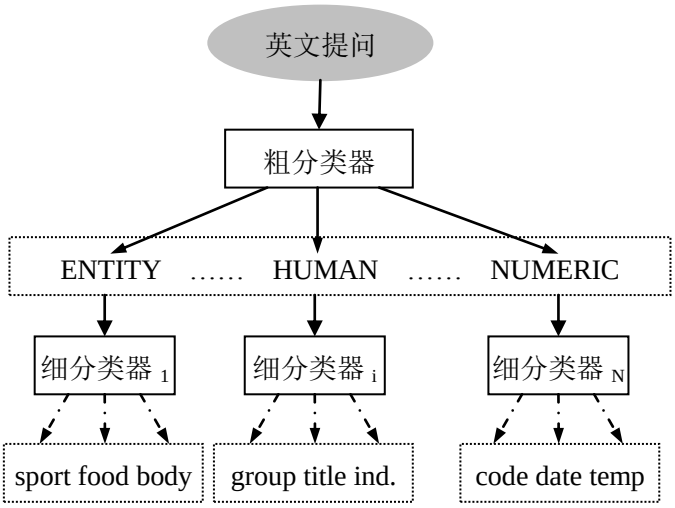

图4-2 英文层级分类体系

4.4 提问分类的特征向量构造

问答系统的提问分类类似于文本分类。不同的是文本分类的输入是一篇文章，类别是文章的话题，而提问分类的输入是一个自然语言的问句，类别是提问的技术类型和答案的语义类型。相对于一整篇文章，提问中包含的信息量更少，分类难度更大。

特征向量的构造是基于特征向量的机器学习算法的核心。选取恰当的特征不仅能够较好地描述提问，还有利于学习效果的提高。所谓恰当的特征即是指与分类相关的，具有较强区分度的特征。

对于提问分类，本文使用的特征主要包括词特征、词性特征、Bi-gram 或者依存结构特征、词汇语义特征等。

4.4.1 特征类型

[词特征] 提问中的单个关键词是最基本的特征，包括名词、动词、疑问词、形容词等。例如提问 7：“我国/n 降雨量/n 最/d 多/a 的/u 是/v 哪个/r 省份/n ？/w”的词特征向量 q0 = {我国, 降雨量, 最, 多, 的, 是, 哪个, 省份}。在文本分类中使用的一些停用词(如疑问词)对于提问分类也是非常重要的，某些提问类型可以直接通过疑问词确定。

[词性特征] 词性是另一种重要的基本特征，也是一种词法特征。提问 7 的词性特征向量 q1 = {n, n, d, a, u, v, r, n}。

[N-gram 特征] 在上述两类特征中，词和词性都是被孤立看待的，没有考虑

其所在上下文信息。引入 N-gram 特征可以避免了这个缺点，且 N-gram 可以被认为是一种最简单的句法结构特征。提问 7 的词和词形 Bi-gram 特征 q2 = {(我国-降雨量), (降雨量-最), (最-多), (多-的), (的-是), (是-哪个), (哪个-省份), (n-n), (n-d), (d-a), (a-u), (u-v), (v-r), (r-n)}。很显然，结构特征(哪个-省份)直接决定了提问的类型，比孤立特征“哪个”、“省份”更加合理。

[**依存句法特征**] 依存句法指出了句子中各个词语在句法上的搭配关系，且这种搭配关系不受距离的限制。为此，本文想通过挖掘句子的依存分析结果来弥补 N-gram 特征无能为力的远距离搭配。提问 8：“哪家/r 石油/n 公司/n 负责/v 中国/LOC 的/u 石油/n 进出口/vn 业务/n ？/w”的依存句法分析树如图 4-3 所示。

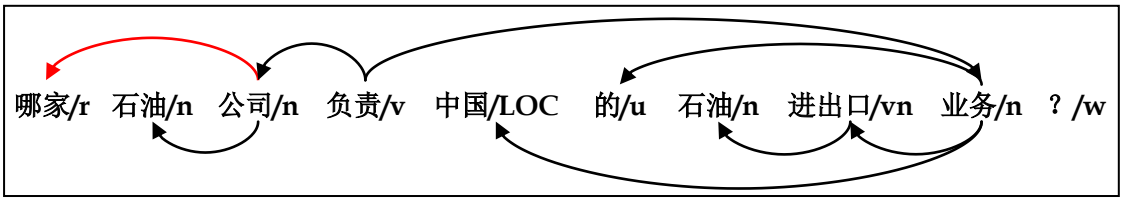

图4-3 提问8的依存句法树

理论上，从依存句法树中得到的二元结构特征向量 q3 = {(哪家-公司), (石油-公司), (公司-负责), (负责-业务), (的-业务), (进出口-业务), (石油-进出口), (中国-业务)}比用 Bi-gram 模型抽取的二元特征 q4 = {(哪家-石油), (石油-公司), (公司-负责), (负责-中国), (中国-的), (的-石油), (石油-进出口), (进出口-业务)}更合理。作者希望通过引入更加合理的依存关系来提高提问分类的正确性。但后续的实验结果却表明，依存关系对提问分类的帮助并没有期望的那么大。论文在 4.5.3 节将详细分析其中的原由。

[**词汇语义特征**] 无论是词、词性，N-gram 结构还是依存关系特征，数据稀疏是它们致命的缺点。引入词汇语义是解决数据稀疏的可行方法之一。提问 9：“央行/ORG 的/u 行长/n 叫/v 什么/r 名字/n ？/w”中的“行长”是决定提问类别的重要词汇，如果训练语料中没有出现过“行长”和类别 PERSON 的共现，无论设计何种算法，将其正确分类的可能性都是非常低的。因此，要获得高性能的提问分类算法，必须引入词汇语义信息。本文主要考虑命名实体和同义词集合这两类词汇语义特征。例如提问 9 的词汇语义特征 q5 = {ORG，的，FeatureNi，叫，什么，名字}，其中 FeatureNi 表示职业身份类同义词集合。

4.4.2 特征的选择和加权

在建立提问特征向量的时候，并不是选择的特征越多分类效果越好。相反，有些特征的加入不但增加了运算复杂度，而且还可能会使分类效果降低，这也就

是所谓的噪音特征。因此，无论从降低计算复杂度、提高效率的考虑，还是从提高分类正确率的角度考虑，都有必要进行特征的选择。特征选择中最重要的是如何量化不同特征对于分类的重要程度，并对量化后的特征进行排序，从而选择重要的特征进行分类。

文本分类中常用的特征选择方法包括文档频率(DF)，信息增益(IG)，互信息(MI)，CHI，领域词频(DWF)等。[Y. Yang, *et al*. 1997]用实验说明了 CHI_{max} 是最好的特征选择方法，故本论文也采用 CHI_{max} 进行特征选择。

不同的特征对于分类的贡献不相等，有的特征贡献较大，有的贡献很小，这即是特征的加权。TF*IDF 是最经典的特征向量权重算法[E. Greengrass. 2001]。但在大多数情况下，一个提问句中出现次数大于 1 的特征还是比较少见的，即 TF 值近似等于 1，TF*IDF 近似等于 IDF，那么有没有其他信息可以修正 TF*IDF 特征权重呢？

实际上，权重计算需要考虑的因素主要包括二种：特征的局部属性(MI、CHI、DWF)和特征的全局属性(IG、DF)。而 TF*IDF 在应用于提问分类时缺少就是特征的局部属性。所以，本文提出了使用 CHI 修正 TF*IDF 的特征向量权重算法，具体如式 4-1 所示。

$$t_{ik} = \sqrt[\lambda]{CHImax} \times TF \times IDF \tag{4-1}$$

其中 t_{ik} 是特征 i 在提问 k 中的权重，CHI_{max} 是反应特征与类别之间关系的局部属性，IDF 则是特征与整个训练语料之间关系的整体属性。

此外，在提问分类中，离疑问词越近的特征往往越重要，但传统的权值算法，并不能体现这个思想。所以，论文使用一种更合理的特征权值算法，即在传统的权值算法中引入距离的信息，如公式 4-2。

$$t'_{ik} = t_{ik} \times \lambda^{d} \tag{4-2}$$

公式 4-2 中 λ 是对距离的惩罚因子，d 表示当前词与疑问词的距离大小。

4.4.3 分类器的选择

过去数年的研究提出过多种分类器，像 SVM、KNN、LLSF(Linear Least Square Fit)、Multi- layered Perceptrons、Na ïve Bayse、Rocchio 等等。不同的研究在不同的环境下得到的实验数据多少有些差距。因此，哪一种分类器效果较好，就有了不太一致的结论。为此，[Y. M. Yang, *et al*. 1999]以 Reuters 文件的 90 个类别比较

了五种分类器的性能。其实验结果发现：{ SVM, KNN} > LLSF > Multilayered perceptrons >> Multinomial >> Naive Bayes。虽然 SVM 数据比 KNN 稍高一些，但是没有统计上的显著差异。而这两种分类器都比 LLSF 好，LLSF 又比 Multilayered Perceptrons 好，Naïve Bayes 是这五种里的分类效果较差的。Yang 只是根据 Reuters 文件得出这样的结果，如果换成其他文件和应用，这个顺序会不会一样？[T. Joachims. 1998]取 Reuters 的前十大类做实验，以 Micro-Average 计算成效，其实验结果同样发现：SVM (0.864)>KNN(0.823) > {Rocchio(0.799),C4.5 (0.794)} > Naïve Bayes (0.72)。同样的报告中，Joachims 对 OHSUMED 文件的 23 个类别做实验，结果为：SVM(0.66) > KNN (0.591) > Naïve Bayes (0.57) > Rocchio (0.566) > C4.5(0.5)。而且在所有的 23 个类别中 SVM 都比 KNN 表现得好，显示出统计上的显著性差异。

所以，本论文采用 SVM 算法实现对提问的分类，并使用由台湾大学 Chih-Jen Lin 教授开发的 SVM 开放源码工具 LIBSVM2.6[21]。

4.4.3.1 支持向量机简介

支持向量机(Support Vector Machine, 简称 SVM)起源于统计学习理论，它研究如何构造学习机，实现模式分类问题[C. J. C. Burges. 1998]。支持向量机使用结构风险最小化(Structural Risk Minimization，SRM 准则)原理构造决策超平面使每一类数据之间的分类间隔(Margin)最大。SRM 准则认为：学习机对未知数据分类所产生的实际风险是由两部分组成的，以 $0<\eta<1$ 满足如下关系：

$$R \le R_{emp} + \sqrt{\frac{h(\log(2n/h)+1) \quad \log(\eta/4)}{n}} \tag{4-3}$$

其中，R 是实际风险，不等式的右边叫做风险边界，R_{emp} 称为经验风险，$\sqrt{\frac{h(\log(2n/h)+1) \quad \log(\eta/4)}{n}}$ 叫做“VC 置信值”，n 是训练样本个数，h 是学习机的 VC 维(h 反映了学习机的复杂程度)。

SVM 的思想就是在样本数目适宜的前提下，选取比较好的 VC 维 h，使经验风险 R_{emp} 和置信值达到一个折中，使每一类别数据之间的分类间隔(Margin)最大，

[21] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

最终使实际风险 R 变小。

对于线性不可分数据，支持向量机将低维空间不可分数据映射到高维空间中。具体说，支持向量机通过核函数把数据由低维空间向高维空间映射，在高维空间为低维数据构造线性可分超平面。这样，在低维不可分情况下，对于每个要判别的未知样本 u，计算高维空间中的最佳超平面分类函数 f 取值：

$$f(u)=\sum_{i=1}^{S}a_i y_i K(u,x_i)+b \tag{4-4}$$

其中，a_i 是支持向量 x_i 对应的拉格朗日乘子，y_i 是支持向量 x_i 的类别标注(为 1 或者-1)，S 是支持向量总数，K 是核函数，$f(x)$ 值的正负分别代表不同的类别。

4.5 提问分类实验

基于前文所述，论文还需要用实验来验证下面几个问题。

1. 求取最佳的特征向量权值算法及其参数，即公式 4-1 的 λ 值；
2. 比较研究不同类型的特征对分类器性能的贡献；
3. Bi-gram 特征与依存句法特征的性能对比研究；
4. 验证多类别输出的系统性能。

本论文的实验是在 7050[吴友政等. 2004]个已经标注技术类别和语义类别的汉语提问句中进行的。随机提取其中 700 个提问作为测试数据，其余 6350 个提问为训练数据。实验使用分类准确率(如式 4-5，4-6)对提问的技术分类器和语义分类器分别进行开放和封闭测试。

$$I_i=\begin{cases}1 & \text{如果第}i\text{个提问的输出类别是正确的}\\ 0 & \text{其他情况}\end{cases} \tag{4-5}$$

$$P=\sum_{i=1}^{N}I_i/N \tag{4-6}$$

式中的 N 表示待测试的提问数。

4.5.1 最佳的特征向量权值算法及其参数

在 4.4.2 节提出的特征向量权重计算公式(式 4-1)中，λ 代表特征全局属性和局部属性的加权因子。λ 值越大，特征局部属性对于特征向量的贡献越大；反之，全局属性对特征向量的贡献越大。本实验目的即是找出最优的特征加权因子 λ*，以获取最优的分类性能。

实验使用的特征包括基本特征(词、词性)、结构特征(Bi-gram)和词汇语义特征(命名实体和同义词集合)。其中，词、词性和命名实体特征的提取是由汉语分词、词性标注、命名实体识别一体化工具 NlprCSegTagNer[Y.Z. Wu, *et al*. 2003; Y.Z. Wu, *et al*. 2005]完成的。同义词集合是在扩展版《同义词词林》基础上由人工抽取。实验的结果如图 4-1。

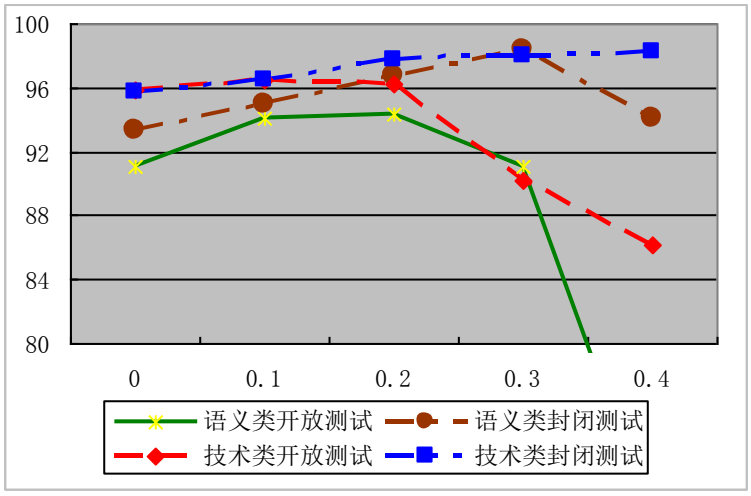

图4-1特征的局部属性和全局属性对分类性能的影响曲线图

分析图 4-1 不难发现，公式 4-1 引入特征局部属性 CHI_{max} 修正 TF*IDF 在应用于提问分类时存在问题的思路是正确的。在开始阶段，随着 λ 的增大，开放和封闭测试的分类准确率均呈上升趋势，但到达某点之后，分类器的性能均在下降，且当 λ=0.2 时，整个系统的分类性能基本达到了最优。这个实验也充分说明了只有综合考虑特征的全局属性和局部属性才能获得最优的分类准确率。

4.5.2 不同特征对分类器的贡献

4.5.1 节的实验选择了四类特征进行实验，获得最优的特征加权因子 λ*。本实验的目的是比较研究这四类不同特征对分类器的贡献。

实验在 λ= 0.2 的情况下进行的。采用不同特征集合，技术类分类器和语义类

分类器的分类性能如图 4-2 所示。

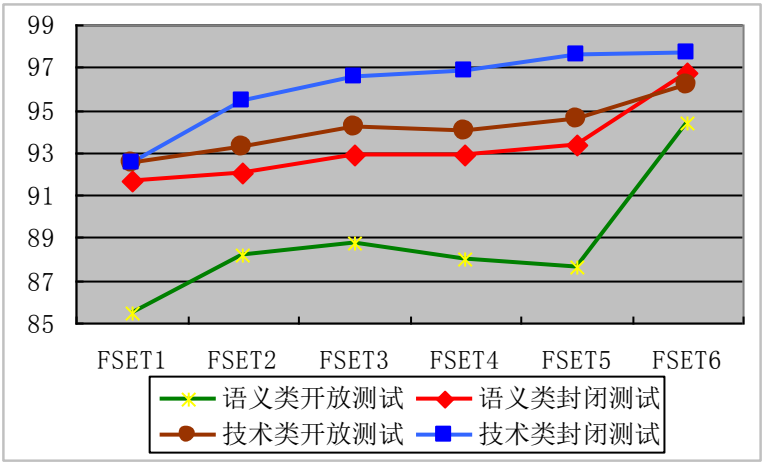

图4-2 不同特征对分类器的贡献

实验中的 *FSET1~FSET6* 代表的特征集合分别是：*FSET1*：词特征；*FSET2*：词特征+词 Bi-gram 特征；*FSET3*：词特征+词 Bi-gram 特征+命名实体特征；*FSET4*：词特征+词 Bi-gram 特征+命名实体特征+词性特征；*FSET5*：词特征+词 Bi-gram 特征+命名实体特征+词性特征+词性 Bi-gram 特征；*FSET6*：词特征+词 Bi-gram 特征+命名实体特征+词性特征+词性 Bi-gram 特征+同义词特征。

图 4-2 充分说明语义特征和结构特征对分类的重要性。提问的语义分类器在引入命名实体特征和同义词特征后分别提高了约 0.6%和 7%。同时，Bi-gram 特征无论是对提问的技术分类器还是提问的语义分类器都有很大的帮助，这和论文在 4.4.1 节的分析是一致，因为结构特征部分体现了隐含在提问中的语义结构信息。但词性对分类器的贡献非常小的，甚至会给语义分类器带来一定的噪声，这是由于有用的词性信息已经包括在命名实体中了，剩余的词性(如名词、动词等)对分类基本没有帮助。

4.5.3 Bi-gram 特征与依存句法特征的性能对比研究

在 4.4.1 节中论文分析了依存句法关系特征可能要优于 Bi-gram 特征，因为依存句法分析可以反映出句子中各成分之间的语义修饰关系，它可以获得长距离的搭配，跟句子成分的物理位置无关[冯志伟著. 2004]。

设计本实验的目的即是为了检验依存句法特征会给分类器带来什么样的影响。表 4-3 给出了依存句法特征和 Bi-gram 特征的性能比较。

标 4-3 Bi-gram 特征与依存句法特征的性能对比研究

	特征集合	开放测试	**封闭测试**
语义分类器	**FSET6'**	**90.5479%**	96.1538%
	FSET6	**94.3681%**	96.773%
	FSET6"	**93.2692%**	96.5666%
技术分类器	**FSET6"**	**92.3967%**	94.5832%
	FSET6	**96.1983%**	97.7133%
	FSET6'	94.0496%	97.0808%

实验中的各特征集合分别表示如下：*FSET6'*：词特征+词性特征+命名实体特征+同义词特征；*FSET6"*：词特征+词性特征+命名实体特征+同义词特征+依存结构特征。

分析表 4-3 可以发现，加入依存结构信息虽然能够较大幅度提高系统的分类准确率(约 2.5%)，但它的表现还不如 Bi-gram 对分类器的贡献(约 4%)。这一点和 4.4.1 节的分析是相违背的。通过对错误的分析，作者发现导致这种情况的原因存在于两个方面：

第一，依存句法分析器的性能目前只有 80%，因而出错的依存关系会给分类器带来一定的噪音。

第二，虽然通过 Bi-gram 提取的特征不如通过依存分析提取的特征合理，但这种"不太合理"的 Bi-gram 特征不会必然导致分类的错误。例如上述提问 8，无论是用 Bi-gram 提取的结构特征还是通过依存关系提取的结构特征都能正确分类，因为提问 8 中对分类起决定作用的是词特征 "哪家"和"公司"，通过依存分析提取的特征"哪家-公司"对分类并不起决定作用。

4.5.4 验证多类别输出的系统性能

对于某些问题，很难确定它们的类别，把它们划分到任何一类都有可能存在问题。例如提问 10："中国国家博物馆/LOC 在/p 哪里/r ？/w"的答案类型既可能 CITY (北京市)，也可能是 OTHER-PLACE(北京天安门广场东侧)。本实验允许分类器对每个提问输出概率最大的两个类别。论文采用同[X.Li, *et al*. 2002; X.Li, *et al*. 2003]的评测指标。

假设 k_{ij} 是提问 i 的第 j 个输出类别，并且$\{k_{ij}\}$按照概率大小降序排列。对于俩类别输出的分类器，无论正确的类别排列在第一还是第二的位置，都算该提问

分类正确。具体公式参加(4-7)和(4-8)。

$$I_{i2}=\begin{cases}1 & \text{如果第}i\text{个提问的第}1\text{个或者第}2\text{个输出类别是正确的}\\ 0 & \text{其他情况}\end{cases} \tag{4-7}$$

$$P_{\le 2}=\sum_{i=1}^{N} I_{i2}/N \tag{4-8}$$

式中 N 表示待测试的提问数。

基于 *FSET6* 的 $P_{\le 2}$ 实验结果如表 4-4 所示。

表4-4 两个类别输出的系统性能

	特征集合	开放测试	**封闭测试**
语义分类器	**FSET6**	**97.5275%**	99.0643%
技术分类器	**FSET6**	**98.5124%**	99.6594%

观察表 4-4 可以发现，相对于单类别分类器，多类别分类器的性能得到了较大幅度的提高。

4.6 错误分析

提问分类看似一个简单的课题，里面却也存在很多的技术难点，要完全解决问答系统的提问分类，必须真正地实现对提问句的理解。下面举几个例子来说明提问分类的难点所在。

■ 类别模糊的提问

某些提问的类别很难确定，要准确确定其类别可能需要了解用户的提问意图或者专家知识。例如提问 10：“吃/v 钉螺/n 容易/ad 得/v 什么/r 病/n ？/w”的类别是 DISEASE(血吸虫病)，而用户提问 11：“风湿性/n 关节炎/n 是/v 一/NUM 种/q 什么/r 病/n ？/w”的类别则不是 DISEASE，而是 DEFINITION 类别。因为“风湿性关节炎”是疾病的一种，提问可能是要求系统给出对该疾病的定义或者介绍。同样的问题还存在于类似于提问 12：“洛克菲勒/PER 的/u 经济/n 来源/n 是/v 什么/r ？/w”(该提问的类别是 DESCRIPTION 还是 OTHER-ENTITY，很难确定)这样类型的提问。

■ 分词、词性标注和命名实体识别错误引起分类错误

分词、词性标注和命名实体识别的错误会给分类器带来负面影响，例如提问13：“张柏芝为什么产品代言？”和提问 14：“哪国产橡胶最多？”，它们的分词、标注结果分别为“张柏芝/PER 为什么/r 产品/n 代/v 言/vg ？/w”和“哪/r 国产/b 橡胶/n 最/d 多/a ？/w”。就因为分词和标注的错误导致分类器错误地把这两个提问分到 REASON 和 OTHER-ENTITY 类别中。

■ 错误地确定提问焦点词带来的分类错误

语义分类器的另外一种错误是由于错误地决定了提问焦点词造成的。例如提问 15：“全世界/n 拥有/v 最/d 多/a 的/u 信徒/n 是/v 什么/r 教/n ？/w”的提问焦点词应该是“教”而不是“信徒”，所以，该提问的语义类型是 OTHER-ENTITY 而不是 PERSON。而提问 16：“那个/r 脾气/n 古怪/a 的/u 隐/v 士/Ng 叫/v 什么/r ？/w”错误分为类型 OTHER-ENTITY 是因为同义词集合中没有把“士”和 PERSON 关联起来。

■ 仅仅通过关键词和一些简单的结构、语义信息很难准确分类

现在的分类方法基本都是基于孤立关键词和一些简单的结构信息、词汇语义信息，无法做到理解的程度。因此，也就无法做到对某些类提问的准确分类。例如，分类器判定用户提问“死海/LOC 在/p 什么/r 位置/n ？/w”和“郝海东/PER 在/p 联赛/n 中/f 踢/v 什么/r 位置/n ？/w”的主要特征是“什么/r 位置/n”，很显然，这个特征在两个提问中的意义却完全不同。所以，还必须引入其它特征。同样的问题还存在于如下的提问对中：

提问对 1：
OTHER-ORG 谁/r 是/v 联通公司/ORG 最/d 主要/b 的/u 竞争者/n ？/w
PERSON 谁/r 是/v 联通公司/ORG 最/d 主要/b 的/u 创立者/n ？/w

提问对 2：
BOOK-MOVIE 《/w 一千零一/NUM 夜/q 》/w 又/d 被/p 称为/v 什么/r ？/w
BODY-OF-WATER 黄河/LOC 又/d 被/p 称为/v 什么/r ？/w

4.7 本章小结

本文从新的角度提出一种新的提问分类体系：提问的技术分类和提问的语义分类。在此基础上，基于词、词性、Bi-gram、依存关系、命名实体和同义词等多种特征，实现了面向汉语提问分类的 SVM 分类器。

实验结果证实了引入语义特征(命名实体、同义词)和结构特征(N-Gram、依存

关系)是提高提问分类的必须途径。实验同时也表明，依存关系特征虽然能够提高分类器的性能，但目前的表现并不如想象中那样好，相反，它的表现还不如Bi-gram。

本文的提问分类体系仅是在第三章收集的 7050 个汉语提问句的基础上建立起来的。因此，该分类体系还不完善，离实用还有一段距离，需要在今后不断地修正、补充。

对于本文提出的基于 SVM 的分类算法，也有需要改进的地方。例如，通过 4.6 节的分析发现，焦点词对提问语义类别的判断具有非常重要的影响，但目前的算法没有单独对焦点词进行分析。所以，今后可以研究提问焦点词的提取方法，并把提取出来的焦点词作为支持向量机的一维特征，以提高语义分类器的性能。

4.8 本章研究成果

Youzheng Wu, Jun Zhao, Bo Xu. Chinese Question Classification from Approach and Semantic Views. In Proceedings of the 2nd Asia Information Retrieval Symposium (AIRS2005), October 13-15, 2005, Korea, pp485~490. (SCI)

第五章 基于多特征的汉语命名实体识别[22]

5.1 引言

命名实体识别是信息提取、句法分析、机器翻译、面向语义网(Semantic Web)的元数据标注等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位，它是目前应用最为广泛的 NLP 工具之一。在问答系统中，命名实体主要用于段落或句子排序以及支撑候选答案的抽取。可以说，命名实体识别的性能直接决定着问答系统的整体性能。

一般来说，命名实体识别的任务就是识别出待处理文本中三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)命名实体。

命名实体识别的过程通常包括两部分：(1)实体边界识别；(2)确定实体类别(人名、地名、机构名或其他)。英语中的命名实体具有比较明显的形式标志(即实体中的每个词的第一个字母要大写)，所以实体边界识别相对容易，任务的重点是确定实体的类别。和英语相比，汉语命名实体识别任务更加复杂，而且相对于实体类别标注子任务，实体边界的识别更加困难。汉语命名实体识别的难点主要存在于：

- 汉语文本没有类似英语文本中空格之类的显式标示词的边界标示符，命名实体识别的第一步就是确定词的边界，即分词。而汉语分词又和命名实体识别互相影响。
- 除了英语中定义的实体，外国人名译名和地名译名是存在于汉语中的两类特殊实体类型。
- 现代汉语文本，尤其是网络汉语文本，常出现中英文交替使用，这时汉语命名实体识别的任务还包括识别其中的英文命名实体。
- 不同的命名实体具有不同的内部特征，不可能用一个统一的模型来刻画所有的实体内部特征。

相对于实体词，时间词和数量词的识别比较容易，通用的方法是有限状态自动机的方法。所以，本文主要针对实体名的识别进行专项研究，并对人名和地名

[22] 本章的工作是和富士通研究开发中心有限公司合作完成的。

进行了细分类。其中人名细分为中国人名(CPN)、日本人名(JPN)、苏俄人名(RPN)、欧美人名(EPN)和简称人名(APN)；地名细分为单字地名(ALN)和多字地名(LN)；机构名细分为简称机构名(AON)和全称机构名(ON)。

针对汉语命名实体识别的难点，本论文提出了基于多特征相融合的汉语命名实体识别模型。该模型是由词形上下文模型、词性上下文模型、词形实体模型、词性实体模型等 4 个子模型组成的。其中，词形上下文模型估计的是给定词形上下文语境中产生实体的概率；词性上下文模型估计的是给定词性上下文语境中产生实体的概率；词形实体模型估计的是给定实体类型的情况下词形串作为实体的概率；词性实体模型估计的是给定实体类型的情况下词性串作为实体的概率。

本文模型主要特点综述如下：

- 为克服数据稀疏问题，模型强调大颗粒度特征(词性特征)和小颗粒度特征(词形[23]特征)相结合
- 为限制候选实体的产生，减小 Viterbi 搜索空间，模型强调统计模型和专家知识相结合
- 为精确刻画不同模型的内部特征，模型强调使用不同细分类实体模型刻画不同实体的内在规律

用 1998 年 2 月份~6 月份人民日报标注语料进行模型训练[24]，1998 年 1 月份人民日报语料进行测试，本文提出的模型对人名、地名和机构名的识别性能(精确率，召回率)分别达到了(94.06%，95.21%)、(93.98%%，93.48%)和(84.69%，86.86%)。

在 2004 年国家 863 评测中，该算法在人名、地名和机构名的识别中获得了令人满意的成绩。

5.2 相关工作

5.2.1 系统评测和系统性能

MUC[N. A. Chinchor. 1998]是对命名实体识别技术影响最大的国际评测会议。MUC 是由美国政府支持的一个专门致力于真实新闻文本理解的例会，在国际上对信息提取领域进展的评测方面起着主导作用。第 6 次和第 7 次 MUC 设立

[23] 5.3 节给出词形的具体定义。

[24] 本章采用的训练语料是由北京大学和富士通研究开发中心有限公司共同开发，由富士通研究开发中心有限公司提供。

的命名实体识别专项评测大大推动了英语命名实体识别技术的发展，许多系统都已经具备了相当程度的大规模真实文本的处理能力。其中由 Language Technology Group Summary 开发的英语命名实体识别算法在 MUC-7 评测中取得第一名，其准确率和召回率分别达到 95%和 92%。这和人对命名实体的识别能力已经非常接近(约 97.0%)。

此外，对命名实体识别进行评测的国际会议还有 CoNLL(Conference on Computational Natural Language Learning)[CoNLL] 、 IEER(Information Extraction-Entity Recognition Evaluation)[IEER] 和 ACE(Automatic Content Extraction)[ACE]。

然而，和英文相比，汉语命名实体识别研究目前还不算成熟，其中还有很多问题没有得到很好的解决。参加 MET-2 (Multilingual Entity Task)评测的汉语命名实体识别系统对人名、地名、机构名识别的最优性能指标(准确率，召回率)也只有(66%，92%)、(89%，91%)和(89%，88%)。因此，有必要对汉语命名实体识别继续开展系统的研究。

5.2.2 代表方法

很多学者对命名实体识别的研究做了大量的研究工作，这些研究大致可以分为以下两种情况：

1. 人工组织规则的方法

人工组织规则方法[R. Grishman, *et al*. 1995; G. R. Krupka, *et al*. 1998; W. J. Black, *et al*. 1998]代价昂贵，系统性能的好坏主要依赖于有经验的语言学家。当系统移植到新的领域或者语种时，规则需要大量的人工修改，甚至需要重新总结和组织。

2. 机器学习的方法

所以，近些年命名实体识别的方法逐渐从初期的人工组织规则的方法向机器学习方法转变。这类的研究工作主要集中在：隐马尔柯夫模型[D. M. Bikel, *et al*. 1997; J. Sun, *et al*. 2002]、最大熵模型[A. Mikheev, 1998; A. Borthwick. 1999]、错误驱动的学习方法[J. Aberdeen, *et al*. 1995]、决策树方法[S. Sekine, *et al*. 1998]、DL-CoTrain 和 CoBoost[M. Collins, *et al*. 1999; M. Collins, *et al*. 2002]等等。

5.2.3 汉语命名实体识别的代表系统

同样地，汉语命名实体识别的解决方案也可以分为个别解决方案和一体化解

决方案两种。

[**个别解决方案**] 所谓个别解决方案，就是针对某一个类型的专有名词，专门设置一对一的解决方法[H. H. Chen, *et al*. 1998; 孙茂松等. 1994; 郑家恒等. 2000]。

[**一体解决方案**] 所谓一体解决方案则是适用于所有类型的命名实体的解决方案[S. H. Yu, *et al*. 1998; J. Sun, *et al*. 2002; H. P. Zhang, *et al*. 2003; T. S. Chua, *et al*. 2002]。

NTU 系统[H. H. Chen, *et al*. 1998]是典型的个别解决方案，使用的是规则和统计相结合的方法。对于中国人名和外国人名使用局部统计特征进行识别，而地名和机构名的识别采用规则方法。

[J. Sun, *et al*, 2002; H. P. Zhang, *et al*. 2003; S. H. Yu *et al*. 1998]提出的算法都可以归结为隐马尔柯夫模型一类，状态转移对应算法中提出的上下文模型，观察对应实体模型，差别在于他们对上下文特征和实体特征的利用方式和利用程度不同。

Sun[J. Sun, *et al*. 2002]提出基于类的语言模型，其定义的语境模型是词类之间的三元语言模型，训练语料是 1998 年 1 月到 6 月的人民日报标注语料(训练语料使用微软 NLPWin 工具自动标注)，实体模型是从训练语料中提取的实体集合训练得到的三元模型。

Zhang[H. P. Zhang, *et al*. 2003]在 2003 年提出角色标注模型实现对汉语命名实体的识别。虽然论文中定义的角色比[D. M. Bikel, *et al*. 1997]多，但使用这种大颗粒度知识不可能使系统获得优良的性能，论文中的实验也说明了这一点。在其改进模型中引入词形语言模型(Bi-gram)的约束，系统的性能才有较大幅度提高。

Yu[S. H. Yu *et al*. 1998]则是从 500,000 词的分词标注语料中学习词性上下文模型 $P(t_k \mid t_k\, t_{k-1})$；人名、地名和机构名实体模型 $P(w_k \mid t_k\, t_{k-1})$分别从 67,616，6,451，6190 个实体列表中学习。

实际上，无论采用何种方法，都在试图充分发现和利用实体所在的上下文特征和实体的内部特征。有的系统使用的特征颗粒度较大(词性和角色级特征)，有的系统使用的特征颗粒度较小(词形特征)。作者认为：大颗粒度特征和小颗粒度特征有互相补充的作用，应该兼顾使用。目前的大部分系统都是只用其一没有做到两者兼顾，尤其是忽略了词性对命名实体影响。在 BaseNP 和 Chunk(命名实体是一种特殊的 BaseNP 和 Chunk)的识别过程中，词性却起了非常重要的作用[E. D.

Xun, *et al*. 2000]。受到 BaseNP 和 Chunk 算法的启发，本论文提出了大颗粒度特征(词性特征)和小颗粒度特征(词形特征)相结合的混合统计模型。

混合模型虽然可以部分地解决词形模型和词性模型的缺点，但数据稀疏仍然存在，且搜索空间的增大降低了算法的效率和系统的性能。对此，本文提出统计模型和专家知识相结合的模型。

同样，为了准确刻画不同实体的内部特征，本文提出使用多个细分类的实体模型来解决这个问题。特别地，把人名分类中国人名、日本人名、苏俄人名、欧美人名和简称人名；地名分为单字地名和多字地名；机构名分为简称机构名和全称机构名等。

5.3 基于多特征的汉语命名实体识别模型

5.3.1 基本思想

设计命名实体识别算法，要考虑的情况无外乎分两种：(1)如何利用实体所在的上下文特征和实体内部特征；(2)使用何种颗粒度大小的特征。

汉语命名实体的内部构成具有非常显著的词形特征[孙茂松等. 1993; 孙茂松等. 1994]，如人名姓氏用字相对集中；地名通常以“去”，“在”等词开头或/和以“县”，“街”，“开发区”等词结尾；机构名通常以“部”，“公司”等词结尾。但词形特征的颗粒度太小，数据稀疏问题严重削弱了对词形特征的利用。所以，应该引入颗粒度较大的特征到命名实体识别算法中，以弥补词形特征的数据稀疏问题。能够方便获取和利用的大颗粒度特征是词性特征。因此，本文提出的是词形特征和词性特征混合的统计模型。

[词形定义] 本文定义的词形包括以下几种情况：字典中任何一个字或词单独构成一类；人名(PER)、人名简称(APER)、地名(LOC)、地名简称(ALOC)、机构名(ORG)、时间词(TIM)和数量词(NUM)各定义为一类。

词形定义如表 5-1 所示。

表 5-1 词形的定义

标记	***标记的描述***
PER	人名
APER	人名简称

LOC	地名
ALOC	地名简称
ORG	机构名
TIM	时间词
NUM	数量词
OTHER	字典中每个词单独成一类

词形语言模型中共定义了$|V|+7$ 个词形，$|V|$表示词典的规模。由词形构成的序列称为词形序列 *WC*。

[词性定义] 词性采用北大汉语文本词性标注标记集(http://icl.pku.edu.cn/nlp-tools/catetkset.html)，另加上人名简称词性和地名简称词性，共 47 个词性。由词性构成的序列称为词性序列 *TC*。

命名实体识别可以看作是一个序列化数据的标注问题。输入是带有词性标记的词序列，如公式 5-1。

$$WT = w_1 / t_1 \quad w_2 / t_2 \quad \cdots \quad w_i / t_i \quad \cdots \quad w_n / t_n \tag{5-1}$$

式中 n 表示句子中词的数量，t_i 是词 w_i 的词性。

在分词和标注基础上进行命名实体识别的过程就是对部分词进行拆分、组合(确定实体边界)和分类(确定实体类别)的过程，最后输出一个最优的“词形/词性”序列 *WC*/ TC**，可以用公式 5-2 表示。

$$WC^* / TC^* = wc_1 / tc_1 \quad wc_2 / tc_2 \quad \cdots \quad wc_i / tc_i \quad \cdots \quad wc_m / tc_m \tag{5-2}$$

式中 $m <= n$，$wc_i = [w_j \quad w_{j+k}]$，$tc_i = [t_j \quad t_{j+k}]$，，$1<=k$，$j+k<=n$。

从公式 5-1 中计算最优“词形/词性”序列 *WC**/TC*，有三种的方法。词形特征模型、词性特征模型和混合模型。

[**词形模型**] 词形特征模型根据词形序列 *W* 产生候选命名实体，用 Viterbi 确定最优词形序列 *WC**。

目前的大部分系统[J. Sun, *et al*. 2002; H. P. Zhang, *et al*. 2003]都是从这个层面来设计命名实体识别算法。

[**词性模型**] 词性特征模型根据词性序列 *T* 产生候选命名实体，用 Viterbi 确

定最优词性序列 *TC**。

目前只有较少的系统[G. D. Zhou, *et al*. 2003; S. H. Yu *et al*. 1998]在命名实体的识别过程中引入词性的知识。

[**混合模型**] 词形和词性混合模型根据词形序列 *W* 和词性序列 *T* 产生候选命名实体，用 Viterbi 寻优，一体化确定最优序列 *WC*/TC**。

这是本论文提出的算法。

5.3.2 基于多特征的汉语命名实体识别模型

为了描述方便，论文把公式(5-1)分开描述成一个词序列(5-3)和一个词性序列(5-4)：

$$W = w_1 \quad w_2 \quad \cdots \quad w_i \quad \cdots \quad w_n \tag{5-3}$$

$$T = t_1 \quad t_2 \quad \cdots \quad t_i \quad \cdots \quad t_n \tag{5-4}$$

从词形层面(5-3)进行命名实体识别的词形特征模型可以用公式(5-5)描述：

$$WC^* = arg\,max_{WC}\, P(WC) \times P(W \mid WC) \tag{5-5}$$

从词性层面(5-4)进行命名实体识别的词性特征模型可以用公式(5-6)描述：

$$TC^* = arg\,max_{TC}\, P(TC) \times P(T \mid TC) \tag{5-6}$$

本文提出的词形和词性混合的汉语命名实体识别模型结合了词形特征模型和词性特征模型的优点，可以描述成公式(5-7)的形式：

$$\begin{aligned}
&(WC^*, TC^*) \\
&= \arg\max_{(WC,TC)} P(WC, TC \mid W, T) \\
&= \arg\max_{(WC,TC)} \frac{P(WC, TC, W, T)}{P(W, T)} \\
&\approx \arg\max_{(WC,TC)} P(WC, W) \times [P(TC, T)]^{\beta} \\
&\approx \arg\max_{(WC,TC)} P(WC) \times P(W \mid WC) \times \left[P(TC) \times P(T \mid TC)\right]^{\beta}
\end{aligned} \tag{5-7}$$

式中 β 是平衡因子，平衡词形特征和词性特征的权重，β>0。

模型(5-7)有四部分组成，论文分别把他们称为：词形上下文模型 *P(WC)*、词

性上下文模型 *P(TC)*、实体词形模型 *P(W|WC)*、实体词性模型 *P(T|TC)*。其中词形上下文模型和词性上下文模型合称上下文模型；实体词形模型和实体词性模型合称实体模型。

5.3.3 词形和词性上下文模型

上下文模型估计的是给定上下文语境中产生实体的概率。本文的词形上下文模型和词性上下文模型均用三元模型近似描述，见公式(5-8)和(5-9)。

$$P(WC) \approx \prod_{i=1}^{m} P\left(wc_i \mid wc_{i-2}\ wc_{i-1}\right) \tag{5-8}$$

$$P(TC) \approx \prod_{i=1}^{m} P\left(tc_i \mid tc_{i-2}\ tc_{i-1}\right) \tag{5-9}$$

5.3.4 实体模型

考虑到每一类命名实体都具有不同的内部特征，因此，不能用一个统一的模型刻画人名、地名和机构名等实体模型。例如，人名识别采用的是基于字的三元模型，地名和机构名更适合用基于词的三元模型等。此外，为提高外国人名的识别性能，本文还把外国人名分为日本人名、欧美人名和苏俄人名三个子类。因为这三类国家人名的内部特征(主要是人名用字集)存在较大的差别。日本人名用字相对较广，具有相对明显的姓氏特征，但姓氏集合却很大(现版本共收集日本人名姓氏 9189)；日本人名姓氏很多和地名重叠；苏俄人名常用斯、基、娃等汉字；而欧美人名常用朗、鲁、伦、曼等汉字。

为计算需要，按照字或词在命名实体内部的位置本文把他们划分成如表 5-2 所示的 19 个子类。

表 5-2 实体模型中的子类定义

标记	***标记的描述***
Sur	中国人名的姓氏
Dgb	中国人名名字首字
Dge	中国人名名字尾字
EBfn	欧美人名首字
EMfn	欧美人名中间字

EEfn	欧美人名尾字
RBfn	苏俄人名首字
RMfn	苏俄人名中间字
REfn	苏俄人名尾字
JBfn	日本人名首字
JMfn	日本人名中间字
JEfn	日本人名尾字
Bol	地名首字
Mol	地名中间字
Eol	地名尾字
Aloc	单字地名
Boo	机构名首字
Moo	机构名中间字
Eoo	机构名尾字

5.3.4.1 人名实体模型

有了表 5-2 的定义之后，基于字的中国人名和外国人名的实体词形模型就可以用公式(5-10)描述：

$$\begin{aligned} P\left(w_{wc_{i_1}}\cdots w_{wc_{ik}} \mid wc_i\right) &= P\left(w_{wc_{i_1}}\cdots w_{wc_{ik}} \mid BNe\ \overbrace{MNe\cdots MNe}^{k-2}\ ENe\right) \\ &\cong P\left(w_{wc_{i1}} \mid BNe\right)\times\prod_{l=2}^{k-1}P\left(w_{wc_{il}} \mid MNe, w_{wc_{i(l-1)}}\right)\times P\left(w_{wc_{ik}} \mid ENe, w_{wc_{i(k-1)}}\right) \end{aligned} \tag{5-10}$$

式中，$w_{wc_{il}}(1\le l\le k)$表示组成人名实体wc_i的单字。*BNe*，*MNe* 和 *ENe* 分别表示实体的首字，中间字和尾字。在具体计算人名的时，分别将其替换成 *Sur*、*Dgb*、*Dge*、*EBfn*、*EMfn* 和 *EEfn* 等。例如，估计典型三字中国人名的实体词形模型可以表示为式(5-11)形式。

$$\begin{aligned} &P\left(w_{wc_{i1}}w_{wc_{i2}}w_{wc_{ik}} \mid wc_i\right) \\ &= P\left(w_{wc_{i1}}w_{wc_{i2}}w_{wc_{ik}} \mid Sur\ Dgb\ Dge\right) \\ &= P\left(w_{wc_{i1}} \mid Sur\right)\times P\left(w_{wc_{i2}} \mid Dgb, w_{wc_{i1}}\right)\times P\left(w_{wc_{i3}} \mid Dge, w_{wc_{i2}}\right) \end{aligned} \tag{5-11}$$

由于人名的词性实体模型的训练语料很难得到，因此，为简化起见，论文使用词形实体模型替代词性实体模型，并乘以一个加权因子。人名实体词性模型如同公式(5-12)。

$$P\left(t_{tc_{i1}}\cdots t_{tc_{ik}} \mid tc_i\right)=\gamma\times P\left(w_{wc_{i1}}\cdots w_{wc_{ik}} \mid wc_i\right) \tag{5-12}$$

在本文的实验部分，γ 取经验值 0.5。

5.3.4.2 地名和机构名实体模型

对于地名和机构名，其模型要复杂得多。这是因为地名中常嵌套人名、地名，如“茅盾故居纪念馆”， “北京经济技术开发区”；机构名中常嵌套人名、地名和机构名，如“富士通(中国)有限公司”，“宋庆龄基金会”等。

基于词的嵌套地名和机构名词形实体模型可以用公式(5-13)描述。

$$\begin{aligned}
&P\left(w_{wc_i-start}\cdots w_{wc_i-end} \mid wc_i\right)=P\left(wc_{wc_{i1}}\cdots wc_{wc_{il}}\cdots wc_{wc_{ik}} \mid BNe\ \overbrace{MNe\cdots MNe}^{k-2}\ ENe\right)\\
&\cong P\left(wc_{wc_{i1}} \mid BNe\right)P\left(w_{wc_{i1}-start}..w_{wc_{i1}-end} \mid wc_{wc_{i1}}\right)\\
&\times\prod_{l=2}^{k-1}P\left(wc_{il} \mid MNe, wc_{i(l-1)}\right)P\left(w_{wc_{il}-start}\cdots w_{wc_{il}-end} \mid wc_{wc_{il}}\right)\\
&\times P\left(wc_{wc_{ik}} \mid ENe, wc_{wc_{i(k-1)}}\right)P\left(w_{wc_{ik}-start}\cdots w_{wc_{ik}-end} \mid wc_{ik}\right)
\end{aligned} \tag{5-13}$$

基于词的嵌套地名和机构名词性实体模型可以用公式(5-14)描述。

$$\begin{aligned}
&P\left(t_{tc_i-start}\cdots t_{tc_i-end} \mid tc_i\right)=P\left(tc_{tc_{i1}}\cdots tc_{tc_{il}}\cdots tc_{tc_{ik}} \mid BNe\ \overbrace{MNe\cdots MNe}^{k-2}\ ENe\right)\\
&\cong P\left(tc_{tc_{i1}} \mid BNe\right)P\left(t_{tc_{i1}-start}..t_{wc_{i1}-end} \mid tc_{tc_{i1}}\right)\\
&\times\prod_{l=2}^{k-1}P\left(tc_{il} \mid MNe, tc_{i(l-1)}\right)P\left(t_{tc_{il}-start}\cdots t_{tc_{il}-end} \mid tc_{tc_{il}}\right)\\
&\times P\left(tc_{tc_{ik}} \mid ENe, tc_{tc_{i(k-1)}}\right)P\left(t_{tc_{ik}-start}\cdots t_{tc_{ik}-end} \mid tc_{ik}\right)
\end{aligned} \tag{5-14}$$

式中的 *BNe*，*MNe* 和 *ENe* 分别表示实体首词，中间词和尾词。在具体计算地名和机构名的时候分别将其替换成 *Bol*、*Mol*、*Eol*、*Boo*、*Moo* 和 *Eoo*。

公式(5-15)是一个具体的例子，说明如何估计嵌套地名和机构名实体词形模型。

$$\begin{aligned}
&P\left(\text{富士通（中国）有限公司} \mid ORG\right)\\
&=P(ORG \mid Boo)\times P\left(\text{富士通} \mid ORG\right)\times P\left(\text{（} \mid Moo, ORG\right)\times P\left(LOC \mid Moo, \text{（}\right)\\
&\times P\left(\text{中国} \mid LOC\right)\times P\left(\text{）} \mid Moo, LOC\right)\times P\left(\text{有限公司} \mid Eoo, \text{）}\right)
\end{aligned} \tag{5-15}$$

5.3.4.3 单字地名实体模型

单字地名词形实体模型和词性实体模型可以分别用公式(5-16)和(5-17)描述：

$$P(w_i \mid ALoc) = \frac{C(w_i, Aloc)}{C(Aloc)} \tag{5-16}$$

$$P(t_i \mid ALoc) = \frac{C(t_i, Aloc)}{C(Aloc)} \tag{5-17}$$

其中，$C(w_i, Aloc)$和 $C(t_i, Aloc)$分别表示在语料中词 w_i 作为单字地名出现的次数以及词性 t_i 作为单字地名出现的次数，$C(Aloc)$为语料中单字地名出现的总次数。

5.3.4.4 简称机构名实体模型

机构名全称是机构名的最正式的叫法。而简称机构名是对机构名全称的缩略叫法。机构名简称的出现形式大致可分为连续简写、不连续简写和混合简写三种方式，如表 5-3 所示。

表 5-3 简称机构名的出现形式

连续简写	上海华联超市股份有限公司	**上海华联**
	上海紫江企业集团股份有限公司	紫江企业
	美国福特公司	福特公司
	香港理工大学	港理工
	北京 25 中学	25 中
不连续简写	**上海证券交易所**	上证/上证所
	北京大学	北大
	电子工业部第六研究所	六所
	福建省绿得罐头饮料有限公司	绿得公司
	北京新唐装饰工程有限公司	新唐公司
	武汉钢铁集团公司	武钢
连续简写与不连续简写混合	东风汽车电子仪表股份有限公司	东风电仪

在表 5-3 中，包括机构名关键词的机构名简称(如福特公司，绿得公司，新唐公司)的识别同机构名全称的识别过程是一样。但对于那些省略了机构名关键词的简称机构名的识别将是非常困难的。

分析发现，简称机构名在文本中的出现基本上包括三种：

[1] 某些简称可以作为常用词收录于词典中，如中共、北约、欧盟等。

[2] 无法收录于词典中简称机构名，但在文本中出现过该简称的全称形式，如华虹 NEC(全称：上海华虹 NEC 电子有限公司，且在文中出现)、海正药业(全称：浙江海正药业股份有限公司，且在文中出现)等。

[3] 文本中直接出现省略了机构名关键词的简称机构名，如百度(省略了关键词”公司”)等。

本文主要集中处理前两种情况。

■ 第一类简称机构名

第一类简称机构名的词形和词性实体模型用公式(5-18)和(5-19)描述：

$$P(w_i \mid Aorg) = \frac{C(w_i, Aorg)}{C(Aorg)} \tag{5-18}$$

$$P(t_i \mid Aorg) = \frac{C(t_i, Aorg)}{C(Aorg)} \tag{5-19}$$

其中，$C(w_i, Aorg)$和 $C(t_i, Aorg)$分别表示在语料中词 w_i 作为机构名简称出现的次数以及词性 t_i 作为机构名简称出现的次数，$C(Aorg)$为语料中简称机构名出现的总次数。

■ 第二类简称机构名的词形实体模型

在真实文本中，简称可能出现在前文，也可能出现在后文，为了完成这类简称机构名的识别，必须把命名实体识别分两阶段。第一阶段识别第一类简称机构名和全称形式的机构名，并将之放入 Cache 中；第二阶段利用第一阶段识别的结果，即 Cache 中的机构名，进行识别。这样做的目的有二个：一是避免了简称机构名和遗漏和限制不必要的简称机构名的产生；二是方便、合理的计算简称机构名的产生概率，即简称的实体模型。

利用 Cache 和对齐技术的简称机构名的词类实体模型和词性实体模型可以统一用(5-20)描述。

$$P(J \mid AORG) = \sum_{A\in\text{所有的队齐}} \beta \times P(A_C)\, P(J \mid A_C, AORG) \tag{5-20}$$

式中的 *J* 表示简称字符串或词性串，*C* 表示 Cache 中的全称机构名，*A* 表示 *J* 是 *C* 的简称和对齐，β 表示对 Cache 中的 *C* 表示一个全称机构名的信任度。图 5-1 是简称机构名和全称机构名的对齐示意图。

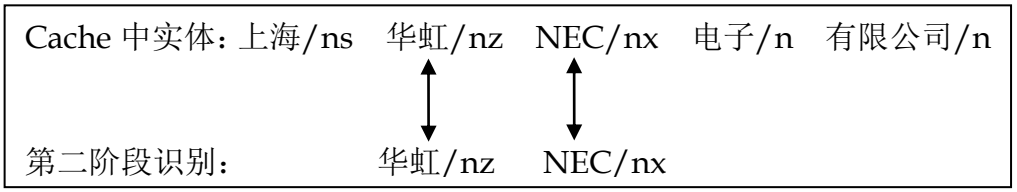

图 5-1 简称机构名与全称机构名的对齐示意图

简称机构名实体模型的具体计算步骤如下：

- 计算当前词串 J= $j_1 \ldots j_m$ 对 Cache 中每个机构名 C_i 的覆盖度 $Overlap_i$，

$$Overlap_i = \frac{J \text{和} C \text{中共有的词}}{J \text{中的词}}$$

- 挑选所有覆盖度 $Overlap_i=1$ 的机构名 C_i 作为当前词串的候选机构名全称
- 根据当前词串 J= $j_1 \ldots j_m$ 和候选全程机构名 C_i 的对齐 A 计算当前词串作为该机构名简称的概率，即实体模型

5.4 专家知识

基于统计模型的命名实体识别，其缺点是数据稀疏严重，搜索空间太大，从而影响系统的性能和效率。本文通过引入专家知识来限制候选实体的产生，从而达到提高性能和效率的目的。专家知识主要包括：

人名识别的专家知识

- 中国人名姓氏列表(476 个)和日本人名姓氏列表(9189 个)：用来限制中国人名和日本人名的候选产生；
- 苏俄人名和欧美人名用字列表：主要用来限制苏俄人名和欧美人名的候选产生；
- 人名长度限制：中国人名长达最大八个字符，外国人名则不受长度限制；

地名识别的专家知识

- 地名关键词列表(607 个)：如果当前词属于地名关键词，如“省、开发区、沙滩、瀑布”等，则触发地名识别；
- 后面常跟地名的介词和动词(如“去、到、在”等)列表：如果前一个词包含在该列表，则触发地名识别；

- 单字地名列表包含(407 个)：单字地名的候选产生使用基于字的触发；
- 地名长度的限制：地名最多包含 12 个汉语字符

机构名识别的专家知识

- 机构名关键词列表(3129 个)：机构名的候选产生使用基于机构名关键词的触发，如果当前词属于该列表，则机构名识别触发；
- 机构名模板：主要用来识别统计模型遗漏的嵌套命名实体，部分模板如下：

ON-->LN D OrgKeyWord*

ON-->PN D OrgKeyWord*

ON-->ON OrgKeyWord

模板中的 *D* 和 *OrgKeyWord* 分别表示机构名中间词和机构名关键词，*D** 表示机构名中可以包含 0 个以上的特征词。

- 机构名长度的限制：机构名最多包含 16 个汉语字符

5.5 模型训练

本文提出的基于多特征的汉语命名实体识别模型(5-7)由四个模型参数组成，他们均通过最大似然估计从不同训练语料中学习得到，其中的词性上下文模型 *P(TC)*和词形上下文模型 *P(WC)*是从 5 个月(1998 年 2 月~1998 年 6 月)的人民日报标注语料中学习的；中国人名、外国人名、地名、机构名的实体词性模型和实体词形模型分别从 156 万，1.4 万，4.4 万，32 万条的实体列表中训练得到的。

尽管使用了这样大规模的训练语料，数据稀疏问题还是非常严重。论文采用的是 Back-Off 模型(5-21)进行参数平滑，并引入逃逸概率计算权值。

$$\begin{aligned}&\hat{p}(W_N|W_1\cdots W_{N\text{-}1})\\&=\lambda_N p(W_N|W_1\cdots W_{N\text{-}1})+\lambda_{N\text{-}1}p(W_N|W_2\cdots W_{N\text{-}1})+\cdots+\lambda_1 p(W_N)+\lambda_0 p_0\end{aligned}\quad(5\text{-}21)$$

式中的 $\lambda_i=(1-e_i)\sum_{k=i+1}^{N}e_k, 0<i<N$，$\lambda_N=1-e_N$。$e_i$ 是各阶逃逸概率，本文使用经验公式(5-22)计算，*N* 是待求的 N-gram 阶数。

$$e_N=\frac{q(W_1W_2\cdots W_{N\text{-}1})}{f(W_1W_2\cdots W_{N\text{-}1})}\quad(5\text{-}22)$$

式(5-22)中 $q(w_1w_2\ldots w_{N-1})$代表对于词序列 $w_1w_2\ldots w_{N-1}$ 后面跟不同词 w_N 的个数，$f(w_1w_2\ldots w_{N-1})$表示词序列 $w_1w_2\ldots w_{N-1}$ 出现的次数。

5.6 模型评测

论文使用准确率(5-23)、召回率(5-24)和 F 值(5-25)对模型进行开放评测和封闭评测。其中，封闭测试的训练集为六个月(1998 年 1 月~1998 年 6 月)的人民日报标注语料，测试集为其中一个月(1998 年 1 月份)的人民日报生语料；开放测试的训练集为五个月(1998 年 2 月~1998 年 6 月)的人民日报标注语料，测试集为其中一个月的人民日报(即 1998 年 1 月份)。

$$精确率=\frac{正确识别的实体数}{总的识别实体数}\times 100\% \tag{5-23}$$

$$召回率=\frac{正确识别的实体数}{总的实体数}\times 100\% \tag{5-24}$$

$$F值=\frac{2*召回率*正确率}{召回率+正确率}\times 100\% \tag{5-25}$$

5.6.1 平衡因子 β 对系统性能的影响

模型(5-7)中的平衡因子的作用是平衡词形特征和词性特征的权值，β 越大，词性特征的作用越强；否则，词形特征的作用越强。下面的实验为了找出词形特征和词性特征的最佳组合，即最佳的 β 值。β 值的变化($0\le\beta\le10$)对人名、地名、机构名识别的影响趋势图分别如图 5-2~5-4 所示。图 5-2~5-4 中横坐标的左端点、中间点、右端点分别是词形特征模型、基于多特征的混合模型、词性特征模型的识别结果。

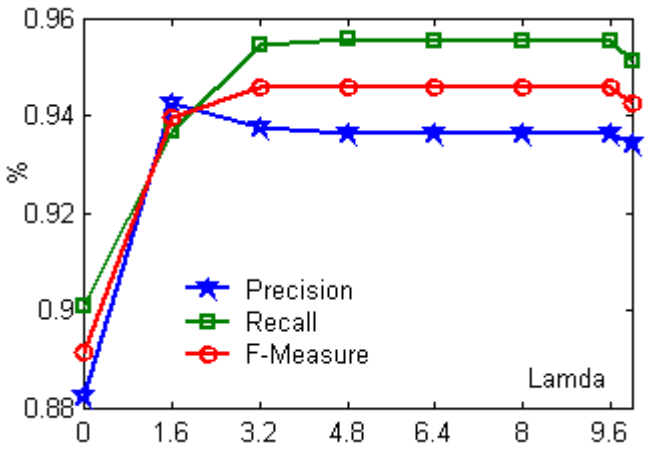

图 5-2 β 值对人名识别的影响

随着 β 值的增加，人名识别的准确率和召回率在有一段提升后，逐渐下降。这说明人名实体识别算法单纯依靠词形特征或者词性特征都不能获得最优的性能，理想的方法是将两者有机的结合起来，扬长避短。

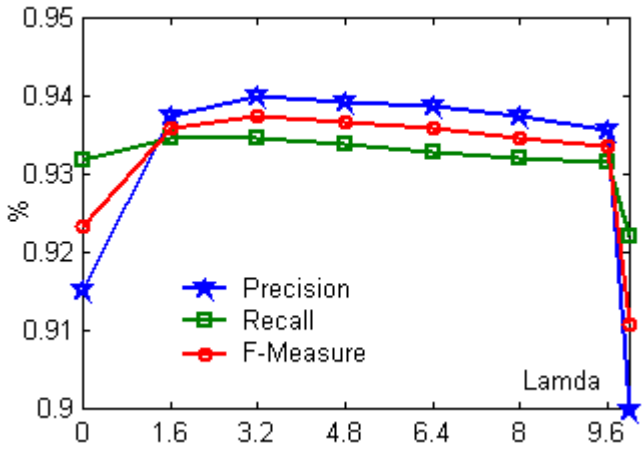

图 5-3　β 值对地名识别的影响

观察图 5-3 发现，同人名一样，地名识别的准确率和召回率也是随着平衡因子 β 的逐步增大在有一段提升后，逐渐下降。这同样也说明只有词形特征和词性特征互相取长补短才能获得优良的地名识别指标。

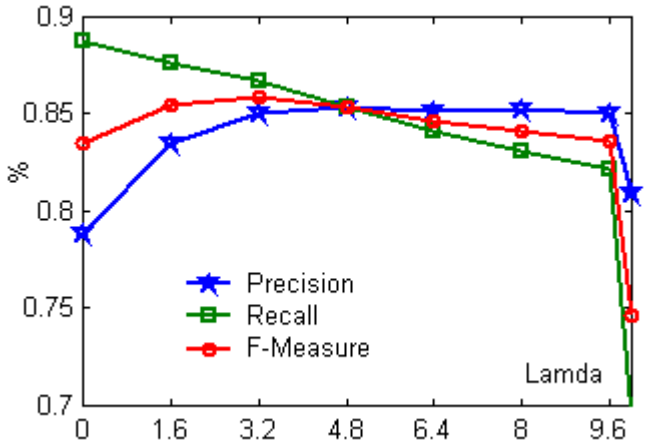

图 5-4　β 对机构名识别的影响

图 5-4 给出平衡因子对机构名识别影响的曲线图，与图 5-2、5-3 不同的是词性特征对机构名识别表现出同人名、地名不同趋势，尤其是机构名识别的召回率表现出非常明显的下降趋势。这说明词性特征对嵌套机构名的识别有非常明显的副作用。这是可能是由于词性在嵌套机构名内部的共现频率要小于其在词性上下文中的共现频率。

综合比较图 5-2~5-4，算法最终采用的平衡因子 β 值是 2.8。表 5-4~5-6 分别给出基于多特征的混合模型(β = 2.8)、词形特征模型、词性特征模型开放测试指

标。

表 5-4 基于多特征的混合模型开放测试性能

模型	**基于多特征的混合模型(β = 2.8)**				
性能	**正确识别数**	**总计识别数**	**P(%)**	**R(%)**	F(%)
人名	**19051**	**20220**	**94.06**	**95.21**	94.63
地名	**20861**	**22159**	**93.98**	**93.48**	93.73
机构名	9390	11094	84.69	86.86	85.76

表 5-5 词形特征模型开放测试性能

模型	**词形特征模型**				
性能	**正确识别数**	**总计识别数**	**P(%)**	**R(%)**	F(%)
人名	**18015**	**20418**	**88.24**	**90.11**	89.16
地名	**20796**	**22687**	**91.50**	**93.17**	92.32
机构名	9603	12172	78.85	88.77	83.52

表 5-6 词性特征模型开放测试性能

模型	**词性特征模型**				
性能	**正确识别数**	**总计识别数**	**P(%)**	**R(%)**	F(%)
人名	**19031**	**20336**	**93.44**	**95.11**	94.27
地名	**20576**	**22837**	**89.97**	**92.20**	91.07
机构名	7476	9254	80.90	69.29	74.65

比较表 5-4~5-6 发现，混合模型的人名、地名、机构名识别性能(F 值)比词形特征模型分别提高约 5.4%，1.4%，2.2%；比词性特征模型分别提高约 0.4%，2.7%，11.1%。

结论 1：表 5-3~5-5 的实验数据表明，本文提出的词形和词性特征相结合的汉语命名实体识别模型的识别性能要优于使用单一特征(词形特征或词性特征)的命名实体识别模型。但由于人名、地名和机构名的内部构成有很大的差别，例如，人名用字相对集中；地名和机构名用字分散；且多数为嵌套的地名和机构名。所以词形和词性特征的混合模型对提高他们的识别性能略有差别。

5.6.2 模型的一致性检验

模型在 MET-2 汉语测试语料上进行评测的目的是为检验算法在不同测试语料上是否具有一致性：结合词形和词性特征的命名实体识别模型是否优于使用单一特征的命名实体识别模型。测试结果(人名、地名和机构名的 F-Measure 值)如表 5-7 所示。

表 5-7　模型在 MET-2 测试语料上的 F-Measure 值

F 值	词形特征模型	词形和词性特征混合模型	**词性特征模型**
人名	**75.21%**	**80.77%**	76.61%
地名	**89.78%**	**90.95%**	89.81%
机构名	76.30%	80.21%	76.83%

比较表 5-7 和表 5-4~5-6 发现，算法在 MET-2 测试数据上的指标要低于在人民日报语料上的测试指标。分析识别错误发现，导致性能下降一个非常重要的原因是 MET-2 对实体的定义和人民日报标注规范对实体的定义有较大的差别。例如，在 MET-2 中，实体“西安卫星测控中心”被定义为地名，而人民日报标注规范定义其为机构名；MET-2 定义“阿利亚娜火箭”中的“阿利亚娜”为普通词，而人民日报标注规范定义其为人名实体；MET-2 定义“亚洲卫星二号”中的“亚洲”为普通词，而人民日报标注规范定义其为地名实体。如果忽略由于规范不同带来的错误，系统的性能将会的较大的提高。

但表 5-7 仍然可以证明词形和词性的混合模型要优于使用单一特征的模型，例如，相对于词性模型，混合模型对人名、地名和机构名识别的 F-Measure 值分别提高为 5.6%，1.2%和 3.9%；相对于词性模型，混合模型对人名、地名和机构名识别的 F-Measure 值提高分别为 4.2%，3.1%和 3.4%：

结论 2：虽然模型在 MET-2 测试语料上表现出的性能指标要低于在人民日报测试语料上的性能指标。但表 5-6 中三个不同模型的 F 值还是说明了本文提出的模型在不同的测试语料能够表现出不错的一致性：结合词形和词性特征的命名实体识别模型要优于使用单一特征的命名实体识别模型。

5.6.3 专家知识对统计模型的贡献

专家知识不仅能够减少统计模型的搜索空间，提高系统的速度，而且还可以

避免产生不必要的候选实体而带来的识别性能的下降。表 5-8 给出了基于多特征的混合模型在引入专家知识前后的系统识别性能。

表 5-8　专家知识对统计模型的影响

模型	纯统计模型			专家知识和统计模型结合		
性能	**P(%)**	**R(%)**	**F(%)**	**P(%)**	**R(%)**	F(%)
人名	**91.81**	**70.65**	**79.85**	**94.06**	**95.21**	94.63
地名	**79.47**	**88.83**	**83.89**	**93.98**	**93.48**	93.73
机构名	64.95	80.63	71.95	84.69	86.86	85.76

从表 5-8 可以看出，结合了专家知识的统计模型对人名、地名和机构名的识别能力(F 值)较纯统计模型分别提高了约 14.8%，9.8%，13.8%。

结论 2：专家知识不仅能够减少统计模型搜索空间，提高系统的识别速度，还大大提高了纯统计模型对汉语命名实体的识别能力。

5.6.4 863 评测

2004 年 863 评测的对象是现代汉语(包含大陆的简体文本和港澳台地区的繁体文本)的命名实体(包含命名实体、时间表达式及数量表达式)识别系统中的核心技术。评测语料来自近期流通广泛的图书、报纸、期刊和网络等载体，以期反映当代汉语的最新面貌，涉及到的主题有政治、经济、体育、交通、旅游、教育等。语料的选择考虑到其平衡性、科学性和代表性。

参加本次评测的有 8 家单位(System1，System2，System3，System4，System5，System6，System7，System8)。表 5-9 是评测系统在简体测试语料上的性能对比情况。本论文是 System4 系统。

表 5-9 863 评测结果

单位 简称			System 1	System 2	System 3	System 4	System 5	System 6	System 7	System 8
总体		召回率	71.34%	81.10%	62.54%	79.56%	78.20%	77.62%	71.44%	70.65%
		准确率	63.25%	83.69%	58.32%	83.79%	81.93%	83.23%	85.21%	77.42%
		F1 值	67.05%	82.38%	60.36%	81.62%	80.02%	80.33%	77.72%	73.88%
分项	地名	召回率	73.49%	79.50%	61.71%	78.43%	76.72%	72.21%	70.22%	68.78%
		准确率	72.23%	81.84%	74.24%	87.02%	79.96%	85.29%	86.45%	84.02%
		F1 值	72.85%	80.65%	67.40%	82.51%	78.31%	78.21%	77.49%	75.64%
	人名	召回率	74.09%	84.34%	67.66%	88.47%	87.48%	81.45%	62.19%	68.83%
		准确率	66.49%	80.71%	49.99%	81.38%	73.13%	89.99%	84.42%	75.57%
		F1 值	70.08%	82.48%	57.50%	84.78%	79.66%	85.51%	71.62%	72.04%
	组织名	召回率	47.32%	39.27%	31.41%	57.41%	33.78%	40.55%	54.67%	45.10%
		准确率	36.31%	61%	30.11%	64.64%	74.35%	36.04%	56.57%	43.51%
		F1 值	41.09%	47.78%	30.74%	60.81%	46.45%	38.16%	55.60%	44.29%
	日期	召回率	56.76%	81.73%	77.50%	76.71%	75.39%	82.12%	75.78%	61.38%
		准确率	74.59%	86.69%	66.67%	81.86%	86.20%	88.56%	88.08%	77.40%
		F1 值	64.47%	84.13%	71.68%	79.20%	80.43%	85.22%	81.47%	68.47%
	时间	召回率	21.48%	74.07%	20.37%	61.48%	53.70%	83.70%	78.52%	34.81%
		准确率	4.51%	81.63%	7.77%	63.60%	73.23%	85.93%	82.17%	16.32%
		F1 值	7.46%	77.67%	11.25%	62.52%	61.97%	84.80%	80.30%	22.22%
	数字	召回率	81.27%	94.12%	66.86%	84.20%	90.15%	91.71%	81.81%	86.36%
		准确率	67.98%	91.14%	59.60%	88.57%	90.52%	92.17%	93.09%	87.93%
		F1 值	74.03%	92.60%	63.02%	86.33%	90.33%	91.94%	87.09%	87.14%

说明：

1. 本次命名实体评测的测试语料简体和繁体约各 40 万字，根据大纲要求，对中文简体和繁体分别打分，并对其中的各个小项也分别打分。

2. 在现场评测过程中，由于有些单位没有能够完全将测试文件运行完毕，可能导致分数明显偏低；这样的单位有：System8(简体 2 个文件)，System3(简体 1 个)，System4(简体 10 个)

5.7 本章小结

在自然处理处理得很多领域，命名实体都起着非常重要的作用。针对汉语命名实体识别中的难点，本文提出了多特征混合的汉语命名实体识别模型。该模型具有以下特点：

■ 强调大颗粒度特征(词性特征)和小颗粒度特征(词形特征)的结合，以克服各自的缺点。词形特征虽然可以较好的刻画实体的内部特征和外部特征，但其颗粒度太小，数据稀疏问题严重削弱了它的作用；词性特征虽然数据稀疏

问题不严重，但其刻画实体内部、外部特征的能力比较差。

■ 提出了统计模型和专家知识相结合的方法，该方法通过限制候选命名实体的产生，减少搜索空间，提高了识别速度。

■ 为准确刻画不同实体的内部特征，设计了多个细分类的实体模型；具体地，论文把人名实体划分为中国人名、日本人名、苏俄人名、欧美人名和人名简称；地名划分为单字地名和多字地名；机构名细分为简称机构名和全称机构名的实体模型。

本论文采用产生式模型，即隐马尔柯夫模型，对汉语命名实体识别进行了研究。正如 Lafferty 在文[J. Lafferty, *et al*. 2001]所指出的，一方面，产生式模型仅利用相邻特征，无法使用更为丰富的特征；另一方面，为了使得模型更容易操作和使用，产生式模型需要严格的假设条件，即观察值出现的概率只和当前状态有关，而与其它信息无关。为解决产生式模型的这些不足，有人提出使用判别式模型(如最大熵马尔柯夫模型)进行命名实体的识别，但判别式模型又容易出现标记偏置问题。近些年提出的条件随机场模型(CRF 模型)是一种新的条件概率模型，它不需要对观察序列建模，而且可以利用丰富的观察特征，同时又可以避免判别式模型的标记偏置问题。理论上，CRF 模型是最合适序列标注的概率模型之一。所以，下一步工作可以尝试采用 CRF 模型识别汉语命名实体，提高实体识别的性能。

5.8 本章研究成果

[1] Youzheng Wu, Jun Zhao, Bo Xu , Hao Yu. Chinese Named Entity Recognition Model Based on Multiple Features. In Proceedings of HLT/EMNLP 2005, October 6-8, Vancouver, B.C., Canada, pp427~434.

[2] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge. In Proceedings of ACL2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, Sappora, Japan, July, 2003, pp. 65-72.

第六章 基于主题语言模型的句子检索技术

6.1 引言

问答系统的检索模块是根据提问处理模块生成的查询关键词，使用某种检索算法，检索出和提问相关的信息。返回的信息可以是段落、也可以是句群或者句子。所以，检索模块在整个问答系统中占有十分重要的地位，检索性能不仅在一定程度上影响问答系统的准确率，而且对系统的运行速度也有一定的影响。如果检索模块的准确率和召回率可靠性很高，问答系统的后续模块(候选答案抽取模块和答案选择模块)只需要在少量的前 N(TopN)个检索结果上进行。

由于大部分问答系统的答案抽取都是在句子的基础上进行的，所以问答系统的检索基本上都是指是句子检索。本文即是针对如何提高问答系统的句子检索质量展开研究，后续章节提到的检索在默认状态下都是指句子检索。

关于英文问答系统的句子检索，已经开展了相当多的研究。例如，Ittycheriach [A. Ittycheriach, *et al*. 2002]和 H. Yang [H. Yang, *et al*. 2002]提出向量空间模型，计算查询向量和文档向量的夹角余弦。Emmanuel [A. C. Emmanuel, *et al*. 2004]和 Vanessa & Croft[V. Murdock, *et al*. 2004]分别提出将语言模型和翻译模型应用到英文问答系统的句子检索中，并取得了一定的成绩。Vanessa 的翻译模型需要一定量的训练语料，这限制了翻译模型的使用。和传统向量空间模型相比，语言模型检索方法具有相对完善的理论框架，且在改善检索性能方面有更加明确的指导作用[J. Y. Nie. 2005]。但是，语言模型检索算法目前还不完善，例如结构信息的引入、平滑算法的改进等方面都还存在空间。本论文主要是针对平滑算法的改进展开了深入研究。

现有的平滑方法主要包括 Jelinek-Mercer、Dirichlet、Two-Stage、Bayesian、Dirichlet、绝对减值法等。这些方法存在的主要缺点是：由于每篇文档的语言模型(建立在一篇文档上的语言模型)均使用同一个全局语言模型(建立在整个文档集合上的语言模型)进行平滑，因而导致对文档的刻画“粒度”过粗。对于同样不包含查询关键词的两篇文档 A 和 B，现有平滑算法认为它们和查询的相关性是一样的，尽管其中的某篇文章是和查询主题密切相关。

现有方法在应用于问答系统中的句子检索中，由于句子包含的信息量更少，所以将语言模型应用于问答系统中的句子检索，数据稀疏将更为严重。为提高问答系统的检索性能，需要对文档集合包含的信息进行更加细致的挖掘。

问答系统的检索算法基本上是从传统的文档信息检索中直接借鉴过来的，没有针对问答系统的特殊性挖掘更多的有用信息。而通过观察发现，问答系统中的某些特殊信息可以用来提高句子检索的质量。这些特殊性主要体现在以下两个方面：

- 相对于传统的文档检索，问答系统的自然语言提问输入更能够刻画用户的需求，且歧义性较小

传统检索系统很难从用户提交的查询中提取用户的真正查询意图，例如查询 QUERY = {发明，电话}，系统无法知道用户需要什么类型的信息：是发明电话的人？是发明电话的时间？是发明电话的过程？还是其它相关信息？而问答系统可以通过对用户提问的分析提取用户需要的信息类型。例如提问 QUESTION = {谁发明了电话？}，系统可以清楚地知道用户需要的是发明电话的人，而非发明电话的时间，地点或者其它信息。问答系统的检索可以利用这一特点提高检索质量。

- 根据用户提问的答案类型，从初检结果中提取的候选答案可以用于主题划分

句子检索结果是一系列和提问相关的句子集合，这些句子或者围绕提问的不同侧面展开，或者描述和提问相关的主题。所以，提问的句子检索结果是由不同侧面的信息组成的，这些不同侧面信息就代表一个主题。因此，同等对待句子中的查询词，这些句子具有不同的主题，显然是不合逻辑的。一个可行的方法是按照主题对检索结果进行聚类，一个句子可以属于一个或者多个主题，即把初检结果转化为以主题为单元的逻辑结构形式。

根据问答系统的特殊性，本文提出了一种新的主题划分方法，即根据问答系统的候选答案对初检结果进行主题划分。例如，提问：谁发明了电话？句子检索的前 10 个结果及其所包含的候选答案参见表 6-1。这样，主题划分可以按照候选答案{贝尔，维·西门子，爱迪生，马丁·库珀，斯蒂芬·福肖，库珀}对原始检索结果进行聚类，从而提高二次检索的性能。

表 6-1 句子检索结果及其包含的候选答案

编号	句子内容	**候选答案**
S1	1876 年 3 月 10 日贝尔发明电话	**贝尔**
S2	维·西门子发明了电机，贝尔发明电话，爱迪生发明电灯。	**维·西门子** **贝尔** **爱迪生**

S3	最近在纪念这一重要发明时，“移动电话之父”马丁 ·库珀再次成为公众焦点。	**马丁 ·库珀**
S4	1876 年，发明家贝尔发明了电话。	**贝尔**
S5	接着，1876 年，美国科学家贝尔发明了电话；1879 年美国科学家爱迪生发明了电灯。	**贝尔 爱迪生**
S6	1876 年 3 月 7 日，贝尔成为电话发明的专利人。	**贝尔**
S7	贝尔不仅发明了电话，还成功地建立了自己的公司推广电话。	**贝尔**
S8	在首只移动电话投入使用 30 年以后，其发明人库珀仍梦想着未来电话技术实现之日到来。	**库珀**
S9	库珀表示，消费者采纳移动电话的速度之快令他意外，但移动电话的普及率还没有达到无所不在，这让他有些失望。	**库珀**
S10	**英国发明家斯蒂芬 ·福肖将移动电话的所有电子元件设计在一张纸一样厚薄的芯片上。**	**斯蒂芬 ·福肖**

基于上述思想，本文提出了基于主题语言模型的汉语问答系统句子检索算法。该算法可以概括为：(A) 依据从初检结果中抽取的候选答案对初检结果进行聚类，即对初检结果进行主题划分；(B) 统计词语在主题上的概率分布以及句子关于主题的概率分布；(C) 通过引入 Aspect Model [A. Berger, *et al*. 2000]将句子所属的主题引入句子语言模型中，从而获得对句子语言模型更精确的逼近。这个新的语言模型较深入地刻画了词汇在不同主题下的分布规律以及文档所蕴含不同主题的分布规律。对初检结果的聚类，本文提出了“一个句子多个主题”和“一个句子一个主题”两种方法。

相对于 PLSI(Probabilistic Latent Semantic Index)[T. Hofmann. 1999]等算法的主题空间维度，本文提出的主题空间具有更加明确的物理意义；由于不需要迭代运算，运行速度更快。对比实验的结果表明，本文提出的基于主题语言模型汉语问答系统的句子检索算法在多个评测指标下，相对于标准语言模型检索算法(BASELINE)均有明显的提高。

整个系统的结构框图如图 6-1 所示。

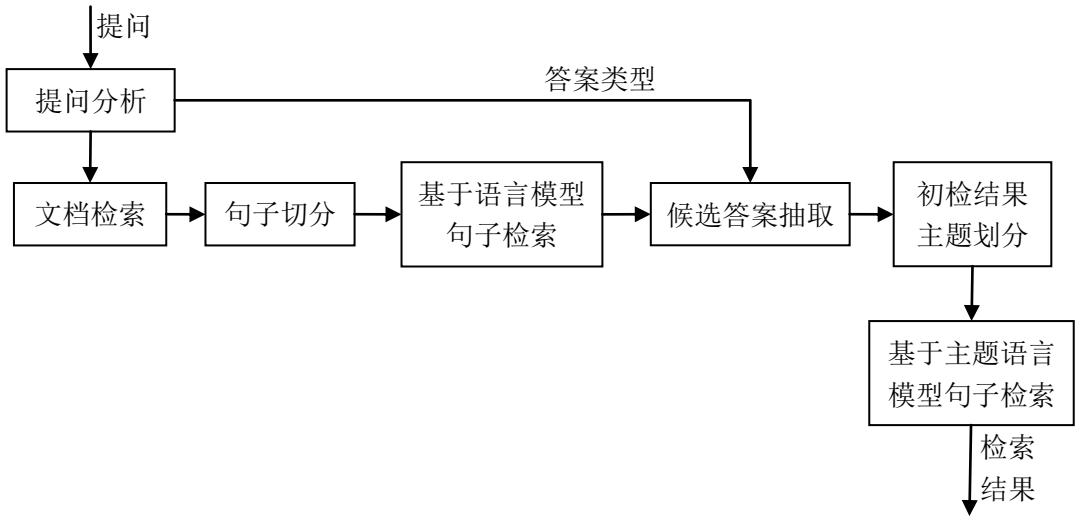

图 6-1 基于主题语言模型汉语问答系统句子检索结构框图

6.2 基于语言模型的信息检索

统计语言模型已经被成功应用于语音识别、词性标注、命名实体识别、句法分析等领域。1998 年 Ponte & Croft 等人首次把语言模型引入到信息检索中来，并取得了一定的成绩。之后，人们又先后提出很多改进的语言模型检索算法。这些算法可以分为文档模型(Document Model) [J. Ponte, *et al*. 1998]、查询模型(Query Model) [V. Lavrenko, *et al*. 2001; J. Lafferty, *et al*. 2001]、距离模型(Divergence Model) [C. Zhai, *et al*. 2001a]和翻译模型(Translation Model) [A. Berger, *et al*. 2000]等四种形式。

6.2.1 文档模型(Document Model)

文档模型的基本思想是：假定查询 q 是由文档 d 的概率模型产生的，并由此对文档进行排列。给定查询 $q=q_1q_2...q_m$ 和文档 $d=d_1d_2...d_n$，文档模型的任务包括 2 个步骤：

- 对文档 d 建模，建立文档语言模型 M_d
- 根据概率 $p(q|M_d)$对文档进行排序

公式 6-1 是文档模型的 Unigram 描述形式：

$$p(q \mid M_d) = \prod_{q_i \in q} p(q_i \mid M_d) \tag{6-1}$$

其中，q_i 是查询词，$p(q|M_d)$表示文档语言模型，$p(q_i|M_d)$反应查询词在文档中 d 的概率分布。

为解决模型中的数据稀疏问题，研究人员提出了很多平滑方法，主要如

Jelinek-Mercer、Dirichlet 和 Two-stage 等算法，如公式 6-2~6-4 所示。

$$p_{JM}(q_i \mid M_d) = a \times p_{ML}(q_i \mid M_d) + (1-a) \times p_{ML}(q_i \mid M_C) \tag{6-2}$$

$$p_{Dir}(q_i \mid M_d) = \frac{tf(q_i,d) + \mu P_{ML}(q_i \mid M_C)}{dl_d + \mu} \tag{6-3}$$

$$p_{TS}(q_i \mid M_d) = (1-\lambda)\frac{tf(q_i,d) + \mu P_{ML}(q_i \mid M_C)}{dl_d + \mu} + \lambda P_{ML}(q_i \mid M_C) \tag{6-4}$$

其中，C 表示整个语料库，$p_{ML}(q_i|M_d)$和 $p_{ML}(q_i|M_C)$分别表示查询词 q_i 在文档 d 和语料库 C 中概率分布，其使用最大似然法估计，如公式 6-5~6-6 所示。

$$p_{ML}(w \mid M_d) = \frac{n(w,d)}{\sum_{w'} n(w',d)} \tag{6-5}$$

$$p_{ML}(w \mid M_C) = \frac{n(w,C)}{\sum_{w'} n(w',C)} \tag{6-6}$$

其中，$n(w,d)$和 $n(w,C)$分别表示词在文档 d 和语料库 C 中的出现次数，$\sum_{w'} n(w',d)$ 和 $\sum_{w'} n(w',d)$ 分别表示文档 d 和语料库 C 的长度，即词数。

6.2.2 查询模型(Query Model)

查询模型的基本思想可以归纳为 2 点：

- 查询 q 和文档 d 均采样自一个未知的相关模型 R，其刻画了 q 和 d 在查询相关文档中概率分布
- 从相关模型 R 中经过 k 采样，观察到查询 $q=q_1q_2...q_m$，估算第 $k+1$ 次采用观察到 w 的概率 $p(w|R)$

所以，查询模型可以用公式 6-7 和 6-8 描述。

$$p(d \mid R) = \prod_{w \in d} p(w \mid R) \tag{6-7}$$

$$p(w \mid R) \approx p(w \mid q_1 \cdots q_m) = \frac{p(w, q_1 \cdots q_m)}{p(q_1 \cdots q_m)} \tag{6-8}$$

其中，w 表示文档 d 中的词。

在没有训练语料的情况下，如何估计查询模型中的 $p(w|R)$是一件极具挑战性的任务。对此，Lavrenko & Croft[V. Lavrenko, *et al*. 2001]在 2001 年提出了 I. I. D Sampling 和 Conditional Sampling 两种方法；Lafferty & Zhai[J. Lafferty, *et al*. 2001]

提出了 Markov Chain 方法。I. I. D Sampling 方法如公式 6-9~6-10 所示。

$$p(w,q_1\cdots q_m)=\sum_{M_d\in U_d}p(M_d)p(w\mid M_d)\prod_i p(q_i\mid M_d) \tag{6-9}$$

$$p(w\mid M_d)=\lambda p_{ML}(w\mid M_d)+(1-\lambda)p(w\mid M_C) \tag{6-10}$$

式中，M_d 代表一篇文档的语言模型，U_d 是相关文档语言模型的集合，$p(M_d)$ 是先验概率。

6.2.3 距离模型(Divergence Model)

距离模型是通过计算文档模型(参见 6.2.1)和查询模型(参见 6.2.2)之间的 Kullback-Leibler(KL)距离对文档集中的文档排序，所以该模型的主要任务包括三个步骤：

- 估计文档模型，$p(w|R)$
- 估计查询模型，$p(w|M_d)$
- 估计文档模型和查询模型之间的 KL 距离

其中文档模型和查询模型可以分别使用前述章节的方法进行估计，KL 的距离计算如公式 6-11。

$$KL(R\mid\mid M_d)=\sum_w p(w\mid R)\log\frac{p(w\mid R)}{p(w\mid M_d)} \tag{6-11}$$

其中，R 表示相关模型，M_d 表示文档模型，$p(w|R)$和 $p(w|M_d)$分别采用公式(6-8)和(6-1)进行计算。

6.2.4 翻译模型(Translation Model)

Berger & Lafferty [A. Berger, *et al*. 1999]把查询看作是文档在同一种语言类的翻译，并根据翻译的概率对文档集中的文档进行排序，其原理可以用公式 6-12 描述。

$$p(q\mid d)=\prod_i p(q_i\mid d)=\prod_i\sum_j p(q_i\mid w_j)p(w_j\mid d) \tag{6-12}$$

公式中的 $p(w_j|d)$代表词 w_j 在文档 d 中的概率分布，$p(q_i|w_j)$表示词 w_j 翻译成词 q_i 的概率。

6.2.5 语言模型工具包-Lemur[25]

Lemur 是由卡内基 梅隆大学和马萨诸塞州大学联合开发基于语言模型的信息检索工具包，其目的在于促进语言模型和信息检索的研究工作。目前最新版本为 4.2 版。Lemur 支持对不同格式的大规模文档建立索引；对文档、查询或文档子集构建简单的语言模型；以及基于语言模型的检索技术。

除此之外，它还支持传统的检索模型，如向量空间模型(VSM)等。Lemur 工具包主要为 Ad Hoc 检索、分布式检索、跨语言检索、文摘、过滤和分类等服务。本文的工作就是在 Lemur 工具包的基础上完成的。

6.2.6 语言模型的检索方法小结

基于语言模型的检索是这几年信息检索领域取得的成果，在很多测试集上的实验结果都说明了该模型要比传统的向量空间模型更有效。但语言模型检索还不完善，很多问题亟待解决，比如如何引入结构信息[J. F. Gao, *et al*. 2004; C. Alvarez, *et al*. 2004]，如何进行数据平滑等等。

本文重点解决语言模型检索方法应用于汉语问答系统句子检索中的数据稀疏问题。现有的数据平滑算法的核心思想都是用全局语言模型 $p_{ML}(q_i|M_C)$线性插值文档语言模型 $p_{ML}(q_i|M_d)$，且通过 a(公式 6-2)、μ(公式 6-3)、λ(公式 6-4)等参数控制每个模型的权重。这些平滑算法可以部分地解决问题，但对文档的刻画仍然存在不足之处。例如，所有的文档语言模型 $p_{ML}(q_i|M_d)$均使用同一个全局语言模型 $p_{ML}\ (q_i|M_C)$平滑，没有针对不同文档区别对待，从而导致对文档的刻画“粒度”过粗。对此，本文采用基于主题语言模型的检索算法，并将之应用于汉语问答系统的句子检索中。

6.3 基于主题语言模型的问答系统句子检索

将语言模型应用于汉语问答系统的句子检索，有个概念需要澄清。在文档检索中，$p(w|M_d)$表示词语 w 在文档 d 中的分布，称之为文档语言模型；而在问答系统的句子检索中，$p(w|M_d)$应该反映词语 w 在句子 S 中的分布，故改写为 $p(w|M_S)$，称之为句子语言模型。

基于主题语言模型检索算法的基本假设是：主题上相关的句子/文档倾向于与

[25] http://www.lemurproject.org/

同一个用户查询相关。所以，按照主题对句子进行聚类，并将聚类信息引入语言模型应该可以提高检索性能。

为了更精确地刻画句子语言模型，本文提出的主题语言模型可以归纳为：

- 通过聚类算法将主题上相关的句子组合在一起
- 通过 Aspect Model 将句子主题信息引入到句子语言模型中
- 用全局语言模型 $p_{ML}(w|M_C)$和主题语言模型 $p_{ML}(w|T)$一起平滑句子语言模型 $p_{ML}(w|M_S)$

论文的模型如公式 6-13~6-14 描述。

$$\begin{aligned} &p(w|M_S) = a \times p_{ML}(w|M_S) \\ &+(1-a)\times(\beta \times p_{ML}(w|T)+(1-\beta)\times p_{ML}(w|M_C)) \end{aligned} \tag{6-13}$$

$$p_{ML}(w|T) = \sum_{t \in T} p(w|t)p(t|M_S) \tag{6-14}$$

公式中的 T 表示主题集合，$p(t|M_S)$表示句子关于主题的概率分布，而词语 w 在主题 t 中的概率分布用 $p(w|t)$表示。

实际上，本文提出的模型也可以看作是一个两阶段的平滑方法：主题语言模型 $p_{ML}(w|T)$首先用全局语言模型 $p_{ML}(w|M_C)$平滑，然后再用平滑后的主题语言模型 $p_{ML}(w|T)$平滑句子语言模型 $p_{ML}(w|M_S)$。

显然，建立主题语言模型关键是如何确定概率 $p(t|M_S)$和 $p(w|t)$。一种典型的做法是通过 K-means 算法自动聚类实现，但 K-means 聚类算法需要通过迭代运算实现，系统的速度会受到影响，且聚类结果没有明确的物理意义。对此，本文提出了两种聚类算法：一个句子多个主题和一个句子一个主题，进行近似估算。

6.3.1 一个句子多个主题

“一个句子多个主题”的聚类思想可以归纳为下面两点：

1. 如果一个句子包含 M 个不同候选答案，则该句描述了 M 个不同的主题
   例如，表 6-1 中的句子 S5 就描述了“贝尔发明电话”和“爱迪生发明电灯”两个不同的主题。
2. 不同的句子，如果包含的候选答案实指同一个实体，则它们陈述同一个主题
   例如，表 6-1 中的句子 S4 和 S5 均包含主题 “贝尔发明电话”。

基于上述思想，表 6-1 的聚类结果如表 6-2 所示。

表 6-2 “一个句子多个主题”主题聚类结果

主题	**涉及主题的句子**
贝尔	S1 S2 S4 S5 S6 S7 S8
维 西门子	S2
爱迪生	S2 S5
马丁 痒珀/库珀	S3 S8 S9
斯蒂芬 福肖	S10

基于“一个句子多个主题”思想的聚类实现步骤具体如下：

- Step 1：提取初检的 TopN 个结果(实验时，TopN = 500)。
- Step 2：根据提问的答案类型抽取每个句子中的候选答案。
- Step 3：所抽取的不同答案数即为 TopN 个句子所包含的和提问相关的主题数，按照候选答案进行初检结果的主题划分。
- Step 4：使用公式 6-15 估算词语在主题 *t* 中的概率分布。

$$p(w|t)=\frac{n(w,t)}{\sum_{w'} n(w',t)} \tag{6-15}$$

式中，*n(w,t)*表示词 *w* 在主题 *t* 中的出现次数，$\sum_{w'} n(w',t)$表示主题 *t* 的长度，即词数。

- Step 5：使用公式 6-16~6-17 估算句子在主题集合上的概率分布。

$$p(t|M_S)=\frac{1/kl_{st}}{\sum_{t=1}^{k}(1/kl_{st})} \tag{6-16}$$

$$kl_{st}=KL(s||t)=\sum_{w} p_{ML}(w|M_S)\times log\frac{p_{ML}(w|M_S)}{p_{ML}(w|t)} \tag{6-17}$$

式中，kl_{st}表示句子与主题之间的 Kullback-Leibler 距离。公式(6-16)的思想是：kl_{st}距离越近，句子在主题上的分布概率 $p(t|M_S)$越大。

6.3.2 一个句子一个主题

“一个句子一个主题”的聚类思想可以归纳为：

1. 无论一个句子包含多少个候选答案，该句也只有一个核心候选答案，并描述一个核心主题
   例如表 1 中的句子 S5，虽然包含“贝尔”和“爱迪生”两个候选答案，但也只有一个核心答案：“贝尔”，讲述一个核心主题：贝尔发明电话。
2. 不同的句子，如果包含的核心候选答案相同，则它们描述的是同一个主题
   例如，表 1 中的句子 S4 和 S5 均属于同一个主题：贝尔发明电话。
3. [核心候选答案定义]：是指距离查询词平均距离最近的候选答案。

基于上述思想，表 6-1 的聚类结果如表 6-3 所示。

表 6-3 “一个句子一个主题”主题聚类结果

主题	**涉及主题的句子**
马丁·库珀/库珀	S3 S8 S9
贝尔	S1 S2 S4 S5 S6 S7
斯蒂芬·福肖	S10

基于“一个句子一个主题”思想的聚类步骤实现具体如下：

- Step 1：提取初检的 TopN 个结果(实验时，TopN = 500)。
- Step 2：根据提问的答案类型抽取每个句子中的候选答案。
- Step 3：对每个句子，根据公式 6-19 计算每个候选答案距离查询词的平均距离，根据公式 6-18 确定核心候选答案。

$$a_i^* = \underset{a_i}{argmin}\{SemDis_{a_i}\} \tag{6-18}$$

$$SemDis_{a_i} = \frac{\sum_j SemDis(a_i, q_j)}{N} \tag{6-19}$$

$$SemDis(a_i, q_j) = \left| Position_{a_i} - Position_{q_j} \right| \tag{6-20}$$

其中，a_i^* 表示核心候选答案，a_i 表示第 i 个候选答案，$SemDis_{a_i}$ 表示第 i 个候选答案距离查询词的平均距离，q_j 表示第 j 个查询词，N 是查询词的个数，$Position_{q_j}$ 和 $Position_{a_i}$ 分别表示查询词 q_j 和候选答案 a_i 在句子中的位置。

- Step 4：根据核心主题对初检结果进行主题划分。
- Step 5：使用公式(6-15)估算词语在主题 z 中的概率分布。
- Step 6：由于每个句子仅描述一个主题，所以取 $p(z|S) = 1$。

6.3.3 和 PLSI 的比较

PLSI 模型是 Hofmann 在 1999 年提出的一种语义模型[T. Hofmann. 1999]。PLSI 模型从概率的原则出发，建立文档 *d* 和词 *w* 之间的概率相似度关系 *p(w, d)*。为了体现两者之间的语义联系以及数据压缩的需要，引入了隐含的多维变量 *z*，也可以理解为主题空间或者语义空间，具体如公式(6-21)所示。

$$p(w,d)=\sum_{z\in Z}p(z)p(w\mid z)p(d\mid z) \tag{6-21}$$

PLSI 通过 EM 迭代算法计算主题空间的先验概率 *p(z)*，主题空间到词的条件概率 *p(w|z)*，主题空间到文档的条件概率 *p(d|z)*。

虽然 PLSI 也引入了主题空间的概念，但它的主题空间维度是靠经验事先决定的，没有明确的物理依据，一般在 20-100 之间。相比之下，本文提出的“一个句子多个主题”和“一个句子一个主题”算法生成的主题空间具有相对明确的物理意义，每个提问的空间维数也不相同。此外，PLSI 模型也是通过 EM 算法迭代估算的，其迭代过程需要一定的时间，而本文提出的算法不需要迭代运算，在速度上更具优势。

所以，在问答系统的句子检索应用中，本文提出的方法可能比 PLSI 算法更加合适。后续的实验部分有具体的性能对比。

6.4 实验部分

本章的实验部分是在第三章建立的汉语问答评测平台上进行的。实验的测试集是从这些提问集合中随机抽取的，且只包括答案类型是命名实体(人名、地名、机构名、时间词、数量词共 5 大类)的提问，共 807 个，如图 6-2 所示。

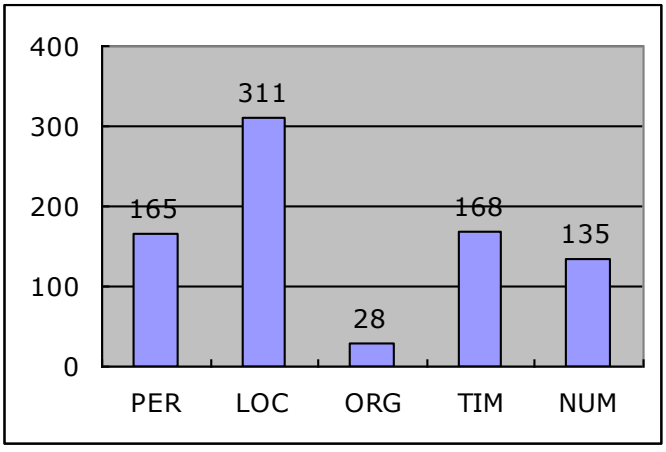

图 6-2 各类型提问的分布情况

表中的 LOC，ORG，PER，NUM 和 TIM 分别表示提问类型是地名、机构名、人名、数量词，时间词。

对于每个测试提问，系统返回前 1000 篇文档，并进行句子切分；再次检索，并返回前 1000 个句子；根据提问的答案类型，提取前 500 个包含候选答案的句子；依据候选答案对句子进行聚类；最后使用基于主题语言模型的检索算法在这 500 个句子的索引上进行最后的检索。

句子检索结果使用 MRR 进行“严格”评测，如公式 6-21 所示。“严格”评测要求系统所检结果不仅要包含提问答案而且要支持该答案才算正确。MRR1, MRR5 和 MRR20 分别表示系统仅返回概率最大的前 1、5 和 20 个检索结果时的 MRR 值。

$$MRR = \frac{\sum_{i=1}^{N} 1/\text{标准答案在系统给出的排序结果中的位置}}{N} \tag{6-21}$$

式中的 N 表示测试提问数。

6.4.1 Baseline 系统

本文以 6.2 节介绍的标准语言模型检索算法(公式 6-1)作为 Baseline 系统。表 6-4 给出 Baseline 在参数 α 在 0.6, 0.7, 0.8 ,0.9 时的 **MRR5** 性能指标。表中的 SUM 表示所有的提问类型。

表 6-4 BASELINE 系统中各类型提问检索性能

α	0.6	0.7	0.8	**0.9**
LOC	**49.95%**	**51.50%**	**52.63%**	54.54%
ORG	**53.69%**	**51.01%**	**50.12%**	51.01%
PER	**63.10%**	**64.42%**	**65.94%**	65.69%
NUM	**48.43%**	**49.86%**	**51.78%**	53.26%
TIM	**56.97%**	**58.38%**	**58.77%**	61.49%
SUM	53.98%	55.28%	56.40%	57.93%

接下来的实验将试图通过回答下列三个问题来检验本文提出的主题语言模型算法的性能。

1. 本文提出的基于主题语言模型的汉语问答系统句子检索算法是否较 Baseline 系统有提高？
2. 本文提出的两种聚类算法对各类型提问的聚类效果如何？
3. 本文提出的主题语言模型和 PLSI 模型在性能上有无差别？

6.4.2 主题语言模型与 Baseline 系统的对比实验

本节实验目的是检验基于“一个句子多个主题”和“一个句子一个主题”两种聚类思想的主题语言模型在应用于汉语问答系统句子检索时是否较 Baseline 有显著提高以及提高的幅度是多少？

6.4.2.1 基于“一个句子多个主题”聚类的主题语言模型

基于“一个句子多个主题”聚类思想的主题语言模型问答式句子检索算法的 **MRR5** 性能指标如表 6-5 所示。表中给出的是主题语言模型(公式 6-13)α 值分别在 0.6, 0.7, 0.8, 0.9 时的最优性能，括号里的数值是相对于 Baseline 的相对提高幅度。

表 6-5 基于“一个句子多个主题”主题语言模型中各类型提问检索性能

α	0.6	0.7	0.8	**0.9**
LOC	**55.57%** **(+11.2%)**	**55.61%** **(+7.98%)**	**56.59%** **(+7.52%)**	57.70% (+5.79%)
ORG	**59.05%** **(+9.98%)**	**59.46%** **(+16.6%)**	**59.46%** **(+18.6%)**	59.76% (+17.2%)
PER	**67.73%** **(+7.34%)**	**68.03%** **(+5.60%)**	**67.71%** **(+2.68%)**	67.45% (+2.68%)
NUM	**52.79%** **(+9.00%)**	**53.90%** **(+8.10%)**	**54.45%** **(+5.16%)**	55.51% (+4.22%)
TIM	**60.17%** **(+5.62%)**	**60.63%** **(+3.85%)**	**62.33%** **(+6.06%)**	61.68% (+0.31%)
SUM	58.14% (+7.71%)	58.63% (+6.06%)	59.30% (+5.14%)	59.54% (+2.78%)

从表 6-5 可以看出，相对于 Baseline 系统，基于“一个句子多个主题”聚类思想的主题语言模型对所有提问类型的句子检索均有不同程度的提高。例如，对于所有类型提问的最高和最低的提高幅度分别约 7.7%和 2.8%。该实验说明，基于“一个句子多个主题”的主题语言模型可以有效地提高汉语问答系统的句子检

索性能。

6.4.2.2 基于“一个句子一个主题”聚类的主题语言模型

基于“一个句子一个主题”聚类思想的主题语言模型问答式句子检索算法的MRR5 性能指标如表 6-6 所示。表中给出的是主题语言模型(公式 6-13)α 值分别在 0.6, 0.7, 0.8, 0.9 时的最优性能，括号里的数值是相对于 BASELINE 的相对提高幅度。

表 6-6 基于“一个句子一个主题”主题语言模型系统中各类型提问检索性能

α	0.6	0.7	0.8	**0.9**
LOC	**53.02% (+6.15%)**	**54.27% (+5.38%)**	**56.14% (+6.67%)**	56.28% (+3.19%)
ORG	**58.75% (+9.42%)**	**58.75% (+17.2%)**	**59.46% (+18.6%)**	59.46% (+16.6%)
PER	**66.57% (+5.50%)**	**67.07% (+4.11%)**	**67.44% (+2.27%)**	67.29% (+2.44%)
NUM	**49.95% (+3.14%)**	**50.87% (+2.02%)**	**52.15% (+0.71%)**	53.51% (+0.47%)
TIM	**59.75% (+4.88%)**	**60.65% (+3.89%)**	**62.71% (+6.70%)**	62.20% (+1.15%)
SUM	56.48% (+4.63%)	57.65% (+4.29%)	58.82% (+4.29%)	59.22% (+2.23%)

观察表 6-6 同样可以发现，本文提出的基于“一个句子一个主题”聚类思想的主题语言模型检索模型也能够有效提高汉语问答系统的句子检索质量。例如，对于所有类型提问的最高和最低的提高幅度分别约 4.6%和 2.2%。但和表 6-5 相比，基于“一个句子一个主题”聚类思想的主题语言模型句子检索的提高幅度没有“一个句子多个主题”好，这主要是因为“一个句子一个主题”的聚类“粒度”偏粗，效果不好。导致聚类效果下降的原因主要体现在 2 个方面：1、根据候选答案与所有查询词的平均距离选择的核心候选答案的方法可能会导致选择的错误；2、有的句子本来就属于多个主题的，把它仅分到任何一个主题也会导致聚类的噪音。

6.4.2.3 不同评测标准下系统的性能

本实验的目的是检验本文提出的主题语言模型方法在不同评测标准下得出

的结论是否具有一致性。表 6-7 是 Baseline 系统的 MRR1，MRR20 性能。

表 6-7 Baseline 系统的 MRR1 和 MRR20 性能

	Baseline	
α	***MRR1***	*MRR20*
0.6	**43.49%**	55.36%
0.7	**44.98%**	56.57%
0.8	**46.84%**	57.76%
0.9	49.07%	59.29%

表 6-8 给出基于“一个句子多个主题”和“一个句子一个主题”的主题语言模型检索系统的 MRR1 和 MRR20 性能。同样，括号里的数值是相对于 Baseline 模型的相对提高幅度。

表 6-8 不同算法的 MRR1 和 MRR20 性能指标对比

	一个句子多个主题		**一个句子一个主题**	
α	**MRR1**	**MRR20**	**MRR1**	MRR20
0.6	**50.00%** **(+14.97%)**	**59.60%** **(+7.66%)**	**48.33%** **(+10.37%)**	57.70% (+4.23%)
0.7	**50.99%** **(+13.36%)**	**60.03%** **(+6.12%)**	**49.44%** **(+9.92%)**	58.62% (+3.62%)
0.8	**51.05%** **(+8.99%)**	**60.68%** **(+5.06%)**	**51.05%** **(+8.99%)**	60.01% (+3.90%)
0.9	51.92% (+5.81) %	61.05% (+2.97%)	51.30% (+4.54%)	60.25% (+1.62%)

表 6-8 证明，在不同的测量标准下，本文提出的方法均较 Baseline 系统有不同程度的改进。例如，在 α=0.9 时，基于“一个句子多个主题”和“一个句子一个主题”聚类思想的主题语言模型的 *MRR1* 值相对于 Baseline 的绝对提高幅度分别约为 5.8%和 4.5%；而 *MRR20* 值绝对提高幅度分别约为 3.0%和 1.6%。

结论 1：实验证明，在多个不同的测量标准下，本文提出的两种聚类方法(“一个句子多个主题”和“一个句子一个主题”)均可以有效地改善汉语问答系统的句子检索质量；由于“一个句子多个主题”的聚类效果要优于“一个句子一个主题”，所以基于“一个句子多个主题”聚类的主题语言模型的系统性能提高的幅

度稍大。

6.4.3 聚类效果的分析

在本文提出的基于主题语言模型的检索算法中(公式 6-13)中，参数 β 反映了主题语言模型在整个模型中的权重。β 越大，主题语言模型所占的比例大。如果此时系统性能越好，则说明聚类的效果好；如果此时的系统性能越差，则说明聚类结果中存在较大的噪音。本实验目的即是为了检验本文提出的两类聚类方法对不同类型提问的聚类效果如何？

在参数 α=0.9 时，变化 β 值，基于“一个句子多个主题”聚类思想的主题语言模型句子检索性能曲线如图 6-3 所示。

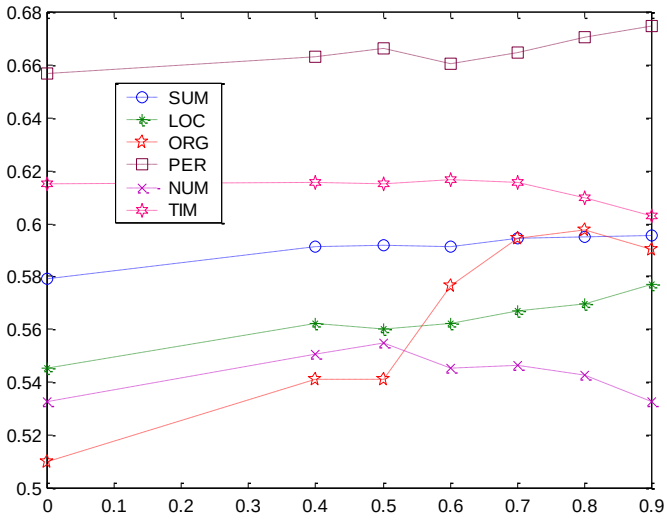

图 6-3 基于“一个句子多个主题”主题语言模型中各类型提问检索性能随参数 β 的变化曲线图

在图 6-3 中，随着 β 值的增加(0.6~0.9 区间段)，曲线呈明显下降趋势的主要是 TIM 和 NUM 提问类型；呈上升趋势的主要是 LOC，PER 和 SUM 提问类型。这说明“一个句子多个主题”聚类思想对时间和数量类型的提问的聚类效果比较差；而对地点和人名类型提问的聚类效果比较理想。这也是可以理解的，因为数量词和时间词在句中出现的相对较多，且本文没有对数量词和时间词进行细分类，所以，不是每个数量词和时间词的出现均表示一个主题。而地名和人名的出现则相对稳定，每出现一个地名和地名大致可以理解为一个新的主题的产生。

在参数 α=0.9 时，变化 β 值，基于“一个句子一个主题”聚类思想的主题语言模型检索性能曲线图如图 6-4 所示。

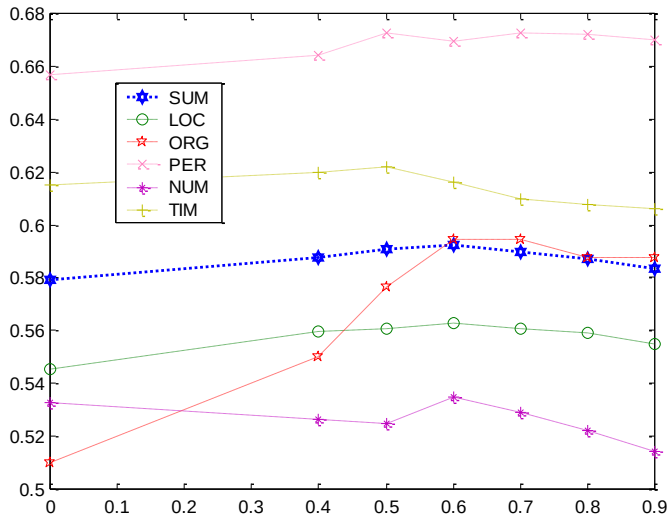

图 6-4 基于“一个句子一个主题”主题语言模型中各类型提问检索性能随参数 β 的变化曲线图

在图 6-4 中，随着 β 值的增加(0.6~0.9 区间段)，曲线呈明显下降趋势的包括 TIM，NUM，LOC 和 SUM。这说明“一个句子一个主题”聚类思想对大部分类型提问的聚类效果均不理想，要差于“一个句子多个主题”的聚类。虽然这种聚类效果不甚理想，但相对于 Baseline 系统，该方法仍然能够在一定程度上提高句子检索的质量。

结论 2：本文提出的“一个句子多个主题”聚类思想对人名、地名、机构名类型提问的聚类效果要优于对时间、数量类型提问的聚类效果，所以在人名、地名、机构名类型提问的“一个句子多个主题”主题语言模型中应该取较大的 β 值(0.8 或 0.9)，而对于时间和数量类型提问应取数值较小的 β 值(0.5 左右)。“一个句子一个主题”聚类对所有类型提问均逊色于“一个句子多个主题”聚类，所以，基于“一个句子一个主题”聚类思想的主题语言模型的 β 值不应该取的偏大，总体性能差不多在 β 值为 0.6 时达到最优。

6.4.4 基于 PLSI 模型的句子检索

本实验目的是对比本文提出的主题语言模型 Aspect Model 与 PLSI Aspect Model 在应用于汉语问答系统的句子检索时的性能差别：即句子检索的质量和速度。本文使用 Lemur 检索工具包中的 PLSI 代码，主题空间维度 z 取值 50。基于 PLSI 聚类思想的主题语言模型如公式(6-22)~(6-23)所示。

$$\begin{aligned} &p(w|M_S) = a \times p_{ML}(w|M_S) \\ &+(1-a)\times\left(\beta \times p_{ML}(w|Z)+(1-\beta)\times p_{ML}(w|M_C)\right) \end{aligned} \tag{6-22}$$

$$p(w|Z)=\sum_{z\in Z}p(w|z)p(z|M_S)\approx\sum_{z\in Z}p(z)p(M_S|z)p(w|z) \quad (6\text{-}23)$$

其中，Z 表示主题空间 z 的集合，$p(z)$表示主题空间的先验概率，$p(w|z)$表示主题空间到词的条件概率，$p(M_S|z)$表示主题空间到文档的条件概率。

基于 PLSI 聚类思想的主题语言模型汉语问答式句子检索算法的 MRR1、MRR5、MRR20 性能指标如表 6-9 所示。

表 6-9 PLSI 和 Baseline 在不同测量指标下性能对比

	Baseline			**PLSI**		
α	**MRR1**	**MRR5**	**MRR20**	**MRR1**	**MRR5**	MRR20
0.6	**43.49%**	**53.98%**	**55.36%**	**50.56%**	**58.94%**	60.27%
0.7	**44.98%**	**55.28%**	**56.57%**	**50.19%**	**58.81%**	60.18%
0.8	**46.84%**	**56.40%**	**57.76%**	**50.19%**	**58.76%**	60.20%
0.9	49.07%	57.93%	59.29%	50.62%	59.00%	60.35%

比较表 6-8、6-9 同样发现，在多种不同的测量标准下，本文提出的方法均稍好于 PLSI 方法，但这样的提高并不显著。然而，本文提出的方法在估算主题语言模型时不需要迭代运算，所以在速度上比 PLSI 方法要更具优势。表 6-10 给出 PLSI 对每种类型提问的迭代次数。

表 6-10 PLSI 进行迭代的次数

提问类型	**平均迭代次数**	
	PLSI	本文方法
LOC	**39**	1
ORG	**32**	1
PER	**29**	1
NUM	**48**	1
TIM	**33**	1
SUM	37	1

结论 3：针对汉语问答系统句子检索这个特殊课题，和 PLSI 相比，本文提出的方法虽然在 MRR 指标上并没有显著性的提高；但是在速度上，要比 PLSI 更

具优势。

6.4.5 基于常规伪相关反馈的句子检索

基于主题语言模型的句子检索算法也可以看作是一种针对问答系统的特殊性提出的伪相关反馈检索方法。所以，论文需要将其和常规伪相关反馈方法进行实验比较。

伪相关反馈通常包括 2 部分：一是统计出扩展词加入原始查询；二是基于扩展查询的二次检索。本文在实现常规伪相关反馈方法时采用 Robertson Selection Value (RSV)作查询词的扩展，如公式(6-24)~(6-26)。

$$\delta(t) = r \times \log \frac{p \times (1 - \bar{p})}{\bar{p} \times (1 - p)} \tag{6-24}$$

$$p = \frac{r}{R} \tag{6-25}$$

$$\bar{p} = \frac{n - r}{N - R} \tag{6-26}$$

其中，N 是整个语料库中总的句子档数；n 是整个语料库中包含该词的句子数；R 是相关句子数，也就是反馈句子数；r 是反馈文档中包含该词的句子数。

由于初始查询词比较少，为减少噪音扩展词，本文实验部分对于初始查询词少于 5 个词的提问，仅扩展一个查询词；对于初始查询少于 10 个词的提问，仅扩展 2 个查询词；其他情况扩展 3 个查询词。

基于扩展查询的二次检索采用公式(6-27)和(6-28)描述的模型。

$$p(eq \mid M_d) = \sum_{eq_i \in eq} p(eq_i \mid eq) \log p(eq_i \mid M_d) \tag{6-27}$$

$$p(eq_i \mid eq) = \begin{cases} 1 & \text{如果} eq_i \text{属于初始查询词} \\ cf & \text{如果} eq_i \text{属于扩展查询词} \end{cases} \tag{6-28}$$

其中，eq 表示扩展后的查询，eq_i 表示第 i 个扩展查询词，$0 <= cf <= 1$。在本文的实验部分，cf 取值 0.3。

表 6-11 给出了基于常规伪相关反馈的汉语问答系统句子检索性能。括号里的数值是相对于 Baseline 模型的相对提高幅度。

6-11 基于常规伪相关反馈的句子检索

	基于常规伪相关反馈的句子检索		
	MRR1	**MRR5**	MRR20
SUM	47.46% (-3.67%)	55.42% (-4.33%)	57.14% (-3.63%)

表 6-11 显示，在多个评测指标下，基于常规伪相关反馈的问答系统句子检索性能较 Baseline 系统均有所下降。导致伪相关反馈句子检索方法性能下降的原因可能在于以下两个方面：(1) 句子本身比较短。因此，从句子中统计的 RSV 值相对于从文本中统计的 RSV 值，缺乏一定的合理性。(2) 从提问句中提取的查询词数量比较少，扩展查询词对性能的影响比较大。因为，对于短查询来说，所有的查询词都很重要，一旦扩展了错误的查询词，性能的下降将是必然的。

6.5 相关工作

关于问答系统中的句子/段落检索，已经开展了很多有意义的工作。这些工作可以大致分为基于向量空间模型、基于统计语言模型和基于统计翻译模型三类。

Hui Yang 等[H. Yang, *et al*. 2002]参加 TREC 的系统算是基于向量空间模型的句子检索算法的代表工作，其根据关键词的重要程度，把查询词分为：普通关键词(O)、扩展关键词(E)、基本名词短语(B)、引用词(Q)和其它关键词(T)等；再使用加权算法把这些关键词的权值进行累加；最后依据加权得分进行句子排序。

基于统计语言模型的问答系统段落或句子检索的代表工作是 Andres & Croft 等人于 2004 年提出的[A. C. Emmanuel, *et al*. 2004]。在标准语言模型检索算法中，假定文档的先验概率 *P(D)*是常数，在排序的过程中被省略了。但 Andres & Croft 提出从问题答案对中获取文档的先验概率，即 Answer Model，并和 Relevance Model 相结合进行问答系统中的段落检索。该方法的特色是将答案抽取模块的一些特征(提问的答案类型，WordNet 同义词等)加入到语言模型(Answer Model)的框架中进而提高段落检索的性能，但 Answer Model 是在有监督的情况下训练得到的，需要一定量的训练语料，从而限制了方法的使用。

文[V. Murdock, *et al*. 2004]借鉴统计翻译模型的思想来解决问答系统中的句子检索，把从提问到答案的查找过程看作是一个翻译过程，并使用 IBM Model 1 来刻画这个过程。同样的做法还有 Echihabi & Marcu[A. Echihabi, *et al*. 2003]和 Berger [A. Berger, *et al*. 2000]。这类方法也是有监督的机器学习，需要大量的训

练语料。

6.6 本章小结

问答系统的自然语言提问输入相对于传统文本检索的关键词输入包含了更多的信息，例如提问的答案类型，提问词之间的结构关系等等。这些丰富的提示信息可以应用于问答系统的各个环节。本文提出的基于主题语言模型的汉语问答系统句子检索利用问答系统中特有的提问分类信息(即提问的答案语义信息)对句子初检结果进行主题聚类，通过 Aspect Model 将句子所属的主题信息引入到语言模型中，从而获得对句子语言模型更精确的描述。

此外，对初检结果的聚类，本文还提出了“一个句子多个主题”和“一个句子一个主题”两种算法。相对于 PLSI 算法的主题空间维度，本文提出的主题空间具有更加明确的物理意义；由于模型的估算不需要迭代，在运行速度上更具优势。对比实验的结果表明，本文提出的方法相对于标准语言模型方法在应用于汉语问答系统句子检索模块有明显的提高。

为了主题划分的方便，本文只对答案类型是命名实体类型的提问进行了研究。对于答案类型为非命名实体的提问，例如，提问：“ftp 的中文全称是什么？”，本文的工作并没有涉及。结构信息和语义信息对于提高检索的性能同样具有非常重要的影响，本文的工作也还没有涉及。

所以，下一步可以对本文提出的方法进行以下几个方面的扩展研究：

- 对于答案类型是其它语义实体的提问进行研究，从而将本文的方法延伸到除 5 大类命名实体提问类型以外的其它提问类型。
- 目前，聚类的多少是完全由候选答案的多少确定的。这不可避免地引入一些噪音主题。可以参考文[H. J. Zeng, *et al*. 2005]中的方法，从候选答案中选择更加恰当的主题，剔出噪音。
- 在主题语言模型的框架下引入结构信息和语义知识。

6.7 本章研究成果

[1] 吴友政, 赵军, 徐波. 基于主题语言模型的中文问答系统句子检索算法. 已于 2005 年 11 月投《计算机研究与发展》.

[2] Youzheng Wu, Jun Zhao, Bo Xu. Cluster-based Language Model for Sentence Retrieval in Chinese Question Answering. 投 SIGHAN 2006，2006/05/12 出评审结果.

第七章 基于无监督学习的问答模式抽取技术

7.1 引言

由于自然语言本身的灵活性和多变性，对同一语义往往存在不同的表述，这使得对问答技术研究面临许多困难。在 TREC 评测的推动下，人们已经提出了一些解决方法，具有代表性有模板的方法[D. Ravichandran, *et al*. 2002; M. M. Soubbotin, *et al*. 2002]，推理的方法[D. Moldovan, *et al*. 2002]等。然而，目前的自动问答技术仍然面临许多难题。例如，在提问句和答案句字面不同的情况下，如何进行匹配，仍然是目前问答技术研究中需要解决的关键问题之一。

为了解决提问句和答案句在字符表面上的不匹配，最直接的方法就是把提问和答案句都表示成统一的语义表示形式，然后进行匹配，抽取答案。然而，在现阶段自然语言处理的各种底层技术仍然不完善和不成熟，对文本进行深层分析，从语义层面来处理语言的灵活性和多变性是一件十分艰难的任务。所以，人们已经开始把注意力从原来的基于深层文本分析方法转移到基于字符的表层文本分析技术上。实际上，语言的灵活性和多样性在一定程度上是可以通过基于字符的表层文本分析技术来获取。模式匹配技术即是这种方法的代表，它可以解决因提问句和答案句的表述不同给问答系统的设计带来的麻烦。已经有一些英文问答系统采用了这种技术，并在 TREC 评测中获得了很好的成绩[M. M. Soubbotin, *et al*. 2002]。本文也希望通过模式匹配技术来转化汉语问答系统答案抽取的难度，把从语义层面进行答案抽取的过程转变成模式匹配的过程。

将模式匹配技术应用于问答系统的代表性工作有 Soubbotin[M. M. Soubbotin, *et al*. 2002]，Ravichandran[D. Ravichandran, *et al*. 2002]， Du[Y. P. Du, *et al*. 2004]，Dumais[S. Dumais, *et al*. 2002]和 Zhang[D. Zhang ,*et al*. 2002]等。

Soubbotin[M. M. Soubbotin, *et al*. 2002]完全采用人工编写规则的方法获取问答模式。这种方法代价昂贵，劳动强度大，速度慢；且模式的扩大很困难，算法可移植性差。

所以，近些年问答模式的获取方法逐渐从人工组织的方法向机器学习的方法转变。2002 年 Ravichandran[D. Ravichandran, *et al*. 2002]提出了通过有监督机器学习从网络文本中自动提取 6 种，即 BIRTHYEAR, INVENTOR, DISCOVER, DEFINITION, WHY-FAMOUS LOCATION 等提问类型的答案模式。例如 INVENTOR 类型提问答案的一个模式为：“the <ANSWER> was invented by

<NAME>”。其中，ANWER 和 NAME 分别表示提问关键词和答案。这种方法使用用户提供的<提问，答案>对进行 Web 搜索，在对 AltaVista 返回的前 1000 篇文章进行后处理后，采用后缀树模型(Suffix Tree)提取字符表层模式。因此，它是一种有监督的机器学习的方法。此外，字符表层模式的缺点是无法解决 ANWER 和 NAME 之间的长距离依存关系以及缺乏良好的泛化性。

Du 等人[Y. P. Du, *et al*. 2004]于 2004 年提出的问答系统的答案模式学习方法类似于 Ravichandran 方法，也是一种基于有监督的机器学习方法，不同之处在于提问分类和模式的表示两个方面。Du 首先把提问关键词定义为 4 大类(Q_Foucs, Q_NameEntity, Q_Verb, Q_BNP)，然后对不同类型的提问学习其答案句的模式。例如 “What Q_BeVerb Q_Focus in Q_LCN” 提问类型的一个答案模式为：“, <A> Q_BeVerb Q_Focus in Q_LCN”。

通过前述分析可以发现，Ravichandran 和 Du 的方法均属于有监督的机器学习算法，算法的性能在很大程度上依赖于用户提供的<提问，答案>对。然而，由于答案表示形式的多样性，用户很难提供答案的所有可能出现形式。这在一定程度上影响着有监督问答模式学习算法的性能。例如，提问“毛泽东同志出生地是哪里？”，它的答案可能是“湖南”、“湖南省”、“韶山冲”、“韶山”和“韶山市上屋场”等等。

对此，本文提出了一种基于无监督学习算法的问答模式抽取技术，从互联网上抽取应用于汉语问答系统的答案模式。该方法和 Ravichandran，Du 等人工作的不同是：本文提出的无监督学习算法无需用户提供<提问，答案>对，只需用户提供每种提问类型两个或以上的提问实例，算法即可通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类等步骤完成该类型提问的答案模式的学习。其中，主题划分采用第六章提出的“一个句子多个主题”聚类方法，模式提取模块抽取字符表层模式和句法模式两类，而垂直聚类和水平聚类是主要思想是：如果从某两个主题中提取的模式具有较高的相似度，则它们均是同一类型提问的答案模式，应该对它们进行聚类。

所以，本文的算法可以很好的避免有监督学习算法的缺点：即因用户很难提供尽可能多的提问答案而造成算法性能的下降。其中，最主要是影响学习问答模式的数量。

当用户提供某提问类型两个提问实例时的算法流程图如图 7-1 所示。

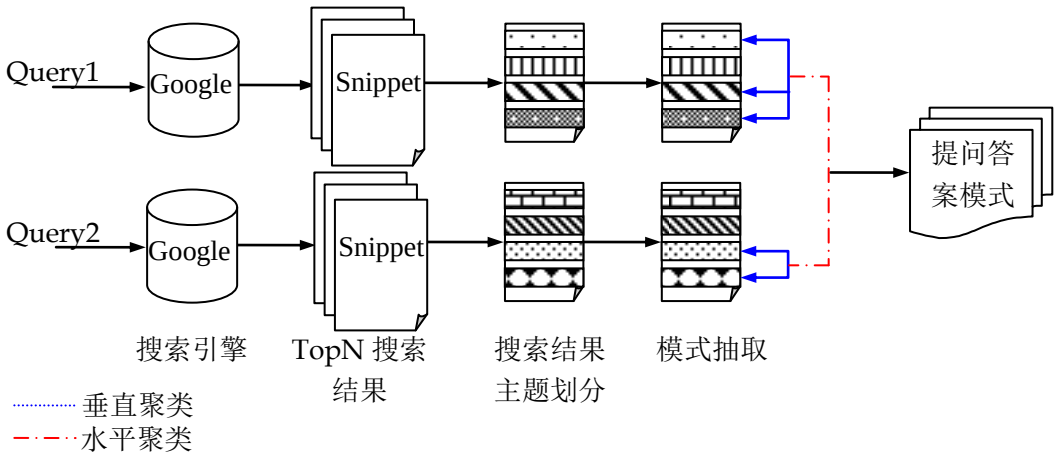

图 7-1 算法流程图

在测试语料上的实验结果表明：本文提出的基于无监督的问答模式抽取方法是有效的，能够较大幅度地提高汉语问答系统的性能。其中，基于字符表层模式的答案抽取系统性能较 Baseline 提高约 9.0%，基于句法模式的答案抽取系统性能较 Baseline 提高约 14.0%。

7.2 基于无监督的问答模式学习算法

本节以 BOOKAUTHOR 类型提问的答案模式抽取为例来说明无监督答案模式学习算法的整个流程，并详细介绍算法的三个核心模块：模式抽取、垂直聚类和水平聚类。

算法输入

Q3：《平凡的世界》是谁写的？

Q4：《西厢记》的作者是谁？

操作步骤：

1. 关键词的提取和分类

   其主要任务是对提问句进行分词；命名实体识别；提取查询关键词以及对查询关键词进行分类。例如 Q3 的查询 Query3=｛平凡的世界/Q_FOUCS，写/ Q_I｝；Q4 的查询 Query4=｛西厢记/ Q_ FOCUS，作者/ Q_I｝。其中，Q_FOUCS 表示提问焦点词，Q_I 表示除提问焦点词之外的其它查询词。

2. 查询词的扩展

   查询扩展的目的为了提高 Web 检索的召回率，本文使用《同义词词林》[梅家驹等. 1996]对 Q_I 类型的查询关键词进行扩展。

---

3. Web 检索

   提交查询 Query3 和 Query4 到搜索引擎(http://www.google.com)；提取 Google 返回的前 1000 个相关网页的片段，分别标记每个查询网页片段集合为 D3 和 D4

4. 句子切分和检索

   剔除 D3，D4 中的 Html 标记和其它标记，并分别对其进行句子切分；使用查询 Query3 和 Query4 进行基于语言模型的句子检索；保留同时包括 Q_FOUCS 和答案类型实体(对于提问 Q3 和 Q4，答案类型实体是人名)的所有句子，标记句子集合为 S3 和 S4

5. 主题划分

   对句子集合 S3 和 S4 采用第六章提出的“一个句子多个主题”方法进行主题划分。对 S1 主题划分后，包括 S1.1, S1.2, …, S1.M 等主题，对 S2 主题划分后，包括 S2.1, S2.2, …, S2.N 等主题，其中，M 和 N 分别表示 S3 和 S4 中的主题数目

6. 模式抽取

   从 S3 = {S1.1, S1.2, …, S1.M} 和 S4 = {S2.1, S2.2, …, S2.N} 的每个主题中分别提取字符表层模式和句法模式，并对应地标记为 P1 = {P1.1, P1.2, …, P1.M} 和 P2 = {P2.1, P2.2, …, P2.N}，其中 P1.i 和 P2.j 分别表示 P1 和 P2 中的一个模式集合。需要说明的是，模式集合中的某些模式可能是错误的。但通过垂直聚类和水平聚类，这些错误的模式均会被过滤

7. 垂直聚类

   如果 P1 ={P1.1, P1.2, …, P1.M}中的某几个模式集合具有较高的相似度，且相似度超过阈值 V1 时，进行垂直聚类；同样地，如果 P2 = {P2.1, P2.2, …, P2.N}中的某几个模式集合具有较高的相似度，且超过阈值 V1，也进行垂直聚类。垂直聚类后，P1 和 P2 中的模式集合将分别变为 P1’ = {P1’.1, P1’.2, …, P1’.m} 和 P2’ = {P2’.1, P2’.2, …, P2’.n}，其中，m 和 n 分别表示垂直聚类后模式集合的数目

8. 水平聚类

   如果垂直聚类后的 P1’ = {P1’.1, P1’.2, …, P1’.m} 和垂直聚类后的 P2’ = {P2’.1, P2’.2, …, P2’.n} 中的某几个模式集合具有较高的相似度，且相似度超过阈值 H1，进行水平聚类。水平聚类后的模式集合标记为 P ={P.1, P.2, …, P.K}，K 表示水平聚类后的模式集合的数目

9. 答案模式的识别

   经过垂直聚类和水平聚类后，如果 P = {P.1, P.2, …, P.K} 中的某个模式集合 P.k 是由最多的原始模式集合[26]组成的，则该模式集合 P.k 即为该类提问的答案模式集合

10. 模式评测

    使用文[D. Ravichandran, *et al*. 2002]中的方法评价 P.k 中每个模式的准确率，以便在答案抽取阶段按照模式准确率从最高到底进行匹配

输出样例(以*字符表层模式*为例)

1.0 ANSWER 成名作 ： Q_FOCUS

1.0 ANSWER 的 小说 Q_FOCUS

0.33 ANSWER 的 处女作 ， 小说 Q_FOCUS

0.44 Q_FOCUS 的 作者 ANSWER

7.2.1 模式抽取

模式的表示主要存在*字符表层模式*和*句法模式*两种形式。字符表层模式是根据词语在文本中出现的先后位置进行提取，主要存在两个缺点：(1) 由于无法处理长距离的依存关系，导致模式在一定程度上和训练语料相关性大，缺乏泛化能力；(2) 由于仅根据锚点词确定模式的长度，使得模式的完整性受到限制。例如，从提问“孙中山是哪一年出生的？”聚类结果“孙中山/PER 名/Vg 文/Ng ，/w 字/Vg 逸仙/PER ，/w 1866 年/TIM 生于/v 广东香山县/LOC”中提取的模式“Q_FOCUS 名 文 ， 字 逸仙 ，ANSWER”就存在这两方面的问题。表 7-1 给出了字符表层模式的部分实例。

表 7-1 字符表层模式的部分实例

Q_FOCUS ， 为 元代 著名 的 戏曲家 ANSWER
ANSWER 的 代表 作品 Q_FOCUS
ANSWER 的 处女作 ， 小说 Q_FOCUS
ANSWER 成名作 ： Q_FOCUS
ANSWER 的 小说 Q_FOCUS
Q_FOCUS 的 作者 ANSWER
……

[26] 原始模式集合是指在未进行垂直聚类和水平聚类时 P1 和 P2 中的模式集合

为了避免字符表层模式的这一缺点，本文尝试在依存句法分析树的基础上进行答案模式的抽取，这就是句法模式。因此，依存句法分析不仅可以反映出句子中各成分之间的语义修饰关系，而且它可以获得长距离的搭配，跟句子成分的物理位置无关[冯志伟著. 2004]。本文使用的汉语依存句法分析工具[X. D. Duan, *et al*. 2003]仅仅给出词语之间存在依存关系，并没有具体指出关系的种类，其准确率在 80%左右。表 7-2 给出了句法模式的部分实例。表 7-2 中的箭头是从被依存词语指向依存词语。

表 7-2 句法模式的部分实例

Q_FOCUS ← 是 → 处女作 → ANSWER
Q_FOCUS ← 代表作 → ANSWER
Q_FOCUS ← 是→ 作品 → ANSWER
Q_FOCUS ← 小说 → ANSWER
……

7.2.2 垂直聚类

提问的答案存在多种不同的表述形式，例如 BIRTHDATE 类型的提问“甘地是何时出生的？”，其答案的表述就包括：1869 年，1869 年 10 月，1869 年 10 月 2 日，一八六九年十月二日等等。所以，在有监督的机器学习算法中，用户必须提供该提问答案尽可能多的表述以提高答案模式的召回率。而本文提出的基于无监督汉语问答系统的问答模式学习算法是通过垂直聚类实现的。因为聚类的过程是在一个提问的检索结果内进行的，不涉及到其它提问实例，所以称之为垂直聚类。

垂直聚类是基于这样的假设：如果某个提问中的某几个聚类结果都包含该提问的答案，那么从这些聚类中抽取的模式集合应该存在一定程度的相似度，当相似度大于某个阈值 V1 时，应该合并这些相似的模式集合。

例如，从提问 Q6“《悲惨世界》的作者是谁？”的某 2 个聚类结果中提取的模式集合分别如表 7-3 和 7-4 所示。

很显然，这两个模式集合都是该提问类型的答案模式，该把它们聚为一类。本文采用的垂直聚类相似度计算公式如公式 7-1~7-2 所示。

$$sim(VC_i, VC_j) = \sum sim(VC_{im}, VC_{jn}) \tag{7-1}$$

$$sim(VC_{im}, VC_{jn}) = \begin{cases} 1 & if \quad VC_{im} = VC_{jn} \\ 0 & else \end{cases} \tag{7-2}$$

其中，VC_i 和 VC_j 分别表示第 i 和第 j 个模式集合；VC_{im} 和 VC_{jn} 分别表示第 i 个模式集合中的第 m 个模式和第 j 个模式集合中的第 n 个模式。

表 7-3 从 Q6 的一个聚类结果中提取的模式集合

<Cluster>
<ClusterNo> 雨果 </ClusterNo>
Q_FOCUS ANSWER
Q_FOCUS 作者 ： ANSWER
ANSWER 的 主要 作品 有 Q_FOCUS
ANSWER 的 小说 Q_FOCUS
ANSWER 代表作 Q_FOCUS
.....
</ Cluster>

表 7-4 从 Q6 的一个聚类结果中提取的模式集合

< Cluster >
< ClusterNo > 维克多·雨果 </ ClusterNo >
ANSWER 的 长篇小说 Q_FOCUS
Q_FOCUS 是 法国 大 文豪 ANSWER
ANSWER 作品集 含 Q_FOCUS
Q_FOCUS ANSWER
.....
</ Cluster >

7.2.3 水平聚类

通过垂直聚类可以获得关于某个提问的很多模式集合，但此时并不知道到底哪个模式集合才是提问答案的模式集合。对此，本文提出了水平聚类方法。因为聚类的过程是在不同提问的检索结果内进行的，所以称之为水平聚类。

水平聚类的基本思想是：从某类型提问的 2 个提问实例中学习到的所有模式集合中，如果存在提问 1 中的某个模式集合和提问 2 种的某个模式集合具有较高

的相似度，即大于阈值 H1，则这两个模式集合应该合并。

经过垂直聚类和水平聚类后，如果某个模式集合是由最多的原始模式集合组成的，则该模式集合即是该类型提问答案的模式集合。

例如，从提问 Q6“《悲惨世界》的作者是谁？”的一个聚类结果中提取的模式集合如表 7-5 所示，从提问 Q7“《平凡的世界》是谁的作品？” 的一个聚类结果中提取的模式集合如表 7-6 所示。

表 7-5 从 Q6 中提取的一个模式集合

< Cluster >
< ClusterNo > 维克多 雨果 </ ClusterNo >
Q_FOCUS ANSWER
Q_FOCUS 作者 ： ANSWER
ANSWER 的 主要 作品 有 Q_FOCUS
ANSWER 的 小说 Q_FOCUS
ANSWER 代表作 Q_FOCUS
ANSWER 的 长篇小说 Q_FOCUS
ANSWER 作品集 含 Q_FOCUS
.....
</ Cluster >

表 7-6 从 Q7 中提取的一个模式集合

< Cluster >
< ClusterNo > 路遥 </ ClusterNo >
Q_FOCUS 作者 ： ANSWER
Q_FOCUS 作者 ANSWER
Q_FOCUS ANSWER
ANSWER 的 小说 Q_FOCUS
ANSWER 的 Q_FOCUS
ANSWER - Q_FOCUS
……
</Cluster>

由于 Q6 和 Q7 是两个相同提问类型的提问实例，因此，如果这两个模式聚

类集合都是各自提问的正确模式集合，则它们之间应该具有较高的相似度；如果有一个模式集合不是对应提问的模式集合，则它们的相似度会比较低。很显然，表 7-5 和表 7-6 都是 BOOKAUTHOR 类型提问的答案模式集合，应该合并。本文采用的水平聚类相似度计算公式如公式 7-3~7-4 所示。

$$sim(HC_i, HC_j) = \sum sim(HC_{im}, HC_{jn}) \tag{7-3}$$

$$sim(HC_{im}, HC_{jn}) = \begin{cases} 1 & if \quad HC_{im} = HC_{jn} \\ 0 & else \end{cases} \tag{7-4}$$

其中，HC_i 和 HC_j 分别表示第 i 和第 j 个模式集合；HC_{im} 和 HC_{jn} 分别表示第 i 个模式集合中的第 m 个模式和第 j 个模式集合中的第 n 个模式。

通过 7.2 节的分析可以发现，通过垂直聚类和水平聚类，学习算法可以自动过滤非提问答案模式的集合，从而提炼出提问答案的模式集合，这就是本文的基于无监督学习方法不需要用户提供<提问，答案>对作为训练集的根本原因。

7.3 实验结果与分析

本文主要针对下列提问类型 (如表 7-7 所示)进行答案模式的自学习，并把学习到的模式应用于汉语问答系统中以验证模式的性能。

表 7-7 学习模式的提问类型

提问类型	训练提问数	测试提问数	提问类型	训练提问数	**测试提问数**
INVENTOR	**4**	**13**	**ADDRESS**	**7**	51
BOOKAUTHOR	**4**	**20**	**BIRTHPLACE**	**11**	19
PERSONSYN	**3**	**13**	**BIRTH**	**5**	7
OLDNAME	**5**	**5**	**DEATH**	**5**	1
POSITION	**7**	**3**	**EVENTDAY**	**7**	7
LOCATIONSYN	**2**	**24**	**LENGTH**	**5**	5
CAPITAL	3	5	POPULATION	4	5

表 7-7 中的训练提问数是指无监督自学习时使用的提问实例个数；测试提问数是指测试阶段的提问实例个数。

7.3.1 评测主题划分

通过前述章节的介绍可以知道，问答模式的提取和聚类都是在主题划分的基础上进行的，可以说，主题划分性能的好坏直接影响模式学习算法的性能。本实验的目的是使用人工评测的方法评测主题划分算法的性能。

对于每个提问的主题划分结果，评测人员只对正确答案的提问聚类结果进行人工评测，而不对其它聚类结果进行判断。因为非正确答案的聚类结果不会影响后续的模式提取和聚类。在正确答案的聚类结果中，只有那些既包含提问答案又能支持提问答案的句子才被判别为正确的聚类结果。例如，句子“韶山的山美，水甜，是一个美丽而神奇的地方。”就不能算作提问“毛泽东生于哪里？”的正确聚类。具体的人工评测性能如表 7-8 所示。

表 7-8 主题划分算法的性能

提问类型	SLM	提问类型	**SLM**
INVENTOR	**0.7769**	**BIRTHPLACE**	0.8179
BOOKAUTHOR	**0.8244**	**CAPITAL**	0.9709
PERSONSYN	**0.9397**	**BIRTH**	0.9496
OLDNAME	**0.7751**	**DEATH**	0.92
POSITION	**0.9288**	**EVENTDAY**	0.9397
LOCATIONSYN	**0.8517**	**LENGTH**	0.7644
ADDRESS	**0.8205**	**POPULATION**	0.8333
SUM	0.8554		

从表 7-8 可以看出，INVENTOR，OLDNAME 和 LENGTH 类型提问的主题划分结果相对较低，其它类型提问的主题划分性能超过 80%。虽然这几类提问的主题划分结果并不理想，但这并不影响答案模式抽取的准确率。不正确的聚类结果中绝大多数是包含答案但不能支持答案的句子，例如，缺少 Q_FOUCS 词，所以基本上是不能从这些句子上抽取答案模式的。

本实验结果也表明，基于“一个句子多个主题”主题划分算法是有效的，在此基础上进行答案模式的抽取和聚类也是可行的。

7.3.2 评测基于模式匹配的答案抽取系统

通过主题划分、模式抽取，垂直聚类和水平聚类后，本文提取的各类型提问的答案模式情况如表 7-9 所示。其中，SUP 和 SYP 分别表示字符表层模式和句法模式。

表 7-9 各类型提问答案模式的数量

提问类型	SUP	SYP	提问类型	SUP	**SYP**
INVENTOR	**148**	**137**	**BIRTHPLACE**	**68**	83
BOOKAUTHOR	**141**	**132**	**CAPITAL**	**205**	322
PERSONSYN	**134**	**153**	**BIRTHTIME**	**43**	22
OLDNAME	**150**	**94**	**DEATHTIME**	**18**	13
POSITION	**233**	**237**	**EVENTDAY**	**128**	176
LOCATIONSYN	**107**	**31**	**LENGTH**	**77**	144
ADDRESS	**108**	**191**	**POPULATION**	**36**	45
SUM	1596	1780			

接下来的实验是将学习到的各种类型提问答案的模式应用于汉语问答系统中，同时使用准确率评测指标(P)验证其性能，如公式 7-5 所示。

$$P = \frac{\text{系统正确回答的提问数}}{\text{所有待测试的提问数}} \tag{7-5}$$

7.3.2.1 基于检索的答案抽取系统

论文在第六章提出了基于主题语言模型的检索算法提高汉语问答系统句子检索性能，从而提高问答系统答案抽取模块的性能和速度。所以，本节采用基于标准语言模型句子检索算法的答案抽取系统(SLM)和基于主题语言模型句子检索算法的答案抽取系统(TLM)为 Baseline。表 7-10 给出了两个 Baseline 系统的准确率性能。

表 7-10　SLM 系统和 TLM 系统的 Precision 性能对照表

提问类型	SLM	TLM	提问类型	SLM	**TLM**
INVENTOR	**0.3077**	**0.3846**	**BIRTHPLACE**	**0.5263**	0.3684
BOOKAUTHOR	**0.55**	**0.5**	**CAPITAL**	**0.8**	1.0
PERSONSYN	**0.5384**	**0.7692**	**BIRTH**	**0.1429**	0.5714
OLDNAME	**0.2**	**0.6**	**DEATH**	**0**	1.0
POSITION	**0.6667**	**1.0**	**EVENTDAY**	**0.4286**	0.2857
LOCATIONSYN	**0.2917**	**0.3333**	**LENGTH**	**0**	0.4
ADDRESS	**0.1373**	**0.1569**	**POPULATION**	**0.2**	0.2
SUM	0.3258	0.3876			

表 7-10 给出的结论再次证明了基于主题语言模型的句子检索算法能够明显提高基于标准语言模型检索的句子检索算法的检索性能。基于 TLM 的答案抽取系统性能相对于基于 SLM 的答案抽取系统有显著的提高，提高幅度约为 6.2%。

7.3.2.2　基于模式匹配的答案抽取系统

本实验的目的是为了验证基于无监督的模式学习算法的性能，并分别对字符表层模式(SUP)和句法模式(SYP)进行对比实验。表 7-11 是基于标准语言模型句子检索算法的两种模式答案抽取系统的准确率性能对比。

表 7-11 基于 SLM 检索算法的 SUP 答案抽取系统和 SYP 答案抽取系统的性能

提问类型	SUP	SYP	提问类型	SUP	**SYP**
INVENTOR	**0.1538**	**0.30777**	**BIRTHPLACE**	**0.3158**	0.2105
BOOKAUTHOR	**0.6**	**0.85**	**CAPITAL**	**0.8**	0.8
PERSONSYN	**1.0**	**0.8461**	**BIRTH**	**0.1429**	0.4286
OLDNAME	**0.6**	**0.6**	**DEATH**	**1**	1.0
POSITION	**0.6667**	**0.6667**	**EVENTDAY**	**1.0**	0.8571
LOCATIONSYN	**0.2917**	**0.375**	**LENGTH**	**0.2**	0.4
ADDRESS	**0.3333**	**0.3333**	**POPULATION**	**0.2**	0.2
SUM	0.4157	0.4719			

对比表 7-10 和表 7-11 发现，相对于基于标准语言模型检索算法的答案抽取系统，基于字符表层模式的答案抽取系统性能和基于句法模式的答案抽取系统性能都得到了较大幅度的提高，其绝对提高幅度分别约为 9.0%和 14.0%。此外，基于句法模式的答案抽取系统性能比字符表层模式的答案抽取系统性能有所提高，其绝对提高幅度约为 4.6%。这说明句法模式要比字符表层模式效果好，更适合用于问答系统的答案抽取环节。

基于主题语言模型句子检索算法的两种模式答案抽取系统的准确率性能对比如表 7-12 所示。

表 7-12 基于 TLM 检索算法的 SUP 答案抽取系统和 SYP 答案抽取系统的性能

提问类型	SUP	SYP	提问类型	SUP	**SYP**
INVENTOR	**0.1538**	**0.3846**	**BIRTHPLACE**	**0.3158**	0.1052
BOOKAUTHOR	**0.55**	**0.75**	**CAPITAL**	**1.0**	1.0
PERSONSYN	**1.0**	**1.0**	**BIRTH**	**0.2857**	0.7143
OLDNAME	**0.6**	**0.4**	**DEATH**	**1.0**	1.0
POSITION	**0.6667**	**0.6667**	**EVENTDAY**	**0.4286**	0.4286
LOCATIONSYN	**0.3333**	**0.3333**	**LENGTH**	**0.4**	0.6
ADDRESS	**0.3333**	**0.3725**	**POPULATION**	**0.2**	0.2
SUM	0.4270	0.4775			

对比表 7-10 和表 7-12 同样发现，相对于基于主题语言模型检索算法的答案抽取系统，基于字符表层模式的答案抽取系统性能和基于句法模式的答案抽取系统性能也得到了较大幅度的提高，其绝对提高幅度分别达到了 4.0%和 8.0%。此外，基于句法模式的答案抽取系统性能比字符表层模式的答案抽取系统性能有所提高，其绝对提高幅度约为 4.5%。

对照表 7-12 和表 7-11 发现，基于 SLM 和 TLM 检索算法的模式答案抽取系统之间的性能没有明显的变化。这是由于基于模式的答案抽取方法具有较高的准确率，句子检索性能对答案抽取的影响相对而言比较小。

7.3.2.3 对比分析字符表层模式和句法模式

7.3.2.2 节已经对字符表层模式和句法模式的性能进行了对比实验，并得出了结论：基于句法模式的答案抽取系统性能要明显优于基于字符表层模式的答案抽

取系统性能。

本实验则从另一个角度对字符表层模式和句法模式进行对比分析，并采用 NotNullPrecision 评测指标，如公式(7-6)所示。

$$NotNullP = \frac{\text{系统正确回答的提问数}}{\text{系统给出的答案非空的提问数}} \tag{7-6}$$

基于 SLM 和 TLM 两种句子检索算法的 SUP 和 SYP 答案抽取系统的 NotNullPrecision 性能如图 7-2 所示。

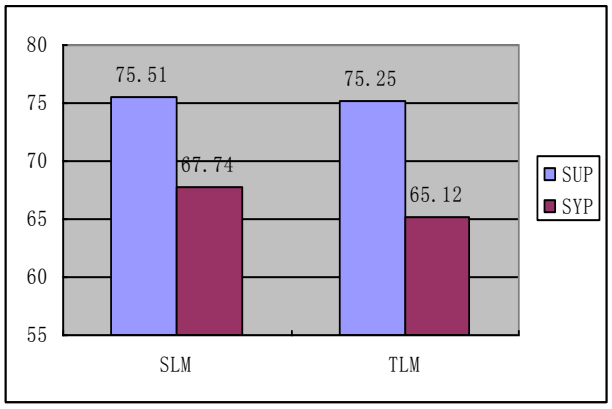

图 7-2 基于 SLM 和 TLM 两种句子检索算法的 SUP 和 SYP 两种答案抽取系统的 *NotNullPrecision* 性能

从图 7-2 可以看出，无论是基于 SLM 句子检索算法，还是基于 TLM 句子检索算法，基于 SUP 答案抽取系统的 NotNullPrecision 性能都优于基于 SYP 答案抽取系统的 NotNullPrecision 性能。导致这种现象的原因正如 3.3 节所分析的，句法模式比字符表层模式具有更好的泛化能力，从而使 NotNullPrecision 性能比较低。

7.4 本章小结

语言本身的灵活性和多变性常导致问答系统的提问和答案的不匹配。然而，从语义层面对这些这一现象进行分析到目前为止还是一件十分艰难的任务。所以，本文希望通过模式匹配技术解决这一问题。

对此，本文提出了一种基于无监督的学习算法从互联网中学习应用于汉语问答系统的问答模式。该方法和有监督机器学习算法的不同在于：无监督学习算法无需用户提供<提问，答案>对，只需用户对每种提问类型提供两个或以上的提问实例，算法即可通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类

等步骤完成该类型提问的答案模式的学习。

在测试语料上的实验结果表明：本文提出的无监督问答模式自学习的方法是有效的，能够较大幅度地提高汉语问答系统的答案抽取性能。

然而，为了主题划分的方便，本文只对答案类型是命名实体类型提问的答案模式抽取进行了研究。对于答案类型为非命名实体的提问，例如提问“ftp 的中文全称是什么？”，将在下一步的工作中展开。

此外，垂直聚类和水平聚类是无监督学习算法的两大核心技术，通过它们可以自动过滤非问答模式的集合，提炼出提问答案的模式集合，但目前的聚类相似度算法(7-1)~(7-4)还过于简单，下一步工作将对这一部分进行更加深入的研究。

7.5 本章研究成果

[1] Youzheng Wu, Jun Zhao, and Bo Xu. Two Novel Clustering Algorithms in Chinese Question Answering. 已投 EMNLP 2006. 2006/05/16 出评审结果.

[2] 吴友政, 赵军, 徐波. 无监督 Paraphrase 学习及其在中文问答系统中的应用. 已于 2005/03 投中文信息学报.

第八章 结论与展望

自然语言问答系统是集自然语言处理技术和信息检索技术于一身的新一代搜索引擎，它的出现旨在提供更有力的信息获取工具，以应对信息爆炸带来的严重挑战。

在 TREC 评测的推动下，英文问答技术已经取得了一定的成绩，并推出了一些成型的问答系统，比如 Ask Jeeves，AnswerBus 和 Start 等。相对于英文，中文问答技术的研究还相对少见，国内外从事中文问答技术研究的科研机构和高校还不多，而且基本没有成型的中文自动问答系统问世。本文就是在这样的情况针对汉语问答技术展开了深入的研究，以下是论文主要工作的概括。

8.1 结论

1、建立了一个具有一定规模并可扩充的汉语问答技术评测平台

系统化、大规模的定量评测对问答系统的研发有巨大的推动作用。然而，缺乏汉语问答系统评测机制已经成为制约汉语问答技术发展的主要障碍。本论文在吸收英文、日文和多语言问答系统评测的成功经验基础上，研发了面向汉语问答系统的评测平台。评测平台的语料来自互联网，规模约 1.8GB；平台测试集通过 4 种不同的渠道(例如，自然语言搜索网站日志、百科知识问答题库、实验室人员对热点问题的提问和对英语提问句的意译)进行收集，现有 7050 个汉语提问句(包括哈尔滨工业大学提供的 2800 个提问)；平台的打分标准主要是借鉴 TREC 的评分标准。

2、提出了面向汉语问答系统的提问分类体系以及基于支持向量机的提问分类算法

依据作者对提问分类的理解，本文从新的角度提出一种提问分类体系，即提问的技术分类和提问的语义分类。在提问分类体系的基础上，研究了基于多特征的提问分类算法。

与英文层级分类体系相比，论文提出的汉语平行分类体系的特点是，既能为提问选择最合适的技术方案，也能确定提问答案的语义类型。

实验数据表明：结构特征和词汇语义特征是提高分类性能的重要途径，但是依存关系结构特征对分类器的贡献并没有料想中的那样好，它的表现还略低于

Bi-gram 结构特征。基于 6150 个训练句和 700 个测试句，提问技术分类器和提问语义分类器的分类准确率分别达到了 96.20%和 94.37%。

3、设计并实现了基于多特征的汉语命名实体识别算法

针对汉语命名实体识别中的难点，本论文提出了基于多特征混合的汉语命名实体识别模型。该模型具有以下特点：

① 强调大颗粒度特征(词性特征)和小颗粒度特征(词形特征)的结合，以克服各自的缺点。词形特征虽然可以较好的刻画实体的内部特征和外部特征，但其颗粒度太小，数据稀疏问题严重削弱了它的作用；词性特征虽然数据稀疏问题不严重，但其刻画实体内部、外部特征的能力比较差。

② 提出了统计模型和专家知识相结合的方法，该方法通过限制候选命名实体的产生减少搜索空间，提高了识别速度。

③ 为准确刻画不同实体的内部特征，设计了多个细分类的实体模型；具体为，把人名实体划分为中国人名、日本人名、苏俄人名、欧美人名和人名简称；地名划分为单字地名和多字地名；机构名细分为简称机构名和全称机构名的实体模型。

模型在人民日报语料上的开放测试结果表明：基于多特征的汉语命名实体识别模型要优于使用单一特征的命名实体识别模型，其中对人名、地名和机构名的识别性能指标(精确率，召回率)分别达到了(94.06%，95.21%)，(93.98%%，93.48%)和(84.69%，86.86%)。在 MET-2 测试语料上的实验结果也证实了本文的模型在不同测试语料上的表现具有一致性。此外，该系统还参加了 2003 年和 2004 年的 863 命名实体识别评测专项，并获得了令人满意的成绩。

4、提出了基于主题语言模型的汉语问答系统句子检索算法

句子检索是问答系统检索模块的核心，其检索性能在一定程度上决定了问答系统的性能。对此，本文提出了基于主题语言模型的汉语问答系统句子检索算法可以归纳为：① 依据从初检结果中抽取的候选答案对初检结果进行聚类，即对初检结果进行主题划分；② 统计词语在主题上的概率分布以及句子关于主题的概率分布；③ 通过 Aspect Model 将句子所属的主题引入句子语言模型中，从而获得对句子语言模型更精确的逼近。这个新的语言模型较深入地刻画了词汇在不同主题下的分布规律以及文档所蕴含的不同主题的分布规律。针对初检结果的聚类问题，本文还提出“一个句子多个主题”和“一个句子一个主题”两种算法。相对于 PLSI 算法的主题空间维度，本文提出的主题空间具有更加明确的物理意义；另一方面，由于不需要迭代运算，运行速度更具优势。

对比实验的结果表明，本论文提出的基于主题语言模型句子检索算法相对于标准语言模型算法有比较明显的提高。其中，基于“一个句子多个主题”聚类思想的主题语言模型对句子检索性能的最大提高幅度约为 7.7%。

5、提出了基于无监督学习的问答模式抽取技术并将之应用于汉语问答系统的答案抽取

提问句和答案句的匹配是问答技术的核心内容，但由于自然语言本身的灵活性和多变性，对同一语义表述形式往往是不同的，这使得问答技术的研究面临许多困难。然而，从语义层面对这些表述进行分析到目前为止仍是一件十分艰难的任务。对此，本文研究通过模式匹配技术绕开语义匹配的难题，并提出了一种基于无监督学习的问答模式抽取技术，从网络文本中提取应用于汉语问答系统的问答模式。该方法和有监督机器学习算法的不同在于：无监督学习算法无需用户提供<提问，答案>对作为训练语料，只需用户对每种提问类型提供两个或以上的提问实例，算法即可通过 Web 检索、主题划分、模式提取(字符表层模式和句法模式)、垂直聚类和水平聚类等步骤完成该类型提问的答案模式的学习。

实验结果表明：本文提出的无监督问答模式学习的方法是有效的，能够较大幅度地提高汉语问答系统的性能。相对于基于检索的答案抽取系统，基于字符表层模式的答案抽取系统性能提高约 9.0%；基于句法模式的答案抽取系统性能提高约 14.0%。此外，基于句法模式的答案抽取系统性能比基于字符表层模式的答案抽取系统性能提高约 4.6%。这说明句法模式要比字符表层模式效果好，更适合用于问答系统的答案抽取环节。

8.2 展望

尽管本论文对汉语问答系统的相关技术进行了较深入的研究，取得了一些成绩，但是，问答技术要在实际应用中得到真正的推广和应用，还有许多工作要做，有待解决的问题和问答系统未来的发展趋势主要包括：

[1] 加强对汉语命名实体识别的研究

命名实体识别是应用很广泛的自然语言处理技术，在问答系统中，命名实体识别的目的主要是限制候选答案的产生，因而其准确率和召回率直接决定了问答系统的整体性能。加强对汉语命名实体识别的研究包括三个方面的内容：

① 提高现有命名实体类型(实体类、时间类和数量类实体)的识别性能

本论文采用产生式模型，即隐马尔柯夫模型，对汉语命名实体识别进行了研究。正如 Lafferty 在文[J. Lafferty, *et al*. 2001]所指出的，一方面，产生式模型仅

利用相邻特征，无法使用更为丰富的特征；另一方面，为了使得模型更容易操作和使用，产生式模型需要严格的假设条件，即观察值出现的概率只和当前状态有关，而与其它信息无关。为解决产生式模型的这些不足，有人提出使用判别式模型(如最大熵马尔柯夫模型)进行命名实体的识别，但判别式模型又容易出现标记偏置问题。近些年提出的条件随机场模型(CRF 模型)是一种新的条件概率模型，它不需要对观察序列建模，而且可以利用丰富的观察特征，同时又可以避免判别式模型的标记偏置问题。理论上，CRF 模型是最合适序列标注的概率模型之一。所以，下一步工作可以尝试采用 CRF 模型识别汉语命名实体，提高实体识别的性能。

② 对命名实体的细分类

对现有命名实体进行细分类将有助于提高汉语问答系统答案抽取模块的性能。比如机构名可以再细分类为公司、研究机构、企业等；时间类实体可在细分为年份、日期、时间等等。

③ 加强其他语义实体识别的研究

目前，对命名实体识别的研究主要还是集中在常规命名实体，如人名、地名和机构名等的识别上，下一步还应该加强其他语义实体识别的研究，比如产品名、创作作品名等等。

[2] 加强问答系统推理能力的研究

自然语言本身的灵活性和多变性使得对问答技术研究面临许多困难。本文是研究利用模式匹配技术来处理这些困难。但并不是所有的类型的提问都合适采用模式匹配技术，下一步应该加强从句法结构层面、语义层面对提问句和答案句进行匹配的研究。例如，LCC 系统通过句法分析、逻辑形式表达(Logic Form)、词汇链等实现答案的推理验证[S. Harabagiu, *et al*. 2000]。

[3] 加强对答案类型是非命名实体的问答研究

在现阶段论文中采用语义标注器只能识别常规的命名实体，因而研究的重点也就是答案类型是命名实体类的事实提问。在下一步的工作中，应该把重点转向答案类型是非命名实体的提问。这包括以下几方面的内容：

① 答案类型是非命名实体的事实提问

例如，对于提问“国际红十字会的旗帜是什么形状？”，语义标注器应该能够识别文本中的形状类实体。所以，这部分工作的重点应该是加强其他语义类别实体的标注工作。

② 非简单答案的提问

例如，定义类(什么是磁偏角？)、程序类(如何办理出国手续？)、描述类(禽流感的病人会出现哪些症状？)等类型的提问。这部分工作的重点应该是，针对每种类型提问的特殊性研究合适的问答技术。

③ 多视角提问

多视角问答技术是问答技术中的重要一类问题，它涉及单文档和多文档的情感、观点的分析计算。如，“哪些国家明确反对 2005 年美国人权报告？”、“网民大都对禽流感爆发持什么态度？”、“法国支持美国对伊拉克的战争吗？”等问题。多视角问答技术具有广泛的应用前景。例如，面向国家安全的舆情分析、信息过滤；面向新闻情报的收集管理；面向商务需求的市场调查、产品反馈；面向金融需求的预测、分析；面向政府管理的民意分析等。

[4] 加强对用户需求的研究

目前，问答系统所研究的是一些相对简单的问题，如事实类、列表类和定义类等。这些提问是用户真正关心的问题吗？这些提问对用户真实需求的覆盖率究竟是多少？在今后必须加强这方面的研究。

为了回答这些问题，本论文做了初步的探讨。从“百度知道[27]”中提取了 2000

表 8-1 各类型提问在百度知道中的分布情况

提问类型	数量	提问类型	**数量**
PERSON/人名	**76**	**PHONE-NUMBER/电话名**	4
PLACE/地名	**159**	**CREATION/创作作品名**	32
CITY/城市	**3**	**OTHER-ENTITY/ 其 他 实体**	125
COUNTRY/国家	**2**	**DEF/定义**	84
PATH/交通路径	**49**	**WHY-FAMOUS/身份**	38
ORG/机构名	**120**	**DES/描述**	209
TEMP/时间	**62**	**HOW/方式**	103
AGE/年龄	**12**	**WHY/原因**	140
NUM/数量	**68**	**YESNO/是非问**	192
ANIMAL/动物名	**4**	**CHOICE/选择问**	57
URL/超链接	**86**	**OPINION/意见类提问**	56
MONEY/货币	**16**	**OTHER-CLASS/其他类型**	421

[27] http://zhidao.baidu.com/

PLANT/植物名	5		

个真实的用户提问，涉及华人明星、自然科学、社会/文化、地区(北京)、体育/运动等五个领域，每个领域 400 个提问。根据第四章的汉语提问分类体系，论文对这 2000 个提问进行了分析，各类型提问的分布情况如表 8-1 所示。

对于表中的 OTHER-CLASS 类中的大部分提问，仅从提问本身是无法确定其类别的。例如提问“关于描写春天的诗”、“北京养花的朋友进来看？”、“土壤专业即相关专业的大师看过来！急！”等等。

百度用户提问的另一个特点是：用户提问中疑问词并不能直接决定提问的类型。例如“谁知道欧倍德在北京的详细店址？”、“谁能告诉我羽泉的官方网站？”等等。对于这些提问，用户并不关心谁会有这些信息，而是关心这些信息的具体位置。所以，对这些类用户提问的分类难度更大。

各类型用户提问所占的百分比如图 8-1 所示。

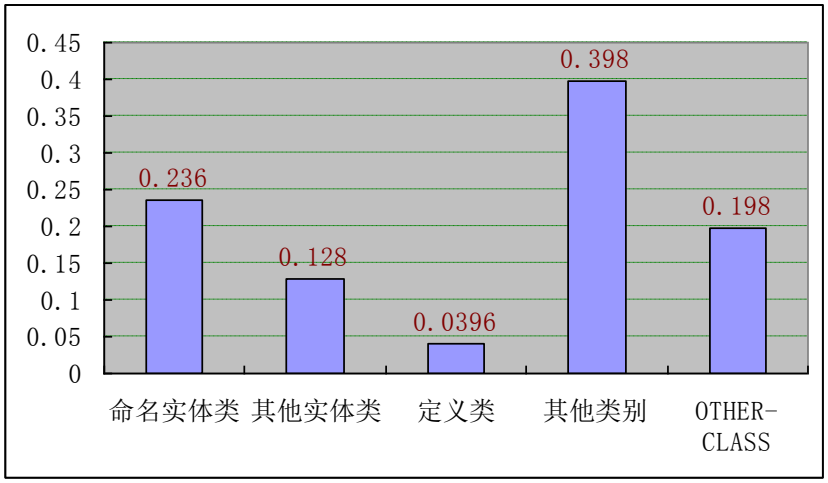

图 8-1 各类型用户提问所占的百分比

通过对“百度知道”的分析发现，目前问答系统的研究对象和用户真正的需求还有一定的差距。所以，对这些问题的回答将能够真正推动问答系统走向实用化。

[5] 问答系统的发展趋势

虽然问答技术和文摘技术各自都发展了很多年，但它们的结合却是未来的发展趋势。近几年的 TREC 和 DUC (Document Understanding Conferences)[28]也正是向这个方面发展的，例如 TREC 推出了 Series Questions，DUC 推出了 Question-focused 文摘都证明了这一点。所以在下一步工作中，应该加强这方面的研究。

[28] http://www-nlpir.nist.gov/projects/duc/

综上所述，对汉语问答技术的研究是一项复杂而充满挑战性的工作，目前的研究水平虽然离实用还存在相当大的差距，但社会的强烈需求将促进该领域研究水平的迅速提高。

希望本论文的研究能够对问答技术的发展做出一定的贡献，并真诚地期盼老师们、同行朋友们提出宝贵的批评和意见。

参 考 文 献

[A. Berger, et al. 1999] A. Berger and J. Lafferty. Information Retrieval as Statistical Translation. In Proceedings of ACM SIGIR-1999, pp. 222—229, Berkeley, CA, August 1999.

[A. Berger, et al. 2000] A. Berger, R.Caruana, D.cohn, D.Freitag, and V.Mittal. Briding the lexical chasm: Statistical approaches to answer-finding. In Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval, pp, 192-199, 2000.

[A. Borthwick. 1999] A. Borthwick. A Maximum Entropy Approach to Named Entity Recognition. PhD Dissertation. 1999.

[A. C. Emmanuel, et al. 2004] Andres Corrada-Emmanuel, W. Bruce Croft, Vanessa Murdock. Answer Passage Retrieval for Question Answering. In Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, pp. 516 - 517, 2004.

[A. Echihabi, et al. 2003] A. Echihabi, D.Marcu. A noisy-channel approach to question answering. In Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 16-23, Sappora, Japan, 2003.

[A. Ibrahim, et al. 2003] Ali Ibrahim, Boris Katz, Jimmy Lin. Extracting structural paraphrases from aligned monolingual corpora. In Proceedings of the Second International Workshop on Paraphrasing (IWP-2003), pp. 57-64 2003.

[A. Ittycheriah, et al. 2002] Abraham Ittycheriah, Salim Roukos. IBM's Statistical Question Answering System-TREC11. In the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 2002.

[A. Mikheev, 1998] A. Mikheev, C. Grover, Moens M. Description of the LTG System Used for MUC-7. In: Proceedings of 7th Message Understanding Conference (MUC-7), 1998.

[ACE] ACE. http://www.itl.nist.gov/iad/894.01/tests/ace/

[B. Magnini, et al. 2003a] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, et al. Creating the DISEQuA Corpus: a Test Set for Multilingual Question Answering. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway, 2003.

[B. Magnini, et al. 2003b] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Penas, V. Peinado, F. Verdejo, M. de Rijke. The Multiple Language Question Answering Track at CLEF 2003. Working Notes for the CLEF 2003 Workshop, pp. 223—228, 21-22 August, Trondheim, Norway, 2003.

[B. Wang, et al. 2001] B. Wang, H. Xu, Z. Yang, Y. Liu, X. Cheng, D. Bu, S. Bai, TREC-10 Experiments at CAS-ICT: Filtering, Web and QA. In The Tenth Text REtrieval Conference (TREC 10), page 109, 2001.

[C. Alvarez, et al. 2004] Carmen Alvarez, Philippe Langlais and Jian-Yun Nie. Word Pairs in Language Modeling for Information Retrieval. In Proceedings of RIAO 2004 Conference, France, 2004.

[C. J. C. Burges. 1998] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. The Fourth International Conference on Knowledge Discovery and Data Mining, 2(2), 1998.

[C. Zhai, et al. 2001a] C. Zhai and J. Lafferty. Model-based feedback in the KL-divergence retrieval model. In Tenth International Conference on Information and Knowledge Management (CIKM 2001), pages 403--410, 2001.

[C. Zhai, et al. 2001b] C. Zhai, J. Lafferty. A Study of Smoothing Techniques for Language Modeling Applied to ad hoc Information Retrieval. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 2001.

[CoNLL] CoNLL. http://cnts.uia.ac.be/conll2004/

[D. Ravichandran, et al. 2002] D. Ravichandran, E. Hovy. Learning Surface Text Patterns for a Question Answering System. In 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002) Conference, Philadelphia, PA, July 2002.

[D. K. Lin, et al 2001] Dekang Lin and Patrick Pantel. Discovery of inference rules for question-answering. In Natural Language Engineering, volume 7, pages 343-360, 2001.

[D. M. Bikel, et al. 1997] Daniel M. Bikel, Scott Miller, Richard Schwartz, Ralph Weischedel. Nymble: a High-Performance Learning Name-finder. In: Fifth Conference on Applied Natural Language Processing, (published by ACL), pp 194-201 (1997).

[D. Moldovan, et al. 2001] D. Moldovan, V. Rus. Logic Form Transformation of WordNet and its Applicability to Question Answering. In Proceedings of 37th Meeting of Association of Computational Linguistics (ACL'2001).

[D. Moldovan, et al. 2002] D. Moldovan, S. Harabagio, R. Girju, P. Morarescu, et al. LCC Tools for Question Answering. NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002).

[D. Moldovan, et al. 2003] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu. Performance Issues and Error Analysis in an Open-domain Question Answering System. ACM Trans. Inf. Syst., 21(2):133-154. 2003.

[D. Mollá 2003] D. Mollá Towards Semantic-Based Overlap Measures for Question Answering (2003). Proc. ALTW03, Melbourne, December 2003.

[D. R. Radev, et al. 2002] D. R. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. Probabilistic Question Answering from the Web. In Proceedings of the 11th World Wide Web Conference, Hawaii, 2002.

[D. Ravichandran, et al. 2002] D. Ravichandran, E. Hovy. Learning Surface Text Patterns for a Question Answering System. In 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002) Conference, Philadelphia, PA, July 2002.

[D. Zhang, et al. 2002] Dell Zhang, Wee Sun Lee. Web based Pattern Mining and Matching Approach to Question Answering. In Proceedings of the 11th Text REtrieval Conference (TREC), NIST, Gaithersburg, MD, Nov 2002.

[D. Zhang, et al. 2003] Dell Zhang, and Wee Sun Lee. Question Classification using Support Vector Machines. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, Page: 26 - 32, 2003.

[E. Brill, et al. 2001] Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais and Andrew Ng. Data-Intensive Question Answering. In Proceedings of the Tenth Text Retrieval Conference (TREC2001), November, 2001.

[E. D. Xun, et al. 2000] Endong Xun, Changning Huang, Ming Zhou. A Unified Statistical Model for the Identification of English BaseNP. In Proceedings of ACL 2000, Hong Kong, 2000.

[E. Hovy, et al. 2001] E. Hovy, U. Hermjakob, and Chin-Yew Lin. 2001. The Use of External Knowledge of Factoid QA. In Proceedings of the 10th Text Retrieval Conference(TREC2001), Gaithersburg, MD, U.S.A., November 13-16, 2001.

[E. Hovy, et al. 2002] E. Hovy, U. Hermjakob, D. Ravichandran. A Question/Answer Typology with Surface Text Patterns. Proceedings of the Human Language Technology (HLT) conference. San Diego, CA. NIST, Gaithersburg, MD, 229-241, 2002.

[E. Greengrass. 2001] Ed Greengrass. Information retrieval: A survey. DOD Technical Report TR-R52-008-001, 2001.

[E. M. Voorhees, et al. 1999] Ellen M. Voorhees, Dawn M. Tice. The TREC-8 Question Answering Track Evaluation. The Eighth Text REtrieval Conference (TREC-8), Spec Pub 500-246, Washington DC: NIST, 1999, 77-82, 1999.

[E. M. Voorhees. 2000] Ellen M. Voorhees. Overview of the TREC-9 Question Answering Track. The Ninth Text REtrieval Conference (TREC-9), Spec Pub 500-249, Washington DC: NIST, 2000, 77-82, 2000.

[E. M. Voorhees. 2001] Ellen M. Voorhees. Overview of the TREC2001 Question Answering Track. The Tenth Text REtrieval Conference (TREC-01), Spec Pub 500-250, Washington DC: NIST, 42-51, 2001.

[E. M. Voorhees. 2002] Ellen M. Voorhees. Overview of the TREC2002 Question Answering Track. The Eleventh Text REtrieval Conference (TREC-02), Spec Pub 500-251,Washington DC: NIST, 2002.

[E. M. Voorhees. 2003] Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), 2003.

[E. M. Voorhees. 2004] Ellen M. Voorhees. Overview of the TREC 2004 Question Answering Track. In Proceedings of the 13th Text REtrieval Conference (TREC 2004), 2004.

[E. Nyberg, et al. 2003] Eric Nyberg, Teruko Mitamura, Jaime Carbonell, Jaime

Callan, Kevyn Collins-Thompson, Krzysztof Czuba, Michael Duggan, Laurie Hiyakumoto, Ng Hu, Yifen Huang, Jeongwoo Ko, Lucian V. Lita, Stephen Murtagh, Vasco Pedro, David Svoboda. The JAVELIN Question Answering System at TREC 2002. In TREC 2002 Proceedings, 2003.

[F. Duclaye, et al. 2003] Florence Duclaye & Francois Yvon. Learning paraphrases to improve a question answering system. In Proceedings of the Natural Language Processing for Question Answering Workshop at EACL (EACL'03), Budapest, Hungary.

[F. Rinaldi, et al. 2003] F. Rinaldi, J. Dowdall, K. Kaljurand, M. Hess, D. Mollá Exploiting Paraphrases in a Question Answering System. Proc. Workshop in Paraphrasing at ACL2003, pp.25-32. July 11, Sapporo, Japan, 2003.

[G. D. Zhou, et al. 2003] G. D. Zhou, and Jian Su. 2003. Integrating Various Features in Hidden Markov Model Using Constrain Relaxation Algorithm for Recognition of Named Entities without Gazetteers. In Proceeding of 2003 International Conference on Natural Language Processing and Knowledge Engineering(NLP-KE). Oct. 26-29, 2003. Beijing, China. Pages 465-470.

[G. R. Krupka, et al. 1998] G. R. Krupka, K. Hausman. IsoQuest, Inc.: Description of the NetOwl TM Extractor System as Used for MUC-7. In Proceedings of the 7th Message Understanding Conference, 1998.

[H. B. Xu, et al. 2002] Hongbo Xu, Hao Zhang, Shuo Bai. ICT Experiments in TREC-11 QA Main Task. In the Eleventh Text REtrieval Conference (TREC 11), 2002.

[H. H. Chen, et al. 1998] Hsin-Hsi Chen, Yung-Wei Ding, Shih-Chung Tsai, et al. Description of the NTU System Used for MET2. In Proceedings of the Seventh Message Understanding Conference, 1998.

[H. J. Zeng, et al. 2004] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma. Learning to Cluster Web Search Results. In Proceedings of SIGIR2004. July 25~29, Sheffield, South Yorkshire, UK.

[H. P. Zhang, et al. 2003] Huaping Zhang, Qun Liu, Hongkui Yu, Xueqi Cheng, Shuo Bai. Chinese Named Entity Recognition Using Role Model. Special issue "Word Formation and Chinese Language processing" of the International Journal of Computational Linguistics and Chinese Language Processing, vol.8, No.2, 2003, pp. 29-60

[H. Yang, et al. 2002] Hui Yang, Tat-Seng Chua. The Integration of Lexical Knowledge and External Resources for Question Answering. In Proceedings of the Eleventh Text REtrieval Conference (TREC'2002), page 155-161, Maryland, USA, 19-22 Nov 2002.

[I. Soboro, et al. 2003] Ian Soboro, Donna Harman. Overview of the TREC 2003 Novelty Track. Text Retrieval Conference (TREC12), NIST, Maryland, USA, 2003.

[IEER] IEER. http://www.nist.gov/speech/tests/ie-er/er99/er99.htm

[J. Aberdeen, et al. 1995] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P.

Robinson, and M. Vilain. MITRE: Description of the Alembic system used for MUC-6. In Proceedings of the 6th Message Understanding Conference (MUC-6), pages 141--1552, November 1995.

[J. Brown. 2003] Jonathan Brown. Entity-Tagged Language Models for Question Classification in a QA System. Available at http://www-2.cs.cmu.edu/ jonbrown/IRLab/Brown-IRLab.pdf.

[J. Burger, et al. 2001] John Burger, Claire Cardie, Vinay Chaudhri, et al. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A) . http://www.ai.mit.edu/people/jimmylin/papers/ Burger00-Roadmap.pdf, 2001.

[J. F. Gao, et al. 2004] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu and Guihong Cao. Dependence Language Model for Information Retrieval. In Proceedings of ACM SIGIR-2004. Sheffield, UK, July 25-29, 2004.

[J. Fukumoto, et al. 2003] Junichi Fukumoto, Tsuneaki Kato and Fumito Masui. Question Answering Challenge (QAC1): An Evaluation of QA Tasks at the NTCIR Workshop 3. In Proc. of AAAI Spring Symposium: New Directions in Question Answering, pp.122-133, 2003.

[J. Kupiec. 1993]Julian Kupiec. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Special issue of the SIGIR FORUM, pages 181-190, 1993.

[J. Lafferty, et al. 2001] J. Lafferty and C. X. Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In Proceedings of ACM SIGIR-2001 Conference on Research and Development in Information Retrieval, 2001.

[J. Lafferty, et al. 2001] J. Lafferty and C. X. Zhai. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of ICML2001. pp.282~289.

[J. Ponte, et al. 1998] J. Ponte, W. Bruce Croft. A Language Modeling Approach to Information Retrieval. In Proceedings of ACM SIGIR 1998, pp 275-281, 1998.

[J. S. Chang, et al. 2004] Jing-Shin Chang, Yu-Tso Lai. A Preliminary Study on Probabilistic Models for Chinese Abbreviations. In SIGHAN2004. Barcelona. July 21-26, 2004.

[J. Sun, et al. 2002] Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, Changning Huang. Chinese Named Entity Identification Using Class-based Language Model. In Proceedings of COLING 2002. Taipei, August 24-25, 2002.

[J. Suzuki, et al. 2003] Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, et al: Question Classification using HDAG Kernel. In Proceedings of 6th Information-Based Induction Sciences. 2003.

[J. Y. Nie. 2005]Jian-Yun Nie. Integrating Term Relationships into Language Models for Information Retrieval. Report at ICT-CAS, 2005.

[K. C. Litkowski. 1999] Kenneth C. Litkowski. Question-Answering Using Semantic Triples. Eighth Text REtrieval Conference (TREC-8). Gaithersburg, MD. November 17-19, 1999.

[K. C. Litkowski. 2000] Kenneth C. Litkowski. Syntactic Clues and Lexical Resources in Question-Answering. Ninth Text REtrieval Conference(TREC-9). Gaithersburg, MD. November 13-16, 2000.

[K. Hacioglu, et al. 2003] Kadri Hacioglu, Wayne Ward, Question classification with support vector machines and error correcting codes. In Proceedings of the HLT-NAACL 2003--short papers, p.28-30, May 27-June 01, 2003, Edmonton, Canada.

[K. Kocik. 2003] Krystle Kocik. Question Classification using Maximum Entropy Models. Available at http://www.it.usyd.edu.au/research/news/kocik_ summary.pdf.

[L. Azzopardi, et al. 2004] Leif Azzopardi, Mark Girolami and Keith van Rijsbergen. Topic Based Language Models for ad hoc Information Retrieval. In Proceeding of IJCNN 2004 & FUZZ-IEEE 2004, July 25-29, 2004, Budapest, Hungary.

[L.D. Wu, et al. 2001] Lide Wu, Xuanjing Huang, Junyu Niu, et al. FDU at TREC-10: Filtering, QA, Web and Video Tasks. 10th Text REtrieval Conference, Gaithersburg, USA, Nov. 2001

[L.D. Wu, et al. 2002] Lide Wu, Xuanjing Huang, Junyu Niu, Yingju Xia, Zhe Feng, Yaqian Zhou. FDU at TREC2002: Filtering, Q&A, Web and Video tasks. 11th Text REtrieval Conference, Gaithersburg, USA, Nov. 2002

[M. Collins, et al. 1999] Michael Collins, Yoram Singer. Unsupervised models for named entity classi cation. In Proceedings of EMNLP, 1999.

[M. Collins, et al. 2002] Michael Collins. Ranking Algorithms for Named Entity Extraction: Boosting and the Voted Perceptron. In Proceeding of ACL 2002, pp489-496, 2002.

[M. M. Soubbotin, et al. 2001]M. M. Soubbotin, S. M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. Tenth Text REtrieval Conference (TREC-10) . Gaithersburg, MD. November 13-16, 2001.

[M. M. Soubbotin, et al. 2002]M.M. Soubbotin, S.M. Soubbotin. Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach. In the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 2002.

[M. Pasca. 2001] Marius Pasca. A Relational and Logic Representation for Open-Domain Textual Question Answering. ACL (Companion Volume), page 37-42, 2001.

[N. A. Chinchor. 1998] Nancy A. Chinchor. Overview of MUC-7/MET-2. In Proceedings of the Seventh Message Understanding Confernece (MUC-7), April 1998.

[P. Cohen, et al. 1998] Paul Cohen, Robert Schrag, Eric Jones, Adam Pease, Albert Lin, Barbara Starr, David Gunning, and Murray Burke. The DARPA high-performance knowledge bases project. AI Magazine, pages 25-49, Winter 1998.

[Q. L. Jin, et al. 2003] Qianli Jin, Jun Zhao, Bo Xu. NLPR at TREC2003 - Novelty and Robust Track. Text Retrieval Conference (TREC-12), NIST, Maryland, USA, 2003.

[Q. L. Jin, et al. 2004] Qianli Jin, Jun Zhao, Bo Xu. Window-based Method for Information Retrieval. The First International Joint Conference on Natural Language Processing. (IJCNLP-04), Hainan Island, China, 2004.

[R. Barzilay, et al. 2003] Regina Barzilay and Noemie Elhadad. Sentence Alignment for Monolingual Comparable Corpora. In Proceedings of EMNLP2003, pages 25-32, Sapporo, Japan.

[R. Grishman, et al. 1995] Ralph Grishman, Beth Sundheim. Design of the MUC-6 evaluation. In: 6th Message Understanding Conference, Columbia, MD, 1995.

[S. Dumais, et al. 2002] Susan Dumais, Michele Banko, Eric Brill, et al. Web Question Answering: Is More Always Better? , In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), August, 2002.

[S. H. Yu, et al. 1998] S. H. Yu, S. H. Bai, P. Wu. Description of the Kent Ridge Digital Labs System Used for MUC-7. In Proceedings of the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[S. Harabagiu, et al. 2000] Sanda Harabagiu, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunescu, Roxana Girju, Vasile Rus and Paul Morarescu. FALCON: Boosting Knowledge for Answer Engines. Proceedings of the Text Retrieval Conference (TREC-9). Gaithersburg, MD. November 13-16, 2000.

[S. J. Li, et al. 2002] Sujian Li, Jian Zhang, Xiong Huang and Shuo Bai. Semantic Computation in Chinese Question Answering System. Journal of Computer Science and Technology, 2002.

[S. Sekine, et al. 1998] S. Sekine, R. Grishman, H. Shinou. A decision tree method for finding and classifying names in Japanese texts. In Proceedings of the Sixth Workshop on Very Large Corpora, Canada, 1998.

[T. Hofmann. 1999] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, 1999.

[T. Joachims. 1998] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of the European Conference on Machine Learning, 1998, Berlin, pp. 137-142.

[T. S. Chua, et al. 2002] Tat-Seng Chua, Jimin Liu. Learning Pattern Rules for Chinese Named Entity Extraction. In Proceedings of AAAI'02, 2002.

[T. Solorio,et al. 2004] Thamar Solorio, Manuel P?erez-Couti?no, Manuel Montes-y-G?omez, Luis Villase?nor-Pineda, and A. L?opez-L?opez. A Language Independent Method for Question Classification. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

page1374-1380, 2004

[T. Tetsuro, et al. 2002] TAKAHASHI Tetsuro, NAWATA Kozo, KOUDA Shinya and INUI Kentaro. Seeking Answers by Structural Matching and Paraphrasing. NTCIR Workshop3 Meeting, Question Answering Task. 2002.

[T. Winograd, et al. 1977] Terry Winograd. Five Lectures on Artificial Intelligence. Linguistic Structures Processing, volume 5 of Fundamental Studies in Computer Science, pages 399- 520, North Holland, 1977.

[U. Hermjakob, et al. 2001] U. Hermjakob. Parsing and Question Classification for Question Answering. In Proceedings of the ACL Workshop on Open-Domain Question Answering, Toulouse, France, 2001.

[V. Lavrenko, et al. 2001] V. Lavrenko and W.B. Croft. Relevance-based language models. In Proceedings of ACM SIGIR-2001 Conference, pp. 120-127.

[V. Murdock, et al. 2004] Vanessa Murdock, W. Bruce Croft. Simple Translation Models for Sentence Retrieval in Factoid Question Answering. In Proceedings of the SIGIR 2004 Workshop on Information Retrieval for Question Answering, pp.31-35, 2004.

[W. A. Woods. 1977] W. A. Woods. Lunar rocks in natural english: Explorations in natural language question answering. Lingusitic Structures Processing, volume 5 of Fundamental Studies in Computer Science, pages 521-569, North Holland, 1977.

[W. Li. 2002] Wei Li. Question Classification Using Language Modeling. In CIIR Technical Report: University of Massachusetts, Amherst. 2002.

[W. J. Black, et al. 1998] W.J. Black, F. Rinaldi, D. Mowart. FACILE: Description of the NE System Used for MUC-7. In Proceedings of the MUC-7, 1998.

[X. D. Zhu, et al. 2003] Xiaodan Zhu, Mu Li, Jianfeng Gao and Chang-Ning Huang. 2003. Single character Chinese named entity recognition. In SIGHAN2002. Sapporo, Japan, 11-12, July, 2003.

[X. Y. Duan, et al. 2003] Xiangyu Duan, Jun Zhao, Bo Xu. 汉语依存句法分析器的构建. 组内报告, 2003.

[X. Y. Li, et al. 2001] Xiaoyan Li, W. Bruce Croft, Evaluating Question -Answering Techniques in Chinese. Computer Science Department University of Massachusetts, Amherst, MA , 2001.

[X.Li, et al. 2002] X.Li, and D. Roth. Learning Question Classification. In Proceedings of the 19th International Conference on Computational Linguistics (COLING2002), Taibai, 2002.

[X.Li, et al. 2003] Xin Li, Dan Roth, and Kevin Small. The Role of Semantic Information in Learning Question Classifiers. In Proceedings of NLPKE2003, Beijing, 2003.

[Y. M. Yang, et al. 1999] Yiming Yang and Xin Liu, "A Re-Examination of Text Categorization Methods". Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999, Pages 42 - 49.

[Y. P. Du, et al. 2004] Yongping Du, Xuanjing Huang, Xin Li and Lide Wu. A Novel Pattern Learning Method for Open Domain Question Answering. In Proceedings of IJCNLP2004, Sanya, China.

[Y. Sasaki, et al. 2005] Yutaka Sasaki, Hsin-Hsi Chen, Kuang-hua Chen, Chuan-Jie Lin. Overview of the NTCIR-5 Cross-Lingual Question Answering Task. In Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan.

[Y. Shinyama, et al. 2002] Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. Automatic Paraphrase Acquisition from News Articles. In Proceedings of Human Language Technology Conference, 2002, San Diego, USA.

[Y. Yang, et al. 1997] Y. Yang, Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of Fourteenth International Conference on Machine Learning Table of Contents, pages 412-420, 1997.

[Y. Z. Wu, et al. 2003] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Combining Statistical Model with Human Knowledge. In: The Workshop attached with 41st ACL for Multilingual and Mix-language Named Entity Recognition: Combining Statistical and Symbolic Models, pp.65-72, Sappora, Japan, 2003

[Y. Z. Wu, et al. 2005] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Model Based on Multiple Features. In Proceedings of HLT/EMNLP2005, Vancouver, B. C., Canada.

[Y. Z. Wu, et al. 2005] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Question Classification from Approach and Semantic Views. In Proceedings of AIRS2005, October 12~15, Korea.

[Y. Zhang, et al. 2003] Yi Zhang, DongMo Zhang. Enabling answer validation by logic form reasoning in Chinese question answering. In Proceeding of 2003 International Conference on Natural Language Processing and Knowledge Engineering. pages 275- 280, Beijing, 2003.

[崔恒等. 2004] 崔恒, 蔡东风, 苗雪雷. 基于网络的中文问答系统及信息抽取算法研究. 中文信息学报, 2004, 3.

[崔桓等. 2001] 崔桓, 蔡东风. 问答系统中疑问句理解的分析研究. 中国人工智能进展, 2001.

[冯志伟著. 2004] 冯志伟. 机器翻译研究. 中国对外翻译出版公司, 2004, 12.

[李斌. 2004] 李斌. 2004. 中文单字国名简称的自动识别. 第二届全国学生计算语言学研讨会论文集. 北京, 2004, 08.

[李鑫等. 2004] 第一届全国内容安全与信息检索学术会议论文集, 上海, 2003.

[孙茂松等. 1993] 孙茂松, 张维杰. 英语姓名译名的自动识别. 计算语言学研究与应用, 144-149 页. 1993

[孙茂松等. 1994] 孙茂松, 黄昌宁, 高海燕, 方捷. 中文姓名的自动辨识. 中文信息学报, Vol.9 No.2. 1994

[吴友政等. 2004] 吴友政, 赵军, 段湘煜, 徐波. 构建中文问答系统评测平台.

第一届全国信息检索与内容安全学术会议， 上海, 2004,11.

[吴友政等. 2005] 吴友政, 赵军, 徐波. 基于主题语言模型的汉语问答系统句子检索算法. 已投计算机研究与发展, 2005.

[吴友政等. 2005] 吴友政, 赵军, 徐波. 问答式检索技术及其评测研究综述. 中文信息学报, 2005 年第 3 期.

[张刚等. 2001] 张刚, 刘挺, 郑实福, 车万翔, 秦兵, 李生. 开放域中文问答系统的研究与实现. 中国中文信息学会二十周年学术会议, 2001,11

[张宇等. 2004] 张宇,刘挺.改进贝叶斯的提问分类技术. 第一届全国内容安全与信息检索学术会议论文集, 上海, 2004.

[郑家恒等. 2000] 郑家恒, 李鑫, 谭红叶. 2000. 基于语料库的中文姓名识别方法研究. 中文信息学报 Vol.14 No.1, 2000.

[郑实福等. 2002] 郑实福, 刘挺, 秦兵, 李生. 自动问答综述. 中文信息学报, 2002, 6.

[梅家驹等. 1996] 梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔. 《同义词词林》. 上海辞书出版社, 1996 年 5 月.

附录 A：命名实体识别结果样例

“/w 超级/b 女声/n ”/w 吸引/v 商家/n 眼球/n
昨晚/TIM ，/w 湖南卫视/ORG 《/w 超级/b 女声/n 》/w 节目/n 年度/n 总决赛/n 5/NUM 进/v 3/NUM 开赛/v 。/w 结果/n 周笔畅/PER 、/w 李宇春/PER 、/w 张靓颖/PER 晋级/v 《/w 超级/b 女声/n 》/w 前/f 三/NUM 名/q ，/w 纪敏佳/PER 、/w 何洁/PER 先后/d 被/p PK/nx 掉/v 。/w 而/c 随着/p 该/r 档/q 节目/n 在/p 市民/n 中/f 影响力/n 越来越/d 大/a ，/w 一些/NUM 精明/a 的/u 商家/n 抓住/v 机会/n ，/w 大/d 做/v “/w 超/h 女/b ”/w 文章/n 。/w
图/n 为/p 8 月 19 日/TIM 下午/TIM ，/w 在/p 浙江省/LOC 杭州市/LOC 建国北路/LOC 某/r 巧克力/n 作坊/n ，/w 两/NUM 位/q 女孩/n 举/v 着/u 自己/r 亲手/d 做/v 的/u 有/v “/w 超/h 女/b ”/w 李宇春/PER 漫画/n 像/v 的/u 巧克力/n 。/w

2005 年/TIM 他们/r 感动/v 中国/LOC (/w 图/n)/w
由/p 中央电视台/ORG 《/w 东方/s 时空/n 》/w 举办/v 的/u “/w 感动/v 中国/LOC ——/w 2005/NUM 年度/n 人物/n ”/w 评选/vn 活动/vn 昨晚/TIM 揭晓/v 。/w 在/p 2005 年/TIM ，/w 温暖/an 和/c 感动/v 我们/r 的/u 是/v ：/w 歌唱/vn 演员/n 丛飞/PER 、/w 农民工/n 魏青刚/PER 、/w 中国工程院/ORG 院士/n 黄伯云/PER 、/w 乡村/n 卫生员/n 李春燕/PER 、/w 青年/n 大学生/n 洪战辉/PER 、/w 留守/v 北大荒/LOC 的/u 上海/LOC 知青/n 陈健/PER 、/w 残疾人/n 舞蹈/n 演员/n 邰丽华/PER 、/w 二炮/AORG 某/r 基地/n 原/b 司令员/n 杨业功/PER 、/w 邮递员/n 王顺友/PER 、/w 航天员/n 费俊龙/PER 和/c 聂海胜/PER 。/w 青藏/ALOC 铁路/n 建设/v 者/r 获得/v “/w 感动/v 中国/LOC ——/w 2005/NUM 年度/n 特别奖/n ”。/w 图/n 为/p 带/v 妹妹/n 求学/v 12 年/TIM 的/u 大学生/n 洪战辉/PER （/w 右/f ）/w 与/p 主持人/n 白岩松/PER （/w 左/f ）/w 及/c 妹妹/n （/w 中/f ）/w 合影/v 。（/w 供/v 图/n CFP/nx ）/w (/w 来源/n ：/w 厦门/LOC 晚报/n)/w

日本/LOC 教练/n 木村昌彦/PER ：/w 柔道/n 是/v 斗智斗勇/l 的/u 项目/n
新华网/ORG 雅典/LOC 8 月 2 0 日/TIM 体育/n 专电/n （/w 记者/n 应强/PER 肖春飞/PER ）/w 奥运会/j 柔道/n 比赛/vn 2 0 日/TIM 在/p 雅典莱奥西亚奥林匹克中心/LOC 结束/v ，/w 在/p 全部/NUM 产生/v 的/u 1 4 /NUM 枚/q 金牌/n 中/f ，/w 日本柔道队/ORG 共/d 获得/v 8 /NUM 枚/q 金牌/n 、/w 2 /NUM 枚/q 银牌/n ，/w 重新/d 占据/v 了/u 世界柔坛/LOC 的/u 霸主/n 地位/n 。/w
8 /NUM 年前/TIM 的/u 亚特兰大/LOC 奥运会/j 上/f ，/w 日本

/LOC 柔道/n 陷入/v 低谷/n ，/w 仅/d 获得/v 3/NUM 枚/q 金牌/n ，/w 多年/NUM 的/u 霸主/n 地位/n 一时/TIM 土崩瓦解/i 。/w 在/p ２０００年/TIM 悉尼/LOC 奥运会/j 上/f ，/w 日本/LOC 柔道/n 选手/n 夺得/v 4/NUM 金/Ng ，/w 虽然/c 名列榜首/l ，/w 但/c 与/p 其他/r 柔道/n 强国/n 相比/v 优势/n 并/c 不/d 明显/a ，/w 在/p 悉尼/LOC 中国/LOC 、/w 法国/LOC 、/w 古巴/LOC 各/r 获/v 2/NUM 金/Ng 。/w 日本/LOC 人/n 用/v 了/u 8年/TIM 时间/n ，/w 终于/d 在/p 雅典/LOC 重振/v “/w 国/n 技/Ng ”/w 辉煌/a 。/w 他们/r 在/p 赛前/TIM 的/u 目标/n 是/v 5/NUM 金/Ng ，/w 结果/n 拿到/v 8/NUM 金/Ng ，/w 实现/v 了/u 历史/n 罕见/a 的/u 大/a 丰收/vn 。/w

日本/LOC 柔道/n 在/p 雅典/LOC 打/v 了/u 个/q 漂亮/a 的/u 翻身仗/n ，/w 原因/n 是/v 什么/r 呢/y ？/w 赛后/TIM 记者/n 采访/v 了/u 日本女子柔道队/ORG 主教练/n 木村昌彦/PER ，/w 他/r 说/v ，“/w 过去/TIM 欧洲/LOC 选手/n 在/p 大/a 级别/n 比赛/vn 中/f 占有/v 优势/n ，/w 他们/r 的/u 力量/n 比较/d 强/a ，/w 但是/c 柔道/n 是/v 一个/NUM 非常/d 讲究/v 技术/n 的/u 项目/n ，/w 光/d 有/v 力量/n 是/v 不行/a 的/u ，/w 我们/r 提高/v 了/u 力量/n ，/w 技术/n 又/d 占据/v 优势/n ，/w 自然/d 会/v 取得/v 好/a 成绩/n 。”/w

最后/f 木村昌彦/PER 还/d 说/v ，/w 柔道/n 是/v 斗智斗勇/l 的/u 项目/n ，“/w 除了/p 技术/n 、/w 力量/n 还/d 需要/v 有/v 头脑/n ”。/w

中国柔道协会/ORG 副/b 主席/n 、/w 国际/n 柔道/n Ａ/nx 级/n 裁判/n 宋兆年/PER 认为/v ，/w 日本队/ORG 取得/v 这样/r 的/u 成绩/n 主要/d 有/v 三/NUM 方面/n 原因/n ：/w 一/NUM 是/v 技术/n 细腻/a 、/w 手法/n 好/a ；/w 其次/c 是/v 他们/r 的/u 拼搏/vn 精神/n ，/w 许多/NUM 场次/q 在/p 落后/a 的/u 情况/n 下/f 反败为胜/i ；/w 第三/NUM 是/v 因为/c 有/v 比较/d 好/a 的/u 心态/n ，/w 日本/LOC 柔道/n 选手/n 的/u 正常/a 水平/n 基本上/d 都/d 能/v 在/p 比赛/vn 中/f 发挥/v 出来/v 。（/w 完/v ）/w

附录 B：北京大学汉语文本词性标注标记集

代码	名称	帮助记忆的诠释
Ag	形语素	形容词性语素。形容词代码为 a，语素代码 g 前面置以 A。
a	形容词	取英语形容词 adjective 的第 1 个字母。
ad	副形词	直接作状语的形容词。形容词代码 a 和副词代码 d 并在一起。
an	名形词	具有名词功能的形容词。形容词代码 a 和名词代码 n 并在一起。
b	区别词	取汉字“别”的声母。
c	连词	取英语连词 conjunction 的第 1 个字母。
Dg	副语素	副词性语素。副词代码为 d，语素代码 g 前面置以 D。
d	副词	取 adverb 的第 2 个字母，因其第 1 个字母已用于形容词。
e	叹词	取英语叹词 exclamation 的第 1 个字母。
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。
h	前接成分	取英语 head 的第 1 个字母。
i	成语	取英语成语 idiom 的第 1 个字母。
j	简称略语	取汉字“简”的声母。
k	后接成分	
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母。
m	数词	取英语 numeral 的第 3 个字母，n，u 已有他用。
Ng	名语素	名词性语素。名词代码为 n，语素代码 g 前面置以 N。

n	**名词**	取英语名词 noun 的第 1 个字母。
nr	**人名**	名词代码 n 和“人(ren)”的声母并在一起。
ns	**地名**	名词代码 n 和处所词代码 s 并在一起。
nt	**机构团体**	“团”的声母为 t，名词代码 n 和 t 并在一起。
nz	**其他专名**	“专”的声母的第 1 个字母为 z，名词代码 n 和 z 并在一起。
o	**拟声词**	取英语拟声词 onomatopoeia 的第 1 个字母。
p	**介词**	取英语介词 prepositional 的第 1 个字母。
q	**量词**	取英语 quantity 的第 1 个字母。
r	**代词**	取英语代词 pronoun 的第 2 个字母,因 p 已用于介词。
s	**处所词**	取英语 space 的第 1 个字母。
Tg	**时语素**	时间词性语素。时间词代码为 t,在语素的代码 g 前面置以 T。
t	**时间词**	取英语 time 的第 1 个字母。
u	**助词**	取英语助词 auxiliary
Vg	**动语素**	动词性语素。动词代码为 v。在语素的代码 g 前面置以 V。
v	**动词**	取英语动词 verb 的第一个字母。
vd	**副动词**	直接作状语的动词。动词和副词的代码并在一起。
vn	**名动词**	指具有名词功能的动词。动词和名词的代码并在一起。
w	**标点符号**	
x	**非语素字**	非语素字只是一个符号，字母 x 通常用于代表未知数、符号。
y	**语气词**	取汉字“语”的声母。
z	状态词	取汉字“状”的声母的前一个字母。

个人简历

吴友政，男，1976 年 12 月 14 日出生，安徽舒城人，中共党员。1995 年 9 月进入武汉水利电力大学动力工程系学习，1999 年 7 月获得工学学士学位。同年考取武汉大学自动化系攻读控制理论与控制工程硕士学位，2002 年 7 月获得工学硕士学位。2002 年 9 月进入中国科学院自动化研究所模式识别国家重点实验室攻读模式识别与人工智能专业博士学位至今，研究方向为自然语言处理。

在学期间发表的学术论文

1. Youzheng Wu, Jun Zhao, Bo Xu, Chinese Named Entity Recognition Model Based on Multiple Features. In Proceedings of HLT/EMNLP 2005, pp.427~434, October 6-8, Vancouver, B.C., Canada.
2. Youzheng Wu, Jun Zhao, Bo Xu. Chinese Question Classification from Approach and Semantic Views. In Proceedings of the 2nd Asia Information Retrieval Symposium (AIRS2005), pp. 485~490, October 13-15, 2005, Korea. (SCI)
3. Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge. In Proceedings of ACL2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, pp. 65-72, Sappora, Japan, July, 2003.
4. 吴友政, 赵军, 段湘煜, 徐波. 构建汉语问答系统评测平台. 第一届全国信息检索与内容安全学术会议, pp. 315~323, 2004.11, 上海.
5. 吴友政, 赵军, 段湘煜, 徐波. 问答式检索技术及其评测研究综述. 中文信息学报, pp. 1~13, 2005 年第 3 期.
6. 吴友政, 赵军, 徐波. 基于主题语言模型的中文问答系统句子检索算法. 已于 2005 年 11 月投《计算机研究与发展》.
7. Youzheng Wu, Jun Zhao, Bo Xu. Cluster-based Language Model for Sentence Retrieval in Chinese Question Answering. 已投 SIGHAN 2006, 2006/05/12 出评审结果.
8. Youzheng Wu, Jun Zhao, Bo Xu. Two Novel Clustering Algorithms in Chinese

Question Answering. 已投 EMNLP 2006. 2006/05/16 出评审结果.

9. 吴友政, 赵军, 徐波. 无监督 Paraphrase 学习及其在中文问答系统中的应用. 已于 2006 年 3 月投《中文信息学报》.

在学期间参加的科研项目

1. 2004.01 至今
   参与国家自然科学基金项目“多语言智能文本处理中基于主题语义空间的文本表示研究”。

2. 2005.08 至今
   参与北京市自然科学基金项目“面向异构 WEB 信息源的汉语问答式检索技术研究”。

3. 2005.6-2005.8
   参与由 2008 北京奥组委技术部支持的奥运项目“奥运综合信息服务系统”的详细规划。

4. 2003.10，2004.10
   参加 2003 年度、2004 年度 863 计划中文信息处理与智能人机接口技术评测-命名识别评测专项，多项评测指标(人名、地名、机构名)获得令人满意的好成绩。

5. 2004.7~2004.9
   参与国际合作项目“产品名识别和实体关系提取工具的研发”(与富士通研究开发中心有限公司合作)。

6. 2003.9~2004.3
   完成国际合作项目“汉语问答系统评测环境的建立以及汉语命名实体识别工具的研发”(与富士通研究开发中心有限公司合作)。

致 谢

在经历了无数的挫折和成功之后，所有的努力和汗水终于凝结成这篇论文。

本论文是在导师徐波研究员和赵军副研究员的悉心指导下完成的，是他们把我带入一个充满了挑战和无上吸引力的研究领域。在我求学期间，两位导师在学习、生活、思想上都给了我无微不至的关怀和照顾，为培养我花费了很大的心血，使我在学业上取得了很大的进步，在思想上更加成熟，在此对两位导师表示崇高的敬意和深深的感谢。

感谢中文信息处理组的宗成庆研究员、刘文举研究员以及模式识别国家重点实验室的所有老师们，是他们给与我最基本的知识和技能，指导我走向知识的殿堂。

感谢中文信息处理组的张艳、解国栋、胡日勒、刘非凡、段湘煜、周玉、金千里、刘丁、陈克利、徐晋、左云存、李幸、吕碧波、吴晓峰、陈钰枫、何彦青、徐昉以及实验室的诸位同学，是他们使我的生活变得丰富多彩。此外，曹文洁、李寿山、刘鹏和柴春光同学对论文的校正提出了很多宝贵的意见，在此一并表示感谢。

最后，我要特别感谢理解、关心和支持我的父母和弟弟，在我成长的过程中一直给予我希望和鼓励。他们的关怀和无私的爱是我不断前进的原动力。

寥寥数语，无以言谢。纸短情长，不尽欲言。谨以此文献给所有帮助过我的人们。