分类号 TP3　　　　　　　　　　　　　　　　密级 ____________

UDC ____________　　　　　　　　　　　　　编号 ____________

中国科学院研究生院

博士学位论文

语言浅层分析与句子级新信息检测研究

张华平

指导教师　白硕　研究员

中国科学院计算技术研究所

申请学位级别 工学博士　学科专业名称 计算机软件与理论

论文提交日期 2005 年 4 月　论文答辩日期 2005 年 6 月

培养单位 中国科学院计算技术研究所

学位授予单位 中国科学院研究生院

答辩委员会主席 ____________

声　　明

本人声明所呈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。就我所知，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的说明并表示了谢意。

作者签名：　　　　　　　　日期：

关于论文使用授权的说明

中国科学院计算技术研究所有权处理、保留送交论文的复印件，允许论文被查阅和借阅；并可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存该论文。

作者签名：　　　　　　导师签名：　　　　　　日期：

摘 要

针对粒度更小、冗余更少的信息需求，本文围绕句子级别的信息检索与新信息检测，进行了深入而又细致的研究，提出了有针对性的浅层语言分析技术，讨论了句子检索的主要建模方法，并阐述了信息新颖度的多种量化手段。在多组对比实验和国际上公开的评测比赛中，依据本文技术方法研制的 Noovel 系统取得了当前最好的性能，超过了所有正式公开的结果，这也表明：本文提出的句子检索方法与新信息检测技术是卓有成效的。

针对新信息检测的英文浅层语言分析主要包括断句、词汇切分、词性标注以及词形还原等自然语言处理过程。作者在已有研究工作的基础上，结合新信息发现的特点，提出了有针对性的改进措施。在中文浅层语言分析方面，本文提出了一种将汉语分词、词性标注、切分排歧和未登录词识别相结合的基于层次隐马模型的理论框架。在语言的分析基础上，查询分析过程通过辅助词过滤与倾向分类，从自然语言表述的主题中理解用户的查询意图，从而抽取出可用于直接建模计算的查询向量。在目前所有能获取的公开数据集合上进行对比实验，基于浅层语言分析的系统性能均超过了目前所见到发表的最好水平。

在句子检索方面，Noovel 采取了三种模型：向量空间模型、概率检索模型与语言模型。为了克服句子的局限性，本文引入了查询扩展的技术，主要包括：借助于 WordNet 的语义衍生扩展、伪相关反馈扩展、采用高频共现词语的局部共现扩展。在 TREC2003 数据集上的实验表明：在浅层语言分析的基础上，简单向量空间模型可以达到目前最好的结果，受到语义资源和分析深度的限制，当前阶段的语义扩展作用有限，而伪相关反馈与局部共现扩展都能够帮助提高句子检索的性能，局部共现扩展是很有潜力的查询扩展与文档扩展的技术。

句子级别的新信息检测是本项研究的最终目标，这是个时序性很强的信息过滤任务，在总结现有方法的基础上，本文提出了三种具有代表性的信息新颖度的量化方法，其中包括：词重叠度及其扩展方法、相似度比较方法与信息增强的评价方法。初衷在于兼顾信息与主题的相关性，同时还要与已有历史的信息进行比对，寻找新信息之所在。

除了非监督条件的新信息检测之外，本文还探讨了在监督条件下如何进行机器自动学习并调整参数的策略，主要的手段包括：进一步的特征选择、真实反馈、调整参数、阈值设置。作者还进一步的提出了基于分类的句子检索与新信息检测方法。

Noovel 系统参加了第 13 届 TREC 比赛新信息发现任务的全部四个子任务，在最关键的任务 1 中，Noovel 的新信息检测结果排名第一；任务 3 的句子检索性能方面，提交的两个结果并列排名第一，其他的子任务也取得不俗的成绩，与参赛的其他 13 支国际研究团队进行综合比较，本文在新信息方面的研究相对较优。

关键词：句子检索；新信息检测；浅层语言分析；信息检索；信息过滤；查询分析；查

询扩展；自然语言处理；汉语分词；词性标注；Noovel

Research on Shallow Linguistic Parsing and Sentence Oriented Novelty Detection

Zhang Huaping (Computer Software and Theory)

Directed By Professor Bai Shuo

The dissertation addresses the research on sentence oriented retrieval and novelty detection. It aims to meet the requirement on concise information with small granularity and less redundancy. It brings up the technique of partial linguistic parsing for novelty detection and discusses the modeling on sentence retrieval. The author also introduces various means to quantify the novelty degree on given information list. Noovel system, developed on the basis of the technique discussed here, achieves better on the final performance than any published result in the international official evaluations and several groups of experiments. It indicates that the technique applied in the system is effective for sentence retrieval and novelty detection.

Shallow English language parsing customized for novelty detection includes sentence boundary detection, tokenization, part-of-speech tagging, and morphological analysis. The author modifies the previous work on related natural language processing to solve the problem of novelty detection. As for shallow Chinese language parsing, the dissertation employs on hierarchical hidden Markov model to incorporate word segmentation, part-of-speech tagging, segmentation disambiguation and unknown words recognition into a unified framework. Based on linguistic analysis results, query analysis filters supplementary words, classifies the query tendency, automatically understands the user's query intention and generates computable query vector from topic description written in natural language. Experiments on all available test data sets are made to testify the function of shallow linguistic parsing. Surprisingly, the sentence retrieval on the basis of shallow linguistic parsing has achieved the best performance.

On sentence retrieval, Noovel applies three modeling approaches: vector space model, probabilistic retrieval model and language model. For the limit of a single sentence, some query expansion strategies have been tried, including semantic query expansion using WordNet, pseudo feedback, and local co-occurrence expansion. Experiments on TREC2003 data set shows that simple vector space model with shallow linguistic parsing could achieve better than any previously published result. Semantic expansion do not affect much on retrieval due to limitation on semantic resource and analysis depth, however, local co-occurrence expansion is helpful on query and document expansion.

Sentence oriented novelty detection is the goal of the research. It is a temporal task. The dissertation summarizes three approaches to modeling on information novelty degree based on previous work. It includes weighted word overlapping, similarity comparison, and information increment. The motivation lies in locating new information by considering relevance to topic and comparison with history.

Besides unsupervised novelty detection, the dissertation also discusses how to perform machine learning and adjust parameters under supervised conditions. The strategies can involve further feature selection, feedback, adjusting parameter and threshold setting. The author also provides supervised classifier to retrieve sentence and detect new information.

Noovel participates all the four tasks in the 13th TREC Novelty track. Task 1 is the most essential and difficult, Noovel rank top on novelty detection in the task. Meanwhile, two runs submitted with Noovel rank top simultaneously. Comparing other thirteen international groups who participated in the track, our work on novelty detection is one of the best.

Keywords: Sentence retrieval; Novelty detection; Shallow linguistic parsing; Information retrieval; Information filtering; Query anlysis; Query expansion; Natural language parsing; Chinese word segmentation; Part-of-speech tagging; Noovel

目 录

图目录

表目录

第一章 引言

随着计算机技术与 WEB (网络)应用的日益成熟，我们迈入了丰富多彩的信息时代。然而，WEB 无疑也是一个信息爆炸和信息混乱的世界：各种类型的网站也在浮浮沉沉中日益庞大，各种形式的数据、文本、语音、视频无穷无尽，信息重复冗余，相互矛盾的情况屡见不鲜。可用的新信息大约每 3 年增加一倍，人们往往迷失在浩瀚如海的信息泥潭之中而无力自拔，永远读不完的各种网页，无底洞般的链接永远追踪不到尽头。最终却往往找不到相关的信息，找到大量的相关信息又往往陈旧过时，人们为此付出了巨大的成本。这也向我们提出了进一步的挑战：如何自动地发现新的、有用的、粒度更小的信息片断。

1.1 句子级新信息检测产生的背景

在这里，我们主要介绍本项研究产生的背景。首先介绍目前信息增长的现状，在第二、三节介绍目前已有的信息检索和自动问答等解决方案，并分析其不足之处。

1.1.1 信息增长的现状与需求特点

根据 2003 年美国加州大学伯克利分校（University of California at Berkeley）发布的报告[Lyman 2003]，我们可以发现：

1. 2002 年产生了 5 万亿兆字节(Exabyte，1 万亿兆=10^{18}) 的以纸、胶片、磁、光电为介质的新信息。这相当于 37,000 个美国国会图书馆（藏书 1700 万本）的信息量。其中 92%的新信息存储在磁介质上，绝大部分存储在硬盘上。
2. 存储在纸、胶片、磁、光电上的新信息量，在三年之内翻了一倍。
3. 2002 年，电话、电台、电视以及因特网等信息频道上出现的信息流总量为 18 万亿兆字节，是已存储信息量的 3.5 倍以上。

截止到 2004 年 12 月，Google 已经索引的网页数量已经超过 80 亿。截止到 2004 年 6 月 30 日，在中国互联网络信息中心注册的 www 站点数目约为 626600 个[CNNIC 2004]。

信息的爆炸式的膨胀，大大地丰富了人们的知识结构，为我们跨进信息时代奠定了坚实的基础。然而，在如此大量繁杂的信息海洋中快速获取有用的相关资源，学习最新的知识，往往面临着巨大的问题。首先，现在的问题不是资源匮乏，而是有用的资源淹没在浩瀚的无关信息之中，即使是借助于当前相对成熟的商业搜索引擎工具（比如 Google, AlltheWeb），我们同样需要在系统返回的成千上万个结果中继续搜寻。其次，信息不断地复制、转载、改写现象非常严重，这成为了我们获取新信息的一大障碍，人们往往在已知的陈旧信息的相互链接中疲于奔命，成本巨大。最后一点和信息的粒度问题

相关，通常信息的粒度往往是文档级别的，带格式文档的平均大小在 12—15K 字节左右，大的文档往往是成百上千页，而用户真正关心的往往是其中不到 1K 字节的小片断，因此，人们还需要再在给出的长篇幅相关文档中，进一步定位真正有关的信息片断。

总之，在信息飞速膨胀的时代背景下，人们需要的信息往往会注重以下三个方面（按照重要程度排列）：

1. 相关的信息；人们往往不愿接受与己无关的“垃圾”信息，大量无关的信息会耗费人们有限的耐心；
2. 新的信息；新的信息往往给人们新的知识，更有助于人们的实践需求，人们不愿意反复接触到自己已经掌握的知识内容，冗余的信息不再具备实用价值。
3. 简明扼要，粒度更小的信息；长篇大论、言之无物的材料很难适应快节奏的信息需求。

1.1.2 文档级信息检索技术

信息检索(Information Retrieval，简称 IR)是一门研究从一定规模的文档库(Document Collection)中找出满足用户需求(User Information Need)的信息的学问[梁 2001]。为了表述方便，在不作明确区分的情况下，信息检索一般都是指文档级信息检索，它起源于图书情报的查询，一开始处理的文档数目和规模极其有限，随着硬件处理能力的提高、大规模数据以及 WWW 的出现，信息检索技术也日益发展。

信息检索和数据库检索存在本质的区别。一方面，传统的数据库是静态的，结构化的，经过严格组织的。而 Web 是自发形成和动态变化发展的，Web 上的页面是动态的，非结构化(Unstructured)或者半结构化(Semi-structured)的，通过超链接彼此缠绕。因此对 Web 的查询和对数据库的查询完全不同；另一方面，IR 的检索结果往往是不精确的，而不象纯粹的关系数据库查询那样正确率一定是 100%。比如，查关于“伊拉克战争”的文章，可能会漏掉有关“巴格达”或者其它城市的战斗。

从处理的技术来说， 包括自然语言处理(NLP)、人工智能、模式识别、机器学习、数理统计、运筹学等等学科和科目在内的技术纷纷被应用于现代信息检索。

WEB 的出现大大地促进了信息检索技术的发展。WEB 上有异常丰富但又充满垃圾的信息资源，其中绝大部分有用的信息还没有被发掘出来。这是因为目前还没有特别好的信息处理和检索工具。搜索引擎（Search Engine）是信息检索技术的集大成者，但是人们常常抱怨搜索引擎表现太差，可又没有其他办法，只能勉强使用它。这一领域的开发仍然处于初级阶段。

我们以最成功的商用搜索引擎Google(www.google.com)为例，来说明目前IR技术存在的不足。

为了了解当前热门的反恐问题，我们以“反恐”为检索词，得到的检索结果如图 1.1 所示，返回的结果有 947,000 条之多，其中绝大多数排名靠前的结果都说的是一种叫做“反恐精英”的游戏，直到第 49 个网页所列举的内容才真正讨论了我们需要的反恐问

题。这说明三个问题：第一、返回的信息太多，达到百万级；第二、不相关的无用信息干扰太厉害；第三、查询意图的理解存在着偏差，人们往往会采用自然语言形式表达自身的查询需求，而现在的搜索引擎基本上都是基于关键字匹配实现的，很难真正理解查询需求，因此高频率的结果成为了首选，也就造成了类似的误解。

图 1.1 “反恐”的检索结果（结果大量无关）

检索“反台独”，在首页得到的 10 个结果中，三条出处和内容完全一致，都是“军事科学院彭光谦少将的讲话”，如图 1.2 所示。结果完全重复冗余，并没有传达任何新的信息。现在的系统一般都根据网页的结构和内容引入了网页去重的技术，但是最终效果还依然不够理想。

一篇文章往往包含多个不同的主题（Topic）或者子主题（Sub-topic），而用户真正关心的往往是某个片断，可能是某个自然段落，甚至是一句话。因此，信息粒度太大是文档级别信息检索的另一个重大的不足，一方面，相关信息片断的权重相对于整篇文档会偏低，检索出来的可能性大大减少了；另一方面，人们还需要进一步地在相关文档中继续定位出相关的片断，进行文档内部的二次检索定位。例如：如果我们关心的是国家关于人才政策的最新信息，而长达 16,835 字的《2004 年政府工作报告》中仅有 51 个字和人才政策相关，现有的一些文档检索的技术往往会降低该篇文档被检索出来的概率（现在也有一些系统考虑到了文档长度的因素），即使检索出来，我们还需要在 15 页的文档中利用关键字匹配定位所需的片断。

最后，目前搜索引擎的查询大都是相关的关键词，这就要求用户能将自己的信息需求简明扼要地抽象成相应的几个词语，比如：当用户需要查找给老师的祝福用语时，比

较合适的关键词往往是“教师节”、“祝词”等。这需要用户具备一定的概括和抽象能力。用户的知识背景千差万别，而大部分没有受过相应理工科训练的人往往很难找到最符合自己需求的资源。

图 1.2 “反台独”的检索结果片断（完全重复冗余的文档）

总之，文档信息检索技术能够在一定程度上满足文档的检索需求，但是往往会包含大量的无关的、重复冗余的信息，同时信息粒度偏大。而且用户需要提炼自己的需求，以适当的关键词表达出来。

1.1.3 自动问答系统

自动问答系统（Q/A：automatic Question Answering）采用自然语言处理技术，一方面完成对用户提问的理解；另一方面实现正确答案的生成。回答的结果要求简明而又准确。这些研究涉及到计算语言学、信息科学和人工智能，是计算机应用研究的热点之一，其核心是自然语言理解技术。目前，虽然离自然语言完全的机器理解尚有很长的距离，但对于一定的领域，采用针对性的方法，人们已经实现出许多成功的应用系统。

自动问答系统有着比较长的研究历史，1950 年，著名的英国数学家图灵(A. M. Turing)发表了里程碑式的论文《Computing Machinery and Intelligence》。在文中，图灵第一次提出“机器智能”的概念，并提出判断计算机是否具备智能的实验方法—“图灵测试”，也就是通过自然语言问答的方式，判断机器是否具备人的智能。“图灵测试”可以看作是问答系统的蓝图。第一个问答系统，是 Jaseph Weizenbaum 在 1956 年实现的“Eliza”。

自动问答系统主要分为几种：聊天机器人、基于 FAQ 的问答系统、基于受限语言的数据库查询系统、基于知识库的问答系统、问答式搜索引擎与具有推理智能机制的自动问答系统[王 2004] 。其中，问答式搜索引擎与传统信息检索不同的是，它根据自然语言表达的用户查询，从系统文档集合或 Web 中，检索出相关文档或文本，并将其提

供给用户。问答式搜索引擎能够弥补信息检索查询表达所存在的缺陷。基于知识库的系统往往围绕一个特定的知识库，取得了一定的成效。同时，中科院计算所的相关研究者[王 2005]进一步提出了问答系统的知识推理机制。

依据自动问答系统所面向的领域，大致可分为两类：一类是开放领域，另一类是受限的专业领域。对于某些特定的用途（如金融、科研、教学等），面向非常专业领域的问答系统具有一定的生命力。而开放的自动问答不是单纯依赖某一项技术，而是需要多种技术的融合，但是其中的大部分技术都不成熟。一方面，目前还不能实现自然语言问题的真正理解，往往是答非所问；另外一方面，在开放领域的信息抽取依然不够理想，问题答案的抽取与组织也不如人意。因此，面向 Web、知识库等开放领域的自动问答系统依然是个遥远的目标，远远不能达到实用的地步。这一点我们可以从每年的 TREC 会议的 QA 任务中得出结论。

1.1.4 句子级新信息检测

一方面，用户需要粒度更小、相关而又新颖的信息；另外一方面，传统的文档检索系统返回大量无关或者冗余的长篇幅文档，自动问答技术结果能符合要求却很难实用化。为此，一个自然而然的中间路线就是：检索出粒度比文档更小的相关信息，并进一步排除冗余、检测出新的信息内容。这里的粒度与文档相对应，我们一般称之为片断(Passage)，其中包括：段落(Paragraph)、句子集(Sentence Cluster)和句子。为了评价与计算的便利，我们将在以句子为单位的粒度上进行信息检索，这一中间路线可称为句子级新信息检测(Novelty Detection at Sentence Level)。

句子级新信息检测内在地包含着两个主要内容：相关句子检索与新信息内容的检测。在不做特别区分的情况下，我们这里一律称之为新信息检测。另外，需要强调的是新信息检测具有时序性，内容相同或相近的句子在一维时间上，最先出现的信息属于新信息而检出，而事后出现的所有相似内容句子均属于过时的信息而被过滤。

下面，针对用户的信息需求，我们分别从信息粒度、相关性检索、新信息过滤、查询表达与技术难度等方面比较句子级新信息检测与现有技术之间的差别。如表 1.1 所示：

技术	信息粒度	相关性检索	新信息过滤	查询表达	技术难度
信息检索	文档、网页	相关	没有排重处理、存在大量冗余	关键词	基本实用化
自动问答	词语、短语	精确相关	唯一答案、无冗余	自然语言	仍不成熟
句子新信息检测	句子	相关	按照报道时序进行新信息检测、基本无冗余	自然语言	正在探索，技术相对成熟

表 1.1　句子级新信息检测与现有技术的比较

1.2 Noovel 系统的总体架构

新信息检测是本文主要的研究核心。为此，我们专门研制了一套 Noovel 新信息检测系统。本文主要围绕着在 Noovel 研制过程中所体现的理论和技术进行深入详细的阐述。在这里，我们首先给出 Noovel 系统的总体架构，如图 1.3 所示。

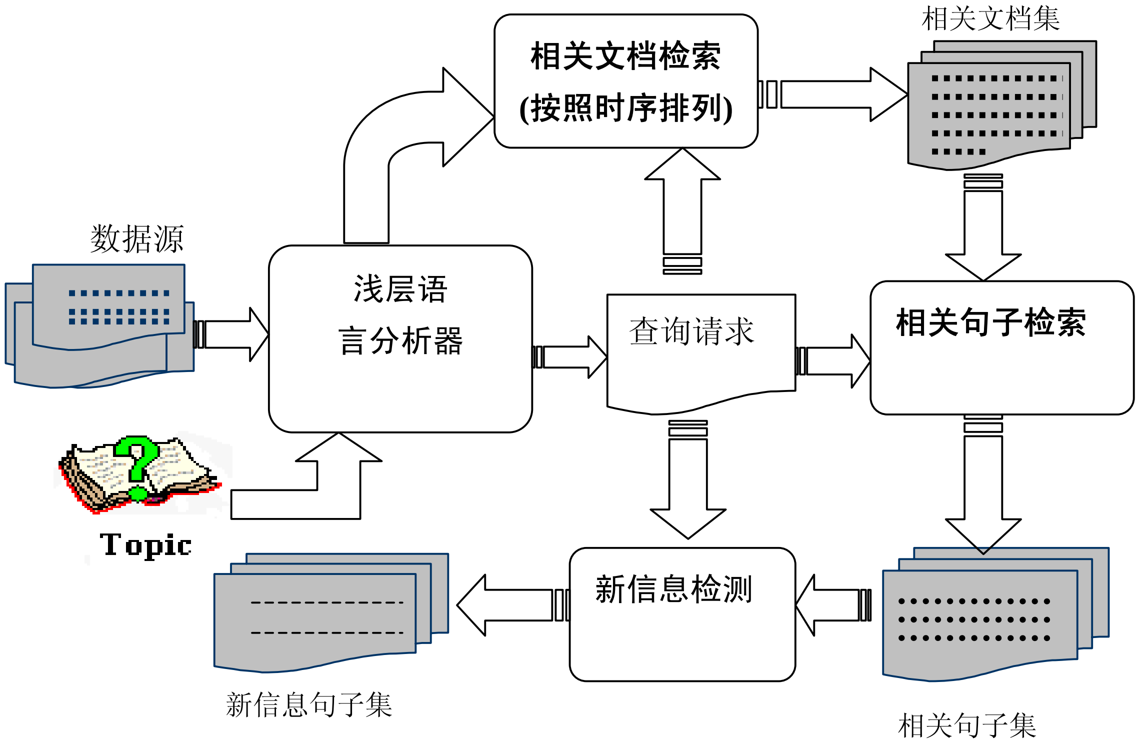


图 1.3　Noovel 系统的总体架构

1.2.1 已知信息

新信息检测给定的信息包括：

- 数据源 (Data Source)

数据源指的是按照时间顺序排列的文档集合，可以是来自于互联网上的网页，也可以是人工整理的半结构化文本。每篇文档包含若干句子。

数据源形式化定义为：$D=<d_1,d_2,\dots d_n>$，其中 d_i 为第 i 篇文档，n 为总的文档数。时序关系排列满足：$timestamp(d_i)<timestamp(d_j)$ 当且仅当 i<j。（*timestamp* 表示的是时间戳，下同）。

文档 d_i 的形式化构成为：$<s_{i1},s_{i2},\dots s_{im_i}>$，其中 s_{ij} 是文档 d_i 的第 j 个句子，m_i 为句子总数，内在地包含着时序关系：$timestamp(s_{ik})<timestamp(s_{il})$ 当且仅当 k<l。

如果按照句子粒度进行组织，数据源又可以表示成为一维向量形式：S=< s_{11}, s_{12},…s_{ij}, … $s_{i\,m_n}$ >。S 是所有的句子集合。

- 主题 (Topic)

主题是用户需求意愿的集中表达，往往采用自然语言自由地表述，呈现形式有标题、关键字、句子与段落描述、乃至示例文本。形式化的定义为 *T*。阐述的一般都是事件的过程与不同的评论意见。我们需要利用自然语言理解的技术自动地抽取出内在的查询请求，定义为 *Q*。*T* 与 *Q* 之间并不完全等同，*T* 为原始的需求描述，*Q* 为查询请求，是 *T* 模型化可计算的内在表示形式。下面分别是中英文的主题描述文字：

<title>General Pinochet Arrested
<toptype>Event
<desc>Description:
Arrest of former Chilean dictator, General Augusto Pinochet, in London. He was charged with murder, torture, genocide, and terrorism during his regime in Chile.
<narr>Narrative:
Information about Pinochet's arrest and evidence of charges of murdering, torturing and the disappearance of people in Chile while he was head of state is relevant. Specifically relevant are mention of charges against him.

图 1.4A 英文的主题表述(标题、类型、简介、详细描述)

我想知道国民党反对台独的评论，中共的意见不需要。

图 1.4B 中文的主题表述(自然语言表述的句子)

新信息检测最终返回给用户的信息为符合要求的句子集合 *N*, *N*= <s^*_1,s^*_2,…s^*_{ns}>, 其中 s^*_i 为第 i 个新信息句子，它们同时满足两个条件：

1. 与主题 T 相关；
2. 包含以前句子< s^*_1,s^*_2,…s^*_{i-1} >所不具备的新信息内容。

1.2.2 子任务

从处理过程来看，Noovel 系统主要包含 4 个子任务。它们分别是：

- 浅层语言分析器(Light Language Parsing, LLP)；
- 相关文档检索(Document Retrieval,DR)；
- 相关句子检索(Sentence Retrieval,SR)；

- 新信息检测(Novelty Detection,ND)。

其中，浅层语言分析器主要是实现语言的浅层分析，主要包括主题的理解并生成查询、文档的解析结果，我们会在相关章节中作详细的介绍。相关文档检索子任务是检索出相关文档，相关句子检索子任务是检索相关的句子，新信息检测子任务是最后一个核心任务，也是最关键的步骤。这些子任务之间前后继承，最后的性能取决于这四个过程相互协调配合。浅层语言分析器与具体的语言相关，其余三个过程基于浅层分析之后的 ID 序列结果，与语言无关，完全适应于各种语言。我们主要处理的是英语，并会简略的介绍中文词法层面的浅层分析。

现在，我们对各个过程进行形式化定义，具体过程如下：

$LLP(T)=Q$，表示从主题 T 生成查询 Q；

$LLP(S)$，表示对文档句子的浅层语言分析；

相关文档检索定义为：$DR(Q, d_x)$=TRUE 当且仅当文档 d_x 与查询 Q 相关；否则 $DR(Q, d_x)$=FALSE；

相关句子检索定义为：$SR(Q, s_x)$=TRUE 当且仅当句子 s_x 与查询 Q 相关；否则 $SR(Q, s_x)$=FALSE；

新信息检测形式化定义为：$ND(Q, <s'_1,s'_2,\ldots s'_{x-1}>, s'_x)$ = TRUE，当且仅当 s'_x 是包含新信息的句子，否则 $ND(Q, <s'_1,s'_2,\ldots s'_{x-1}>, s'_x)$ = FALSE，其中 $s'_1,s'_2,\ldots s'_{x-1}>$为 s'_x 以前出现的新信息句子或者相关句子。

1.3 新信息检测的相关技术

从 Noovel 系统的总体架构中的四个子任务，我们可以发现新信息检测实际上是一项综合的技术，其中主要包含信息检索、信息过滤、文本分类、自动文摘以及自然语言理解。我们一一作简单的介绍。

1.3.1 信息检索

信息检索主要应用于新信息检测的文档检索与句子检索。换个角度来看，新信息检索实际上是信息检索基础上的排重和粒度细化。信息检索得出的文档与句子是新信息最终检测的数据来源，它直接影响着最终的性能。

信息检索技术中的文档表示、长度归一化、检索模型与相关度计算同样是这两个阶段需要应对的问题。

1.3.2 信息过滤

信息过滤是一种非常典型的信息获取技术。它根据用户给定的一个比较稳定的信息需求，采用一定的技术从动态到来的庞大信息流中选择出满足用户需求的信息[许 2003]。新信息的检测最终目的是排除重复冗余的内容，过滤出包含新信息的句子。最后也可以归结为信息过滤的问题。我们需要研究的内容包括：过滤条件的建模、阈值设置与参数调整、信息的各种反馈。日本明治大学[Ryosuke 2003]就采用了基于信息过滤的技术来实现新信息检测的任务。

1.3.3 文本分类

文本分类就是将一篇文本自动地按照先验的类别进行匹配，确定其归属。从文本分类的角度来看，相关句子的检索与新信息检测均可以看成是句子的分类问题，类别分别对应于相关与不相关、新信息与旧信息。

在监督学习条件下，即给定了部分结果的情况下，文本分类方法往往行之有效。文本分类中使用到的特征抽取与选择、贝叶斯分类器、支持向量机等方法均可以适应新信息检测任务。

1.3.4 自动文摘

新信息检测最早是在自动文摘中提出的，自动文摘的目标就是提取能够概括当前文档、没有冗余的句子集合，这一点与新信息检测是类似的。很多自动文摘方面的研究成果都可以在新信息检测当中得到验证。Maryland 大学[Ganesh 2003]Michigan 大学[Qi 2002]的研究者就研制出了基于自动文摘的系统。

与新信息检测不同的是：自动文摘不存在查询主题，没有相关的查询过程。我们可以将新信息检测看成是带有查询需求、有指导的自动文摘过程。

1.3.5 自然语言理解

实际上，前面提到的信息检索、信息过滤、文本分类与自动文摘或多或少都需要自然语言理解的技术。不过，自然语言理解在新信息检测中担当着更加重要的角色，对查询需求的理解直接决定最终结果的相关性。我们处理的基本单元是句子，一个句子排除虚词等停用词之后，实际包含的信息量十分有限，短到一两个实词，最长也不过十来个词语，纯粹的字面信息太少，我们必须借助于自然语言理解技术扩展内涵，帮助信息的检索与新信息的发现。

在解决新信息检测问题的同时，我们这项研究的另外一个目的还在于具体地研究自然语言理解与信息检索过滤的内在联系。尽管我们都知道：自然语言的成功理解有助于信息检索的正确性，但是并不是所有的自然语言理解技术都能改进信息检索与信息过滤，我们希望研究出专门针对信息检索与信息过滤特定的语言分析方法，并对具体的相

互作用进行量化比较。在现有的技术条件下，同时兼顾效率，我们提出了浅层语言的特定分析方法，具体的关系如下图所示。

图 1.5　特定的浅层语言分析

1.4 评测方法与测试平台

新信息检测的结果具有一定的主观性，其测试评价是一个难题。一方面，我们需要相对客观公正的评价标准；另外一方面，我们需要一定规模的自由文本数据进行开放测试，少量或者封闭数据的测试都不具备实际的意义。

本文主要采用目前国际上通行的 TREC 评测标准和测试数据集合。下面，先介绍国际上新信息评测发展的历程，然后给出目前通行的评测方法。

1.4.1 发展历程

新信息检测技术最早起源于 Jaime Carbonnell 教授在 2001 年 5 月召开的 NAACL 自动文摘工作组会议（the Automatic Summarization workshop at NAACL）上的一次演讲。Jaime Carbonnell 教授提到“除了单纯的依靠相关度排序之外，我们应该还有其他的方法来优化检索结果，比如说文章的时序性、信息源的可靠性以及所传达信息对用户的易学性与新颖性。时序性排序需要每篇文章具有时间戳，比较琐碎；信息源是否可靠，用户是否容易理解，评判本身就十分困难。而是否具备新颖性就很好操作，我们可以假定用户在看第一篇文档之前对主题一无所知，相关的知识都来源于检索到的文档”。

2002 年 9 月，第 11 届国际文本检索会议(Text Retrieval Text Retrieval Conference, TREC)首次将新信息检测作为其中的一个正式比赛任务(Novelty Track)。TREC会议由美国国防部的高级研究发展署(Advanced Research and Development Activity, ARDA)与NIST(国家标准技术研究所)共同主办，是国际上文本检索方面的最高学术论坛。每年 8 月到 10 月举行的各类文本检索与过滤等比赛，吸引了大量的知名大学、科研院所以及相关企业参加，11 月召开的学术会议公布比赛结果，并交流各自的研究成果。在最近几年的比赛中，国内的中科院计算所、清华大学在Web检索、信息过滤方面均取得了不俗的成绩。目前，TREC已经成为信息处理方面的权威测试平台。

第一届 Novelty 比赛只有一个任务，即：给定 TREC 主题与相关文档（按照时间顺序排列组织），要求输出所有相关的句子集合，同时给出相关结果中包含新信息的句子。测试数据包括：50 个 TREC 格式的主题、每个主题给定了相应的 25 篇相关文档。当时参

赛的共有 13 支研究队伍，其中包括清华大学、卡耐基梅隆大学、哥伦比亚大学等。

2003 年召开的第 12 届 TREC 会议对 Novelty 任务作了进一步的细化，在原来任务(作为任务 1）的基础上，又增加了三个子任务，它们分别是：

- 任务 2：给定所有相关的句子，要求给出所有包含新信息的句子；
- 任务 3：对于每个主题，给定前五篇文档中包含的相关句子与新信息句子，要求给出剩余 20 篇文档中的相关句子与新信息句子。
- 任务 4：对于每个主题，给定所有文档中相关的句子，同时给出前五篇文档中包含新信息的句子，要求给出剩余 20 篇文档中的新信息句子。

新增任务逐步放宽条件，有助于进一步评价新信息检测的性能。它们属于监督学习的范畴，可以衡量出综合利用信息反馈的能力。

2004 年，第 13 届 TREC 会议又进一步地增加了 Novelty 任务的难度，即不再给定相关文档，每个主题给出的文档集合包括 25 篇相关文档，而其他文档则为不相关的噪声数据集。这种情况更符合实际应用场景，客观上增加了文档检索的过程，给新信息检测尤其是监督条件下的检测带来了新的挑战课题。具体的情况可以参见附录 1“TREC 2004 Novelty Track Guidelines”。

1.4.2 评价方法

在这里，我们借用 TREC 会议 Novelty 任务的评价方法。

对于每个主题，TREC 事先给定了人工标注的相关句子与新信息句子，作为系统的参考答案。针对句子检索与新信息检测，均引入召回率 Recall、准确率 Precision 与 F 值三个评价指标。其中：

$$\text{Recall} = \frac{\#\text{正确的结果数}}{\#\text{参考答案的结果数}}$$

$$\text{Precision} = \frac{\#\text{正确的结果数}}{\#\text{提交的结果数}}$$

$$F = \frac{\text{Recall} \times \text{Precision} \times (1+\beta^2)}{\text{Recall} + \text{Precision} \times \beta^2}$$，其中 β 为权重因子。

F 值作为综合指标，对 Recall 与 Precision 进行了权衡。对于给定的每个主题及系统提交的结果，我们都可以计算出提交系统在该主题上的 Recall、Precision 和 Recall。然后，在所有主题的基础上，TREC 将这些指标的平均值定义为系统的综合性能。即：

$$\overline{\text{Recall}} = \frac{\sum_{i=1}^{N} \text{Recall}(i)}{N}$$

$$\overline{\text{Precision}} = \frac{\sum_{i=1}^{N} \text{Precision}(i)}{N}$$

$$\overline{F} = \frac{\sum_{i=1}^{N} F(i)}{N}$$

其中：N = 主题总数，在 TREC 中 N=50

实际上，这是一个宏平均的评测方法。借鉴文本分类的评测方法，我们还可以定义出微平均的评测指标。即：

$$\text{Micro\_Recall} = \frac{\#\text{所有主题中正确的结果数}}{\#\text{所有主题中参考答案的结果数}}$$

$$\text{Micro\_Precision} = \frac{\#\text{所有主题中正确的结果数}}{\#\text{所有主题中提交的结果数}}$$

$$\text{Micro\_F} = \frac{2 \times \text{Micro\_Precision} \times \text{Micro\_Recall}}{\text{Micro\_Precision} + \text{Micro\_Recall}}$$

在不做特别强调的前提下，我们采用的都是 TREC 公认的评测标准，以 50 个主题的 F 宏平均值作为系统的综合评价指标。除非特别提示，以后提到的准确率、召回率、F 值都是相应的宏平均指标。

1.4.3 测试数据集

目前，我们可以公开得到的相关测试数据集合主要来自于 TREC 三年的测试数据。同时，麻省大学的研究者还另外整理了 48 个话题及相应的训练文档。

TREC 主题包括六部分，分别为<num>（编号）、<title>（标题）、<toptype>（主题类型）、<desc>（简述）、<narr>（详细描述）、<relevant>（相关文档编号，出现在 TREC2004 以前）或者<doc>（文档编号，不一定相关）。其中<toptype>有两种：事件类(Event)与评论类(opinion)。这两种类型的主题以及文章风格特点迥异，事件类偏重于对重大事件发生过程等细节的叙述，评论类偏重于对特定观点的倾向性分析与评价。各年话题类型以及相应文档的统计分布信息如下：

年份	事件类主题数	评论类主题数	主题总数	文档篇数
TREC2002	50	0	50	1105
TREC2003	28	22	50	1250
TREC2004	25	25	50	1808
合计	103	47	150	5163

表 1.2　TREC 各年类型主题分布情况

1.5 论文的组织结构

论文针对实际的信息需求，从现有的解决方案出发，在第一章中给出了新信息检测的产生背景，进行了形式化定义，并与现有的技术作了详尽的比较。随后，给出了依据本文思想开发出的 Noovel 系统的总体架构，介绍了相关技术，并阐明了具体的评测方法与测试平台。

第二章简单介绍了现在主要的句子检索与新信息检测算法，并对主要涉及的理论与技术作了总括性的综述。

从第一章的总体架构出发，第三章详细描述了我们在 Noovel 系统中专门为新信息检测任务研制的浅层语言分析技术，依次介绍了相关的理论与实际处理环节。

接着，在随后的第四章中，我们着重研究了针对句子级别检索的模型与算法，其中包括向量空间模型及其扩展变种、概率模型与语言模型。在如何提高检索综合性能方面进行了各种有益的试验，并提炼出了相关的规律性结论。

第五章主要研究了各种新信息检测的技术，总结并提出了各种评价新信息程度的衡量指标。

第六章针对监督学习条件，我们有针对性地进行了模型的机器学习与自适应调整。提出并试验验证了各种监督条件下的新信息检测技术。

第七章在 2004 年度数据集合的基础上，给出了我们在 TREC2004 上的各种试验及其数据分析。

最后，对整个论文进行了总结，并给出了将来的研究方向。

第二章 句子检索与新信息检测的主要算法模型

2.1 概述

新信息检测作为一个新兴的应用技术，从 2001 年提出到最后一届 TREC 大会的新信息检测大赛，总共只有 3 年的历史。但是，新信息检测技术的潜在价值却纷纷引起了国际上的大学、科研机构、商业公司的广泛关注。

我们从 TREC 的新信息检测大赛的参与情况就可以了解大致的情况。2002 年，参与新信息检测比赛的队伍有 13 支，2003 年和 2004 年均为 14 支队伍，总共 41 人次，分别来自 25 个单位，主要是大学与科研机构（包括：中科院计算所、爱荷华大学、卡耐基梅隆大学、明治大学、都柏林城市大学、清华大学等），另外还有 NTT 通讯，LexiClone 等商业公司，爱荷华大学等 5 个研究团体三年来一直在这方面进行了深入的研究。

图 2.1 [Harman 2002]和图 2.3[Soboroff 2003]分别给出了 TREC2002 和 TREC2003 新信息检测任务所有参赛队伍的评价结果。

	Relevant	New
Second human judges	0.371	0.353
Random sentences	0.040	0.036
thunv1	0.235	0.217
thunv2	0.235	0.216
thunv3	0.235	0.216
CIIR02tfnew	0.211	0.209
thunv4	0.225	0.206
CIIR02tfkl	0.211	0.196
pircs2N02	0.209	0.193
pircs2N01	0.209	0.188
pircs2N04	0.197	0.184
ss1	0.186	0.183

图 2.1 TREC2002 新信息检测任务的结果

图 2.1 中给出了第二个人（相对于主题设计者来说）进行人工判断的得分结果，其中句子检索的平均 F 值仅为 0.371，新信息检测的性能为 0.353，这反映出新信息检测结果本身就是一个“仁者见仁，智者见智”的主观评价，人的判断本身就存在很大的争议性，没有绝对的评价标准。与人工判断相比，自动系统的性能差距很大，最好的系统性能也比人工判断低了差不多 40%。

在图 2.2 中，我们可以看到：有 5 个提交的系统结果综合性能均超过了人工的判断（水平虚线），这反映了系统机器学习的成功，基本反映了新信息检测技术的进步。和 TREC2002 的综合性能比较，平均性能提高了 2 倍多，而TREC2004 的新信息检测的最

好结果仅仅只有 0.239①，这表明了新信息检测技术的不稳定性，其最终性能更大程度依赖于给定的主题和文档集合。

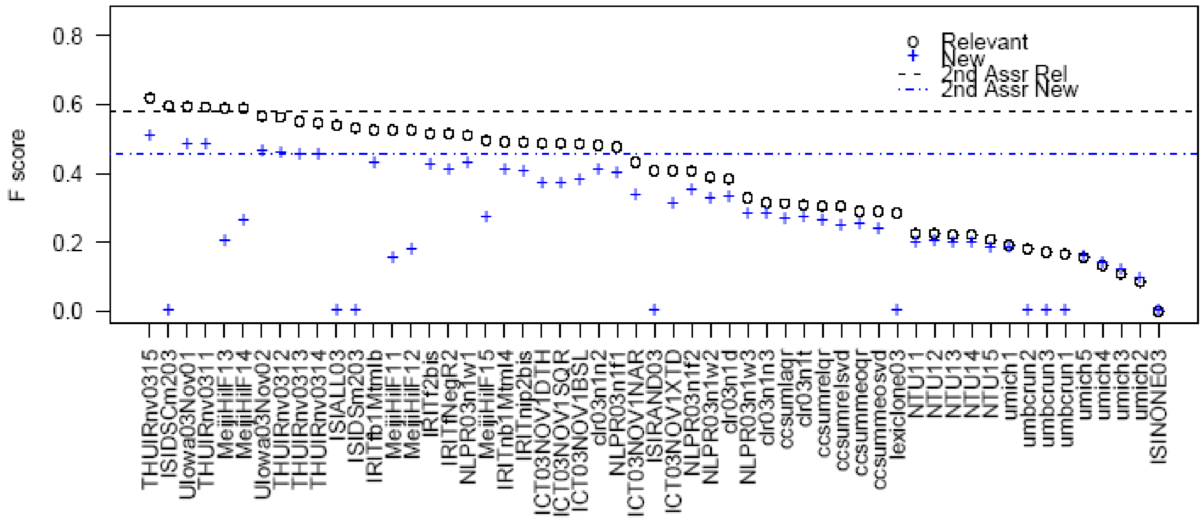


图 2.2　TREC2003 新信息检测任务的结果

随着互联网络的蓬勃发展，新信息检测牵涉到的信息检索、信息过滤、文本分类、自动文摘以及自然语言理解等一直是信息技术的一个研究热点。

但是，我们还应该看到，作为一个综合性的技术，新信息检测受到各方面技术水平的限制，至今还没有成熟的科学理论或者技术框架可以从根本上解决所有的新信息检测问题。无论是从目前的性能评测结果来看，还是从理论总结来分析，新信息检测仍然处在低水平的阶段。

本章主要从历届 TREC 比赛中的 41 个系统出发，总结新信息检测技术的研究进展，进行一个相对系统的综述。

2.2 句子检索方法综述

在句子检索方面，研究者一般都将句子看成是小型的文档，采用基于文档的信息处理办法，同时进行适当的改进。按照采用的核心技术，我们将句子检索方法划分为：传统的文档检索方法、信息过滤方法、分类方法、语义比较方法、隐马模型方法与自动文摘方法等。

2.2.1 传统的文档检索方法

单个句子本身就可以看成是一个微型的文档，因此，在句子检索方面，研究者普遍采用的往往是传统的文档检索方法，其中常用的方法主要有经典的 tf*idf 方法、语言模型方法。[Leah 2002]中扩大了训练样本，同时尝试了这两种句子检索的方法。

给定查询 q 与句子 s，经典的 tf*idf 可以采用下面的公式计算出该句与查询的相关

① TREC2004 系统性能比较低的另一个重要因素在于：给定的文档集合包含了非相关文档。这一点，我们会在第七章作进一步分析。

度：

$$score(s) = \sum_{t \in q} \log(tf_{t,q} + 1) \log(tf_{t,s} + 1) \log\left(\frac{n+1}{.5 + sf_t}\right)$$

其中，tf 为词语 t 在句子中出现的频率，sf 为出现 t 的句子数目。

在语言模型中，采用 KL 距离的方法计算句子 s 与查询的相关度。

$$score(s) = KLD(q \| s) = \sum_{w \in V} p(w|q) \cdot \log \frac{p(w|q)}{p(w|s)}$$

实际上，句子和文档仍然存在着较大的差别，我们不能够简单地照搬文档检索的方法，还需要针对句子的特点进行重大的改进。人们纷纷引入了各种查询扩展、文档扩展以及参数调整策略。我们会在第四章简单地介绍相关的理论及其已有的研究状况，并详细阐述 Noovel 系统所采用的各种技术方法。

2.2.2 信息过滤方法

信息过滤方法的基本思想为：将主题作为初始的用户兴趣背景（Profile），而将句子看成是文档集，需要作出保留或者过滤的决定，其中过滤的条件为句子与用户兴趣的相关度，过滤条件采用的阈值在训练集上机器学习得到，目标是综合指标 F 值的最大化。

和文档信息过滤不一样的是，句子信息量太少，往往难以运算，需要做进一步的扩展。明治大学[Ryosuke 2003]在具体的信息过滤实现过程中，用户兴趣主要来源于主题的简介部分。初始的权重由 tf*idf 计算得到，同时，还引入了概念模糊集（Conceptual Fuzzy Sets）[Takagi 1995]对用户兴趣和句子进行概念扩展，如图 2.3 所示。

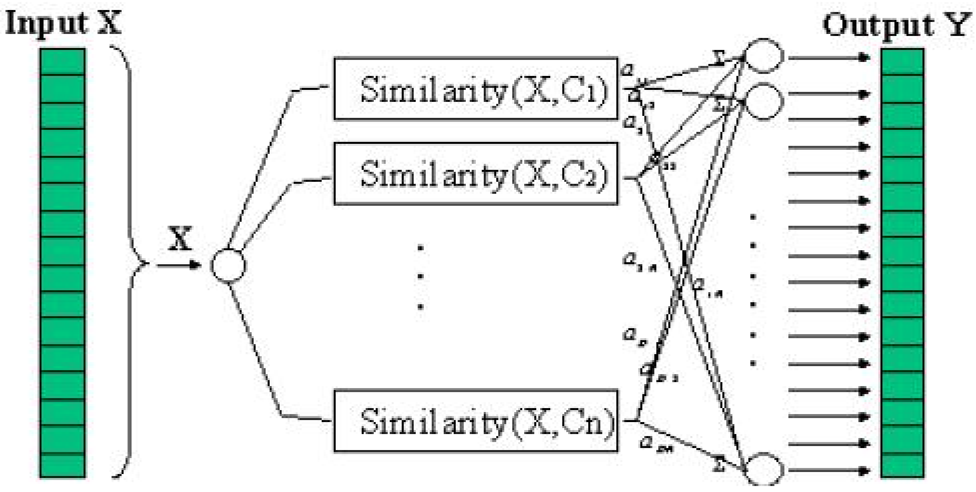


图 2.3　明治大学引入概念模糊集扩展信息过滤输入的示意图

其中：X 为输入，$C_1, C_2, \ldots, C_n$ 为概念集合中所有的概念，分别计算它们之间的相似度后，乘以不同的加权系数 a_{ij} 最后得到扩展之后的输出概念。

2.2.3 分类方法

分类的句子检索方法主要根据词语的分类特征属性来判断所在句子与主题的信息

相关度 [Taoufiq 2004] [Taoufiq 2003] [Taoufiq 2002]。

[Taoufiq 2004] 依据主题分析的结果定义了四种类型的词汇：高相关词（highly relevant terms, HT）、低相关词（scarcely relevant terms, LT）、不相关词（non-relevant terms, iT、类似于停用词）与逆相关词（negative terms, IT）。其中，逆相关词从主题详细描述中不相关文档的叙述文字中抽取，而其他类别的词均从剩余部分中抽取，相应权重根据其出现频率和在主题中的位置来综合评估。最后，根据设置的阈值和权重的大小，得到不同类别的词汇集合。

给定一个主题 Q_k 和句子 S_j，句子 S_j 与主题 Q_k 相关的充要条件为：

$$Score(S_j, Q_k) > f\left(\frac{|LS_j|}{|LS_j| + |HS_j|}\right) \cdot |HT_k| + g\left(\frac{|HS_j|}{|LS_j| + |HS_j|}\right) \cdot |LT_k| + \alpha$$

其中：

$$Score(S_j, Q_k) = \sum \left(weight(t_i, S_j) \cdot weight(t_i, Q_k) \right)$$

LT_k 为主题 Q_k 中的低相关词集合；

LS_j 为同时出现在句子 S_j 和主题 Q_k 中的低相关词集合；

HS_j 为同时出现在句子 S_j 和主题 Q_k 中的高相关词集合；

| X |表示的是集合 X 中元素的个数；

而 $weight(t_i, S_j)$为词语 t_i 在句子 S_j 中出现的频率，$weight(t_i, Q_k)$由词语 t_i 在主题 Q_k 中出现的频率及其位置综合加权得到。

除此之外，还有另外一种基于文本分类思想的句子检索方法，即将相关和不相关视为两种不同类别，分别采取已有的数据进行训练，最终得到一个句子检索的分类器，NTT 通讯公司就尝试了 SVM 的方法 [Hideto 2002]，密歇根州立大学在 2003 年采用了最大熵的分类器完成了所有的四个子任务[Jahna 2003]，也有的研究者在任务三利用给定的部分结果训练分类器，本文会在第六章监督条件下的句子检索部分介绍相关的细节。

2.2.4 语义比较方法

语义比较方法主要是借助于现有的语义体系来实现不同词语之间的相似性计算，然后综合计算得到句子与查询之间的相关度。对于句子来说，按照词形进行精确匹配往往过于严格，无法发现句子与主题之间的潜在关系。

台湾国立大学在 2002 年提出了一种比较简单的语义比较的句子检索方法[Tsai 2002]。首先，对于两个不同形的词语 w_1 与 w_2，将其对应的概念在 WordNet 层次结构中的最短路径长度定义为两个词的语义距离 $dist(w_1,w_2)$，例如，在图 2.4 中，“sky”和“universe”的语义距离为 4。设定一个阈值，如果语义距离小于该值，则两个词相似，否则两者没有关系。这里只考虑动词和动词。

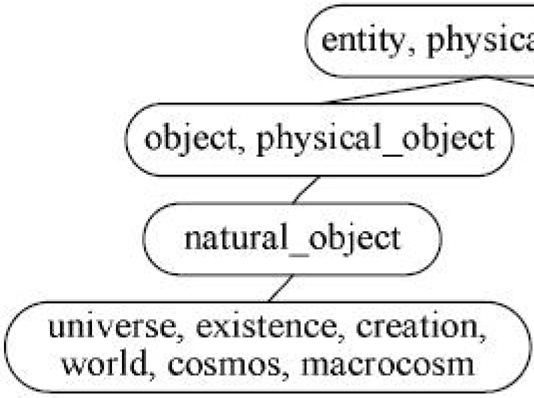


图 2.4　语义距离示意图

对于任意两个句子，其相似度计算模型为：

- 一个句子中的名词或动词和另外一个句子精确匹配，则相应的匹配度加上 1；
- 如果不存在精确匹配，则计算两个词语之间的语义距离 $dist(w_1,w_2)$，如果 $dist(w_1,w_2)<4$，则相应的匹配度加上 0.5；
- 每个词语只能匹配一次

句子之间的相似度由名词相似度和动词相似度来表示：

$$sim(s_1,s_2) = noun\_sim(s_1,s_2) + verb\_sim(s_1,s_2);$$

$$noun\_sim(s_1,s_2) = \frac{m}{\sqrt{ab}}$$

$$verb\_sim(s_1,s_2) = \frac{n}{\sqrt{cd}}$$

其中，s_1 和 s_2 分别表示两个句子；m 和 n 分别表示名词和动词匹配的值；a 和 b 分别表示 s_1 和 s_2 中包含的名词数；c 和 d 分别表示 s_1 和 s_2 中包含的动词数。

2.2.5 隐马模型 (HMM) 方法

和简单贝叶斯方法相比，隐马模型独立性的假设更少，它不会假设第 i 个句子的相关概率与第 i-1 个句子是否相关完全无关。

马里兰大学的研究者[Conroy 2004][Conroy 2003]借鉴自动文摘[Conroy 2001]的做法，提出了一种基于隐马模型的句子检索方法，观测的是和句子中词语相关的特征，其中包括：

- 句子中标志词（signature terms）的数目 n_{sig}，其观测值为：$o_1(i) = log(n_{sig}+1)$；标志词指的是那些相对于通用语言环境来说，更能代表当前主题、在当前主题中更可能出现的词语。我们可以采取互信息等统计方法发现这些标志词。
- 句子中主题词（subject tokens）的数目 n_{sub}，其观测值为：$o_2(i) = log(n_{sub}+1)$；主题词属于标志词的一个子集，指的是那些出现在标题或者头条（Headline）中的标志词。
- 句子中的词数（tokens）n_{tok}，其观测值为：$o_3(i) = log(n_{tok}+1)$；
- 句子在当前文档中的位置。

在隐马模型中，总共有 2s+1 个状态，其中 s 个相关状态，s+1 个不相关状态。图 2.5 给出的是 7 个状态的马尔科夫链。

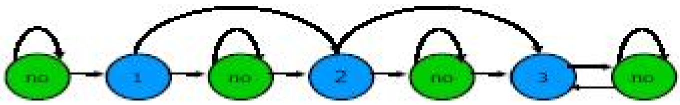


图 2.5 7 个状态的马尔科夫链（其中 4 个为不相关状态）

状态之间的转移概率在训练集上采用最大期望估计的方法学习得到，最后形成马尔科夫过程的状态转移矩阵。每个状态上的输出概率 $b_i(O) = P\,(O|$状态 i)，其中 O 为观测到的特征向量，为了简化计算，个个特征之间假定相互独立，直接采用联合概率估计。

最后，句子检索的过程就转化为搜索一条概率最大的路径，路径中的节点对应的就是检索出来的相关句子结果集。

2.2.6 自动文摘方法

新信息检测最早是在自动文摘中提出的，自动文摘的目标就是提取能够概括当前文档、没有冗余的句子集合，这一点与新信息检测是一致的。自然而然，人们想到了可以将一些成熟的自动文摘方法应用于新信息检测，不过，人们还需要进一步依据句子的特点进行改进。

Michigan大学的研究者三年来一直尝试着将他们研制的多文档自动文摘系统MEAD②应用于句子的检索与新信息过滤[Erkan 2004] [Jahna 2003] [Qi 2002]。在句子检索方面，他们在自动文摘系统当中主要融入如下特征：

- **Centroid**：句子到句子集合中心向量的距离；
- **LexRank**：LexRank特征主要用来表示句子的显著程度或者说是重要程度，最早是在多文档自动文摘中提出来的[Erkan & Radev 2004]。为了计算句子的LexRank，我们先要构造一个无向图，其中每个节点是一个句子。如果任意两个句子的Cosine相似度大于某个预设的阈值，则将对应的两个节点用一条边联接。边的权重需要进行归一化处理，使得每个节点出边的权重之和为1。图采用矩阵进行表示，各个数据都是一个统计量，因此，我们可以将该图看做是离散的马尔科夫链，一个句子的LexRank值就是它在马尔科夫链中相应的静态分布概率。直观地说，如果一个句子和大量的其他句子相似，那么其LexRank值会比较高，同时，和它相似的句子也会有一个较高的LexRank值。
- **Length**：句子的词数；
- **QueryTitleCosine**：句子与主题中标题部分的Cosine相似度；
- **QueryDescriptionCosine**：句子与主题中简介部分的Cosine相似度；

在自动文摘的结果基础上，可以进一步组织，构建出相关的句子结果。

2.3 新信息检测方法综述

在新信息检测方面，核心的问题在于如何在句子序列基础上度量某个句子所含信息

② 参见http://www.summarization.com

的新颖度，换个角度说，我们需要量两个不同句子之间的信息冗余度。信息冗余度量化之后，我们就可以根据预设的阈值，检测出新的信息。

但是，信息冗余度很难严格地量化，研究者提出了各种技术手段并在实际数据集上进行了测试实验，下面，我们介绍几种常用的信息新颖度的计算方法。

2.3.1 词重叠度

将两个句子分别看成词集合 A 和 B，其中 A 在 B 之前。最简单的信息冗余度可以量化为$\left|A \cap B\right|$，[Zhang 2002]定义了一种相对的词重叠度概念：

$$OverlapB_A = \frac{A \cap B}{B}$$

因此，B 的词重叠度就可以定义为：$\max\{OverlapB_i \mid i为B之前的句子\}$。

2.3.2 最大区间相关度（Maximum Marginal Relevance）

最大区间相关度最早是[Carbonell 1998]提出的。在新信息检测方面，我们主要兼顾当前句子与查询和以前句子的相关度，寻找两者区间最大的结果。[Sun 2003]定义如下：

$$MMR(s_i) = \lambda Sim_1(V_{si}, V_{tp}) - (1-\lambda)\max_{S_j \in R} Sim_2(V_{si}, V_{sj})]$$

其中：s_i为当前句子，V_{si}、V_{tp}、V_{sj}分别为当前句子、主题和以前句子的向量表示形式，R 为 s_i之前的相关句子集合，Sim_1为句子和主题相似度计算函数，而 Sim_2为句子之间的相似度计算函数，λ为两者之间的调节参数，$0 \le \lambda \le 1$，一般设为 0.7。

2.3.3 Cosine 冗余度

两个句子 s_i与 s_j可以采用 Cosine 计算向量之间的相似度，即：

$$\cos(s_i, s_j) = \frac{\sum_{k=1}^{n} v_{i,k} \times v_{j,k}}{\left\|s_i\right\| \cdot \left\|s_j\right\|}$$

换个角度来看，句子之间的相似度实质上也是两个句子之间的信息冗余度[Tsai 2002]。因此，可以采用 Cosine 来计算当前句子与以前句子的信息冗余度，排除信息冗余度大于一定阈值的句子，就可以最终获得新信息结果。

2.3.4 命名实体触发方法

在新信息检测中，我们还可以另外采用一些启发式的规则，一般来说，新的人物、新的机构或者新的地方往往意味着有新的信息内容。因此，可以得到一条启发式规则：出现了新的命名实体的句子往往包含了新的信息。

[Abdul-Jaleel 2004]提出：如果句子中出现的人名、机构名或者其他命名实体是以前没有见过的，则断定该句包含了新的信息。他们在新信息检测系统中引入了 20 多种命

名实体，其中包括：人名（PERSON），地名（LOCATION），机构名（ORGANIZATION），日期（DATE）以及金钱货币（MONEY）等。

命名实体识别是经典的信息抽取问题，也是国际 MUC(Message Understanding Conference)大会主要的评测任务，计算语言学方面的研究者有过详尽地研究和探讨。目前已经拥有了相对成熟的技术， BBN 公司的 IdentiFinder[Bikel 1999]就是一个比较优秀的英文命名实体识别系统。

2.3.5 统计机器翻译模型

CMU 大学提出了一种基于统计机器翻译模型的新信息检测方法[Thompson 2002]。基本思想为：将冗余句子视为第一个句子的统计机器翻译结果，在这种语言上，如果可以构建一个好的翻译模型，那么，系统就可以检测出这两个句子是否翻译了同样的内容，从而发现新的信息内容。在新信息检测的应用中，源语言（待翻译的语言）与目标语言（翻译结果的语言）都是同一种语言，因此，问题就大大简化了。在这里，我们主要借助于 WordNet 来估计词语与短语的相似度，同时采用浅层句法分析来抽取并比较不同句子的结构。

给定两个待比较的句子，该算法主要包括下面几个阶段：首先，我们先获得每个句子的句法分析树；其次，将每棵句法分析树转化为图的表示形式，要求该图可以表达出各个词语修饰成分的结构；再次，实现一个简单的图匹配算法，从而比较句子中的各个词语，其中词语的权重主要表示可能的重要度；最后，计算出一个句子在同一语言上翻译为另一个句子的概率，这个概率可以视为两个句子之间的相似度。

下面是两个从 TREC 中训练文档中抽取的例句：

Sentence A:
Some of the best shots, **released** this month by the US space agency **Nasa**, **show** parts of the universe billions of light years away - and therefore **billions** of **years** in the **past**.

Sentence B:
The images sent back this **year**, after astronauts repaired the telescope’s defective mirror, **show** a **myriad** of astronomical objects too distant to be seen with the most **powerful** Earth-bound observatories.

采用统计机器翻译模型，按照上面的过程计算句子 A 和 B 的相似度：先抽取相似度最高的六个词对，表 2.1 给出了计算的中间结果。其中第三列给出的是句子 B 中与对应词相似距离最小的词，例如：与句子 A 中的“past” 相似度距离最小的词为“year”。相似距离计算主要采用 WordNet，并结合统计信息。我们会在后面的章节做深入地讨论。最后，这两个句子的相似距离为这六个词对的平均权重，结果为：14.821，小于阈值 15，因此可以判定：B 为冗余信息。

Sentence A	Graph weight A_i	Sentence B (Most similar word)	Graph weight B_i	Similarity Distance S_i
past	6	year	4	2.456
years	3	year	4	0.0258
released	4	show	2	53.631
Nasa	5	powerful	1	68.240
show	2	show	2	0.000
billions	3	myriad	1	0.152
Weighted mean: 14.821				

表 2.1 采用统计机器翻译模型计算的句子 A 与 B 的信息相似度

2.3.6 LexRank 方法

2.2.6 中，我们介绍了 LexRank 的概念。[Erkan 2004]采取了 LexRank 来对当前句子与以前句子的冗余度进行建模。

首先，按照相关句子向量之间的 Cosine 值建立相似句子有向图，边的权值为节点之间的相似度，图中边的方向表示的是句子之间的时序关系。图 1 给出了一个相似句子有向图的示例，其中句子 1 与句子 2 和句子 4 相似，而句子 2 与句子 3 相似。

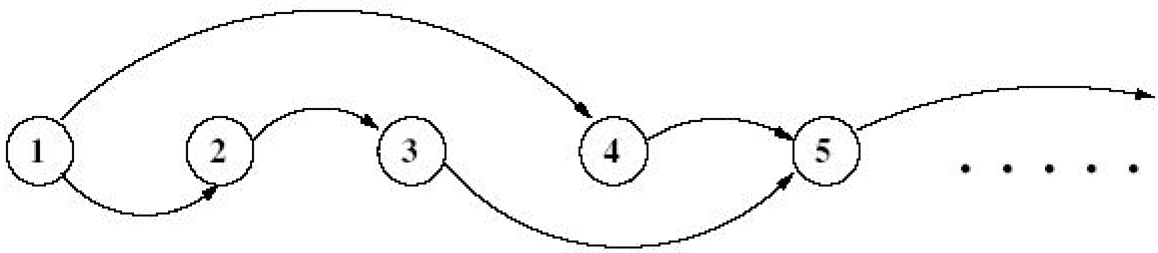


图 2.6　采用 LexRank 表示的相似句子有向图

其次，为了评估每个句子节点的新信息量，我们在有向图上计算 LexRank 值，理想的情况下，LexRank 值较小的句子应当包含新信息，因为 LexRank 值较小意味着句子基本上没有入边或者入边很少，也就是说它和以前的句子基本上都不相似，反之亦然。因此，在图 2.6 中，句子 5 包含新信息的概率非常小，基本可以断定这是一个冗余的句子。而第一个句子没有入边，它的 LexRank 值总是很小，因此它肯定是新信息句子。

最后，我们仅仅依靠 LexRank 值，采取一个简单的决策算法就可以预测出所有的新信息结果。

2.4 本章小结

本章主要总结了新信息检测技术的研究进展，并综合阐述了句子检索与新信息检测的方法。

新信息检测历史很短，但是具有潜在的应用前景，已经引起了国际上的大学、科研

机构、商业公司的广泛关注。新信息检测本身是一个主观性很强的应用，目前仍然处在研究的初级阶段，性能尚有一定的提升空间，仍然是一个有待进一步深入研究的课题。

句子检索方法的方法主要有：传统的文档检索方法、信息过滤方法、分类方法、语义比较方法、隐马模型方法与自动文摘方法等。

新信息检测的核心问题在于如何度量两个不同句子之间的信息冗余度。常用的评估手段有：词重叠度、最大区间相关度、Cosine 冗余度、命名实体触发方法、统计机器翻译模型、LexRank 方法。

第三章 Noovel 特定的浅层语言分析

自然语言作为人类进化的产物，是信息的高级载体，是人类发展过程当中自然产生、约定俗成的用于人类社会交际的工具，如英语、汉语、日语等。自然语言是人际交互的基本语言，是社会文化的传承载体。人们往往采用自然语言来传递自己的思想以及需求，而计算机系统需要对自然语言表达的内在内容进行建模，转化为可计算的数据结构，尽可能地进行表层的理解并进一步提取其深层的语义信息。

在这一章中，我们主要介绍 Noovel 特定的浅层语言分析，特定指的是这里并不做通用的语言分析，而是针对新信息检测的特定问题进行有目的性的研究。浅层指的是分析只涉及语言的段落、词法等层面，不进行句法乃至语义层面的深层挖掘。本文需要特别说明的是：浅层分析并不是无结构的，浅层并不意味着表层，它需要针对待分析的对象，采用综合的语言分析技术，并同时权衡应用系统的特性，把握好其中的分析尺度，适可而止。这对语言分析技术和信息处理过程来讲都是一大挑战。太宽泛的表层分析达不到应用的要求，太精细的语言解析又需要耗费过多的资源，而且在性能上也会有过犹不足的负面作用。一个有针对性的浅层语言分析直接影响着信息检索与过滤的质量。本文中会在实际的处理环节过程中，从理论分析和实验数据两个方面展开论述。

我们主要介绍英文的浅层分析过程，停用词处理与特征提取算以及查询分析。另外，在最后一节介绍我们在 ICTCLAS 系统中所采取的汉语浅层语言分析模型，主要包括汉语的分词、命名实体识别与词性标注等。

3.1 自然语言的特点与语言计算分析

语言大概可分为以下几种：动物本能语、自然语言（约定俗成语、法制语）、程序语言（比如机器语言、汇编语言、高级语言、函数语言、面向对象语言）、形式语言（比如巴柯斯范式、逻辑语言）和抽象语言（比如数学语言）。语言在物理世界的主要实现方式是语音方式（通过听觉）和文字方式（通过视觉），对应的语言层次就是语音和文字]。通过语音识别和文字识别，语言赖以实现的物理信号转化为符号，其最小的单位是字符。[白 2001

3.1.1 自然语言的特点

顾名思义，“自然语言”具备两个属性特点：语言属性与自然属性。

首先是语言属性，即存在着为大家所公认的某些约定俗成的内在规律性，任何一种语言的构成在形式上都具有一定的规范体系，说话者依此通过语言来表达自己的思想，听众根据相同的规范从语言中解析出内在的信息，这样便于各种思想在不同群体之间的

传达与沟通。这一点类似于信道模型，输入对象是说话者，输出对象是听众，信道上传输的物理信号就是语言，而信号的加工与解析方法就是语言的内在规律。比如在汉语里面，“把车开回北京去”是一个符合这种内在规定性的构成形式，而“车回开北京把去”就是一种不符合这种内在规律性的构成形式。这种内在规律性也就构成了语言自身的本质特点，各种不同语言之间的内在规律特点大相径庭，而这种规律性就是区分不同语言的主要标准。比如，英语与汉语之间的差异性就非常大，只有同时掌握其内在的语言特点，人们才可以顺利地完成中英文之间的相互翻译。

其次是自然属性，也就是说并不存在某个人为制造的、严格的语法规则体系来约定人们的语言表达方式，它是随着历史的进程自然而然形成的。自然语言的语言特点在随着历史的进化而变化；而在同一阶段时期，各个人的表述也各具特点，这也极大地丰富了人类的自然语言与客观世界。程序语言与形式语言就大不相同，这些都是人们抽象并逻辑化的“人工”语言，必须遵循严格的语法规则，它们往往采用正则文法体系就可以完全表达，其中任何正确的语句都可以进行正确地解析，而且只有一种正确的解释，肯定不存在二义性，这样大大地方便了自动的计算与推理。而自然语言并不存在什么通用的语法体系，语言学家往往试图对语言进行概括，但任何一部语法体系几乎毫无例外地都存在着特例。

总之，自然语言需要遵循一定的内在规律，但更大程度上是“存在即合理”。一方面，自然语言现象只要能传情达意，就能存在；另一方面，只有错误的语法体系或者语言表达模型，而不存在错误的语言现象。比如：汉语中的“你先走”符合人们的正常使用习惯，后来随着《大话西游》在网络上的流行，“你走先”、“给个理由先”都逐渐成为了合理的语言新现象。

3.1.2 自然语言的计算分析

传统的语言研究是为语言教学、文献整理、社会历史研究服务的，完全是面向人的，这样的研究搞了近两千年，已经取得了可观的成绩。电子计算机出现以来，人与计算机之间要进行信息的传输和交流，因此，除了继续进行面向人的语言研究之外，还要开展面向计算机的语言研究。学者们开始采用计算机技术来分析并计算自然语言。语言的计算分析首先是从机器翻译系统的研究开始的。1946 年电子计算机刚一问世，人们在把计算机广泛地应用于数值运算的同时，也想到了利用计算机把一种或几种语言翻译成另外一种语言或另外几种语言。四十年来，这项研究取得了长足的进展，成为了一门重要的新兴学科：自然语言处理。

计算机对自然语言的研究和处理，一般应经过如下三个方面的过程：[冯 1995]

第一，把需要研究的问题在语言学上加以形式化（linguistic formalism），使之能以一定的数学形式，严密而规整地表示出来；

第二，把这种严密而规整的数学形式表示为算法（algorithm），使之在计算上形式化（computational formalism）；

第三，根据算法编写计算机程序，使之在计算机上加以实现（computer implementation）。

因此，为了研究自然语言处理，我们不仅要有语言学方面的知识，而且，还要有数学和计算机科学方面的知识，这样自然语言处理就成为了一门界乎于语言学、数学和计算机科学之间的边缘性的交叉学科，它同时涉及到文科、理科和工科三大领域。

3.1.3 自然语言分析的不同层次知识

一个自然语言系统必须考虑许多语言自身结构方面的知识，其中包括：什么是词、词如何组成句子、词的意义是什么、词的意义对句子意义有什么贡献等等。而这还是远远不够的，比如说，一个系统如果要回答提问或者直接参与对话，它不仅需要知道很多语言结构的知识，而且还要知道人类世界的一般性知识，并具备人类的推理能力。

下面，我们给出了与自然语言分析有关的不同层次知识[刘 2005]。

- 语音和音韵知识

关注的是词语与其发音的关系。这种知识对于语音相关的系统是至关重要的。

- 词语形态学知识

关注的是词素如何构成词语。词素是语言中一种最基本的意义单位（例如，单词“friendly”的语义可以由名词“friend”的语义以及将名词转换为形容词的后缀“-ly”推导出来）。

- 句法知识

关注的是如何将词语放到一起组成正确的句子，并判定每个单词在句子中所充当的结构角色，以及短语之间的构成关系。

- 语义知识

关注的是词语的语义、以及词语语义在句子中是如何互相结合并形成句子的整体语义。这是上下文无关的语义研究，即不考虑句子所处的上下文。

- 语用知识

关注的是如何在不同的情形下使用句子，以及这种使用如何影响句子的解析。

- 篇章知识

关注的是前面的句子如何影响下一个句子的解释。这种信息对于代词解析以及时态解释特别重要。

- 世界知识

即人类世界的一般性知识，这种知识对语言的使用者来说是必须的，比如说，要进行一个对话，系统就需要利用这种知识。每一个谈话者都必须了解其他谈话者的观念和谈话目的。

以上这些定义不是一种精确的定义，事实上，它们更多的是这些知识的一些特征，而不是对这些知识进行严格的分类。任何一个特定的事实都可能涉及到多个不同的层面，而且一个算法也可能需要同时从几个不同的层面进行提炼。

3.2 新信息检测与浅层语言分析

新信息检测实质上是一个具备排除冗余信息机制的片断检索技术，这和传统的文档检索大不相同。一篇文档一般都包含数十个句子，150 个以上的单词，而新信息检测处理的基本单元是一个句子，包含的单词一般只有 10 个左右，去除虚词等停用词之后，包含的实词往往不到 5 个。因此，很多用于文档检索的方法，比如纯粹的文档词频统计，在新信息检测方面往往很难取得满意的效果。单词词形本身中蕴含的信息非常少，因此，我们必须借助自然语言理解的技术，从自然语言表达的主题中提炼出用户真正的查询意图，并从单个句子中抽取内在的语言信息。

在 Noovel 的系统架构中，我们不难发现：无论采取什么样的句子检索模型，无论利用什么样的新信息检测技术，针对新信息检测的浅层语言分析都是首要环节。浅层语言分析器是 Noovel 的基础与前提，语言分析的质量制约着新信息检测的最终性能。相对于文档级信息检索而言，新信息检测更倚重于自然语言的处理。这是一个自然语言处理与信息检索过滤的综合技术。

在以前的新信息检测研究过程中，研究者们几乎都采用了自然语言处理的相关技术，常用的处理过程包括词干还原、词性标注、命名实体识别、查询分析等。还有的研究者进一步引入了一些先验的本体语义知识，如 WordNet[Fellbaum 1998],[Ganesh 2003], [Sun 2003], [Tsai 2003], [Zhang 2003], [Zhang 2002], [Ryosuke 2003]、同义词典[Zhang 2002]、概念模糊集[Ryosuke 2003]、语义树[Jin 2003]等资源。

然而，到目前为止，研究者更多的是关注句子检索方法，尤其是新信息的检测建模技术，往往是采取简单的语言分析工具。还没有研究者研制出有针对性的语言分析技术，也没有进一步通过实验来验证语言分析的作用。我们的研究以及试验表明，前期的语言分析往往能起到更关键的作用。在有针对性的语言分析结果基础上，Noovel 采取了简单的向量空间模型进行检索并检测新的信息，最终的性能可以超过以往最好的评测结果。在简单的语言分析中间结果基础上，无论采取怎样复杂的新信息检测技术，最终性能都非常有限。这表明：浅层语言分析决定新信息检测性能的上限，它在整个框架中起着关键的基础性作用。

在新信息检测系统 Noovel 中，我们引入了浅层语言分析过程包括：断句、词汇切分、词性标注、词形还原、查询分析、停用词、索引与特征选择，如图 3.1 所示。其中在预处理阶段，英文需要进行断句、词汇切分与词形还原，中文对应的浅层处理过程是汉语分词。查询分析过程从主题中抽取可以直接建模计算的查询词语，停用词与特征选择的目的在于剔除干扰信息并保留有代表性的内容。索引过程主要采用倒排索引，便于词汇的检索统计，并作进一步的模型计算。

接下来的章节中，我们主要介绍与新信息检测相关的英文浅层分析过程；随后，集

中介绍特征选择与停用词处理，以及查询分析的技术，这两个过程所体现的方法均与语言无关；最后一节介绍汉语相关的浅层分析方法。

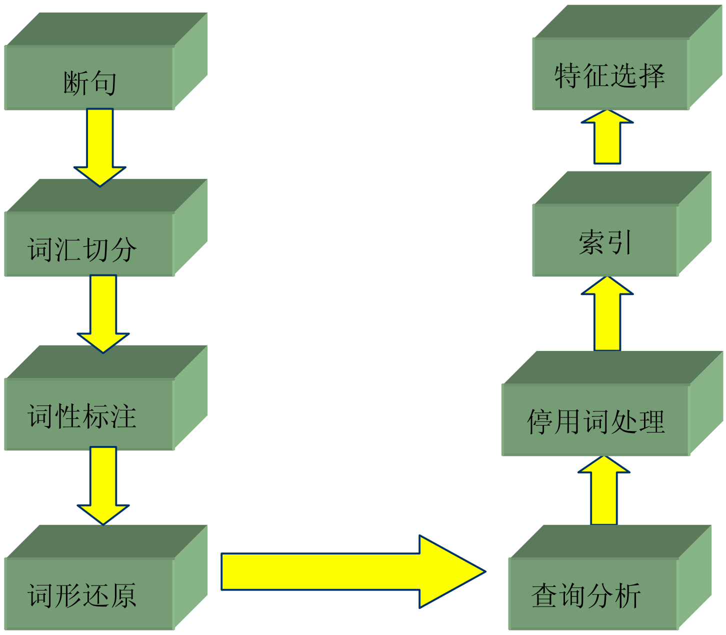


图 3.1　Noovel 中的浅层语言分析过程

3.3 英文浅层分析

这一节，我们主要集中介绍英文浅层语言分析的几个环节：英文断句、词汇切分、词性标注以及词形还原。

3.3.1 英文断句(Sentence Boundary Detection)与词汇切分(Tokenization)

新信息检测的信息粒度是一个具有独立意义的单句，因此，语言处理的第一个环节就是将一个完整的句子从文本中的字符流中切分开来，即断句。中文的句子标记很少存在歧义，如句号（。)、问号（？)、感叹号（！)，因此，断句相对简单。在这里，我们主要介绍英文的处理。TREC的新信息检测任务中，给定的文档事先都进行了自动断句，不过主题的描述部分都是自由文本，如图3.2所示。断句对单个查询意图的分析非常重要，我们会在查询分析中作进一步的阐述。

英文的句子主要是以问号（?)、感叹号(!)、引号（′）和句点(.)作为结束标记。在没有上下文的情况下，单个的词语或者短语也可以作为一个句子。而有的句子不带标点，直接用空格或者回车换行来表示句子的结束，但是图3.2中的回车换行符号仅仅等同于连接的空格，断句时需要根据下一个词语进行区分。问号（?)、感叹号(!)作为句子的结束符号，基本上不存在歧义，但需要注意的是：并不是所有的句点(.)都是句子的结束符号，

它还可以是多种词汇的组成部分，如缩略名、数字等。在布朗语料库[Francis&Kucera 1982]中，总共包含了48,885个句子，其中就有3490个句子中至少包含了一个作为非终结符号的句点。因此，我们需要在词汇的切分基础上进行断句，断句的歧义往往是由词汇切分歧义引起的。

```
<num>Number: N53
<title>Dragging Death of James Byrd, Jr.
<toptype>Event
<desc>Description:
Dragging death of black man, James Byrd, Jr.
<narr>Narrative:
To be relevant, a document contains any opinion of
the family, the public, the police, the judicial or
even those of the news reporter as to the reason
for the dragging. Also relevant is the ongoing
investigation into the crime, the suspects, the
juror selection, and the trial results regarding the
dragging death of James Byrd, Jr.   Documents that reflect
only on the incident without elaboration are not relevant.
```

图 3.2 Noovel 中主题的原始文字示例

词汇切分（Tokenization）的目标就是从原始的字符流中分离出类似于词语的独立单元序列[Grefenstette 1994]，如字符串“I’m a Ph.D. candidate.”，词汇切分后的结果为“I#’m#a#Ph.D.#candidate#.”①。英语词汇切分的任务类似于汉语的分词，但难度要小得多，其关键在于考虑分隔符在词汇中的特殊使用，主要包括下面几类情况[刘2003]:

1. 日期与数字

例如：2/14/2005，2005-2-14，Feb. 14, 2005，2005.2.14，12:18
123,456.78，-90.7%，3/8

2. 缩略名

缩略名主要包括两种情况：

1）字母－点号－字母－点号组成的序列，比如：U.S. i.e. Ph.D. 等等；

2）字母开头，最后以点号结束，比如：Mr.，Dr.，eds.，Prof.

3. 包含非字母字符

比如：AT&T Micro$oft

① 为了呈现方便，我们这里引入了“#”分隔各个单元

4. 带杠的词串

   比如：three-years-old，one-third，so-called

5. 带撇号的词串

   比如：I'm，can't，dogs'，let's

6. 带空格的词串

   比如："and so on"，"ad hoc"

7. URL、Email、数学表达式等其他情况

   比如：http://lcc.software.ict.ac.cn/~zhp/，zhanghp@software.ict.ac.cn，a+b=c

对于日期、数字、URL 等规则明确的情况，我们采用正则表达式[杜 1990]进行表示，其中：[0-9]+(\/[0-9]+)+表示分数与日期，([+\-])?[0-9]+(\.)?[0-9]*% 表示百分数，([0-9]+,?)+(\.[0-9]+|[0-9]+)* 表示十进制数字。对于其它常见的特殊词汇，我们大部分都收录在了一部 15000 条词条的英文词典中。

图 3.3 给出了英文词汇切分与断句的实现算法，最后的实验结果表明该算法能够满足新信息检测的需要。

（1）对一个待分析的字符串（S），如果当前位置 P 已经是 S 的末尾，转至（END）；否则，从当前位置 P 开始，由左至右进行扫描，读取一个子串 W 直到空格或者字符串结束符号为止，同时更新 P

（2）如果 W 在词典 LEXICON 中存在，作为词汇单位输出，转至（1）；否则：

  1）P 在 S 的末尾，输出 W 并设置句子分隔标记，转至（1）。

  2）如果 W 的最后一个符号属于集合{“.”、“!”、“?”}；或者 W 的最后一个符号为{“"”、“}”、“}”}，而且下一个子串 W 的首字母大写，在 W 的最后一个字符前增加词汇分割符号，依次输出这两个词汇单元，并设置句子分隔标记。转至（1）。

  3）如果 W 的首字符属于集合{“(”、“{”、“[”、“<”、“"”}，在 W 的首字符之后增加词汇分割符号，依次输出这两个词汇单元。

  4）否则：采取相关正则表达式进行校验 W，输出 W 这个词汇单元，转至（1）。

（3）如果不是字符流尾部，转入（1）；

（END）

图 3.3 Noovel 中的英文词汇切分与断句算法

3.3.2 词性标注（Part-Of-Speech Tagging）

词性是词汇最重要的特性之一，是连接词汇到句法的桥梁。自动的词性标注是自然语言浅层理解的一个重要环节，它可以帮助系统自动地判定词语所属的语法范畴(grammatical category)，为进一步的语言处理提供更高层面上的支持。词性标注主要任

务是消除词性兼类歧义，是自然语言处理中重要而又成熟的技术，对于新信息检测来说，它的实际意义还在于：

(1) 能够很大程度上实现词义的歧义；

例如：下面的两个句子

“The pill-induced abortion <u>*can*</u> be painful, causing bleeding and nausea for days”；

“Food in <u>*cans*</u> is called <u>*canned*</u> food.”

第一个句子中的 can 是情态动词，仅仅是起到了词汇连接作用，并不具备实际的语义。而在第二个句子中，第一个“cans”表示的是名词“罐头”，第二个“canned”表示的是形容词“罐装的”，它们都表达了具体的实体对象。仅从词形出发，系统很难判别它们之间的差别，往往会作为停用词统一处理。

(2) 具有提高句子检索性能的潜力

词性标注可以帮助我们保留表征实际意义的名词、动词、形容词以及数词等实词，同时过滤掉对检索不具备实际影响的介词、代词、连词与冠词。相对于纯粹从词形出发的过滤手段，词性更能有效地消除噪声，在句子层面的检索上更具有潜力。这一点，我们会在停用词部分作进一步的说明。

(3) 提高区分信息新颖程度的能力

新信息本质上是寻找与以前不同的相关信息，词性标注结果有助于甄别同形异用的词汇，同时还可以帮助对词汇进行语法范畴的分类。

按照学习过程与标注过程的关系，我们可以将词性标注方法分为监督型与非监督型两种。监督型标注器一般都采取事先已经标注好的语料库作为知识来源，并从中学习抽取出标注过程需要的资源，比如：标注词典、词语/词性频率、词性序列的概率以及规则集合。而非监督词性标注模型并不需要已经标注好的语料库， 往往采取比较复杂的计算方法自动地进行词聚类，然后根据聚类结果来估算相关概率，或者推导出上下文规则。两种方法各有利弊，可以满足不同情况的实际需求。

按照采用的技术方法来划分，我们还可以将词性标注分为三类：规则方法、统计方法与神经网络方法。传统的规则方法往往根据上下文信息来标记未知词与歧义词的词性。比如：规则“det - X - n = X/adj”说的就是“如果歧义词 X 在限定词与名词之间，则其词性为形容词”。因此，“a good boy”中的“good”为形容词；而“I'm telling you this for your good.”（我告诉你这件事是为你好）中的“good”为名词。 除了上下文信息之外，构词规则也往往起到关键的作用，比如字母大小写、前后缀、标点等。在英文中，后缀为“-ness”的词往往是名词。规则往往是语言学家手工整理，有的研究者还尝试自动地从语料库中自动地抽取规则[Brill 1995]。典型的系统有 TAGGIT[Greene&Rubin 1971]、Brill Tagger[Brill 1992]。统计方法可以是任何一种结合了概率信息的标注模型，其目标就是为句子中的词串选择一个最可能的词类序列。80 年代初由英国 Lancaster 大学的研究小组设计的 CLAWS(Constituent-Likelihood Automatic Word-Tagging System)系统则第一次在词性自动标注中运用了统计语言学的模型。该系统利用带有词类标记的

Brown 语料库，通过统计分析获得一个反映任意两个邻接标记同现频率的转移概率矩阵，根据这种统计信息进行词性标注。该系统的正确率达 96-97%。在 CLAWS 之后，词性自动标注多采用统计的方法，典型的统计方法有 n 元语言模型、最大熵方法、隐马模型[Baum 1972] [Lawrence 1989][刘&张 2005]。实现 n 元模型标注最常用的方法是 Viterbi 算法[Viterbi 1967][Brill& Marcus 1994][刘&张 2005]。80 年代中期后，人工神经网络(artificial neural network) 即连接机制 (connectionism)兴起，运用神经网络[Helmut 1994]的方法进行词性自动标注的研究也取得了一定的成果。

在 Noovel 的英文词性标注中，我们主要在 Eric Brill 标注器[Brill 1992]的源代码基础上作了进一步的改进，采用的词性标注集为宾州大学树库的标注集合，参见附件 2。Brill 标注器采用的是基于错误驱动的词性标注方法，其基本过程如下：

(1) 初始词性赋值；

(2) 对比正确标注的句子，自动学习结构转换规则；

(3) 利用转换规则调整初始赋值。

Eric Brill 词性标注器除了存在严重的内存泄漏等技术缺陷之外，它不能实际应用的问题还在于它只能处理一个单句的词汇串，而不能适应实际的自由文本。比如原来的系统只能接受“I#’m#a#Ph.D.#candidate#.”，而不能正确处理实际句子“I’m a Ph.D. candidate.”。为此，我们做了如下改进：

(1) 在词性标注之前，增加了断句与词汇切分预处理过程；

(2) 针对新信息检测的需要，增加了新的词条以及部分针对主题分析的规则集合。

改进后的词性标注器能较好的适应各种文本，并能有针对性的分析主题与文档内容。

3.3.3 词干抽取（Stemming）与词形还原（Morphplogical Normalization）

信息检索关的键问题是判断数据集合中的哪些文档能够满足用户的信息需求，检索与否的决策取决于查询词与文档中索引词的比较。然而，与中文不同的是，文档和查询中的英文词汇（Term）往往有多种多样的变种，如果不进行自然语言处理，“computing”与“computation”就不会被看作是同等的。在绝大多数情况下，同一词语的各种词形变换形式在语义解释上往往相同，在信息检索的角度来看，它们可以认为是无差别的。

在大多数文档级的信息处理应用中，词干抽取（Stemming）是最常用的一种词形变换手段。词干抽取可以处理词形的变化形式，把所有同根词转变为单一形式。例如：“computing”、“computation”、“compute”、“computer”、以及“computable”等词语统一转换为词干“comput”。对于信息检索来说，词干抽取减少了不同词语的数量，帮助识别了相似的词语，在不进行语言分析的同时，以较小的代价提高了检索性能。比较有影响力的词干抽取工具有：Porter Stemmer[Porter 1980]、Paice/Husk Stemmer [Paice 1990]、Lovins Stemmer [Lovins 1968] 等。

词干抽取基本上能满足大规模文档级信息处理的要求。然而，对于句子检索与新信

息检测来说，它还远远不能达到要求。其主要缺陷在于：

（1） 词干抽取本身就存在欠处理（understemming）与过处理（overstemming）问题。自然语言本身的结构并不存在一个绝对的规则，因此词干抽取不可避免地会出现错误。一方面，本来应该合二为一的词在词干抽取之后仍然不同，即欠处理问题。例如“adhere”与“adhesion”词干抽取的结果分别为“adher”与“adhes”。实际上，它们的语义是对等的。另外一方面，本来存在很大差别的词对往往在词干抽取之后变成了同形词，即过处理问题。例如：“experiment”与“experience”词干抽取的结果都是“experi”; 动词“computing”与名词“computer”同样会被看成是同一个结果“comput”。

（2） 一个句子中包含的词汇数目本身就非常很有限，词干抽取的欠处理降低了相关句子被检索到的概率，从而降低了句子检索的召回率；同样，一个词汇的过处理往往会导致非相关句子被错误检索，从而降低了句子检索的准确率。

（3） 新信息检测主要是从词汇之间的差异判断信息的新颖程度，而词干抽取更多的是将不同词性、不同词形的词汇融合为同一个词干，这在一定程度上抹杀了信息之间的差别，从而降低了新信息检测的召回率。

基于以上的考虑，在新信息检测中，我们采用了更为精细的词形还原，而没有采用常用的词干抽取。词形还原的目标是还原词语的词根、去除在实际文本中出现的各种形态，主要包括：名词的单复数、动词的各种时态与语态变化、形容词和副词的比较级与最高级形式。与词干抽取不同的是，词性还原的结果是一个实际的词汇，并且与原词的词性保持一致，能在去除形态变换的同时，最大限度地保留原词的意义。

Noovel 中的词性还原是在词性标注的基础上实现，其主要过程如图 3.4 所示。

输入：词语 W、其词性 P

（1） 根据词语 W 的词性 P，查询英文词典 Lexicon，如果存在，输出 W。否则转至（2）;

（2） 根据词语 W 的词性 P 查找对应的特例列表 ExceptList(P)，如果属于 ExceptList(P)中的特例情况，直接输出对应的词根 R，返回。否则转至（3）

（3） 遍历对应词性的转换规则表 TransRule(P)，依次尝试每条转换规则，转换后的结果为 R’，查询 R’在英文词典 Lexicon 中是否存在，如果存在，输出对应词根 R’，返回；否则尝试下一条规则，直到遍历完成为止；

（4） 输出 W

图 3.4 Noovel 中的英文词形还原算法

3.4 停用词处理与特征选择

文档主要是通过其文本内容中的词语相互区分，不过，不同词语的分辨力千差万别。

分辨力指的是一个词作为特征将它所在的文档与其它文档区别开来的能力。在新信息检测中，我们需要过滤掉没有分辨力或者分辨力很低的词语，并保留具有一定分辨力的特征词语，这样可以排除干扰因素，提高最终的性能。在 Noovel 的研究过程中，我们主要采取两种选择策略：停用词处理与特征选择。

3.4.1 停用词处理

常用的停用词处理非常简单，首先人工整理收集一份类似于黑名单的停用词表，一旦文档中出现了停用词表中的词，则将它删除，最后只保留停用词表之外的词语。最终的性能和停用词表的覆盖面息息相关，英文常用的停用词表大约有 540 条词语。Noovel 系统主要依据标注的词性去除停用词，最后再采用停用词表排除剩下的词语。具体过程如下：

（1） 保留所有的名词 N、动词 V、形容词 J 以及副词 R；

（2） 出现在文档标题或者主题描述中出现的数词 CD 予以保留，除此之外的所有词语均予以过滤；

（3） 剩余的结果如果出现在停用词表中，将其过滤；

在很多情况下，数词并不影响信息检索的结果，很多研究者往往将其视为停用词不予保留，或者不参与相关性计算。但是，在文档标题或者主题描述中出现的数词往往具有很强的区分能力，例如：TREC2004 年主题 N56 中的主题描述（词性标注结果）：

“Woodstock/JJ 99/NN music/NN festival/NN reunion/NN in/IN Rome/NNP ,/, NY/NNP”

在这里的数词“99”强调的是 Woodstock 音乐节的年份，以示和其他年份的音乐节的区分。因此，我们保留了有区分能力的数词。

3.4.2 特征选择

停用词处理仅仅是从语言功用的角度筛选出具有实际语义的词语，还不足以选择出对句子检索具有分辨力的特征词语。文档往往采用高频的词语来强调需要表达的特征涵义，而句子中往往是单一的词语，因此，句子检索需要进一步选择更有区分能力的特征词。为此，我们进一步引入了文本分类过滤中常用的特征选择过程。句子集合的初始特征中可能存在很多噪声，通过特征选择舍弃一些不太重要的词，将有效地消除噪声词语的影响。特征选择具有降低向量空间维数、简化计算、消噪、防止过分拟合等作用。

Noovel 中采用了互信息与 χ^2 统计相结合的思路。

3.4.2.1 互信息方法

互信息是计算语言学模型分析的常用方法，它度量两个对象之间的相关性。在新信息检测中，Noovel 采用互信息度量特征词语与主题的关联度。特征 w 对于主题 T 的互信息量 $MI(w,T)$的计算公式如下：

$$MI(w,T) = \log \frac{P(w \mid T)}{P(w)}$$

其中：$P(w|T)$表示的是特征词语 w 在与主题 T 相关的文档集合中的分布概率，$P(w)$表示的是特征词语 w 在所有文档集合中的分布概率，其中包括与主题 T 相关和不相关的文档集合。在 Noovel 系统中，我们将主题 T 相关的文档作为 $P(w|T)$的概率空间，而将所有 50 个主题的文档集作为 $P(w)$的概率空间。对于具有相同条件概率 $P(w|C_i)$的特征，低频特征的 $P(w)$较小，因此互信息比中频特征要高，因此，出现频率差异很大的特征的互信息大小不具有可比性。

选择 $MI(w,T)$较大的特征组成特征空间，因为对于每一主题来讲，特征 w 的互信息越大，说明它与该主题的共现概率越大，因此，以互信息作为特征提取的评价标准时应选择互信息最大的若干个特征。我们在实验的基础上，Noovel 系统采用的阈值为 0.1。

3.4.2.2 χ^2统计量方法

χ^2统计量用于度量特征词语 w 和主题 T 之间的独立性。$\overline{w}$表示除 w 以外的其它特征，$\overline{T}$表示除 T 以外的其它主题，那么特征 w 和主题 T 的关系有以下四种情况：(w,T), $(w,\overline{T})$, $(\overline{w},T)$, $(\overline{w},\overline{T})$，用 A, B, C, D 分别表示这四种情况的文档频次，总的文档数 $N=A+B+C+D$，χ^2统计量的计算公式如下：

$$\chi^2(w,T) = \frac{N \times (AD - BC)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

当特征 w 和主题 T 之间完全独立的时候，χ^2统计量为 0。χ^2统计量和互信息的差别在于它是归一化的统计量，但是它对低频特征的区分效果也不好。在 Noovel 系统中，我们根据已有的研究结果[Manning& Schütze 1998]，将经验阈值设置为 3.841。即选取χ^2统计量大于 3.841 的特征词组成特征空间。

在 Noovel 系统中，特征选择的结果是互信息特征空间与χ^2统计量特征空间的交集，目的在于相互弥补两种方法之间的不足，选取区分能力更强的特征词。

3.4.3 浅层语言分析的中间结果

到目前为止，我们已经阐述了除查询分析之外的所有浅层语言分析任务。为了进一步的实验并跟踪中间的结果，Noovel 将浅层语言分析结果作为中间结果输出到一个 XML 格式的文件中。图 3.5 给出了其中的一个句子片断。

```
<SENTENCE>He/PRP will/MD also/RB seek/VB the/DT general/JJ 's/POS extradition/NN to/TO
Spain/NNP ./.    </SENTENCE>
<SENTENCE_Query>seek/VB general/JJ extradition/NN spain/NNP </SENTENCE_Query>
<SENTENCE_QueryID> 24358/NN 28581/JJ 63288/VB 66983/NNP</SENTENCE_QueryID>
```

图 3.5　浅层语言分析的中间结果片断

其中：<SENTENCE>部分为词性标注后的原始词串结果；<SENTENCE_Query>部分为停用词处理与特征选择之后的结果；<SENTENCE_QueryID>部分将特征词串转换为唯一对应的 ID，便于进一步的快速计算。接下来的文档检索、句子检索与新信息检测过程都是基于浅层分析之后的 ID 序列结果。ID 序列与具体的语言无关，完全适应于各种语言的处理。

3.5 查询分析

查询请求是句子检索与新信息检测的需求与根据。对查询的理解直接决定最终的性能，错误的理解不可能产生正确的检索结果。与传统检索应用不一致的是：新信息检测中并不存在直接的查询请求，查询内容都在自然语言表达的主题描述之中，如图 3.2 所示。这一点和自动问答有相似之处，在真正检索响应之前，我们需要从自然语言描述中抽取出用户真正的查询意图。

在主题描述中，停用词处理之后的词语可以分为辅助词与查询词两大类。查询词就是那些真正表达用户查询意图的词语；辅助词尽管能够独立地表达实际的意义，但是在主题中仅仅起到语气连贯、礼貌客套等辅助作用，和实际要查询的内容并不存在本质的联系，其功能等同于英语中的“by the way”（顺便问一下）之类的插入语，没有这些词，并不影响人们对真正语义的掌握。例如：自然语言查询语句“麻烦帮我查查《中国农民调查》这本书。”真正的查询请求是“中国农民调查”，而“麻烦”与“查”等词都是辅助词。针对 TREC 的主题描述，我们专门收集整理了 100 多个频繁出现的辅助词词表，比如 “document”、“mention”、“find”、“describe”以及“opinion”等。查询分析过程中，辅助词往往会干扰对用户真实意图的理解，因此，我们只保留辅助词之外的查询词。

除此之外，我们还要根据查询的倾向，将查询分为正向查询与逆向查询。正向查询指的是用户表示需要的信息请求；反之，逆向查询就是那些用户指定不需要的信息过滤要求。比如：“气功摧残身体的信息不相关”就属于逆向查询。正向查询与逆向查询的区分处理，可以帮助我们进一步甄别出不同的查询请求，简单地混杂在一起进行查询，往往会检索到用户并不需要的结果。

在这里，我们假定单个英文句子只能表达一种查询倾向，也就是说一个句子不可能

同时表达正向查询与逆向查询。它要么是正向查询，要么是逆向查询，要么就没有表达任何查询内容。因此，那些包含了“irrelevant”、“irrelvance”或者同时出现“not”与“relevant”等表示否定意义的句子可以认为是逆向查询。而其他的句子均可以视作正向查询请求。

> To be *relevant*, a *document contains* any *opinion* of the family, the public, the police, the judicial or even those of the news reporter as to the reason for the dragging. Also *relevant* is the ongoing investigation into the crime, the suspects, the juror selection, and the trial results regarding the dragging death of James Byrd, Jr. <u>*Documents* that *reflect* only on the incident without elaboration are not *relevant*.</u>

图 3.6 查询分析结果示例

图 3.6 给出了 TREC2004 主题 N53 的部分片断。其中斜体表示的词语都属于辅助词，下划线表示的句子为逆向查询。

经过词性标注、停用词处理、特征选择以及查询分析等浅层语言过程之后，主题就可以解析为能直接用于检索计算的正向查询与逆向查询序列。图 3.7 给出了 TREC 主题 N53 的解析结果。其中<queryid>部分给出的是用于正向查询的特征词语 ID 序列、<stop_queryid>给出的是逆向查询的特征词语 ID 序列。它们的知识来源主要是主题描述的标题、简介以及详细描述字段。这三个部分的重要程度很不一样，在具体计算的时候，我们经验地将它们的权重比设置为 4:2:1，从而表征出主题各个不同部分的重要性。

```
<top>
<num>Number: N53
<title>Dragging/VBG Death/NN of/IN James/NNP Byrd/NNP ,/, Jr./NNP
<title_query>drag/VBG death/NN jame/NNP byrd/NNP jr./NNP </title_query>
<desc_query>drag/VBG death/NN black/JJ man/NN jame/NNP byrd/NNP jr./NNP </desc_query>
<desc_queryid> 7411/JJ 9968/NNP 17609/NN 20697/VBG 37598/NNP 42684/NN
84386/NNP</desc_queryid> <stop_query>incident/NN elaboration/NN </stop_query>
<stop_queryid> 21991/NN 34921/NN</stop_queryid>
<queryid> 7411/JJ 9968/NNP 9968/NNP 9968/NNP 16303/NN 17609/NN 17609/NN 17609/NN
20697/VBG 20697/VBG 20697/VBG 20697/VBG 24744/NN 36469/NN 37598/NNP 37598/NNP
37598/NNP 37634/JJ 37767/NN 42684/NN 47901/NN 50268/VBG 54400/NN 56697/NN 58291/NN
58948/VBG 59580/NN 59961/NNS 63393/NN 69926/VBZ 73647/NN 84386/NNP 84386/NNP
84386/NNP</queryid>
</top>
```

图 3.7 主题解析结果示例

3.6 汉语浅层分析与ICTCLAS②

词是最小的能够独立活动的有意义的语言成分[朱 1982]。在汉语中，词与词之间不存在分隔符，词本身也缺乏明显的形态标记，因此，汉语浅层分析的特有问题就是如何将汉语的字串分割为合理的词语序列，即汉语分词。汉语分词是句法分析等深层处理的基础, 也是机器翻译、信息检索和信息抽取等应用的重要环节。

从 1983 年第一个实用分词系统CDWS[梁 1987]的诞生到现在，国内外的研究者在汉语分词方面进行了广泛的研究，提出了很多有效的算法。我们可以粗略地将这些方法分为两大类：第一类是基于语言学知识的规则方法，如：各种形态的最大匹配、最少切分方法、以及综合了最大匹配和最少切分的N-最短路径方法 [张 2002]，还有的研究者引入了错误驱动机制[Hockenmaier 1998]，甚至是深层的句法分析[Wu 1998]。第二类是基于大规模语料库的机器学习方法 [Palmer 1997][Dai 1999][高 2001]，这是目前应用比较广泛、效果较好的解决方案。用到的统计模型有N元语言模型，信道-噪声模型，最大期望[Peng 2001]，隐马模型等。在实际的分词系统中，往往是规则与统计等多类方法的综合。一方面，规则方法结合使用频率，形成了可训练的规则方法[Palmer 1997]；另一方面，统计方法往往会自觉不自觉地采用一些规则排除歧义、识别数词、时间及其他未登录词。同时，我们也注意到国外同行提出的一些探索性算法，如：基于压缩的方法[Teahan 2001] ，分类器的方法[Xue 2002]，无词典的自监督学习方法[Peng 2001]。目前，对不含歧义和未登录词 ③的文本进行分词，已有方法和系统表现相当出色，其准确性已经达到相当高的水平。

实际上，汉语分词的主要瓶颈是切分排歧和未登录词识别。切分歧义和未登录词降低了自身正确切分的可能性，同时还干扰了其相邻词的正确处理。更糟糕的是，未登录词往往和切分歧义交织在一起，进一步增加了分词的难度。如：在 “克林顿对内塔尼亚胡说”中，“内塔尼亚胡”是一个词典中没有收录的译名，实际切分的时候，“对”与“内”，“胡” 与“说”往往会粘在一起，最终导致错误的切分结果：“克林顿/对内/塔

② 本节主要的研究内容，笔者以第一作者身份最早发表在“Hua-Ping ZHANG, Qun LIU, Xue-Qi CHENG, Hao Zhang, Hong-Kui Yu. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, Second SIGHAN workshop affiliated with 41st ACL; Sapporo Japan, July, 2003, pp. 63-70”，在这里，我对论文的合作者刘群老师、程学旗老师等表示感谢。

③我们这里所说的未登录词指的是核心词典中没有收录而又不能用正则表达式描述的词，如没有被收录的人名、地名。

尼亚/胡说/”。

文献[Yu 2001]对切分歧义进行了较好的形式化描述，并引入了嘈杂度的概念对歧义的程度进行了定量的描述。目前切分排岐的研究路线基本上以规则为主[张 1998][Kit 2002][Zheng 1999]，还有的针对某一类歧义，引入了一些成熟的模型作专门处理，如引入向量空间解决组合歧义问题[Luo 2002]。在未登录词识别方面，主要的出发点是综合利用未登录词内部构成规律及其上下文信息。未登录词识别处理的对象主要是人名、地名、译名和机构名等命名实体。在语料库不足的条件下，未登录词识别唯一的出路是采取精细的规则[Luo 2001][Luo 2001(2)][Sun 1993][Tan 1999]，规则一般来源于观察到的语言现象或者是大规模的专名库。目前比较成功的解决方案大都是从大规模的真实语料库中进行机器学习，解决方案有隐马模型[张 2004][ZHANG 2002]、基于 Agent 的方法[Ye 2002]、基于类的三元语言模型[Sun 2002]等。

经过二十余年的努力，研究者们在分词算法、切分排歧和未登录词识别方面均取得了较大的进展。然而，现有方法和系统往往缺乏一个相对统一的模型框架将三者进行有机的融合。排歧、未登录词识别往往和分词相对独立，排歧的结果和识别出的未登录词缺乏科学的可信度计算，即使量化，往往流于经验，很难在量值上与普通词作真正意义上的比较。一般都倾向于假定排歧和未登录词结果正确无误，忽略具体的分词算法，直接修正分词结果。现有的分词方法更大程度上是专门切分出词典中收录的词，基本上没有将未登录词和歧义纳入到同一个算法体系当中，一旦遇到歧义或者未登录词就作为特例进行专门处理，因此使用的模型和方法都没有贯彻到底，缺乏统一的处理算法，对切分结果也缺乏统一的评估体系。最终导致分词的准确率在开放测试的条件下并不像宣称的那样理想，处理含有未登录词、歧义字段的真实文本时，效果更是不尽人意。

我们在这一节介绍我们提出的一种基于层次隐马模型的方法，旨在将汉语分词、切分排歧、未登录词识别、词性标注等浅层语言分析任务融合到一个相对统一的理论模型中。首先，在预处理的阶段，采取 N-最短路径粗分方法，快速的得到能覆盖歧义的最佳 N 个粗切分结果；随后，在粗分结果集上，采用底层隐马模型识别出普通无嵌套的人名、地名，并依次采取高层隐马模型识别出嵌套了人名、地名的复杂地名和机构名；然后将识别出的未登录词以科学计算出来的概率加入到基于类的切分隐马模型中，未登录词与

歧义均不作为特例，与普通词一起参与各种候选结果的竞争。最后在全局最优的分词结果上进行词性的隐马标注。该方法已经应用到了中科院计算所汉语词法分析系统 ICTCLAS 中，取得了较好的分词和标注效果。ICTCLAS 在 973 专家组机器翻译第二阶段的评测和 2003 年 5 月 SIGHAN 举办的第一届汉语分词大赛中，取得了不俗的成绩，是目前最好的汉语词法分析系统之一。

在下面的小节里,我们将概述层次隐马模型和汉语浅层语言分析的总体框架，随后介绍基于类的切分隐马模型；然后分别叙述基于角色隐马模型的未登录词识别方法，以及切分排岐的 N-最短路径粗切分策略，最后给出各种条件下的对比测试结果，以及国家 973 开放评测和国际分词大赛的测试结果，并给出简单分析。

3.6.1 层次隐马模型和汉语浅层语言分析

3.6.1.1 层次隐马模型概述

隐马模型(Hidden Markov Model; HMM)[Lawrence 1989]是经典的描述随机过程的统计方法，在自然语言处理中得到了广泛的应用。然而，相对于复杂的自然语言现象来说，传统的 HMM 仍然略显简单，为此，文献[Shai 1998]对 HMM 进行了扩展和泛化并提出了层次隐马模型(Hierarchical Hidden Markov Model; HHMM)。

在文献[Shai 1998]的工作基础上，我们将 HHMM 形式化为六元组 M=< Ω_X , Ω_O , A, B, Π, D>，其中：Ω_X 为有限状态集合；Ω_O 为观测结果有限集；A 为状态转移矩阵；B 为状态到观测值的概率矩阵；Π 为初始状态分布；D 为 M 的深度。HHMM 与 HMM 的主要区别在于：

1) Ω_X 中状态的表示为：q_i^d ($d\in\{1,\ldots,D\}$)其中 i 表示的是该状态在当前层 HMM 状态中的编号；d 是该状态在 M 中的层次深度，所有状态形成了一个深度为 D-1 的树型结构，其中根的深度为 1，最深的叶子深度为 D。$d<D$ 的状态我们称之为内部状态。

2）每一个内部状态 q^d ($d\in\{1,\ldots,D\text{-}1\}$)存在子状态，子状态数记为$|q^d|$，所有的子状态构成一个隐马尔科夫链，d+1 层的状态输出可视为该层 HMM 的状态序列，该层 HMM 的状态转移矩阵为 $A(q^d)=(a_{ij}(q^d))$,其中$a_{ij}(q^d)=P(q_j^{d+1}|q_i^{d+1},q^d)$ 同时，各个状态的初始

分布为： $\Pi(q^d)=\pi^d(q_i^{d+1})=P(q_i^{d+1}|q_i^d)$，其物理意义可以理解为第 d 层 HMM 的某一内部状态 q^d 到第 d+1 层 HMM 的激活概率。

3)各层 HMM 中，只有第 D 层才有真正能观测到的终结符，即状态到观测的输出概率 $B(q^D)=(b_k(q^D))$,其中$b_k(q^D)=P(o_k|q^D)$；o_k为观测值，属于一个有限终结符集合。

因此，HHMM 的参数集合可以表示为：

$$\lambda=\{(\lambda(q^d)\}_{d\in\{1,...,D\}}$$
$$=\{\{A(q^d)\}_{d\in\{1,...,D-1\}},\{\Pi(q^d)\}_{d\in\{1,...,D-1\}},\{B(q^D)\}\}$$

实际上，HHMM 在 D=1 时就会退化成简单的 HMM。

3.6.1.2 基于 HHMM 的汉语浅层语言分析框架

针对汉语浅层语言分析各个层面的处理对象及问题特点，我们引入 HHMM 统一建模，该模型包含原子切分、普通未登录词识别、嵌套的复杂未登录词识别、基于类的隐马切分、词类标注共五个层面的隐马模型，如图 3.8 所示。

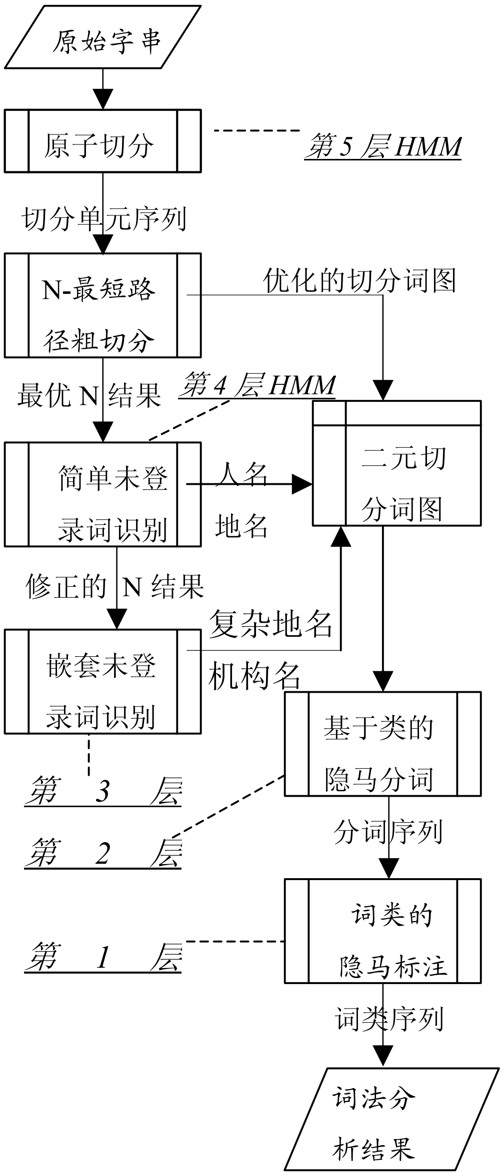


图 3.8 基于 HHMM 的汉语词法分析框架

其中，N-最短路径粗切分可以快速产生 N 个最好的粗切分结果，粗切分结果集能覆盖尽可能多的歧义。在整个词法分析架构中，二元切分词图是个关键的中间数据结构，它将未登录词识别、排歧、分词等过程有机的进行了融合，在分词模型中会详细地介绍。

原子切分是词法分析的预处理过程，主要任务是将原始字符串切分为分词原子序列。分词原子指的是分词的最小处理单元，在分词过程中，可以组合成词，但内部不能做进一步拆分。分词原子包括单个汉字，标点以及由单字节、字符、数字等组成的非汉字串。如“2002.9，ICTCLAS 的自由源码开始发布”对应的分词原子序列为“2002.9/，/ICTCLAS/的/自/由/源/码/开/始/发/布/”。在这层 HMM 中，终结符是书面语中所有的字

符，状态集合为分词原子，模型的训练和求解都比较简单，本文就不赘述。词法分析的最高层隐马模型是词性标注过程，和第 5 节中的角色标注过程本质一样，在这里不重复论述。下面主要介绍汉语浅层语言分析中其他层面的隐马过程。

3.6.2 基于类的隐马分词算法

本算法处于 HHMM 的第二层，也就是在所有的未登录词识别完成后进行。首先，我们可以把所有的词按照图 3.9 分类，其中，核心词典中已有的每个词对应的类就是该词本身。这样假定核心词典中收入的词数为|Dict|，则我们定义的词类总数有：|Dict|+6。

$$c_i=\begin{cases} w_i & \text{iff } \; w_i\text{在核心词典中收录} \\ \text{PER} & \text{iff } \; w_i \text{ 是人名 and } w_i \text{ 是未登录词;} \\ \text{LOC} & \text{iff } \; w_i \text{ 是地名 and } w_i \text{ 是未登录词;} \\ \text{ORG} & \text{iff } \; w_i \text{ 是机构名 and } w_i \text{是未登录词;} \\ \text{NUM} & \text{iff } \; w_i \text{是数词 and } w_i \text{是未登录词;} \\ \text{TIME} & \text{iff } \; w_i \text{是时间词 and } w_i \text{是未登录词;} \\ \text{BEG} & \text{iff } \; w_i \text{是句子的开始标记} \end{cases}$$

图 3.9 词的分类

给定一个分词原子序列 S，S 的某个可能的分词结果记为 $W=(w_1,\ldots w_n)$, W 对应的类别序列记为 $C=(c_1,\ldots c_n)$，同时，我们取概率最大的分词结果 $W^{\#}$作为最终的分词结果。则：

$$W^{\#}=\arg\max_{W} \mathrm{P}(W)$$

利用贝叶斯公式进行展开，得到：

$$W^{\#}=\arg\max_{W} \mathrm{P}(W|C)\mathrm{P}(C)$$

将词类看作状态，词语作为观测值，利用一阶 HMM 展开得：

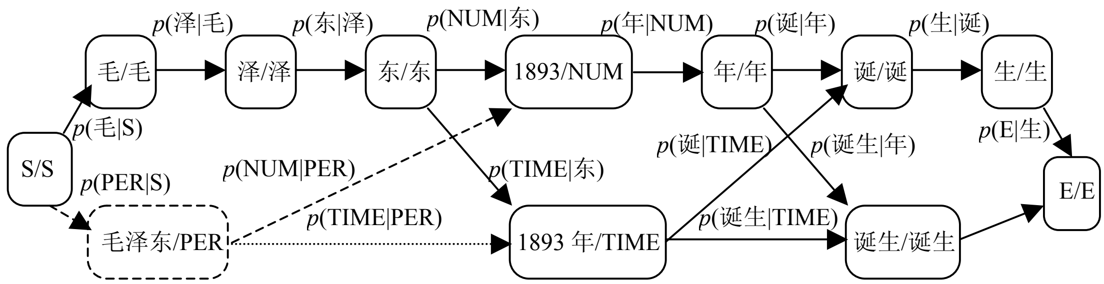
$$W^{\#}=\arg\max_{W} \prod_{i=1}^{n} p(w_i \mid c_i)p(c_i \mid c_{i-1})$$ (其中 c_0 为句子的开始标记 BEG，下同.)

为计算方便，常用负对数来运算，则：

$$W^{\#}=\operatorname*{argmin}_{W} \sum_{i=1}^{n}[-\ln p(w_i \mid c_i)-\ln p(c_i \mid c_{i-1})]$$

根据图 2 中类 c_i 的定义，如果 w_i 在核心词典收录，可以得到 $c_i=w_i$，因此 $p(w_i|c_i)=1$。NUM 和 TIME 两类词的构成符合正则文法，ICTCLAS 可以采用确定性的有限状态自动机进行识别，基本上不存在歧义组合的问题，我们将之视为确定性的词类，和标点符号一样处理，不作为未登录词对待，将相应的 $p(w_i|c_i)$概率值作为一个常数，从而将问题进一步简化。因此，在分词过程中，我们只需要考虑未登录词的 $p(w_i|c_i)$。在图 3.10 中，我们给出了“毛泽东 1893 年诞生”的二元切分词图。最终所求的分词结果就是从初始节点 S 到结束节点 E 的最短路径，这是个典型的最短路经问题，可以采取贪心算法，如 Dijkstra 算法快速求解。

在实际应用基于类的分词 HMM 时，切分歧义能否在这一模型内进行融合并排解是一个难题；另外一个关键问题还在于如何确定未登录词 w_i、识别其类别 c_i 并计算出可信的 $p'(w_i|c_i)$；本文的随后两节将依次阐述这两类问题的解决思路。



说明：

1. 节点中表示的是“词语/类”(即 w_i / c_i),节点的权值为类到词语的概率 $p(w_i | c_i)$;
2. 有向边的权值为相邻类的转移概率 $p(c_i | c_{i-1})$; S 为初始节点；E 为结束节点
3. “毛泽东/PER”相关的虚线部分是人名识别 HMM 作用过之后产生的。

图 3.10 基于类的二元切分词图（原始字串为“毛泽东 1893 年诞生”）

3.6.3 N-最短路径的切分排歧策略 ④

从构成形态上划分，切分歧义一般分为交叉歧义和组合歧义。“结合/成/分子/时”是个典型的交叉歧义，“这/个/人/手/上/有/痣”中的“人/手”构成了一个组合歧义字段。

从排歧的角度看，切分歧义可以分为全局歧义和局部歧义。全局歧义指的是必须结合当前句子的上下文才能准确排除的歧义。如“乒乓球拍卖完了”，在缺乏语境的情况

④这里的主要研究内容，笔者最早以论文的形式发表在“张华平，刘群.基于 N-最短路径的中文词语粗分模型. 中文信息学报. 2002.9, Vol.16(5):pp.1-pp.7; ”，对第二作者表示感谢。

下，可以合理地切分为“乒乓球/拍卖/完/了”和“乒乓球拍/卖/完/了”。与此相反，局部歧义完全可以在句子内部进行排除，本节开头列举的两个歧义句子均属于局部歧义。根据我们对大规模语料的统计发现，局部歧义占绝大多数，全局歧义几乎可以忽略不计。

我们采取的是 N-最短路径的切分排歧策略[张 2002]。其基本思想是在初始阶段保留切分概率 P(*W*)最大的 N 个结果，作为分词结果的候选集合。在未登录词识别、词性标注等词法分析之后，再通过最终的评价函数，计算出真正最优结果。实际上，N-最短路径方法是最少切分方法和全切分的泛化和综合。一方面避免了最少切分方法大量舍弃正确结果的可能，另一方面又大大解决了全切分搜索空间过大，运行效率差的弊端。

该方法通过保留少量大概率的粗分结果，可以最大限度地保留歧义字段和未登录词。常用切分算法往往过于武断，过早地在初始阶段做出是否切分的判断，只保留一个自己认为最优的结果，而这一结果往往会因为存在歧义或未登录词而出错，这时候，后期补救措施往往费时费力，效果也不会很好。表 3.1 给出了 8-最短路径与常用算法在切分结果包容歧义方面的对比测试结果。

方法	切分最大数	切分平均数	正确切分覆盖率
最大匹配	1	1	85.46%
最少切分	1	1	91.80%
最大概率	1	1	93.50%
全切分	>3,424,507	>391.79	100.00%
8-最短路径	8	5.82	99.92%

表 3.1 N-最短路径与常用算法对比

说明：

1)切分最大数指的是句子可能的最大切分结果数。

2)切分平均数指的是单个句子平均的切分结果数。

3)正确切分覆盖率=正确切分被覆盖的句子数/句子总数

4)测试语料大小为 200 万汉字

同时，我们对最终选择出的唯一切分标注结果进行了开放歧义测试，测试集合是北大计算语言所收集的 120 对常见组合歧义、99 对常见交叉歧义，最终组合歧义和交叉歧

义排除的成功率分别为 80.00%和 92.93%。

3.6.4 未登录词的隐马识别方法

未登录词识别的任务有：1）确定未登录词 w_i 的边界和类别 c_i ；2）计算 $p(w_i|c_i)$。我们在N个候选切分结果的词类序列基础上，引入了高层HMM来实现未登录词的识别。

3.6.4.1 未登录词识别角色表

和基于类的隐马分词模型类似，我们对初始切分得到的各个词按照其在未登录词识别中的作用，进行分类，并将词所起的不同作用称为角色。表 3.2 是人名识别的角色表。与隐马分词中定义的类相比，角色不同的是：类和词是一对多的关系，而角色与词是多对多的关系，即：一个词可以充当多个角色，而一个角色也可以对应多个词。

角色	意义	示例
A	人名的上文	又/*来到*/于/洪/洋/的/家
B	人名的下文	新华社/记者/黄/文/*摄*
C	中国人名的姓	*张*/华/平/先生；*欧阳*/修
D	双名的首字	张/*华*/平/先生
E	双名的末字	张/华/*平*/先生
F	单名	张/*浩*
G	人名的前缀	*老*/刘、*小*/李
H	人名的后缀	王/总、刘/老、肖/氏
L	译名的首部	*蒙*/帕/蒂/•/梅/拉/费
M	译名的中部	蒙/*帕*/*蒂*/*•*/*梅*/*拉*/费
N	译名的末部	蒙/帕/蒂/•/梅/拉/*费*
O	日本人名末部	小泉/纯/一/*郎*
X	连接词	邵/钧/林/*和*/稽/道/青/说
Z	其它	*人民*/*深切*/缅怀/邓/小/平

表 3.2　人名识别角色表

3.6.4.2 角色标注与未登录词识别

对于一个给定的初始切分结果 $W=(w_1,\ldots w_n)$,在一个角色集合的范畴内，假定 $R=(r_1,\ldots r_n)$为 C 的某个角色序列。我们取概率最大的角色序列 $R^{\#}$作为最终的角色标注结

果。和第 3.6.2 节隐马分词的推导过程类似，我们最终可以得到：

$R^{\#} = \arg\min_{R} \sum_{i=1}^{n} [-\ln p'(w_i \mid r_i) - \ln p(r_i \mid r_{i-1})]$ (其中 r_0 为句子的开始标记 BEG，下同.)

$R^{\#}$可以通过 Viterbi 算法[Viterbi 1967] [Lawrence 1989]选优得到；图 4 给出了在词类序列“毛/泽/东/TIME/诞生”的 Viterbi 算法标注人名角色的过程。（这里 TIME 是在原子切分阶段通过简单的有限状态自动机识别出来的。）

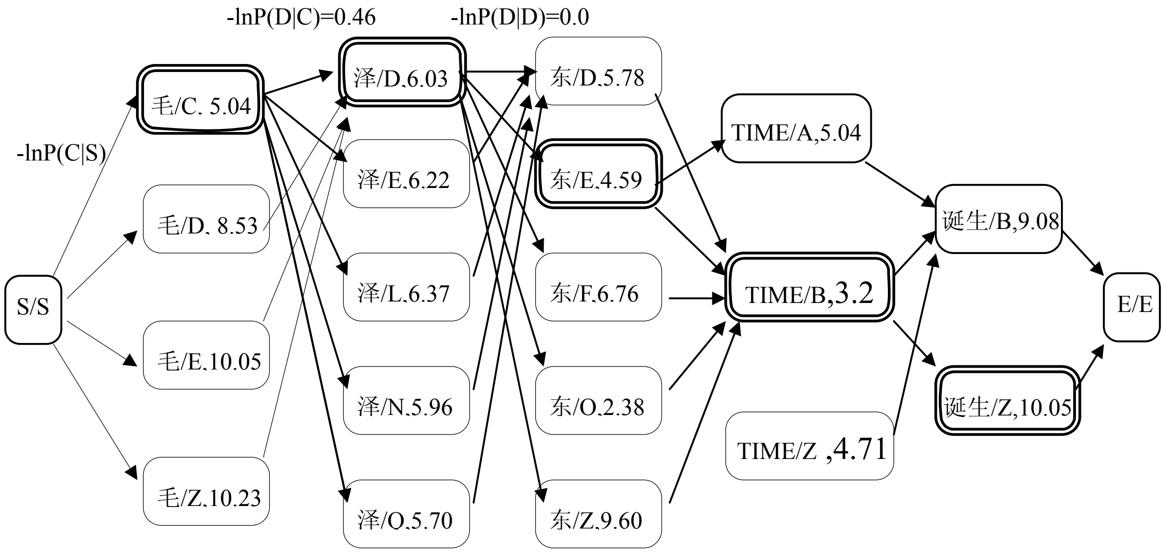
在最大概率角色序列的基础上，我们可以简单的通过模板匹配实现特定类型未登录词的识别。在图 3.11 中，我们可以求解出最优的角色标注结果：“毛/C 泽/D 东/E TIME/B 诞生/Z”,而 CDE 正好构成一个典型的汉语人名，因此“毛泽东”被识别为人名 PER。

识别出来的未登录词为 w，类别为 c，利用隐马过程可以得到:

$p(w|c) = \prod_{j=0}^{k} p(w_{p+j} \mid r_{p+j}) p(r_{p+j} \mid r_{p+j-1})$；

其中 w_i 由第 p，$p+1...p+k-1$ 个初始切分单元组成。

最后，识别结果及其概率加入到二元 HMM 切分图中，和普通词一样处理，竞争出最佳结果，如图 3 中的虚线部分所示。



说明：1. 图中结点的格式为：词类 c_i /角色 r_i, -logP(c_i | r_i)，双线节点 Viterbi 选优结果。

2. 图中的有向边权值为相邻角色的转移概率–logP(r_i | r_{i-1})，这里没有全部列出。

图 3.11 角色标注的 Viterbi 算法选优过程

3.6.4.3 未登录词的识别

复杂地名和机构名往往嵌套了普通无嵌套的人名、地名等未登录词，如“张自忠路”、“周恩来和邓颖超纪念馆”。对于这种嵌套的未登录词，我们的做法是：在低层的 HMM 识别过程中，先识别出普通不嵌套的未登录词，然后在此基础上，通过相同的方法采取高层隐马模型，通过角色标注计算出最优的角色序列，在此基础上，进一步识别出嵌套的未登录词。以切分序列片断“周/恩/来/和/邓/颖/超/纪念馆”为例，我们先识别出“周恩来”和“邓颖超”为人名 PER，得到新的词类序列“PER/和/PER/纪念馆”，最终就可以识别出该片段为机构名。这样的处理优点在于能够利用已经分析的结果，并降低数据的稀疏程度。

我们用来训练HMM角色参数的语料库是在北大计算语言所切分标注语料库的基础上，甄别出各种类型的未登录词之后，自动转换得到的。

3.6.5 实验与分析

采取HHMM的方法，我们研制出了计算所汉语词法分析系统ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System，该系统全部的源码和文档，均可在中文自然语言处理开放平台www.nlp.org.cn中自由下载，免费使用)。下面我们给出ICTCLAS在不同条件下的测试结果，提供ICTCLAS在国家973英汉机器翻译第二阶段评测中开放测试的结果，并介绍我们在第一届国际分词大赛中的比赛情况。在这里，我们按照惯例引入如下评测指标：切分正确率SEG,上位词性标对率TAG1,下位词性标对率TAG2，命名实体(NE)识别的准确率P和召回率R，以及F值。它们的定义分别如下：

SEG=切分正确的词数/总词数×100%；

TAG1=上位词性标注正确数/总词数×100%；

TAG2=下位词性标注正确数/总词数×100%；

P=正确识别该类NE数/识别出该类NE总数×100%

R=正确识别该类NE数/该类NE总数×100%

$F=\frac{R\times P\times(1+\beta^2)}{R+P\times\beta^2}$ ，这里我们取β=1，称为F-1值。

3.6.5.1 词法分析与 HHMM

我们使用北京大学计算语言学研究所加工的《人民日报》语料库进行了训练和测试。在人民日报 1998 年一月份共计 1,108,049 词的新闻语料库上，我们进行了如下四种条件下的性能测试：

1）BASE: 基准测试，即仅仅做隐马分词和词性标注，不引入其他层面的 HMM；

2）+PER: 在 BASE 的基础上引入人名识别 HMM。

3）+LOC: 在+PER 的基础上引入地名识别 HMM。

4）+ORG: 在+LOC 的基础上引入机构名识别 HMM。

图 3.12 给出了四种条件下，词法分析的分词正确率 SEG、上位词性标对率 TAG1、下位词性标对率 TAG2，人名识别的 F-1 值 FP、地名识别的 F-1 值 FL 以及机构名识别的 F-1 值 FO。

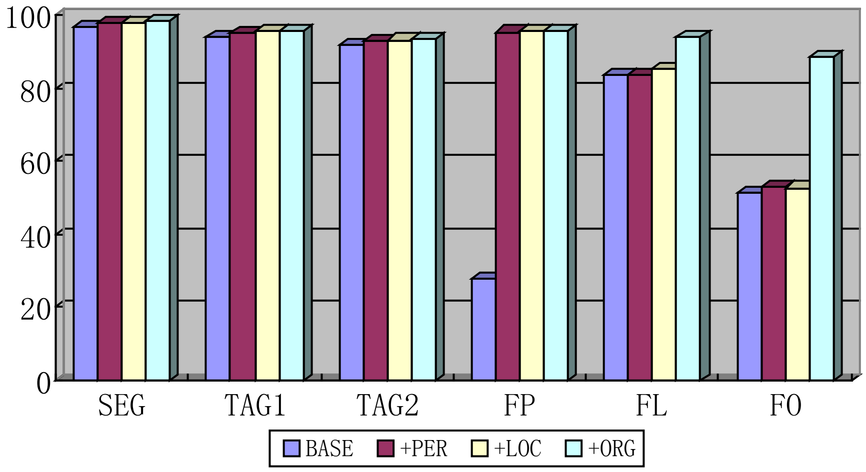


图 3.12 四种条件下的词法分析的性能指标

从图 3.12 中我们可以发现：

1) 随着各层隐马模型的逐层加入，词法分析的效果逐步提升。其中，人名识别引入后，切分正确率 SEG 在 96.55%的基础上，增加到 97.96%，增幅最大。人名、地名、机构名等识别 HMM 均加入后，切分正确率 SEG、上位词性标对率 TAG1、下位词性标对率 TAG2 分别达到了 98.38%,95.76%,93.52%。这表明：各层 HMM 对最终词法分析均发挥了积极作用。

2）同时，随着各层 HMM 的加入，不仅极大的提高了本层 HMM 的最终性能，还改进了低层 HMM 处理精度。人名识别 HMM 加入后，人名识别的 F-1 值，立即从 27.86%

提升到 95.40%，低层的分词 HMM 的正确率提高了 1.41%；机构名识别 HMM 引入后，机构名识别的 F-1 值提高了 35.59%，同时低层的地名识别也提高了 8.49%，人名识别的 F-1 值也达到了最高点 95.58%。其原因在于：高层 HMM 的成功应用在解决当前问题的同时，也消除了低层 HMM 部分的歧义，排除了低层 HMM 的错误结果。例如：在人名识别 HMM 中很容易将“刘庄的民风很纯朴” 中的“刘庄”错误识别为人名，然而，高层的地名识别 HMM 会正确地将“刘庄”作为地名召回，因此达到了排歧的作用。

3.6.5.2 ICTCLAS 在 973 评测中的测试结果

2002 年 7 月 6 日，ICTCLAS 参加了国家 973 英汉机器翻译第二阶段的开放评测，测试结果如下：

领域	词数	SEG	TAG1	RTAG
体育	33,348	97.01%	86.77%	89.31%
国际	59,683	97.51%	88.55%	90.78%
文艺	20,524	96.40%	87.47%	90.59%
法制	14,668	98.44%	85.26%	86.59%
理论	55,225	98.12%	87.29%	88.91%
经济	24,765	97.80%	86.25%	88.16%
总计	208,213	97.58%	87.32%	89.42%

表 3.3. ICTCLAS 在 973 评测中的测试结果

说明：

1) 数据来源：国家 973 英汉机器翻译第二阶段评测的评测总结报告；

2) 标注相对正确率 RTAG=TAG1/SEG*100%

3) 由于我们采取的词性标注集和 973 专家组的标注集有较大出入，所以词性标注的正确率不具可比性。

专家组的开放评测结果表明：基于 HHMM 的 ICTCLAS 能实际的解决汉语浅层语言分析问题，和兄弟单位的类似系统对比，ICTCLAS 的分词结果表现出色。

3.6.5.3 第一届国际分词大赛的评测结果

为了比较和评价不同方法和系统的性能，第四十一届国际计算语言联合会(41st Annual Meeting of the Association for Computational Linguistics, 41th ACL)下设的汉语特

别兴趣研究组(the ACL Special Interest Group on Chinese Language Processing, SIGHAN；www.sighan.org) 于 2003 年 4 月 22 日至 25 日举办了第一届国际汉语分词评测大赛(First International Chinese Word Segmentation Bakeoff) [Richard 2003]。报名参赛的分别是来自于大陆、台湾、美国等 6 个国家和地区，共计 19 家研究机构，最终提交结果的是 12 家参赛队伍。

大赛采取大规模语料库测试，进行综合打分的方法，语料库和标准分别来自北京大学（简体版）、宾州树库（简体版）、香港城市大学（繁体版），台湾“中央院”（繁体版）。每家标准分两个任务(Track)：受限训练任务(Close Track)和非受限训练任务(Open Track)。

ICTCLAS 分别参加了简体的所有四项任务，和繁体的受限训练任务。其中在宾州树库受限训练任务中综合得分 0.881[Richard 2003]，名列第一；北京大学受限训练任务中综合得分 0.951[Richard 2003]，名列第一；北京大学受限训练任务中综合得分 0.953[Richard 2003]，名列第二。值得注意的是，我们在短短的两天之内，采取 ICTCLAS 简体版的内核代码，将层次隐马模型推广到繁体分词当中，同样取得了 0.938[Richard 2003]的综合得分。

3.7 本章小结

在这一章中，我们首先介绍了自然语言的一般特性，并探讨了自然语言计算的一般规律，然后，详细阐述了针对新信息发现的中英文浅层语言分析技术，论证了浅层语言分析在 Noovel 系统中的基础性作用。

针对 Noovel 的英文浅层语言分析主要包括断句、词汇切分、词性标注以及词形还原。我们一一介绍了各个环节存在的问题与困难，对现有方法进行了简单综述。并在已有工作的基础上，结合新信息发现的特点，提出了有针对性的改进措施。

中文浅层语言分析的特有问题是汉语的分词，我们提出了一种基于层次隐马模型的汉语浅层语言分析方法，旨在将汉语分词、词性标注、切分排歧和未登录词识别集成到一个完整的理论框架中。不同层面的实验表明，层次隐马模型的各个层面对汉语词法分析都发挥了积极的作用。

停用词处理、特征选择以及查询分析属于和语言无关的独立过程。

在新信息检测过程中，停用词处理与特征选择能够过滤掉缺乏分辨力的词语，并保留具有一定分辨力的词语特征，这样可以排除干扰因素，提高最终的性能。停用词处理采用了词性过滤与词形过滤两个过程，有针对性地排除了对信息检索没有实质意义的词

汇。特征选择采用了互信息与χ^2统计相结合的思路，进一步地选取了更有区分能力的特征词。

查询分析是对用户查询意图的理解，直接决定了最终的检索结果是否符合需求。我们首先将主题中的词汇分为查询词与辅助词两大类，这样可以帮助过滤与查询无关的干扰内容。其次，我们将查询倾向进行区分处理，分离出正向查询与逆向查询。

最后，所有语言浅层分析的中间结果输出到 XML 格式的文件，其中主题与文档中的句子分别生成特征词语 ID 序列，最终可以直接用于句子检索与新信息检测的计算。

第四章 Noovel 句子检索算法与分析

4.1 概述

在浅层语言分析的中间结果基础上，我们可以直接对句子或文档与查询的相关性进行建模，并根据计算结果检索出相关的文档或者句子。在新信息检测任务中，句子的检索结果是 Noovel 系统返回给用户的第一批输出信息。新信息只在相关句子中出现，因此，句子检索结果是新信息检测的数据来源，直接影响着新信息检测的召回率和准确率。

相关句子只出现在相关文档中，因此文档检索是句子检索的前期处理阶段，属于句子检索不可分割的一部分。在 1.2 节，相关文档检索定义为：$DR(Q, d_x)$=TRUE 当且仅当文档 d_x 与查询 Q 相关；否则 $DR(Q, d_x)$=FALSE；相关句子检索定义为：$SR(Q, s_x)$=TRUE 当且仅当句子 s_x 与查询 Q 相关；否则 $SR(Q, s_x)$=FALSE。在这里，DR 与 SR 分别对应于文档检索模型与句子检索模型。

文档或者句子检索的一般步骤为：首先，依据检索算法计算出文档或者句子与正向查询的相关度，同时兼顾逆向查询的相关度，最后，我们会选择正向查询相关度超过一定阈值、逆向查询相关度比较小的文档或者句子作为检索结果返回。为了行文的方便，下面我们探讨的检索模型只介绍正向查询的处理。

为了简化模型的复杂度，我们将文档视为句子集合，或者将单个句子看成是一篇小型的文档，因此，文档检索和句子检索存在互换性。这样，我们就可以针对文档和句子采取统一的检索模型，不过，我们还会针对句子的特点，适当地改进文档检索的模型方法。这种改进更适合于句子检索，而不能简单扩展到文档检索中。在表述过程中，本章主要从文档的角度来阐述通用的检索模型，并会详细介绍专门应用于句子检索的技术方法。

本章主要阐述三种常用的句子检索算法：向量空间模型、概率模型与语言模型。最后，我们分别在 TREC2002 与 TREC2003 的数据集上进行实验，并给出分析结论。

4.2 向量空间模型及其扩展

在这一节，我们依次介绍向量空间模型的基本思想、表示方法、查询相关性计算方法、特征权重估计与规格化方法。并针对句子的特点，探讨句子检索的各种查询扩展方法。

4.2.1 向量空间模型基本思想

向量空间模型（VSM：Vector Space Model）将文档视为向量。向量的每一维代表一个信息表达元素，它可以是汉语中单个的字、也可以是任何语言中的一个词或者短语，也可以是没有实际意义的 n 元组。这些元素通过对数据集合中的每一个文档进行处理而得到[许 2003]。

在 Noovel 中，每一个文档或者句子经过浅层语言分析处理之后，我们得到了一系列的特征词语 ID，它们就可以构成一个“特征空间”，每个词语 ID 就是特征空间中的一维。文档就可以用来自文档中的那些词语 ID 来表示，也就是说可以把文档看成是这个空间中的向量，每一维对应一个词语 ID，文档被转化为一个点。这个空间因而可以被称为文档向量空间，或者向量空间。向量空间模型的方法并不局限于这个模型自身。它实际上是一种通用的文档信息表示方法，任何一种检索模型都可以利用它的概念来表示文档，利用它的方法来辅助检索。

经典的向量空间模型[Salton, 1983][Salton, 1989]由 Salton 等人于 60 年代末提出，并成功地应用于著名的 SMART 文本检索系统。

4.2.2 向量空间表示法

文档（Document）：由基本的语言符号组成的字符串，也泛指一般的文本或文本中的片断（句子、句群或段落）。文档也可以是多媒体对象，本文只讨论字符串对象，并且不对文本和文档进行区别。在新信息检测应用中，我们将句子视为较短的特殊文档。

特征（Term）：文档内容用它所包含的基本语言单位来表示，基本语言单位包括字、词、词组、短语、句子、段落等，它们统称为特征词项（Term），为了表述方便，直接成为特征。则文档 D 可以用特征集合表示成（t_1，t_2，…，t_N ），其中 t_k 是特征，$1 \le k \le N$ 。

特征权重（Term Weight）：不同的特征对于文档 D 的重要程度不同，用特征 t_k 附加权重 w_k 来进行量化，文档 D 可表示为（t_1，w_1；t_2，w_2；…；t_N，w_N），简记为 D（w_1，w_2，…，w_N ）。这时我们说特征 t_k 的权重为 w_k，$1 \le k \le N$ 。

向量空间模型（Vector Space Model）：给定一个自然语言文档 D（t_1，w_1；t_2，w_2；…；t_N，w_N）， t_k 在文档中既可以重复出现，又存在着先后次序的差别，分析起来仍有一定的难度。为了简化分析，我们忽略 t_k 在文档中的先后顺序，同时假定不同特征之间相互独立而忽略其依赖性，这时可以把（t_1，t_2，…，t_N ）看成一个 N 维的坐标系，而（w_1，w_2，…，w_N ）为相应的坐标值，因而 D 被看成是 N 维空间中的一个向量。我们称 D（w_1，w_2，…，w_N ）为文档 D 的向量表示或向量空间模型。

通过向量空间模型，一篇文档或者一个句子可以映射到相应特征空间上的唯一一点，然而反过来，我们不可能通过这一点重构原文。因此，向量的表示是有损的简化过程，损失了很多原有的语义内容。一方面特征项不能涵盖原来所有的内容；另一方面向量空间模型忽略了特征的位置，并做出了特征之间的独立性假设。不过，对于新信息检

测来说，在浅层语言分析阶段，经过停用词处理、特征选择之后，我们将这种损失降低到了最低限度，对于信息检索与新信息检测来说，这种损失是在接受范围之内的。

4.2.3 查询相关性计算

查询相关性度量（Similarity）：*Sim*（*Q*，*D*）用于度量文档 *D* 和查询 *Q* 之间内容的相关程度。当文档 *D* 和查询 *Q* 被表示为特征空间的向量时，我们就可以通过计算向量之间的相似性来度量文档与查询的相关度。

最简单的方法就是计算文档 *D* 和查询 *Q* 之间的向量内积，即：

$$sim(Q,D)=\sum_{t\in Q\cap D} w_{d,t} * w_{q,t}$$

文本处理中最常用的相似性度量方式是向量的夹角余弦函数，即：

$$Sim(Q,D)=\frac{\sum_{t\in Q\cap D} w_{q,t} * w_{d,t}}{\sqrt{\sum_{t\in Q} w_{q,t}^2 * \sum_{t\in D} w_{d,t}^2}}$$

除此之外，我们还可以采用距离计算函数，如Minkowski距离函数：

$$Sim_r(Q,D)=\left(\sum_{t\in Q\cap D}\left|w_{q,t}-w_{d,t}\right|^r\right)^{1/r}$$

其中：*r* 为 Minkowski 参数：

当*r*＝1 时，称为“city-block[①]”距离；

当 *r*＝2 时，为典型的欧式(Euclid)距离；

当 *r*＝∞时，为切比雪夫(Chebyshev)距离。

4.2.4 特征权重估计与规格化

在确定了特征空间以后，对文档和句子向量化面临着特征权重的估算问题，即如何估计向量每维的值。基本原则是：特征权重需要能真实地反映出其意义上的重要性，同时需要具备比较性，即越重要的特征，其权重越大。

第一种方法是由专家或用户根据自己掌握的领域专业知识与系统使用经验，人工地赋上不同权值。这种办法随意性很大，效率也不高，很难适用于大规模真实文本的自动处理。另一种实际的方法是运用统计学的知识，根据文档中的词频、词之间的共现频率等统计信息来估计各个特征的权重。特征权重估计和具体的应用息息相关，似乎并不存在“放之四海而皆准”的“最优公式”。

4.2.4.1 “tf*idf”估计

Noovel 中采用目前通行的“tf*idf”（tf: term frequency; idf: Inverse Document

① City-block 指的是城市由单位方块组成，旅行者每步都不能跨越对角线，只能是水平或者竖直行走，这样两点之间的行使长度，称之为 City-block 距离

Frequency）公式来计算权重，tf*idf 存在很多的变种，我们采用的权重估算公式为：

$$w(t,\vec{d}) = \log(tf(t,\vec{d})+1)\times\log(N/n_t+1.0)$$

其中，$w(t,\vec{d})$ 为特征词语 t 在文档 $\vec{d}$ 中的权重，而 $tf(t,\vec{d})$ 为特征词语 t 在文档 $\vec{d}$ 中出现的频率，N 为文档总数，n_t 为训练文档集合中出现 t 的文档数。

关于这个公式需要说明一下，其中：$\log(tf(t,\vec{d})+1)$反映的是特征 t 在文档中的重要程度，反映的是局部影响，不同文档中的同一特征 t 的 $\log(tf(t,\vec{d})+1)$值不同；而 $idf(t)=\log(N/n+1.0)$ 反映的是特征 t 在整个向量空间中相对于其他维的重要程度，它反映的是一种全局影响，每个特征的 idf 值是在样本训练时计算出来的，只要训练集相同，不同的文档同一个特征的 idf 值相同。

同样，我们可以得到句子中的权重估算公式如下：

$$w(t,\vec{s}) = \log(tf(t,\vec{s})+1)\times\log(N_s/n_t+1.0)$$

其中，$w(t,\vec{s})$ 为特征词语 t 在句子 $\vec{s}$ 中的权重，而 $tf(t,\vec{s})$ 为特征词语 t 在句子 $\vec{s}$ 中出现的频率，N_s 为主题中所有句子的总数，n_t 为主题中所有句子中出现 t 的句子数。

文档检索与句子检索时，我们可以分别采用 $w(t,\vec{d})$ 与 $w(t,\vec{s})$ 的计算公式计算查询词汇的权重。

4.2.4.1 权重的规格化（Normalization）

简单的“tf*idf”权重计算有时不能取得令人满意的效果。例如，假设一个有关某个主题的句子特别长，它的质量并不比短句高，但是计算出来的相关度会和实际情况恰恰相反。因为在长句中，检索词在其中出现的频率会比其它短文档中高，因此用 tf*idf 计算得到的权重也很高，短的文档得到的权重无法与之相比。为了解决这个问题，引入了权重规格化（Normalization）的概念。

我们主要采用了三种规格化方法：单一长度规格化、Cosine规格化、转轴文档长度规格化（Pivoted document length Normalization）[Singhal 1996]。

给定一个文档向量$D(w_1, w_2,\ldots,w_n)$，单一长度规格化公式为：

$$w_i^* = \frac{w_i}{\sum_{j=1}^{n} w_j} \qquad 1 \le i \le n$$

Cosine规格化公式为：

$$w_i^* = \frac{w_i}{\sqrt{\sum_{j=1}^{n} w_j^2}} \qquad 1 \le i \le n$$

转轴文档长度规格化作用于词频之上，其公式为：

$$tf_i^* = \frac{\dfrac{1+\log(tf_i)}{1+\log(tf_{avr})}}{slope * \#unique\_term + (1-slope) * pivot}$$

其中：tf_{avr}是文档中所有词的平均词频，*#unique_term*是文档中唯一特征词的种类。*slope*为转换轴的斜率，*pivot*为交错点，对于特定的数据集，它们是相对稳定的调节参数。Singhal 等人在TREC 数据集上的文档检索实验表明，这个计算模型的效果远好于cosine 规格化方式。

4.2.5 句子检索的查询扩展

句子包含的内容十分有限，仅仅依靠简单的向量空间模型，句子检索很难达到文档检索的效果。原因在于：单个句子中，具有实际意义的词语往往只有 4-5 个，任何一个词语几乎都不可能重复的出现，而文档往往包含数十倍的实词，它可以通过高频出现的词语来强调其内在的含义，同时也增加了其被检索到的可能。

通过下面两组例子，我们来看一下简单的向量空间检索句子存在的局限性。

(1a) Daily we read news stories about *dissatisfaction* with managed care, Medicare fraud and overbilling.

(1b) Eighty percent *agreed* with this.

(2a) However, the Scottish team was the first to make a clone from adult *animal* cell.

(2b) The seminar was held in the context of a recently reported *sheep* cloning case in Britain.

第 1 组、第 2 组的句子均来自于同一个主题的句子。其中第 1 组分别标明了对医疗改革的两种对立态度：*dissatisfaction*（不满）与 *agreed*（支持）；第 2 组讨论的其实都是克隆羊的问题，(2b)中的“sheep” （绵羊）属于“animal” （动物）的一个种类。从语义上来说：两者都存在着很强的相关性。然而，仅仅根据词形计算，第一组的相关性为 0，第二组仅有“clone”相同，向量空间模型相关度计算的结果不能正确地反映它们之间的关系。

相关词语不能匹配（word mismatch）的问题在文档检索中也存在，但是句子检索要严重得多。为此，我们需要针对句子的特点，对主题的查询作进一步的扩展。查询扩展（Query Expansion）指系统根据知识库中的知识对用户查询进行扩展，以提高检索效果的技术。查询扩展技术一般需要利用外部的语义词典及其它包含概念之间相关性信息的知识库。

在句子检索阶段，我们主要采用了三种查询扩展办法：语义扩展、伪相关反馈扩展

与局部共现扩展。

4.2.5.1 语义扩展（**Semantic Expansion**）

一个很直观的想法就是引入语义体系，抽取出词形不同的相关词语之间共有的语义特性。比如：上面例子中的“*dissatisfaction*”与“*agreed*”同属于情绪和态度的范畴，而“sheep”是“animal”的下位概念。因此，

我们引入了免费的在线英文词汇数据库 WordNet[Fellbaum 1998]。WordNet 由普林斯顿大学心理语言学实验室研制，其初衷是作为研究人类词汇记忆的心理语言学成果，在自然语言处理中得到广泛的应用。WordNet 用一组同义词的集合 Synset 来表示一个概念，每一个概念有一段描述性的说明，有多个语义的词分别对应不同的 Synset。Synset 之间存在上下位关系（hyponymy，troponymy）、同义反义关系（synonymy，antonymy）、部分整体关系（entailment，meronymy）等关系。

Noovel 主要考虑名词的上下位以及同义关系，并兼顾形容词的近义与反义关系。如图 4.1a，图 4.1b 所示。

借助 WordNet 的语义体系，我们提出了两种语义扩展方法：语义映射与词语衍生。

语义映射指的是将词语组成的向量空间，通过语义关系转换映射到概念空间之上，原来的词语向量最终转换为概念向量，具体做法是将词语的各个语义按照一定的加权策略映射到概念空间，同时将概念空间上的同义概念、上下位概念、近义概念等进行归并。基本思想如图 4.2 所示。最后在概念空间上进行向量运算，并检索结果。

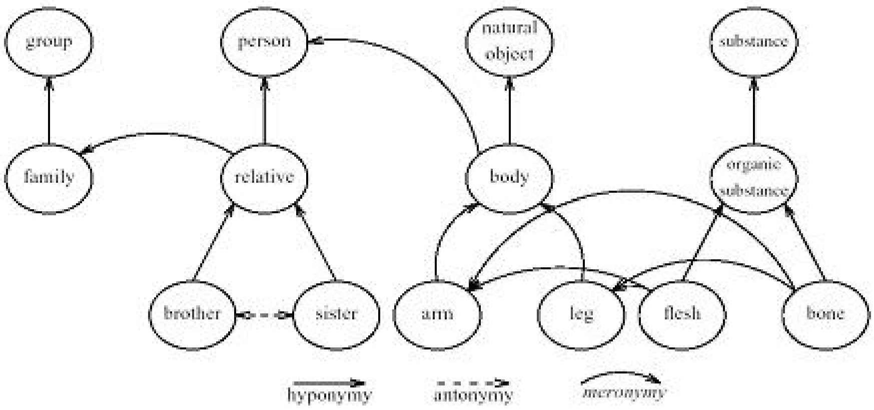


图 4.1a WordNet 名词的上位、下位与反义等语义关系

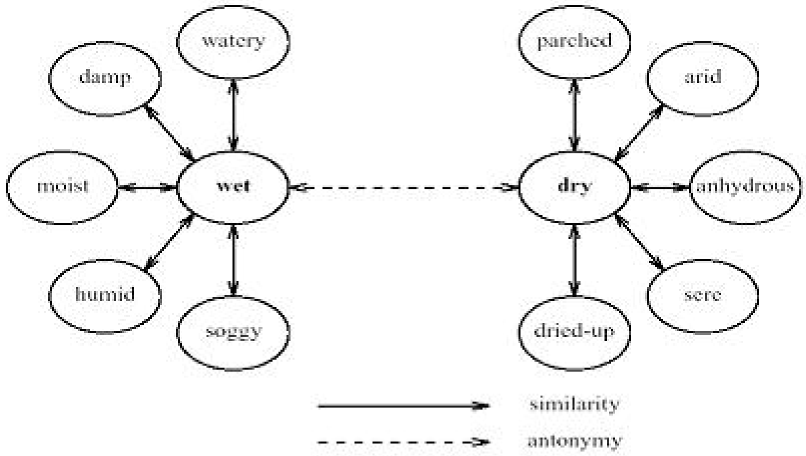


图 4.1b WordNet 中形容词的近义、反义等语义关系

语义映射的目的在于将词语转化为语义上更一般性的概念，避免同义相关词语词形匹配不一致的问题，从而可以扩展句子查询相关度的计算。但是，语义映射也存在着本质的缺陷。首先，它过分依赖于一份完备的语义体系，实际上，目前的语义体系往往难以达到这种完备的高度；另外，语义概念之间的关系往往千丝万缕，很难满足向量空间模型中的独立性假设；最后，词语到概念的映射存在歧义，而且计算代价非常高。

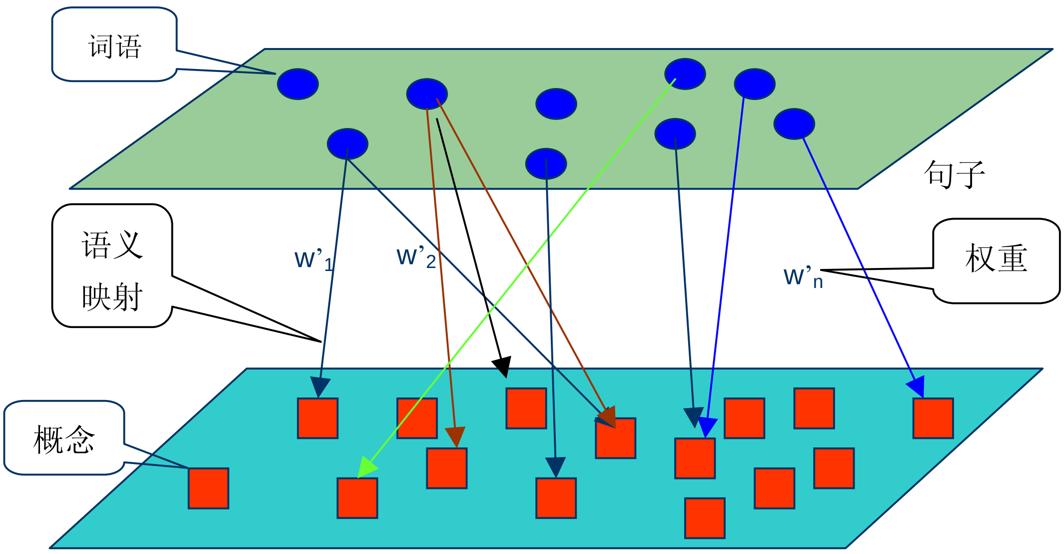


图 4.2　语义映射的扩展示意图

词语衍生方法则相对实用简便，主要的思路是：给定某个词语 w，通过 WordNet 的体系，我们可以得到 w 对应的所有概念 Synset 集合，将所有属于同一个 Synset 或者上下位 Synset 中的词语组成一个扩展词集合 S_{ext}，所在主题的特征词集合为 T_{feature}，则我们可以将词语 w 衍生为词语集合 DS_{w}：

$$DS_w = \{w' \mid w' \in S_{ext} \cap T_{feature}, w' \neq w \wedge sf(w') \geq 2\}$$

其中 $sf(w')$表示的是包含 w' 的句子数目。即：w' 衍生出来的词语是那些在当前主

题的句子集内出现过 2 次以上的同义词或者上下位词。接下来，我们就可以将向量中的 *w'* 扩展为（*w'*，DS_w）。DS_w 相应的权重有对应的 Synset 分布概率结合 *w'* 的特征权重估计而得，最终的向量需要进行二次规格化处理。通过这种方式，我们就可以相对简单的解决词语不匹配的问题。

4.2.5.2 伪相关反馈扩展（Pseudo Relevance Feedback Expansion）

相关反馈（Relevance Feedback）是一种利用查询结果中的相关结果增强查询效果的技术。早期的相关反馈是通过人工进行的。用户提交原始查询后，逐个浏览查询结果，给出相关性判定。系统根据相关性判定，将查询中的相关信息加强，非相关信息减弱，然后将修改后的查询再次提交检索系统。如此反复进行，可以得到非常好的检索效果。从这个过程可以看出，相关反馈必然包括对查询的修改，但不一定属于查询扩展的范畴，因为反馈之后的查询有可能不包括初始查询中的关键词。

即使人工干预的相关反馈能取得很好的效果，但是，它仍然是一种不实用的技术，因为实际的检索过程一般都是一次性的，用户并不愿意也不会提供更多的反馈。因此，研究者们引入了伪相关反馈技术，不再需要人工进行相关性判定，从而提高了自动化程度，使实用成为可能。

伪相关反馈的基本思想是：自动地选取查询相关度比较高的句子作为“相关”的检索结果，将这些句子按照一定的加权比例对查询进行正向反馈，采用它们的向量进一步修正查询向量，从而实现查询的扩展。

具体的算法流程如下所示：

输入： 查询 Q、相关文档中的所有句子集合 S、反馈比例 θ，加权系数 λ。

（1） 采用简单的向量空间模型，计算出查询 Q 与相关文档中的所有句子集合 S 的查询相关度；

（2） 按照相关度从大到小的顺序，对句子集合 S 重新排序。

（3） 选取 S 的前 $|S|*\theta$ 个句子向量，作为正向伪相关反馈向量集合 S_{pseudo}。

（4） 按照下面的公式修正查询 Q 的向量：

$$\overrightarrow{Q} = (1-\lambda)\overrightarrow{Q} + \lambda \sum_{\vec{s} \in S_{pseudo}} \vec{s}$$

图 4.3 句子查询的伪相关反馈扩展算法

在这里，句子检索主要的困难在于检索到那些并不包含查询词语的句子，应此，伪相关反馈只考虑了正向的伪反馈，而没有加入逆向的伪反馈。

4.2.5.3 局部共现扩展（LCE: Local Co-occurrence Expansion）

从信息检索的角度来看，在局部范围内，频繁地共同出现的词对存在某种内在的关联性，因此，我们可以利用高频共现的词对来实现查询向量与句子向量的进一步扩展。

按照词语 ID，设词对为<ID_1,ID_2>。在新信息检测中，我们将局部范围窗口设置为

一个单句内部，选取一个句子中共现的高频词 ID_2 来扩展 ID_1 的条件是：

$$SF(ID_1,ID_2)>\theta_1 \text{ 且 } P(ID_2 \mid ID_1)>\theta_2$$

其中 $SF(ID_1,ID_2)$为同时出现 ID_1 与 ID_2 的句子数目，$P(ID_2 \mid ID_1)$为共现条件概率，θ_1 与 θ_2 为阈值，可用于调节扩展词的质量与数量。Noovel 中，θ_1=3，而 θ_2=0.05。

以 TREC 的主题 N1 为例，这是一个讨论“partial birth abortion ban”（反堕胎）方面的评论型主题，从 25 篇相关文档的浅层分析结果中，我们统计了所有词对的共现概率，按照刚才的条件，最终可以得到如表 4.1 所示的高频局部共现词对表。

词语 ID_1（词语 $_1$）	**词语 ID_2（词语 $_2$）**	**共现频率 $SF(ID_1,ID_2)$**	**共现概率 $P(ID_2\|ID_1)$**
52104 (partial)	7333 (birth)	26	0.619048
52104 (partial)	5501 (ban)	19	0.452381
7333 (birth)	52104 (partial)	26	0.433333
52104 (partial)	482 (abortion)	17	0.404762
7333 (birth)	5501 (ban)	18	0.3
7333 (birth)	482 (abortion)	16	0.266667
5501 (ban)	482 (abortion)	24	0.186047
52104 (partial)	15879 (court)	7	0.166667
5501 (ban)	52104 (partial)	19	0.147287
5501 (ban)	7333 (birth)	18	0.139535
52104 (partial)	56001 (procedure)	5	0.119048
52104 (partial)	69802 (supreme)	5	0.119048
7333 (birth)	15879 (court)	7	0.116667
69802 (supreme)	15879 (court)	6	0.107143
69802 (supreme)	7333 (birth)	5	0.089286
69802 (supreme)	52104 (partial)	5	0.089286
7333 (birth)	56001 (procedure)	5	0.083333
7333 (birth)	69802 (supreme)	5	0.083333
69802 (supreme)	5501 (ban)	4	0.071429
482 (abortion)	5501 (ban)	24	0.066667

表 4.1 TREC 主题 N1 中的高频局部共现词对表

通过上表，我们可以发现很多有意思的现象，比如堕胎与最高法院（supreme）、法院（court）、法律程序（procedure）的内在关联性。尽管它们具有完全不一样的语义，但是，从内在的主题相关性来说，它们又具备潜在的关联性。这是复杂的语义分析都很难解决的现象。

本文需要特别指出的是：局部共现扩展和具体统计采样的文档集合密切相关。在不同立场色彩的文档集合上统计高频共现词对，最终的词表很可能大不相同，扩展的结果也会受到影响，它们会更多地代表采样文档集的特性。直观地来说，如果我们在非相关文档集合上进行采样统计分析，那么形成的高频共现词对表会更多地偏向于非相关文档，最终干扰信息的正确检索。局部共现扩展存在被滥用甚至错用的可能，因此，我们

需要有选择性地选取采样文档集合，Noovel 系统利用伪相关反馈的思想，选取查询相关度排名靠前的句子集合，在此基础上进行局部共现统计。

利用高频的局部共现词对表，我们可以扩展句子向量中的词语。假定句子的向量为 s，其中有一个特征词语为 *t*，其权重为 *w*。*t'*是 t 的高频局部共现词，那么，我们可以将 *t'*加入到 s 中，其权重为 *w**P(*t'*| *t*)，而 *t* 的权重相应修改为 *w**(1-P(*t'*| *t*))，其他的高频共现词以此类推，最终实现句子检索的扩展。

4.3 概率检索模型

给定一个用户查询 *Q* 与文档集合中的文档 *D*，概率模型通过估计用户查询 *Q* 与文档 *D* 相关的概率来判定 *D* 是否与 *Q* 相关。概率模型假设这种概率只取决于查询串和文档 [刘 2004]。这样就将检索问题转化为条件概率 Prob(R|*D*,*Q*)的求解问题，即：

if Prob(R|*D*,*Q*)>Prob(NR| *D*,*Q*) then：

D 属于相关文档集合 R，即检索结果，

else

D 属于不相关文档集合 NR，即不是检索结果。

图 4.4 给出了概率检索模型的算法原理图。

概率检索模型最早由 Maron 和 Kuhns 于 1960 年提出[Maron, 1960]。经过 Maron、Cooper、Robertson、Jones、van Rijsbergen、Croft、Turtle 等人的发展，概率检索模型已经从概念走向实际应用。基于概率检索模型的 OKAPI 系统[Robertson 1999]在多届 TREC 中取得了优异成绩。概率检索模型与其它统计模型并没有很清楚的界限。本文曾经论及，在向量空间模型中得到的相关度数值实际上就是一种概率。而概率检索模型也要借鉴向量空间模型的文本表示方法与某些统计量。

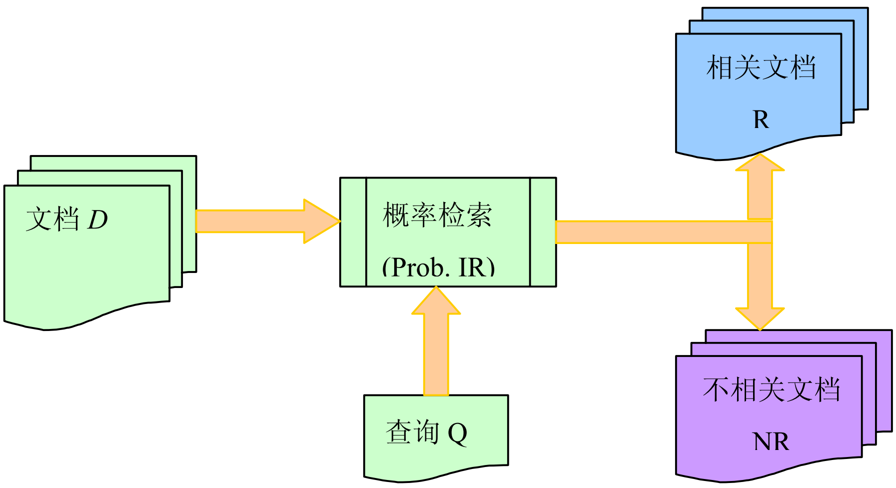


图 4.4 概率检索模型的算法原理图

OKAPI 系统是Robertson 等在伦敦城市大学研究和开发的一个著名的概率检索系统。它有较完整的理论基础和严格的理论推导。在它出现之后，经过10多年的发展，OKAPI系统健壮性越来越强，检索精确度也越来越高，已经在综合检索效果上超过了其

它系统，连续几年在TREC 的检索任务中名列前茅，其理论成果已经被广为接受[Robertson, 1994] [Robertson 1996]。它基于概率检索模型设计，使用著名的bm25公式及其变形（bm2500、bm250）等计算特征词语的权值，这种权重计算公式在概率模型中已经成为公认成熟的标准计算方法。

与向量空间模型中的相似度的概念不同的是，概率检索模型从一开始就把相关性当成了一种概率事件。当我们用一个查询Q在某个文档集合上进行查询时，我们没有任何先验知识，只能假设利用文档的摘要（文档向量和索引）可以猜测文档与查询的相关程度，也就是相关的概率。

在新信息检测中，我们的概率检索模型主要是沿用OKAPI系统的计算公式。即文档d与查询q的相关度为：

$$sim(q,d) = \sum_{t \in q \cap d} w_{d,t} * W_{q,t}$$

其中：

$$w_{d,t} = \frac{(k_1 + 1) * f_{d,t}}{k_1 * [(1-b) + b * \frac{W_d}{avr\_W_d}] + f_{d,t}}$$

$$W_{q,t} = \frac{(k_3 + 1) * f_{q,t}}{k_3 + f_{q,t}} * \log \frac{N - f_t + 0.5}{0.5}$$

W_d为当前文档的词数，avr_W_d为所有文档包含词数的平均值，f_t为包含特征词语t的文档数，而$f_{d,t}$为文档d中特征词语t的词频，$f_{q,t}$为查询q中特征词语t的词频。k_1，k_3，b均为调解参数，在Noovel系统中，参数设置为：

k_1=1.2；

k_3=1000；

b=0.75

4.4 语言模型检索（Language Modeling IR）

我们首先介绍语言模型检索的基本思想，然后阐述在句子检索方面的改进。

4.4.1 语言模型的基本思想

语言模型（Language Modeling Approaches）是1998年开始兴起的一种新的检索模型[Ponte, 1998] [Zhai, 2002]，实际上也是一种基于概率的检索模型。它的思想别出心裁，与OKAPI的思路刚好相反[杨 2003]。它认为：每个文档对应一个统计语言模型，称为文档的语言模型(Language Model)。它主要描述了该文档中各个单词的统计分布特征。因此每个文档看作是由其语言模型抽样产生的一个样本。查询 q 也可以看作是由该文档

的语言模型抽样产生的一个样本。因此，我们可以根据每个文档语言模型抽样生成查询的概率来排序，概率值越大，则该文档就越满足查询要求，作为检索结果的可能性就越大。

传统的概率检索模型将文档 d 与查询 q 的相关度排序函数定义为事件R(文档是否满足检索要求)的概率，即：$f(q,d)=P(\mathrm{R}|d, q)$；相关度排序函数定义虽然比较直观，但相关性是一个抽象的概念，该定义本身没有也无法具体给出R的定义，所以该模型在理论上存在很大的模糊性。因此，我们一般需要在检索中，首先给定带有相关性标记的文档作为建立模型的基础。在实际中，要针对每个检索给定训练样本数据，这几乎不可能。这也是传统概率检索模型存在的主要问题。

基于语言模型的检索相关度排序函数则定义为由文档 d 的语言模型生成查询 q 的概率，即：$f(q,d) = P(q|d)$。建立在统计语言模型理论基础上，定义明确，便于操作。它可以在文档的词语概率分布基础上直接计算出相关度排序函数，从而有效地避免这个问题。

与向量空间模型、概率检索模型等类似，这里也假设查询词之间是相互独立的关系（即一元语言模型），因此有：

$$P(q \mid d) = \prod_i P(q_i \mid d)$$

所以，最后将$P(q|d)$的求解转化为如何估计概率$P(q_i \mid d)$。

特征词 q_i 既可能出现在文档 d 中，也可能不出现。但是，我们不能简单地认为 w 不出现在文档 d 时它的概率就是0。这是个典型的数据稀疏问题，统计方法的核心是利用已知事件的发生频率来估计该事件客观存在的概率，我们往往需要根据已知事件的发生频率来尽可能地“逼近”其内在的发生概率。某个事件概率是客观存在的，观测到的频率符合一定的规律，但是具有很大的随机性。另外，受客观条件的限制，人们也不可能观测到所有的现象。我们估计出来的概率往往和实际的概率存在一定的差距，在样本数量很有限的情况下，简单最大期望估计甚至可能完全背离客观实际。例如，抛硬币2次，很可能2次都是正面，仅仅根据观测到的小规模结果，我们会得到抛硬币得到正面的概率为100%，而反面的概率为0，这显然不合理。

为了解决因数据稀疏引起的零概率问题，我们需要采用数据平滑的技术进行处理，其基本思想是：从已发生事件的概率中扣除一部分，然后把这些“多余”的概率重新分配给那些未发生的事件和已发生的低频事件中，数据平滑的好坏直接影响到语言模型的性能[丁 2004]。[Stanley&Joshua 1996]对数据平滑方法进行了很好地综述。

一般来说，估计概率$P(q_i \mid d)$的方法是对它出现和不出现分别处理。如果q_i出现在文档中，可以按照它的词频信息计算$P(q_i \mid d)$；如果q_i不出现，在没有其它先验知识的情况下，可以按照 q_i 在所有文档集合中出现的情况来估计它的概率。Zhai 等人[Zhai 2001]得到的$P(q \mid d)$表达式为：

$$\log P(q \mid d) = \sum_i \ln P(q_i \mid d)$$

$$\log P(q \mid d) = \sum_{TF_{d,qi} > 0} \ln P(q_i \mid d) + \alpha_d \sum_{TF_{d,qi} = 0} \ln P(q_i \mid C)$$

其中：C为当前主题的所有文档集合。

如果q_i在C上仍然没有出现，我们还可以进一步引入大的真实语料库，来估计q_i在整个通用语言环境的概率。

4.4.2 句子级语言模型的改进

在句子检索中，语言模型需要进一步的进行平滑处理，在文档检索的基础上，我们采用下面这个公式来计算句子 s 的查询相关度：

$$Sim(q,s) = \ln P(q \mid s) = \sum_{q_i \in q} \ln[P(q_i \mid s) + \alpha_s \cdot P(q_i \mid d) + +\alpha_d \cdot P(q_i \mid C)]$$

其中：d 为当前句子所在文档，C 为主题所有的文档集合，目的在于将句子检索与文档检索进行结合，利用所在文档来平滑句子 s 生成查询 q 的概率。

4.5 句子检索实验与分析

在 TREC2002 与 TREC2003 新信息检测任务的数据集基础上，我们设计了三组实验：

（1） 浅层语言分析的贡献度实验，主要是通过实验来评估浅层语言分析对新信息检测的潜在贡献度。

（2） 三种句子检索模型的基准实验，主要是测试向量空间模型、概率检索模型与语言模型的基本性能。

（3） 查询扩展实验，在简单向量空间模型的基础上，测试各种查询扩展技术的对比性能。

同时，我们要说明的是：TREC2002 与 TREC2003 给定的文档都是与主题相关的，也就是说，我们可以忽略文档检索过程，可以排除不相关文档的干扰。我们可以更好地衡量出句子检索的实际性能。最后，我们也会在第七章给出 Noovel 参加 TREC2004 比赛的官方测试结果，其中就包括了文档检索过程，这种情况考虑了文档检索等环节，更符合实际的应用场景。

4.5.1 浅层语言分析的贡献度实验

在第三章，我们详细阐述了浅层语言分析的重要作用，讨论了具体的语言分析过程。接下来，很自然而然的一个问题是：我们如何证明语言分析的作用？进一步的来说，如何通过实验来量化浅层语言分析对新信息检测的贡献程度？

新信息检测是一个错综复杂的问题，它涵盖了浅层语言分析、文档检索、句子检索、

新信息检测等四个有机的过程，而且每个环节都存在着多种变化因素。各种各样的因素都会直接或者间接地影响着新信息检测的最终性能，甚至说，某个技术环节的小技巧都很可能提高或者降低最终性能。

为了评估浅层语言分析的作用，我们尽可能地减少其他因素的干扰。因此，我们只考虑句子检索的性能，因为句子检索结果直接决定后期的新信息检测，从句子新信息检测的角度出发，最大的影响因素是句子检索，我们很难直接衡量浅层语言分析的作用。同时，我们直接给定相关文档，去除句子检索之前的文档检索环节。所以，最终决定句子检索性能的只有两大因素：浅层语言分析与句子检索模型。

在句子检索模型方面，我们采取的是简单的向量空间模型：即采取简单的 tf*idf 来度量特征词语的权重，如果句子的查询相关度超过一定的阈值 θ，则判定为相关句子，并作为检索结果返回。为了进一步的简化，θ 设为 0.0。应该说，我们采取的句子检索模型是最基准的一个检索方法，并没有引进更复杂的处理技术。

通过这样的种种限制，浅层语言分析成为了句子检索唯一的决定性因素。在 TREC2002 与 TREC2003 的数据集上，我们分别作了两组实验。与此同时，我们引入当年最好的比赛结果及其方法作为对比，从而验证浅层语言分析对句子检索的贡献度。

在 2002 年的数据集上，我们首先采用简单向量空间模型在浅层语言分析的基础上，进行了句子检索的实验，结果见表 4.2。

RunID	句子检索模型	相关技术	平均准确率P②	平均召回率 R	平均 F-measure
NOOVEL	简单向量空间模型	浅层语言分析	0.19	0.68	0.257
Thunv1	OKAPI	查询扩展、文档扩展、主题分类、动态阈值调整	0.23	0.34	0.235
NOOVEL2	OKAPI	浅层语言分析	0.19	0.68	0.257

表 4.2　TREC2002 数据集上的浅层语言分析贡献度评测实验

在上表中，Thunv1[Zhang 2002]是 TREC2002 年新信息检测比赛发布的排名第一的结果。我们发现，[Zhang 2002]采用的是 OKAPI 方法，并采取了复杂而又比较有效的改进措施，其中包括查询扩展、文档检索、主题分类以及动态阈值调整等。这代表了当时最好的一个研究水平。为了测试的公平性，其他条件不变，我们同样采用 OKAPI 进行了第二个实验 NOOVEL2，发现结果与 NOOVEL 保持一致。通过测试数据的分析，我们不难发现，采用浅层语言分析技术的优势所在，它的最终性能比历史最高成绩提高了 9.4%。

采取同样的条件，我们在 TREC2003 的数据集上进行了第二组实验。结果如下：

② 我们这里提到的所有指标都是在 50 个主题基础之上的平均性能。

RunID	句子检索模型	相关技术	平均准确率 P	平均召回率 R	平均 F-measure
NOOVEL	简单向量空间模型	浅层语言分析	0.59	0.79	0.614
THUIRnv0312	PFM 模型与 bm2500 相结合	WordNet 的概念扩展结合局部上下文扩展	0.62	0.67	0.564

表 4.3 TREC2003 数据集上的浅层语言分析贡献度评测实验

在上表中，THUIRnv0312 [Zhang 2003]是 TREC2003 年正式发布的最好结果。无论是采用简单向量空间模型还是 OKAPI 检索方法，Noovel 系统的平均 F 值均可以达到 0.614，超过最好水平 8.9%。

通过这两组实验，我们可以得出一个很有意思的结论：在句子级的检索乃至新信息检测，前期的语言分析预处理比检索算法本身对最终性能的影响还要显著，而人们往往忽视前期的语言预处理，在不准确的语言处理结果基础上，无论在检索的建模上进行怎样的努力都很难取得理想的效果，语言处理的中间结果决定了最终性能的上限，更精确的语言处理结果结合一个相对简单的检索方法往往就可以达到甚至超过前者的最好的性能。从我们的两组实验来看，在 Noovel 系统中，浅层语言分析至少提高了 9%的最终性能。

另外，从准确率与召回率的指标来分析，我们还可以发现：语言有针对性的准确分析主要贡献在于：它更大程度地提高了句子检索的召回率。由于句子自身的特点，已有系统最大困难就在于难以达到更高的召回率。例如，TREC2002 的数据集上，所有系统的召回率均不超过 0.4，而在 TREC2003 数据集上，召回率依然局限在 0.7 以内。浅层句法分析大幅度的突破了这些上限。同时，我们还可以看到不同检索方法的准确率差别并不大。

4.5.2 三种句子检索模型的基准实验

下面，我们针对各种检索模型进行实验，同时调节参数与阈值，但不引入查询扩展技术。从而测试出各种方法的基准性能。

4.5.2.1 向量空间模型

在向量空间模型中，我们尝试不同的权重规格化方法，并调节查询相关度的阈值 θ。在 2003 年的数据集上，测试结果如下表所示：

规格化方法	阈值	平均准确率 P	平均召回率 R	平均 F-measure
余弦/长度	0.0	0.59	0.79	0.614
余弦	0.04	0.59	0.78	0.611
余弦	0.1	0.60	0.51	0.482
长度	0.04	0.57	0.29	0.312
长度	0.1	0.37	0.01	0.023

表 4.4 向量空间模型的句子检索实验（TREC2003 数据集）

从上表中，我们发现典型的余弦规格化方法比较适合句子的检索。另外，从相关度阈值来看，随着阈值的提高，准确率的提升相当有限，而召回率却急剧下降，最终性能受到影响。这说明：经过浅层语言分析，某个句子结果中只要包含了查询的特征词语，哪怕是相关度很小，它就很可能就是检索结果。

在 TREC2002 的数据集上，我们同样作了几组实验，如表 4.5 所示。结论与上面的一致。

规格化方法	阈值	平均准确率 P	平均召回率 R	平均 F-measure
余弦	0.0	0.19	0.68	0.257
余弦	0.04	0.18	0.43	0.225
余弦	0.1	0.18	0.40	0.216

表 4.5　向量空间模型的句子检索实验（TREC2002 数据集）

另外，对比 TREC2002 与 TREC2003 的测试结果，我们还可以发现句子级新信息发现的最终性能非常依赖于具体的主题和文档。采用同一种方法，在不同的数据上，最终的性能指标几乎相差三倍之多。

4.5.2.2 概率检索模型

在 TREC2003 的数据集合上，采用 OKAPI 的方法，采取不同阈值，我们做了三个实验，结果如下：

阈值	平均准确率 P	平均召回率 R	平均 F-measure
0.0	0.59	0.79	0.614
2.0	0.59	0.79	0.613
8.0	0.61	0.60	0.544

表 4.6　概率检索模型的句子检索实验（TREC2003 数据集）

OKAPI 方法计算出来的相关度没有归一化，结果都比较大，在句子检索方面的性能与向量空间模型相当。不过，已有的文献表明 OKAPI 方法在文档检索方面比简单的向量空间模型更具有优势。这一点我们会在第七章作进一步的实验。

4.5.2.3 语言模型

我们采用语言模型计算所有句子生成查询的一元语言概率，然后依据这个概率对句子进行排序，依次取排名靠前的句子作为检索结果。在 TREC2003 的数据集上，经过各种对比实验，我们发现取前 70%的句子集合作为检索结果，能够取得最优的性能，各项指标如表 4.7 所示：

阈值	平均准确率 P	平均召回率 R	平均 F-measure
70%	0.48	0.63	0.498

表 4.7　语言模型的句子检索实验（TREC2003 数据集）

文档级的语言模型检索能够达到与向量空间模型相当的性能，但是，在句子检索方面，不管怎样对语言模型进行改进，F-measure 都很难超过 0.5，与向量空间模型概率和模型相比，相差甚远。

从理论上，我们不难发现：利用一个句子来构建语言模型存在着本质的缺陷，经过浅层语言分析之后，单个句子包含的词数一般都在 10 个以内，这么少的样本空间来估计语言的概率，数据稀疏问题相当严重，无论采取什么样的数据平滑技术，都很难弥补少量数据所带来的不足。因此，我们可以最终断定：语言模型不适宜于句子级的检索，实验也验证了语言模型在句子检索方面的不足。

4.5.3 查询扩展实验

在这一部分，我们设计了一组对比实验来测试查询扩展技术在句子检索方面的作用。

在 TREC2003 数据上，采用向量空间模型，我们分别完成了如下六个实验：

（1）**NONE**：采用简单的向量空间模型，不引入任何其他的查询扩展技术；

（2）**WordNet**：在 NONE 的基础上，引入 WordNet 进行语义衍生；

（3）**PSEUDO5**：在 NONE 的基础上，引入伪相关反馈，其中相关度排名排名前 5%的句子集合作为伪反馈结果；

（4）**PSEUDO15**：在 NONE 的基础上，引入伪相关反馈，其中相关度排名排名前 15%的句子集合作为伪反馈结果；

（5）**PSEUDO20**：在 NONE 的基础上，引入伪相关反馈，其中相关度排名排名前 20%的句子集合作为伪反馈结果；

（6）**LCE**：在 NONE 的基础上，引入局部共现扩展。

测试结果见表 4.8。

实验代码	平均准确率 P	平均召回率 R	平均 F-measure
NONE	0.59	0.79	0.614
WordNet	0.59	0.79	0.614
PSEUDO5	0.56	0.88	0.624
PSEUDO15	0.55	0.92	0.631
PSEUDO20	0.55	0.93	0.633
LCE	0.57	0.98	0.643

表 4.8 查询扩展技术对比实验（TREC2003 数据集）

从上面的测试结果，我们可以发现：

（1） WordNet 的语义衍生对句子检索没有明显的提高，主要原因在于：我们进行语义衍生的时候，进行了限制，因此，并没有衍生出很关键的特征词语。另外，还有一个原因是：WordNet 语义衍生出来的词语更多已属于查询的一部分，它增强了高相关句子的相关度，但是对那些相关而相关度很低的句子帮

助甚微，我们可以形象地形容它是一种“锦上添花”而不是“雪中送炭”式的“近亲繁殖”。

（2） 伪相关反馈是一种比较好的查询扩展方法，在不特别损失检索准确率的前提下，它极大地提高了检索的召回率，从而提高了整体性能。另外，随着反馈句子数目的增加，性能也在逐渐地提高，但增涨幅度在缩小。同时，我们的进一步研究还发现超过 20%的反馈比例就很难再提高系统性能，继续增加反馈比例，只会起到副作用。当然，具体的百分比例和具体的数据集有相当大的关系，不能简单的一概而论。

（3） 局部共现扩展尽管缺乏很严格的理论依据，但是它可以通过统计的方法挖掘出特征词语之间的内在关联性，并能够较好地改进查询的结果。在我们的实验中，它取得了最好的性能，这是一种很有潜力的查询扩展思路。

最后，按照综合指标平均 F-measue 值，我们将各种方法进行综合排序，结果如下图所示：

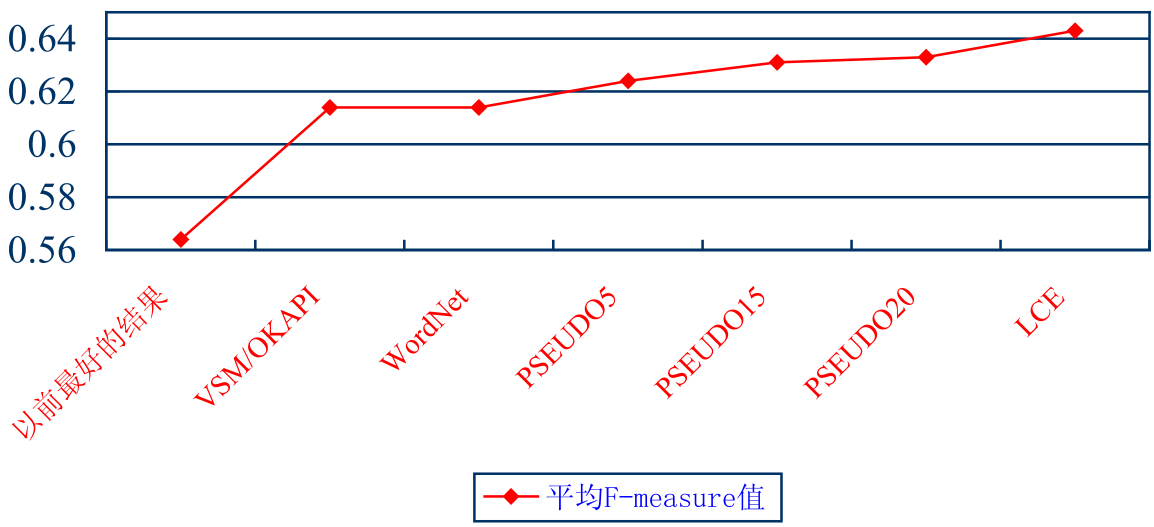


图 4.5 各种句子检索技术的性能对比

4.6 本章小结

本章详细介绍了文档与句子检索的各种模型方法，其中包括：向量空间模型、概率检索模型与语言模型。并针对句子的特点，引入了相应的处理技术。

在向量空间模型中，我们阐述了相关的概念与原理，介绍了特征权重的估计办法，并给出了相应的规格化处理手段。针对句子的局限性，引入了查询扩展的技术，主要是借助于 WordNet 的语义衍生扩展、伪相关反馈扩展、采用高频共现词语的局部共现扩展。最后的实验表明：在浅层语言分析的基础上，简单向量空间模型可以达到目前最好的结

果，受到语义资源和分析深度的限制，当前阶段的语义扩展作用有限，而伪相关反馈与局部共现扩展都能够帮助提高句子检索的性能，局部共现扩展是很有潜力的查询与文档扩展的技术。

概率检索模型主要是借助现有的 OKAPI 的思想，运用概率计算的方法实现检索，它比较有利于文档的检索，在句子检索方面，能达到的性能与向量空间模型相当。

语言模型作为信息检索的一种方法，具备很好的理论基础，最大的问题在于数据的稀疏，核心在于如何进行平滑，从而保证能更“逼真”地模拟出文档内在的语言分布特点。在文档检索方面，它能够达到与向量空间模型差不多的性能。但是，通过实验与理论的分析，语言模型并不适用于句子的检索。

最后，我们还设计了三组不同的实验。首先，证实了浅层语言分析的作用，其次给出了各个检索方法的基准性能，另外，还对比测试了各种查询扩展技术。

第五章 Noovel 新信息检测技术

5.1 概述

新信息句子具有两个主要特点：内容相关性与时序新颖性。内容相关性指的是新信息句子的内容必须与主题密切相关。而时序新颖性指的是它传达了以前句子所没有涵盖的新信息。我们假定用户在检索之前对当前主题的信息一无所知，也就是说，他不具备相关的背景知识，所有的信息都来源于检索结果，因此，时序意义上的第一个相关句子默认为带有新信息的句子。

新信息检测子任务是最后一个核心任务，也是最关键的步骤。新信息检测形式化定义为：$ND\ (Q, <s'_1,s'_2,\ldots s'_{x-1}>, s'_x) =$ TRUE，当且仅当 s'_x 是包含新信息的句子，否则 $ND(Q, <s'_1,s'_2,\ldots s'_{x-1}>, s'_x) =$ FALSE，其中 $<s'_1,s'_2,\ldots s'_{x-1}>$ 为 s'_x 以前出现的新信息句子或者相关句子。*ND* 对应具体的新信息检测技术方法。在本章中，我们所讨论的句子都是浅层语言分析之后已经去除噪音的结果，而且均与主题相关。

我们引入新颖度的概念来量化一个句子在特定时间序列上的新颖程度，从而根据新颖度来判别句子是否含有新信息。在已有研究成果的基础上，我们提出了三种新颖度的评价方法：词重叠度及其扩展方法、相似度比较方法与信息增强的评价方法。随后，我们还阐述了一些不常用的其他计算方法：语言模型与语义距离计算的方法，最后，我们在 TREC2003 的数据集上进行新信息检测实验，并给出了分析结论。

5.2 词重叠度及其扩展（Word Overlapping）

5.2.1 基于词重叠度的句子新颖度计算

词重叠度是一种简单而有效的句子新颖度衡量方法。给定某个句子 S_i，则其前面句子 S_j 与 S_i 的词重叠度可以定义为：

$$Overlap\ (S_i \leftarrow S_j) = \frac{|S_i \cap S_j|}{|S_i|}$$

在这里，我们将句子看成是词的集合，词重叠度表示的是两个句子共同使用的特征词数在当前句子中所占的比重。

Overlap 的计算具有时序方向性，即：$Overlap\ (S_i \leftarrow S_j) \neq Overlap\ (S_j \leftarrow S_i)$

Overlap 的取值范围为[0.0, 1.0]，取值越大，则句子 S_j 与 S_i 越接近，拥有的共同信

息越多，也就是说新信息越少。*Overlap* 为 0.0 时，两个句子没有任何相同的词语，则句子 S_i 包含的都是新的信息；若 *Overlap* 为 1.0，则表示两个句子完全相同，不存在新的信息内容。

基于词重叠度，我们提出了一种句子新颖度的衡量方法。

$$OverlapNov(S_i) = 1 - \max_{0<j<i}\left\{Overlap(S_i \leftarrow S_j)\right\}$$

计算过程是：先计算当前句子与以前所有相关句子的词重叠度，选择重叠度最大的结果作为当前句子与历史结果的信息重叠度，最后，通过减法运算计算出句子的信息新颖度。

同样基于词语重叠度，我们还可以推出另外一种计算形式：

$$OverlapNov2(S_i) = 1 - Overlap(S_i \leftarrow \bigcup_{0<j<i} S_j)$$

这种方法将以前所有的相关句子视为一个整体，便于计算。但是，信息并不是词语的简单叠加，在语言学上缺乏必要的理论依据。以下面三个时序上连续的句子为例：

句子 S_1： Tom likes Jerry.

句子 S_2： Jerry likes Jack.

句子 S_3： Jack likes Tom.

按照 *OverlapNov2* 的定义，计算过程如下：

$$OverlapNov2(S_3) = 1 - Overlap(S_3 \leftarrow \bigcup_{0<j<3} S_j)$$

$$= 1 - \frac{|\{Jack, like, Tome\} \cap \{\{Tome, like, Jerry\} \cup \{Jerry, like, Jack\}\}|}{|\{Jack, like, Tome\}|}$$

$$= 1 - \frac{|\{Jack, like, Tome\} \cap \{\{Tome, like, Jerry, Jack\}\}|}{|\{Jack, like, Tome\}|}$$

$$= 1 - \frac{|\{Jack, like, Tome\} \cap \{\{Tome, like, Jerry, Jack\}\}|}{|\{Jack, like, Tome\}|}$$

$$= 1 - \frac{|\{Jack, like, Tome\}|}{|\{Jack, like, Tome\}|}$$

$$= 0$$

信息新颖度的计算结果为 0，而实际上句子 S_3 传达了 S_1 和 S_2 所没有的新信息内容，*OverlapNov2* 的缺陷在于它将信息内容进行简单的叠加，另外一个问题是没有考虑词语位置的先后顺序。*OverlapNov* 也忽略了位置信息，但是它将一个单句作为研究的对象，避免了上面例子中存在的问题。

5.2.2 带权重的词重叠度计算

词重叠度简单、直接而有效，很多研究者采用词重叠度都达到了很好的效果。不过，在 *Overlap* 的计算中，我们发现它将句子中所有的词语都平等对待，没有进行区分，实际上这是不科学的。

以 TREC 主题 N51 的句子“Pinochet cast a long and wide shadow in economic affairs as well , launching a privatized social security system and other free market policies that set examples that are still models from Argentina to Mexico.”为例，“long”，“wide”等词语对新信息的区分度就远远小于“economic”、“security”、“policies”等词。

为此，我们引入了词的 χ^2 统计值 chi_w 作为词的权重，对词重叠度进行扩展，即：

$$Overlap^*(S_i \leftarrow S_j) = \frac{\sum_{w \in S_i \cap S_j} chi_w}{\sum_{w \in S_i} chi_w}$$

当词语 χ^2 统计值 chi_w 相等的时候，$Overlap^*$ 就退化为 $Overlap$。

5.3 相似度比较方法 (Similarity Margin)

词重叠度方法主要考虑了当前句子与历史信息的词重叠度，而没有兼顾当前句子与主题的相关性。为此，我们需要对当前句子与主题和历史信息的相关性进行比较，目的在于选择与主题相关度更高，而与以前句子相关性更低的句子作为承载新信息的结果。

依据历史信息处理的不同策略，我们提出了三种比较方法来衡量句子的新颖程度：

$$MaxPrevNov(S_i) = \lambda Sim(S_i, T) - (1-\lambda)\max_{k<i \wedge S_k \in R} Sim(S_i, S_k)$$

$$AverPrevNov(S_i) = \lambda Sim(S_i, T) - (1-\lambda) AVG_{k<i \wedge S_k \in R} Sim(S_i, S_k)$$

$$PrevAverNov(S_i) = \lambda Sim(S_i, T) - (1-\lambda) Sim(S_i, \overline{SP})$$

其中，λ 为主题相关度与历史相关度的调节参数，取值范围：$0 \le \lambda \le 1$；R 为相关句子集合，T 为表示主题的查询向量，$\overline{SP}$ 为 S_i 以前所有句子的平均向量。

MaxPrevNov 表示的是当前句子 S_i 与历史相关度最大值的差值。

AverPrevNov 表示的是当前句子 S_i 与历史相关度平均值的差值。

PrevAverNov 表示的是当前句子 S_i 与历史平均句子向量相关度的差值。

目前，研究者常使用的最大区间相关度（Maximal Marginal Relevance）就属于这类方法，和 *MaxPrevNov* 的思路一致。

5.4 信息增强评价方法（Information Increment）

词重叠度与相似度都是直接对一个单句进行计算，单句的长度有限，实际计算的时候往往容易发生大的偏差，而在句子集合或者文档上计算就要好得多。为此，我们提出了一种信息增强的评价方法：比较引入当前句子前后信息的增强情况，从而间接的估算出该句子的新颖程度。假如当前句子没有新的信息，则增加该句子并不会对整体信息带

来收益。反之，则丰富了原来的语义。这和直接计算单个句子的信息要宽泛得多，理论上更有说服力。

基于相关度，我们可以得到信息增强评价方法如下：

$$\mathrm{IGN}ov(S_i) = \mathrm{Re}\,l(T, \mathrm{Pr}\,evSentences + S_i) - \mathrm{Re}\,l(T, \mathrm{Pr}\,evSentences)$$

其中，*PrevSentences* 是 S_i 以前的所有相关句子的并集， T 为表示主题的查询向量，*Rel* 为主题与句子集合的相关度，可以采用第四章提到的检索模型进行计算。

同样，我们还可以采用语言模型条件概率的方式进行估计，即：

$$\mathrm{IGN}ov'(S_i) = P(T \mid \mathrm{Pr}\,evSentences + S_i) - P(T \mid \mathrm{Pr}\,evSentences)$$

最后，我们还可以采取类信息熵的办法来估计，即：

$$\mathrm{IGN}ov''(S_i) = P(T \mid \mathrm{Pr}\,evSentences)\log P(T \mid \mathrm{Pr}\,evSentences)$$

$$- P(T \mid \mathrm{Pr}\,evSentences + S_i)\log P(T \mid \mathrm{Pr}\,evSentences + S_i)$$

在这里，需要作进一步解释的是，我们提到的信息增强和机器学习方面的信息增益不是同一个概念，采用的计算方法也不一样，但是出发点是一致的。

5.5 其他方法

除了上面的信息新颖度评价的方法以外，还有多种方式可以借鉴。我们主要介绍两种很有意思但不常用的计算方法：语言模型与句子语义距离计算。

5.5.1 语言模型（Language Model）

在 4.4 中，我们介绍了语言模型的信息检索技术。我们可以采用 Kullback-Leibler [Kullback 1951]交叉熵来度量两个语言模型Θ_1与Θ_2之间的距离，即：

$$KL(\Theta_1 \| \Theta_2) = \sum_w p(w \mid \Theta_1)\log\frac{p(w \mid \Theta_1)}{p(w \mid \Theta_2)}$$

通过 KL 距离，我们可以定义句子 s_i 与以前所有相关句子的新颖度：

$$KLNov(s_i \mid s_1,\ldots,s_{i-1}) = KL(\Theta_{s_i} \| \Theta_{s_1 \ldots s_{i-1}}) = \sum_w p(w \mid \Theta_{s_i})\log\frac{p(w \mid \Theta_{s_i})}{p(w \mid \Theta_{s_1 \ldots s_{i-1}})}$$

同样，语言模型估计时存在数据稀疏问题，采用最大似然估计与插值平滑技术，我们可以计算其中的参数[Leah 2002]：

$$p(w \mid \Theta_{s_i}) = \lambda_1 p(w \mid \Theta_{MLs_i}) + (1-\lambda_1) p(w \mid \Theta_{MLs_{1,\ldots,s_i}})$$

$$p(w \mid \Theta_{s_1},\ldots,_{s_{i-1}}) = \lambda_2 p(w \mid \Theta_{ML_{s_1,\ldots,s_{i-1}}}) + (1-\lambda_2) p(w \mid \Theta_{MLs_{1,\ldots,s_i}})$$

也可以采取退化的平滑技术来估计参数[James 2003]，即：

$$p(w|\Theta_{s_i}) = \lambda_s p(w|\Theta_{MLs_i}) + \lambda_t p(w|\Theta_{MLt}) + \lambda_e p(w|\Theta_{MLe})$$

其中 Θ_{MLs_i} 为句子 s_i 的最大似然语言模型，Θ_{MLt} 为句子 s_i 所在的主题范围内的最大似然语言模型，Θ_{MLe} 为英语通用语言环境下的最大似然语言模型。

经过实验，语言模型的方法能够取得的性能与相似度比较方法相当。

5.5.2 句子语义距离计算方法（Sentence Semantic Distance）

句子语义距离计算旨在量化语义或者概念表达式之间的不同程度，从而间接地计算出句子语义相似度。 采用 WordNet 作为语义知识库，我们提出了一种语义距离计算的方法来度量句子信息的新颖程度。

仅考虑语义的上下文关系，WordNet 是一种树型结构，语义的孩子节点（也就是一个概念 Synset）c 与它的父亲节点 p 之间的语义距离可以简单地估计为：

$$Dist(c,p) = IC(c) - IC(p)$$

其中，$IC(c)$表示的是概念 c 的信息量，即：

$$IC(c) = -logP(c) = -log\{\Sigma freq(w)/N\}$$

而 w 属于集合 c 或者 c 所有的下位 Synset，$freq(w)$ 为 w 的频率，N 为总频率，这些都可以从 WordNet 提供的平衡语义语料库中训练得到。

基于此，我们可以计算出任意两个语义节点 c_1 与 c_2 之间的语义距离，如下：

$$Dist(c_1,c_2) = IC(c_1)+IC(c_2)-2\times IC(LSuper(c_1,c_2))$$

其中，$LSuper(c_1,c_2)$表示的是两个语义节点 c_1 与 c_2 最近的共同上位概念，其意义类似于最小公倍数。

如图 5.1 所示，*LSupper*(“brother”, “arm”)=“Person”，而 *Dist*(“brother”, “arm”)实质上就是图 5.1 中所有红色虚线的权重之和。根据 $IC(c)$的定义，我们可以推算出 $Dist(c_1,c_2) \geq 0$ 。

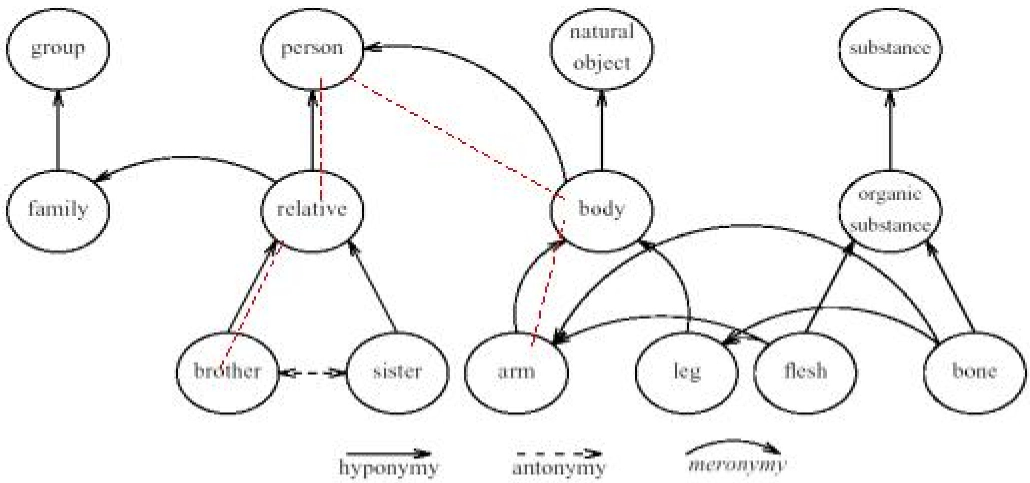


图 5.1 WordNet 中语义距离计算示意图

某个词语往往存在多个语义 Synset，我们计算词语之间的距离时，并不进行词义的排岐，采取下面的公式：

$$\mathrm{Dist}(w_1, w_2) = \max_{w_1 \in c_1 \wedge w_2 \in c_2} \{\mathrm{Dist}(c_1, c_2)\}$$

仅考虑词语之间的上下位关系还是不够的，我们还进一步考虑了同义、部分整体等关系。

在词语距离的基础上，我们借助于几何学上关于点到直线距离的原理，提出了词 w 与句子 S 距离 $WSSD(w, S)$的计算方法。即：

$WSSD(w, S)$=min { $Dist(w,w_i)$ | $w_i \in S$, 其中 w 与 w_i 是词，而 S 为句子}

在此基础上，我们定义一个句子 S_1 与另一个句子 S_2 之间的语义距离 $SSSD$。即：

$$SSSD(S_1, S_2) = \frac{\sum_{w_i \in S1} WSSD(w_i, S_2) + \sum_{w_j \in S2} WSSD(w_j, S_1)}{|S_1| + |S_2|}$$

在这里，$|S_1|$与$|S_2|$为句子中包含的词数，根据表达式，我们可以知道 $SSSD$ 满足交换律，即 $SSSD(S_1,S_2)=SSSD(S_2,S_1)$。图 5.2 给出了句子语义距离计算的原理图。

最后，我们可以定义基于语义距离计算的句子新颖度：

$$SDNov(S_i, S_{1,\ldots,i-1}) = \lambda_1 SSSD(S_i, T) - \lambda_2 SSSD(S_i, S_{1,\ldots,i-1})$$

其中：

$SSSD(S_i, T)$ 表示的是句子 S_i 与主题 T 的语义距离；

$SSSD(S_i, S_{1,\ldots,i-1})$ 表示的是句子 S_i 与历史信息 $S_{1,\ldots,i-1}$ 的语义距离；

λ_1与λ_2为调节参数。

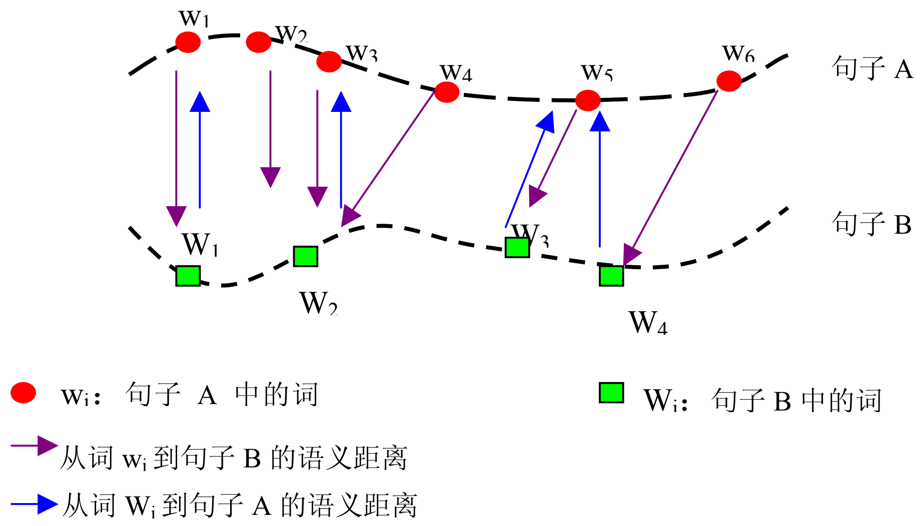


图 5.2 句子语义距离计算的原理图

5.6 新信息检测试验与分析

在 TREC2003 的 50 个主题数据上，我们采用不同的新信息检测技术，经过多次参数调节，进行了多组实验，表 5.1 给出了 11 组有代表性的实验结果。

新信息检测方法	调节参数	平均准确率 P	平均召回率 R	平均 F-measure
MaxPrevNov	0.7	0.50	0.16	0.185
AverPrevNov	0.7	0.39	0.75	0.462
PrevAverNov	0.7	0.23	0.02	0.031
MaxPrevNov	0.9	0.40	0.71	0.462
AverPrevNov	0.9	0.39	0.79	0.470
PrevAverNov	0.9	0.40	0.59	0.419
OverlapNov	0.0	0.47	0.55	0.443
OverlapNov2	0.0	0.39	0.79	0.470
IGNov	0.5	0.37	0.66	0.426
KLNov	0.0	0.39	0.61	0. 410
SDNov	0.9	0.38	0.58	0. 389

表 5.1 新信息检测实验（TREC2003 数据集）

从表中，我们总结了如下规律性的结论：

（1） 前面六组都是相似度比较方法，调节参数λ由 0.7 升为 0.9，各种相似度比较方法都提高了性能，尤其是 *PrevAverNov* 的平均 F-measure 值由 0.031 跃升为 0.419。这意味着一个句子是否传达新的信息，一方面和历史信息的比较有关，其本身与主题的相关度更为关键。一个句子和主题相关度很高的时候，它往往就是一个新信息的载体。

（2） 相比较而言，在相似度比较方法中，*AverPrevNov* 较优。

（3） *OverlapNov* 和 *OverlapNov2* 是简单直接的词重叠度计算方法，但是都取得了较优的结果，这表明：词重叠方法简单而有效。

（4） 对比 *OverlapNov* 和 *OverlapNov2*，*OverlapNov2* 提高了 6.1%，这说明：带权重的词重叠度更能客观地反映新信息的实际情况。

（5） 信息增强评价方法 *IGNov* 取得了一定的效果，比较性能偏低，这说明，尽管计算更加便利，但是引入了全部历史信息，增大了误差，导致了最终结果没有更好地反映信息的新颖程度。

（6） *KLNov* 和 *SDNov* 作为一个试验性质的方法，有一定的理论依据，也可以取得一定的效果，但是仍然无法实用。*KLNov* 仍然受到数据稀疏的困扰；而 *SDNov* 过分依赖于语义体系，而且计算代价非常高，在我们的实验中，简化之后的句子语义距离的计算复杂度依然很高，在三台 CPU 为 P4 1.6GHz 的机器上并行计算 50 个主题的数据，平均花费的时间大约为 26 个小时。

5.6 本章小结

本章主要阐述了最终新信息检测的技术方法，其中包括：词重叠度及其扩展方法、相似度比较方法与信息增强的评价方法。随后，5.5 节还阐述了一些不常用的计算方法：语言模型与语义距离计算的方法。

最后，我们分别在 TREC2003 的数据集上进行新信息检测实验。

第六章 监督学习条件下的句子检索与新信息检测

6.1 概述

在第四章、第五章，我们分别介绍了句子的检索与新信息的检测，给定的信息只有待分析的主题与文档集合，如图 6.1a 所示。这是一种非监督学习条件下的句子检索与新信息检测，没有任何先验的知识或者结果可以用来学习和训练，非监督学习环境是新信息检测绝大多数的应用场景，也是技术实用化必须面对的挑战，也是我们研究的核心所在。

在某些情况下，用户可能会给出部分的提示或者反馈，因此，我们可以事先获得部分结果。依据已知的部分结果，我们可以采取机器学习的方法进行适当的调整，从而进一步提高新信息检测的最终性能。TREC 会议的新信息检测任务从 2003 年起，除了保留核心的非监督子任务外，还有针对性地设计了三种监督学习子任务，如图 6.1b，6.1c，6.1d 所示，分别对应新信息检测任务的子任务二、子任务三与子任务四，其中虚线范围给定了已知的信息，带点号的图表示相关句子结果，方格图为新信息结果。

监督学习环境具有如下优点：

（1） 给定部分结果，可以减少中间环节，直接评价并测试后续技术的性能。例如，给定了所有的相关句子，我们就可以减少中间的句子检索环节，减少了关键的影响因素，利于新信息检测技术的直接研究与评测。

（2） 部分结果的反馈，有利于机器学习与系统的自适应。通过前面的实验，我们可以发现新信息检测技术的性能很大程度上依赖于具体的主题与文档集，性能千差万别，波动范围极大。监督条件下，在知道部分检索结果或者新信息结果的前提下，我们可以从中抽取出可利用的一般性信息，用于调整系统的参数，并对不同主题进行自适应。

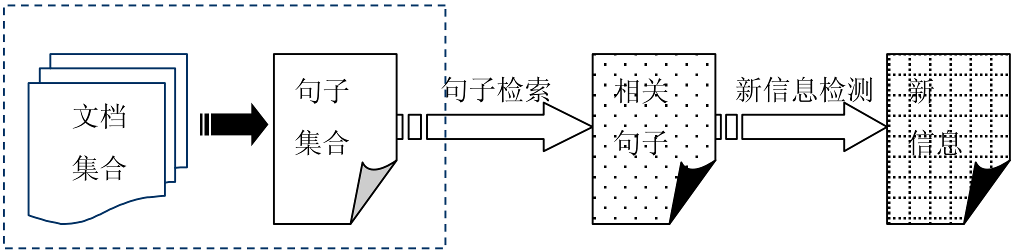


图 6.1a 非监督学习条件下的句子检索与新信息检测

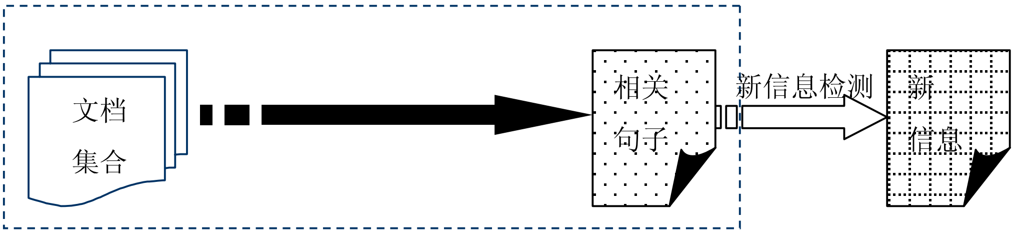


图 6.1b 监督学习条件下的新信息检测（给定所有的相关句子）

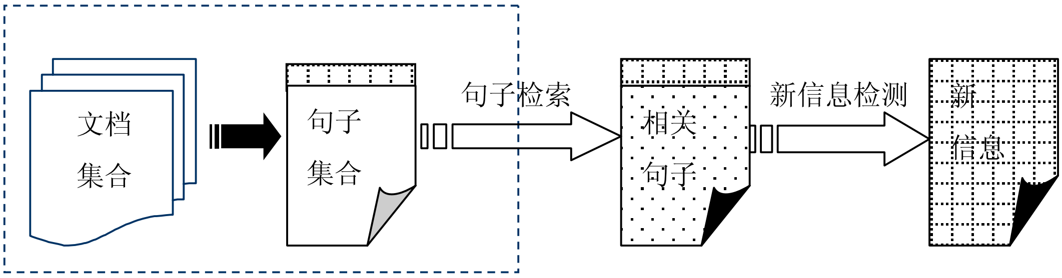


图 6.1c 监督学习条件下的新信息检测（给定部分文档中的相关句子与新信息句子）

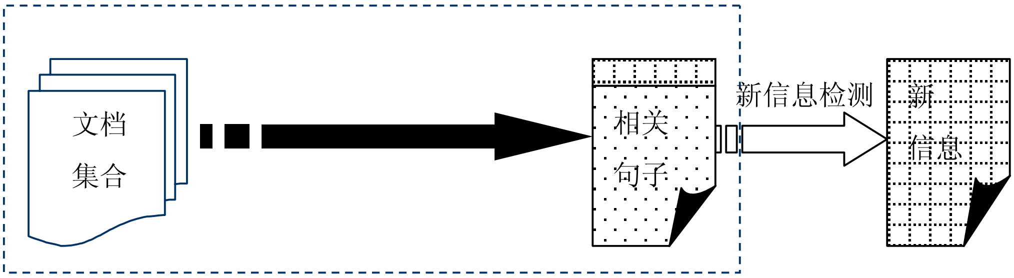


图 6.1d 监督学习条件下的新信息检测（给定所有相关的句子与部分文档中的新信息的句子）

本章主要简要地介绍监督学习条件下的参数调整与阈值设置机制，并阐述分类方法进行句子检索与新信息检测的基本思想。最后，进行实验并给出分析结论。

6.2 监督学习环境下的参数调整与阈值设置

在监督学习条件下，我们可以利用已有的知识或者部分结果，通过机器学习的方法，进一步的提高新信息检测的性能。主要环节包括：

- 进一步的特征选择

  以图 6.1b 所示的监督学习条件为例，我们可以将已知的相关句子作为一个结果集合，而其他的不相关句子作为对立的结果集合，利用第三章提到的互信息或者 χ^2 统计方法，作进一步的特征选择，再次排除噪音数据，保留更有区分能力的特征词语。这为后续的句子新信息检测构造了更优的特征空间。

- 真实反馈

  4.2.5.2 节介绍了非监督条件下的伪相关反馈的查询扩展方法，而在监督条件下，给定的结果真正相关，是真实的反馈。图 6.1c 给出了相关部分结果，对应部分

剩余的句子就是不相关的句子集合。我们可以分别视之为正向反馈和逆向反馈，采取经典的 Rocchio 公式[Rocchio 1971]对原始查询进行扩展，可以得到：

$$Q^{\tau}=\alpha*Q+\beta*\frac{\sum S_{rel}}{N_{rel}}-\gamma*\frac{\sum S_{nrel}}{N_{nrel}}$$

其中 N_{rel} 为已知的相关句子数，N_{nrel} 为已知的不相关句子数，α，β，γ 分别为各部分的权值，属于可调参数。最终的查询实现了部分结果的反馈，更好的适应了当前的应用环境。

我们也可以采取类似的手段，实现部分新信息结果的反馈。

- 参数调整

句子检索与新信息检测的算法模型往往都存在部分可以调节的参数。在非监督条件下，参数更多的是某个数据集合上实验基础上得到的经验值，而新信息检测往往依赖于具体的主题与文档集，因此，原来的参数并不一定适用于新的数据集。而在监督条件下，给定具体主题的部分结果，我们可以据此有目的地调节参数，使之尽可能地符合当前主题的实际情况，从而更好地指导剩余结果的分析。

- 阈值设置

阈值的设置直接决定最终的结果集合。和参数调整类似，阈值设置存在同样的问题。在监督条件下，阈值存在一般性的规律。

以图 6.1d 所示的情况为例，此时，给定了所有的相关句子，并已知每个主题前 5 个文档具有新信息的句子。在某个范围的句子集合内，我们定义一个参量 *Ratio_N_R* 如下：

$$Ratio\_N\_R=\frac{\text{新信息句子数}}{\text{相关句子数}}$$

对于每个主题，在前 5 个文档范围内，逐步扩大句子集合，每增加 1 个带新信息的句子，我们都记录此时的 *Ratio_N_R*，对于每个不同的主题来说，*Ratio_N_R* 千差万别，不存在一般性的规律。但是，对于同一个主题来说，随着文档叙述文字的逐渐深入与展开，相关句子中包含新信息的几率愈来愈小，*Ratio_N_R* 存在逐渐变小的趋势。通过曲线拟合的方法，我们可以通过早期的 *Ratio_N_R* 来估计出内在的变化趋势，最后计算出整个文档集合范围内的 *Ratio_N_R*。给定了所有的相关句子数目，我们就可以大致估算出最后新信息句子的总体数目。

图 6.2 给出了 TREC2003 数据集合上 *Ratio_N_R* 的拟合曲线，其中 x 轴表示的是前 5 个文档范围内的 *Ratio_N_R*，y 轴表示的是整个文档范围内的 *Ratio_N_R*。我们分别采用了一次函数、二次函数、对数函数、指数函数、幂函数进行模拟，最后发现，对数函数 $y=0.1168\ln(x)+0.2995$ 的平方误差最小， 50 组数据平方误差的总和为 $R^2=0.4567$。

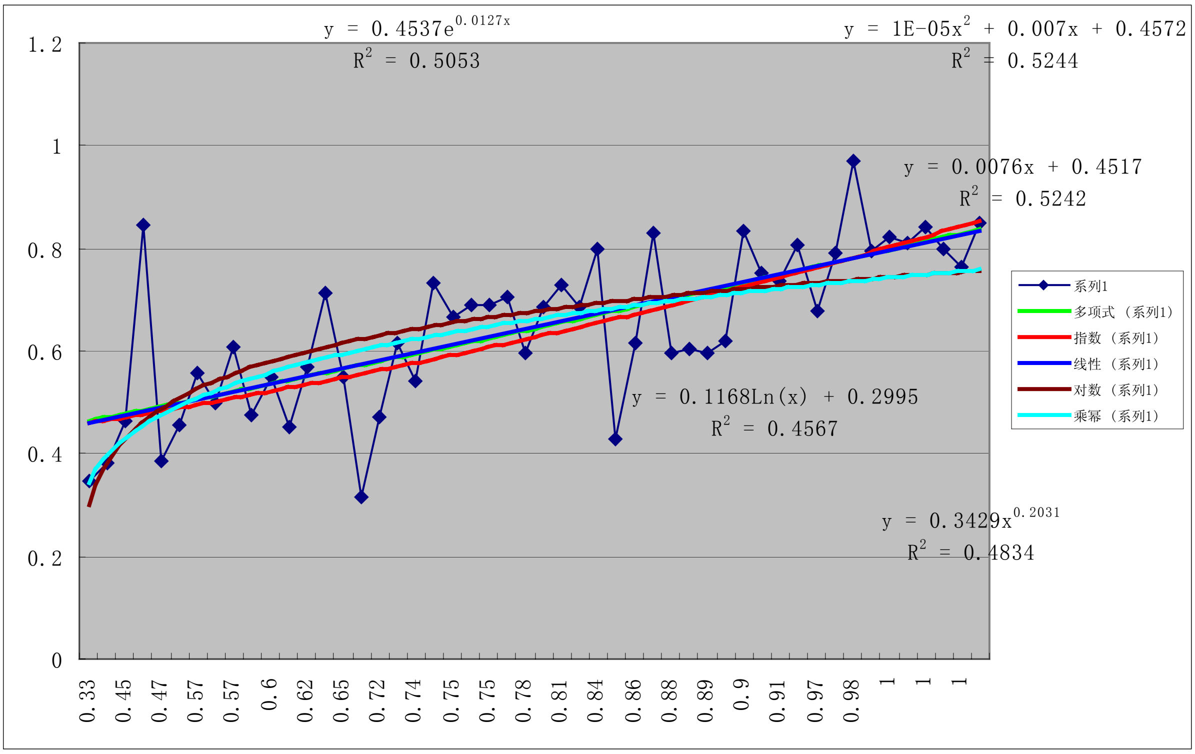


图 6.2 TREC2003 数据集上 *Ratio_N_R* 的拟合曲线

通过这种方法，我们最后能大致的估计出输出的结果数目，因此，我们可以按照新颖度进行排序，从高到低输出所需数量的句子结果，从而避免了非监督条件下阈值设置的盲目性。

6.3 基于分类的句子检索与新信息检测方法

在监督条件下，我们可以采取分类的思想实现句子的检索或新信息检测。基本思想为：在给定的部分文档结果中，将相关句子视为类别 REL，而将不相关的句子视为 NON-REL，训练出一个二元分类器，对剩余文档中的句子依次分类，将它们归结为相关的 REL 类或者不相关的 NON-REL。

同样，在新信息检测环节，我们可以将新信息的句子视为 NEW 类，而相关而不包含新内容的句子归为 NON-NEW 类，具体的训练过程和分类过程一致。

我们以 TREC2003 子任务三的句子检索为例，这里举例的数据都来自第一个话题 N1。给定了前 5 篇的相关句子信息，因此，我们可以将其中的 52 个相关句子作为 REL 类，而其他的 106 个句子为对应的 NON-REL 类，为此，我们可以得到 158 个训练样本，表 6.1 给出了部分的训练样本与测试数据片断。

句子 ID	向量表示	类别
NYT19980629.0465 8	482:0.329448 30418:0.768812 85128:0.548081	REL
NYT19980629.0465 9	482:1.000000	NON-REL

NYT19980629.0465 10	482:0.226595 5501:0.287315 14602:0.513173 51265:0.431760 66307:0.436277 76766:0.367625 82458:0.301422	NON-REL
NYT19980629.0465 11	482:0.489471 485:0.457421 5501:0.310316 44565:0.421367 55411:0.405021 79280:0.336597	REL
APW19990924.0219 6	3446:0.342615 3848:0.425703 12612:0.401669 15879:0.250162 36578:0.434427 47512:0.309547 82290:0.439235	UNKNOWN（未知，待分类）
APW19990924.0219 7	56001:0.445543 82428:0.771948 82458:0.453418	UNKNOWN（未知，待分类）

表 6.1 相关性分类的训练样本与测试数据片断

在表 6.1 中，句子向量每维的格式表示为“特征词语 ID:权重”，前四行属于训练样本，后两条是待分类的测试数据。

如何根据已有的训练结果，来推断测试数据的归属，这是一个典型的分类问题，Noovel 采用了支持向量机（SVM: Support Vector Machine）的二元分类方法。

支持向量机由前苏联数学家 Vapnik 首先提出来的[Vapnik 1995]，是统计机器学习的代表方法。它基于有序风险最小化归纳法(Structural Risk Minimization Inductive Principle)，通过在特征空间构建具有最大间隔的最优分类面，得到两类问题的划分准则，使期望风险的上界达到最小。所谓“最优分类面”就是要求分类面不但能将两类无错误地分开，而且要使两类的分类间隔最大，图 6.3 给出了最优分类面的示意图。支持向量机在解决小样本、非线形及高维模式识别问题中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中[许 2005]。

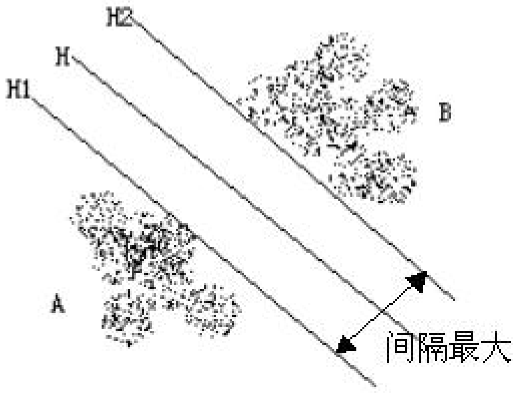


图 6.3 最优分类面示意图

在具体的技术实现上，我们采用了国际上非常通行的一套开放源代码系统 SVM^{light}（Version 5.00）[Joachims, 2002] [Joachims, 1999]。针对句子检索与新信息检测，我们都训练了两种 SVM 分类器。其中一种直接采用已知的部分结果训练分类器（称之为：“Inductive SVM”），另外一种将测试数据作为未知分类结果也加入到训练样本中，一起

参与机器学习训练（称之为："Transductive SVM"）。后者对小样本分类很有帮助，比较适合于测试数据分布特性和训练样本相似的分类问题。

6.4 实验与分析

依据不同的监督条件，在 TREC2003 的数据集上，我们做了两组监督条件下的实验：

（1） 监督实验一：给定所有的相关句子，测试此时的新信息检测性能；

（2） 监督实验二：给定前 5 篇文档中所有的相关句子，在剩下的 25 篇文档中，测试此时的句子检索性能。

下面给出实验结果，并进行简要地分析。

6.4.1 监督实验一

给定了所有的相关句子，我们可以直接在相关句子上检测出带有新信息的句子，实验结果如表 6.2 所示。

新信息检测方法	平均准确率 P	平均召回率 R	平均 F-measure
MaxPrevNov	0.67	0.73	0.684
AverPrevNov	0.65	0.79	0.697
OverlapNov	0.71	0.74	0.704
OverlapNov2	0.65	1.00	0.774
IGNov	0.65	0.99	0.770
IGNov+反馈	0.65	1.00	0.774

表 6.2 给定所有相关句子的新信息检测实验（TREC2003 数据集）

对比非监督条件下的新信息检测（见表 5.1），给定相关句子的条件下，新信息检测的性能至少提高了 45.5%（表 5.1 最好的结果为 0.470，而这里最差的性能为 0.684），由此可以估计出句子检索所带来的损失。

将非监督条件下的新信息检测 F-measure 除以任务 2 对应的指标，大致可以推算出句子检索的性能指标。如我们同样采取 *OverlapNov2* 方法，不同条件下的性能比为 0.470:0.774=0.607，这正好是句子检索 F-measure 大致的性能范围。

最后两组数据能够表明反馈发挥的作用，方法之间的对比与非监督条件类似，这里不再赘述。

6.4.2 监督实验二

给定了每个主题前 5 篇文档中的所有相关句子，表 6.3 给出了句子检索的实验结果。

句子检索方法	平均准确率 P	平均召回率 R	平均 F-measure
VSM	0.61	0.82	0.664

OKAPI	0.60	0.81	0.657
支持向量机分类	0.59	0.71	0.627
VSM+反馈	0.58	0.96	0.701

表 6.3　给定前 5 篇文档相关句子的句子检索实验（TREC2003 数据集）

对比非监督条件下的句子检索（见第 4.5 节），给定部分结果的前提下，句子检索的性能均得到了提高，参数调整与阈值设置起到了一定的作用。

支持向量机分类的方法能够部分地解决句子检索的问题，但并不是最优的方案。一方面，支持向量机的训练过程非常的耗费时间和空间，Transductive 方式的支持向量机往往需要 5-6 个小时才能完成小样本的训练。另外一方面，支持向量机的最终性能并不如普通的向量空间模型。主要原因在于：分类器主要是依据训练样本分布来计算一个最优的分类面，而忽略了句子检索过程中很重要的查询需求，而向量空间模型直接计算句子与查询向量的相关度，更能反映句子检索的内在规律。

6.5 本章小结

本章阐述了监督条件下的句子检索与新信息检测的应对策略，主要包括：进一步的特征选择、真实反馈、调整参数、阈值设置。进一步地提出了基于分类的句子检索与新信息检测方法。实验结果表明：相应的应对策略具有一定的效果，分类方法能够部分地解决问题，但存在着自身的缺陷，实验性能也不是很理想。

第七章 Noovel 系统在 TREC2004 新信息检测任务中的公开评测

7.1 概述

2004 年 7 月，第 13 届 TREC 会议开始发布新信息检测任务数据集和比赛指南（参见附录 1“TREC 2004 Novelty Track Guidelines” ）。和 TREC2003 一样，它主要包括四个子任务，它们分别是：

- 任务 1：要求直接从给定的文档中，检索出所有的相关句子，并给出包含新信息的句子；这是最重要也是最根本的非监督条件下的任务；
- 任务 2：给定所有相关的句子，要求给出所有包含新信息的句子；
- 任务 3：对于每个主题，给定前五篇文档中包含的相关句子与新信息句子，要求给出剩余文档中的相关句子与新信息句子。
- 任务 4：对于每个主题，给定所有文档中的相关句子，同时给出前五篇文档中包含新信息的句子，要求给出剩余文档中的新信息句子。

与以往不一样的是，TREC2004 不再给定相关文档，每个主题给出的文档集合包括 25 篇相关文档，而其他文档则为不相关的噪声数据集。这种情况更符合实际的应用场景，客观上增加了文档检索的过程，给新信息检测尤其是监督条件下的检测带来了新的挑战课题。

不相关文档带来的困难主要体现在以下两个方面：

- 增加了文档检索过程，给句子检索和新信息检测带来了人为的干扰因素；

  TREC2004 总共包含了 50 个主题，其中给定的文档数目有 1808 个，平均每个主题 36.16 个文档，不相关文档所占比例高达 30.09%。编号为 N75 的主题给定的文档数目为 70 个，不相关文档占了接近 2/3。文档检索过程需要能从中正确地挑选出相关的文档，一旦错误地将不相关文档检索出来，那么，以后该文档中所有计算出来的相关句子或者新信息结果都是错误的。这无疑给新信息检测增加了新的不确定因素，必然降低最终的性能。

- 每个主题前 5 个文档的信息有限，减少了监督条件下的可训练样本数，增加了监督条件下的机器学习困难。

  每个主题前 5 个文档可能与主题相关，也可能不相关，不相关的文档只能提供反面的训练样本，加剧了数据的稀疏程度。根据我们对 50 个主题的统计，前 5 个文档中，平均只有 3.14 个为相关文档，其中也只有 2.76 个文档包含了新信息的句子。甚至有 9 个主题的前五个文档均与主题不相关，因此，子任务 3、子任务 4 可以利用的信息相当有限。

国际上总共有 14 支队伍参加了 Novelty2004 的比赛，其中包括中国科学院计算技术研究所、清华大学、国立台湾大学、美国的哥伦比亚大学（Columbia University）、麻省大学（University of Massachusetts）、密歇根州立大学（University of Michigan）、爱荷华大学（University of Iowa）、南加州大学（University of Southern California-ISI）、爱尔兰的都柏林城市大学（Dublin City University）、日本的明治大学（Meiji University）、法国的 Universit_e Paris-Sud / LRI、Institut de Recherche en Informatique de Toulouse、IDA / Center for Computing Science、CL Research。

按照规定，每个子任务可以提交 5 个结果（TREC 中定义为 Run）。最后，任务 1 提交了 60 个 Run，任务 2 提交了 55 个 Run，任务 3 提交了 40 个 Run，任务 4 提交了 28 个 Run。只有 7 支队伍参加了全部四个子任务。

Noovel 系统代表中国科学院计算所参加了 TREC2004 新信息检测任务全部四个子任务，提交了 19 个 Run。和国际同行的对比，Noovel 在最关键的任务 1 中，最终的新信息检测结果排名第一；在任务 3 的句子检索方面，我们提交的两个 Run 并列排名第一。综合全部四个子任务，Noovel 名列前茅。

本章主要汇报 Noovel 在各个任务中的评测结果，并进行简单的对比分析。

7.2 任务 1 测试结果与对比

下面给出了 Noovel 在 TREC2004 任务 1 中提交的 5 个 Run 的结果。

Run ID	主要方法描述	句子检索平均 准确率 P/召回率/F 值	新信息检测平均 准确率 P/召回率/F 值
ICTOKAPIOVLP	OKAPI 方法检索；*OverlapNov2* 新信息检测	0.32/0.73/0.415	0.17/0.57/0.239
ICTVSMCOSAP	VSM 检索；*AverPrevNov* 新信息检测	0.31/0.68/0.397	0.08/0.47/0.130
ICTVSMLCE	VSM 检索，LCE 扩展；*AverPrevNov* 新信息检测	0.29/0.73/0.392	0.12/0.71/0.199
ICTVSMFDBKL	VSM 检索，前 10%的结果进行伪相关反馈；*AverPrevNov* 新信息检测	0.28/0.77/0.385	0.13/0.71/0.202
ICTVSMFDBKH	VSM 检索，前 20%的结果进行伪相关反馈；*AverPrevNov* 新信息检测	0.28/0.77/0.389	0.12/0.77/0.198

表 7.1　Noovel 在 TREC2004 任务 1 的评测结果

在任务 1 所有提交的 60 个 Run 中，Noovel 提交的 ICTOKAPIOVLP 结果在句子检索性能方面排名第 5，比第 1 名的 UIowa04Nov11 仅差 0.005。而在新信息检测方面，

ICTOKAPIOVLP 排名第一，如图 7.1 所示[Soboroff 2004]。

根据我们对 50 个主题新信息检测结果的统计分析，最终发现 ICTOKAPIOVLP 在 8 个主题上排名第一、2 个主题上排名第二、5 个主题上排名第三，其中总共有 22 个主题的新信息检测性能均在前五名，接近一半的主题数目。这表明，该方法有一定的普适性。

总结 ICTOKAPIOVLP 所采用的算法模型，我们可以得出一些规律性的结论：OKAPI 方法在文档检索方面比其它模型更有效，这很大程度上间接地提高了句子检索的准确率和召回率，我们的评测结果也佐证了 OKAPI 在以前 TREC Web 检索比赛上的优越表现。*Noovel* 在新信息检测方面的优势也表明：浅层语言分析为新信息的检测提供了高质量的中间结果；另外，和我们提交的其他结果相比，基于权重的词重叠度方法 *OverlapNov2* 能够更好地度量信息的新颖度。

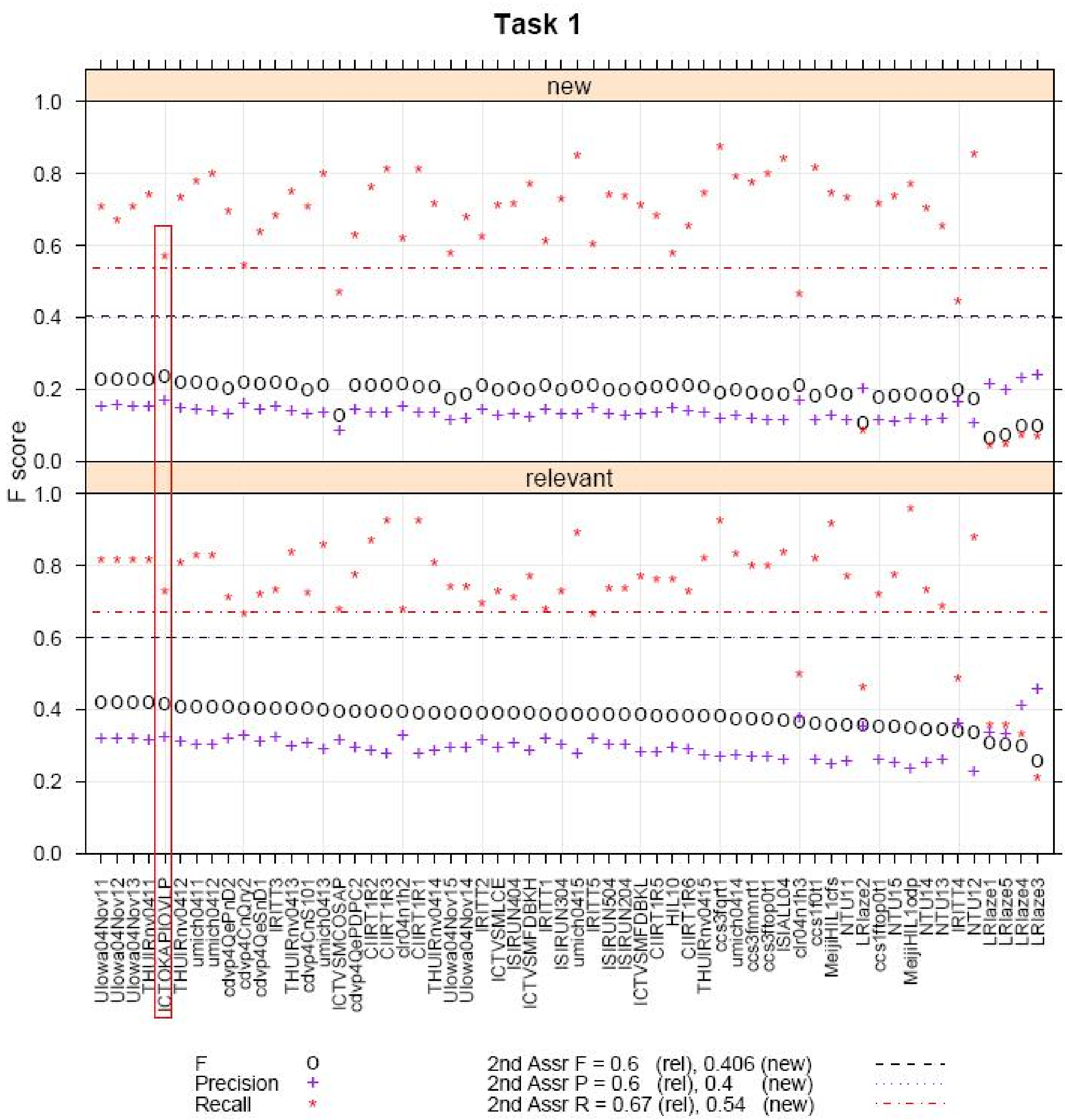


图 7.1 TREC2004 任务 1 中所有 60 个 Run 的综合对比图（方框内为 ICTOKAPIOVLP）

不过，我们还发现，几乎所有系统的平均召回率均高于人工判断的结果（红色的虚线），但是其准确率与 F 值均远远低于人工判断，最好系统的 F 值（ICTOKAPIOVLP：0.239）比人工判断的 F 值整整低了 41.1%。而 TREC2003 最好的系统结果略高于人工判断，这种损失主要是由不相关文档引起的，这也反映了不相关文档对系统判断的巨大干扰。

7.3 任务 2 测试结果与对比

下表是 Noovel 在 TREC2004 任务 2 中提交的 5 个 Run 的结果。

Run ID	主要方法描述	新信息检测平均准确率 P/召回率/F 值
ICT2VSMLCE	*IGNov* 进行新信息检测，动态调整阈值	0.43/0.89/0.559
ICT2VSMIG95	*IGNov* 进行新信息检测，返回排名前 95%的结果	0.42/0.77/0.523
ICT2OKAPIAP	VSM 检索，LCE 扩展；*AverPrevNov* 新信息检测	0.42/0.81/0.534
ICT2OKALCEAP	OKAPI 建模；*AverPrevNov* 新信息检测	0.42/0.83/0.539
ICT2VSMOLP	VSM 建模，*OverlapNov* 新信息检测	0.47/0.89/0.599

表 7.2　Noovel 在 TREC2004 任务 2 的评测结果

对比任务 1 提交的结果，给定所有相关句子的条件下，新信息检测的综合指标提高了差不多 3 倍，主要原因在于它避免了文档检索与句子检索的损失。在我们提交的五个 Run 中，*OverlapNov* 的新颖度评价方法依然占据优势，*IGNov* 也取得了较好的效果。

图 7.2 给出了参赛的所有 55 个 Run 的结果[Soboroff 2004]，我们最好的结果 ICT2VSMOLP 相对靠前，但是并不占据绝对的优势。与任务 1 中 ICTOKAPIOVLP 的性能相比，我们在监督条件下的新信息检测仍然有可以提升的空间。

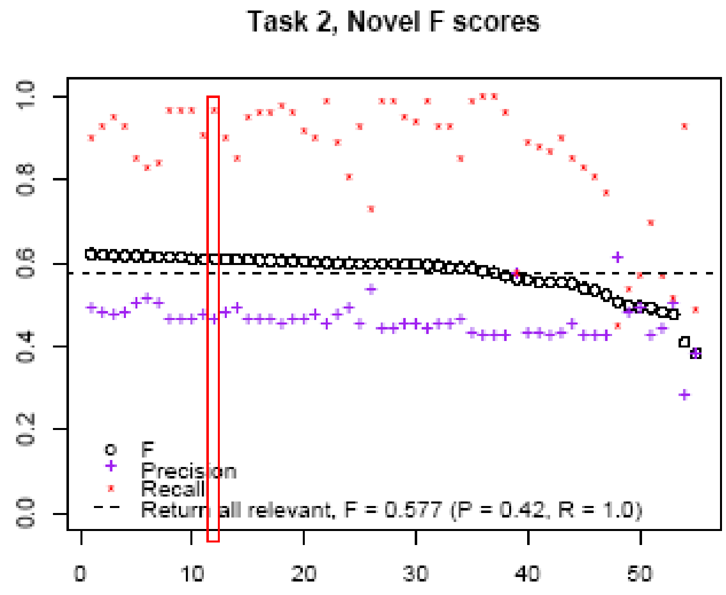


图 7.2 TREC2004 任务 2 中所有参赛结果的综合对比图（方框内为 ICT2VSMOLP）

7.4 任务 3 测试结果与对比

表 7.3 给出了 Noovel 在 TREC2004 任务 3 中 4 个 Run 的评测结果。

Run ID	主要方法描述	句子检索平均 准确率 P/召回率/F 值	新信息检测平均 准确率 P/召回率/F 值
ICT3OKAPFDBK	OKAPI 方法进行句子检索，引入反馈；*OverlapNov2* 新信息检测	0.38/0.68/0.441	0.12/0.46/0.212
ICT3OKAPIIG	OKAPI 方法进行句子检索；*IGNov* 新信息检测	0.37/0.72/0.459	0.15/0.50/0.216
ICT3OKAPIOLP	OKAPI 方法进行句子检索，学习并调整阈值和参数；*OverlapNov2* 新信息检测	0.37/0.76/0.464	0.15/0.52/0.217
ICT3VSMOLP	VSM 方法进行句子检索，学习并调整阈值和参数；*OverlapNov2* 新信息检测	0.370.76/0.464	0.14/0.63/0.213

表 7.3 Noovel 在 TREC2004 任务 3 的评测结果

在任务 3 中，每个主题都给定了前五篇文档中所有的相关句子与新信息结果，我们采取了监督条件的机器学习策略，有针对性地调整了参数和阈值。对比任务 1 提交的结果，我们在这里提交的 Run 在句子检索性能方面均有显著性的提高，新信息检测有一定的进步但是并不明显，相对最好的结果来说，新信息检测稍微有一点下降，这和训练过程的过学习有关。

图 7.3a 给出了参赛的所有 40 个 Run 的句子检索对比结果[Soboroff 2004]，我们提交的两个结果 ICT3OKAPIOLP 和 ICT3VSMOLP 并列第一。根据我们对 50 个主题句子检索结果的统计分析，最终发现它们在 11 个主题上排名并列第一、6 个主题上排名并列第二、2 个主题上排名并列第三，其中总共有 19 个主题的句子检索结果均在前三名。这表明：在监督条件下，Noovel 句子检索的机器学习和调整是有效的，在国际上处于领先的技术水准。

图 7.3b 给出了新信息检测的综合对比，Novel 的结果处在前列但优势有限，还需要作进一步的研究分析。

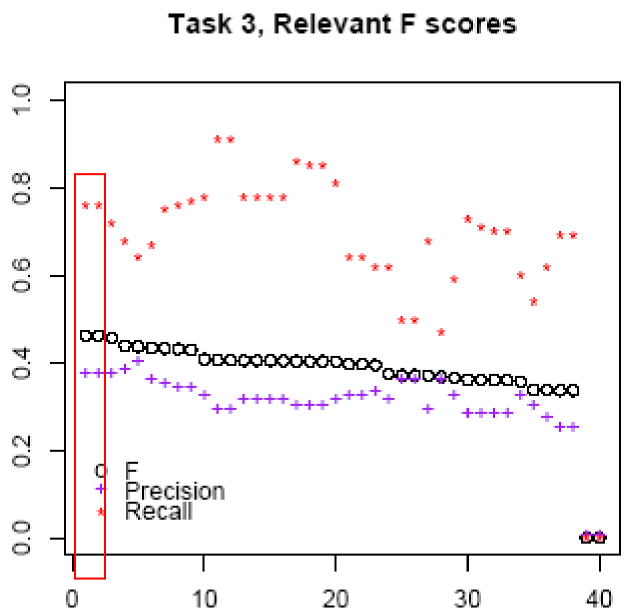


图 7.3a TREC2004 任务 3 中所有参赛结果的句子检索性能综合对比图（方框内为 ICT3OKAPIOLP 和 ICT3VSMOLP）

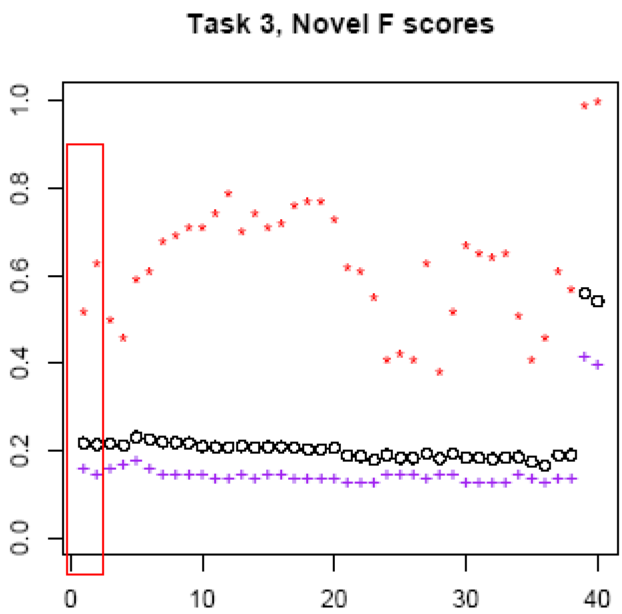


图 7.3b TREC2004 任务 3 中所有参赛结果的新信息检测性能综合对比图（方框内为 ICT3OKAPIOLP 和 ICT3VSMOLP）

7.5 任务 4 测试结果与对比

下表给出了 Noovel 在 TREC2004 任务 4 中 5 个 Run 的结果。

Run ID	主要方法描述	新信息检测平均 准确率 P/召回率/F 值

ICT4OVLPCHI	OKAPI 方法建模；*OverlapNov2* 新信息检测	0.41/0.63/0.469
ICT4OVERLAP	VSM 检索；*OverlapNov* 新信息检测	0.41/0.65/0.476
ICT4OKAPIIG	OKAPI 方法建模；*IGNov* 新信息检测	0.37/0.85/0.496
ICT4OKAAP	OKAPI 方法建模；*AverPrevNov* 新信息检测	0.37/0.73/0.464
ICT4IG	VSM 建模；*IGNov* 新信息检测	0.37/0.88/0.504

表 7.4　Noovel 在 TREC2004 任务 4 的评测结果

对比提交的 5 个结果，*IGNov* 的评价方法在 OKAPI 和向量空间建模的情况下，均取得了较好的效果。这表明，它是一种相对有效的评价手段。

在所有 28 个参赛结果当中，我们的结果不太理想，见图 7.4[Soboroff 2004]。这反映了 Noovel 在监督条件下的新信息检测仍然有一定差距。

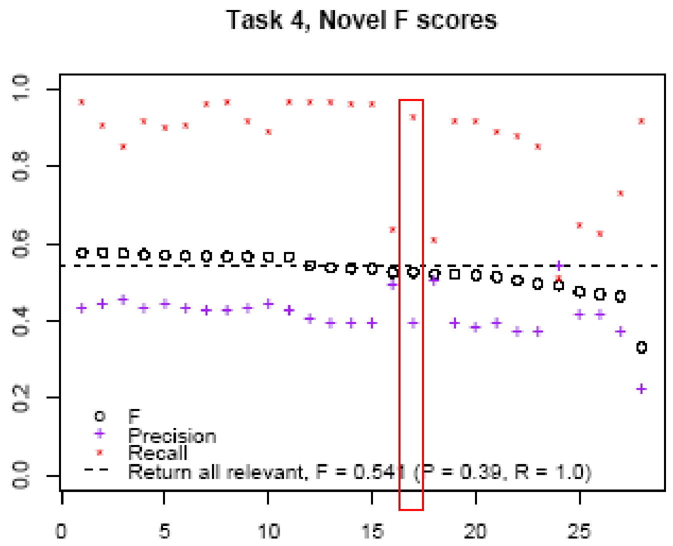


图 7.4 TREC2004 任务 2 中所有 28 个 Run 的综合对比图（方框内为 ICT4IG）

7.6 本章小结

本章首先介绍了第 13 届 TREC 比赛新信息发现任务的主要情况，总共参赛的有 14 支队伍，提交了 183 个 Run。更大的新挑战在于给定的文档不全是相关文档，每个主题平均有 30.09%的不相关文档作为干扰因素。这增加了新信息检测的难度，同时也给监督条件下的机器学习带来了挑战。

接下来，我们依次介绍了 Noovel 系统参加 TREC2004 四个子任务的情况，Noovel 在最关键的任务 1 中，新信息检测结果排名第一；在任务 3 的句子检索方面，我们提交的两个 Run 并列排名第一，其他的子任务也取得不俗的成绩，和国际同行进行综合对比表明：Noovel 在新信息检测方面的算法研究和技术实现处于领先的地位，基于浅层语言分析的句子检索与新信息检测技术是非常有效的。

第八章 结束语

8.1 本文主要贡献与创新

本文介绍了新信息检测的需求背景，主要围绕句子级别的信息检索与新信息检测的总体构架，展开了深入地研究，提出了针对新信息检测的浅层语言分析技术，详尽地介绍了英文浅层语言分析的各个环节，并对中文浅层语言分析特有的分词问题，给出了具体的解决思路；在句子检索方面，本文提出了三种有针对性的方法：扩展的向量空间模型、概率检索模型与语言模型；在相关句子结果基础上，我们对信息新颖度进行建模量化，主要提出了带权重的词相关度、相似度比较以及信息增强方法。另外，第六章还专门讨论了监督条件下的机器学习方法，从而能够进一步提高新信息检测的性能。最后，针对浅层语言分析、句子检索与新信息检测，在开放测试数据集合上，作者进行了大量的实验，并从理论上对实验结果进行了各种对比分析。

本文主要的贡献与创新表现在以下几个方面：

- 特定的中英文浅层语言分析

  在句子级别的检索与新信息检测的研究过程中，通过实验发现了信息检索和过滤与语言分析之间的内在关系，并将这些内在的影响作进一步的量化比较。为此，我们研究并实现了针对句子检索与新信息检测的中英文浅层语言分析技术，极大地提高了最终的句子检索与新信息检测的性能，语言分析质量对信息检索的影响甚至超越了检索模型本身。

  在针对信息检索与新信息过滤的浅层语言分析方面，本文还作了一般性的理论分析并进行了对比实验。比如：如何从自然语言表述中生成并扩展用户真正的查询意图；英文的各种形态究竟是采用词形还原还是采用传统的词干抽取。

- 基于浅层语言分析的句子检索

  我们主要分析了句子检索存在的特定问题，进一步深入地研究了各种适合于句子检索的模型算法，并与文档级别的检索进行了比较。在实验中，我们发现，句子检索的困难主要在于可计算的信息粒度太小，很多在文档级处理使用的模型算法不能简单地推广到句子检索，影响性能的更大因素在于句子信息的进一步挖掘分析与查询扩展。

- 句子级别的新信息检测

  这是新信息检索的最终目标，我们总结了现有的种种方法，并进一步提炼升华，提出了三种有一定代表性的新信息度量的指标。句子级别的新信息检测建立在相关检索的基础之上，同时，它具有自身的特点。这是时序性很强的信息过滤任务，需

要兼顾信息与主题的关联程度，同时还要与已有历史的信息进行比对，寻找新信息之所在。

- 一个可以灵活配置、健壮实用的新信息检测实验系统 Noovel

  最终，我们将研究的理论与技术综合地集成到了 Noovel 系统中，这是一个可以灵活配置，融入了多种技术的试验平台。其中，我们将各个阶段的分析结果保留到中间结果文件，便于各个研究阶段在较高的相同水准上，进行各种对比实验。

8.2 下一步研究方向

本文对浅层语言分析技术、句子检索建模、信息新颖度的量化方面进行了卓有成效的研究，在查询分析与扩展、监督学习、参数调节以及阈值设置等方面取得了一定成果，最终的实验效果和公开评测均取得了不俗的成绩。但是，新信息检测仍然是一个很新的应用需求，总共只有 3 年的研究历史。Noovel 目前还是一个实验性质的系统，离最终的实用化尤其是在中文环境下的实用化仍然存在一定的距离。

新信息检测方面的进一步研究工作可以围绕以下几个方面来开展：

- 信息粒度的进一步分析

  传统的信息检索与信息过滤的粒度一般都是一篇文章或者网页，新信息检测的信息粒度是一个单句。整篇文章和单句都比较容易抽取，信息处理的可操作性较强。但是，在人们的实际写作或者浏览过程里，整篇文章往往传达多种信息，甚至涵盖多个主题思想，而用户往往只关注某种单一的信息内容，整个篇幅的信息粒度往往偏大。比如，只关心姚明的球迷可能不会关心火箭队之外的赛事，而一篇体育新闻往往会报道包括火箭队之内、所有当时发生的 NBA 赛事。另外一方面，单个句子的信息粒度又偏小，多个句子的有机组合才能有效地传达出一个事件的完整信息。段落似乎是一个折衷的方案，但并不是最合适的。我们需要进一步研究能够独立表述一个完整信息、可计算的最小单元。

- 信息的演化、追踪与再组织

  围绕特定的主题，人们会采取多种方式，逐渐地展开并一步步地传达信息。我们需要检测出新的信息，但新信息并不是简单地累加。在同一篇文章之内，信息之间存在着起承转合，而伴随着人们认识的变化，不同时期的信息也在不断地演化发展，甚至会出现前后矛盾的情况。实际上，从一个历史阶段来看，语言作为信息的载体也在不断的演变和进化。因此，新信息检测不仅要能发现新的信息，还需要追踪这种变化，并对信息进行再组织，从而客观地反映信息的变化和发展过程。

- 中文新信息检测的研究与评测

  本文给出了中文浅层语言分析的过程，句子检索乃至新信息检测都独立于具体的语言表达。但是，中文新信息检测仍然具有自己特有的问题和困难，比如：中文句子粒度的切分问题，和英文不一样的是，早期的中文本身就没有标点符号，行文

风格相当随意，有的很长一段话就一个句号，逗号很大程度上就分割了一个完整的句子。因此，我们不能仅仅根据句号来划分句子，但是也不能完全依据逗号。另外，中文的分词对后续的分析也存在着阻碍，需不需要汉语分词？汉语分词对新信息检测的影响有多大？这些都是我们需要研究的特定问题。

中文新信息检测研究的瓶颈还在于：目前，我们还没有一个相对科学的评测机制，缺乏一定规模的测试数据集。

8.3 前景与展望

新信息检测技术可以针对自然语言表达的需求，为用户提供粒度更小、冗余更少的相关信息。随着互联网的日益壮大和人类信息需求的增加，这项技术会给我们的生活带来是实实在在的便利，促进信息的进一步整合和挖掘。

下面，我们以三个潜在的应用场景来展望新信息检测的未来前景，并进一步展望浅层语言分析的应用前景。

8.3.1 可排重、更精细的信息检索与过滤平台

（Fine-grained IR/IF with redundancy eliminator）

如图 8.1 所示，在这个平台上，用户可以采用自然语言自由地表达信息查询需求，通过文档检索系统 Google 和新信息检测系统 Noovel 的处理，人们可以直接看到相关的句子片断，并可以继续过滤出只包含新信息的句子列表。

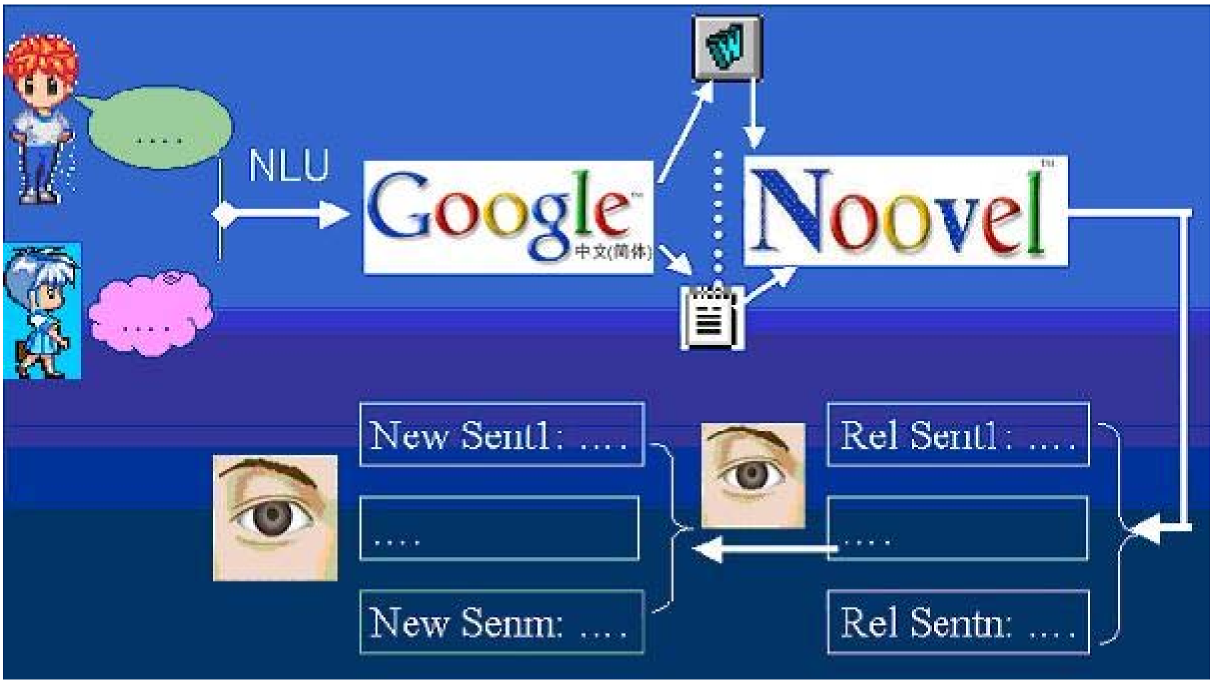


图 8.1 可排重、更精细的信息检索与过滤平台

8.3.2 可订制的新闻摘要（Customized News Abstraction；CNA）

人们可以将自己的需求，通过自然语言告诉新信息检测系统 Noovel，它会根据需要按照时间顺序自动地检索出相关新闻的摘要，如标题或者新闻导语。并进一步排除冗余，最后推送给信息需求者。图 8.2 给出了一个球迷订制的新闻摘要示例。

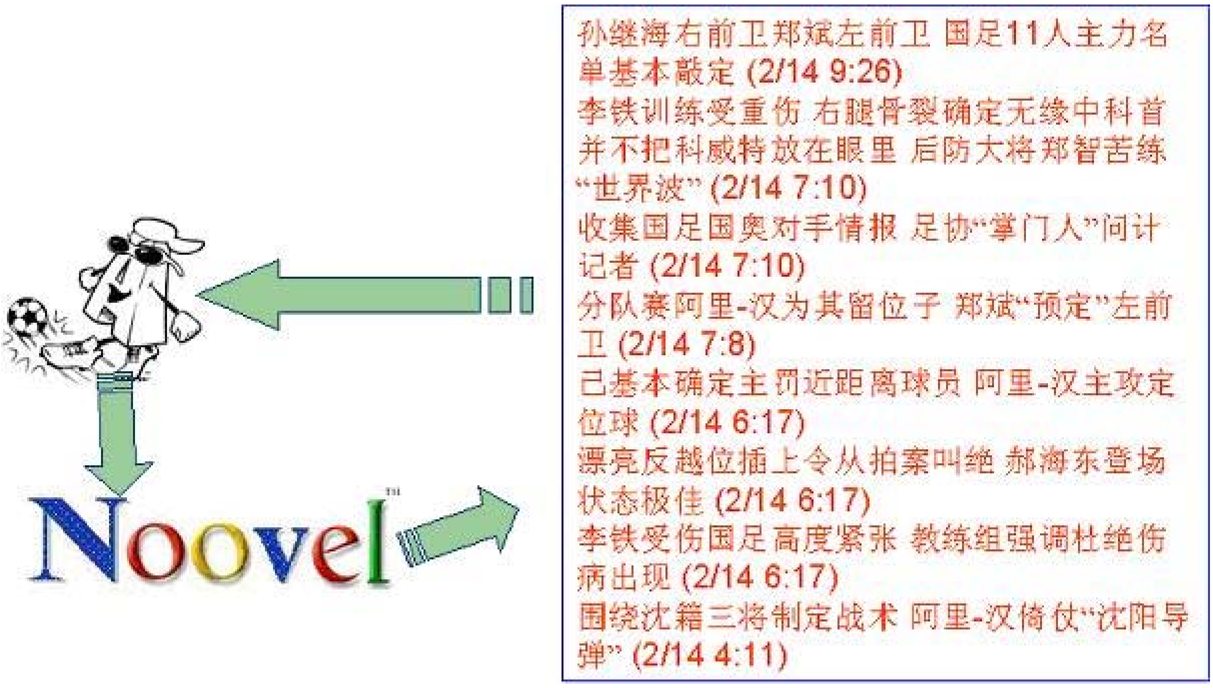


图 8.2 球迷订制的新闻摘要示例

8.3.3 新信息检测辅助阅读器（Noovel Aided Reader; NAR）

人们在阅读或者检索长篇幅文献的时候，往往需要尽快地定位所需信息，并掠过已知的冗余信息。为此，我们可以利用新信息检测技术，为读者提供一个辅助查询框，如图 8.3 所示。和常用的查找界面不一样的是，这里不要求完全一致的精确匹配，人们不仅可以采取多种方式表达查询定位需求，还可以选择阅读只包含新信息的内容片断。

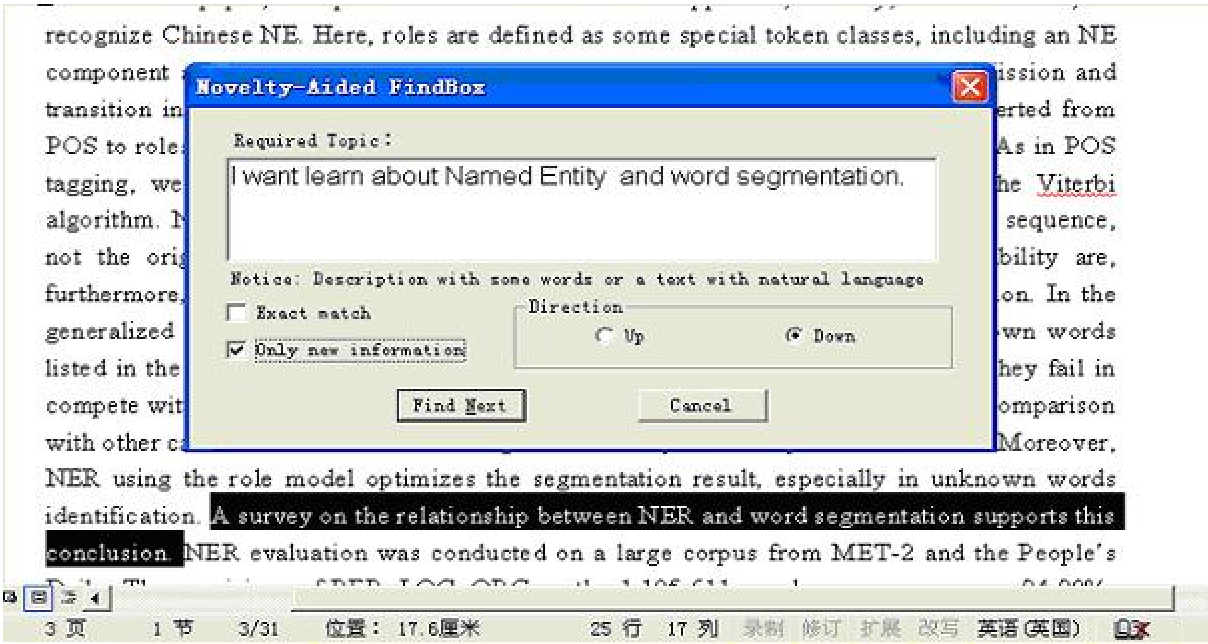


图 8.3 新信息检测辅助阅读器

在新信息检测的研究过程中，本文通过理论分析和具体的实验发现：浅层语言结构在信息处理中扮演着重要的作用，往往直接影响着信息检索和过滤的最终性能，甚至超过其本身的建模方法。

作为信息的载体，自然语言表层的呈现形式多种多样，文本内容存在着丰富的内涵，相互之间又存在着广泛而又复杂的联系。而现有的信息处理技术更多地停留在文本的表层，并没有涉及到语言本身的内在规律性。源数据的分析处理过程中已经损失了语言潜在的信息内容，因此，纯粹的信息处理建模往往陷于“力不从心”的困境。

另外一方面，自然语言的多样性和复杂性决定了计算机不可能真正地实现自然语言的完全理解。在语言的深层挖掘上，目前的语言学研究和计算语言学研究尚不能充分地理解语言的内在规律。而精细的语言分析并不能充分解决问题，而大量的系统消耗也限制了语言深层分析在信息处理中的实际使用。因此，浅层语言分析是一条目前可行的、相对最优的中间路线，一方面它能够针对需求抽取语言表层之下特定的关键信息，另外一方面，它兼顾了分析结果的准确性和系统的实用性能。本文探索的新信息检测特点的浅层语言分析，实际上提供了另一种进一步提升当前信息处理技术的研究方法。

附录 1. TREC 2004 Novelty Track Guidelines

Summary

The Novelty Track is designed to investigate systems' abilities to locate relevant AND new information within a set of documents relevant to a TREC topic. Systems are given the topic and a set of relevant documents ordered by date, and must identify sentences containing relevant and/or new information in those documents.

For information on past Novelty Tracks, see the overviews:

- Overview of the TREC 2002 Novelty Track
- Overview of the TREC 2003 Novelty Track

This year, the tasks and topic structures remain largely the same as in TREC 2003. The main differences include:

There will be exactly 25 event and 25 opinion topics, and

Each topic will include zero or more irrelevant documents in addition to 25 relevant documents.

These changes are detailed below.

Due dates:

Test data released:	July 1, 2004
Results due date:	September 1/15/22, 2004
Runs allowed:	maximum of 5 runs per group per task

Goal

Currently systems return ranked lists of documents as the answer for an information request. The TREC question-answering track takes this a major step forward, but only for direct questions and only for short, fact-based questions. Another approach to providing answers would be to return only new AND relevant sentences (within context) rather than whole documents containing duplicate and extraneous information.

A possible application scenario here would be to envision a smart "next" button that walked a user down the ranked list by hitting the next new and relevant sentence. The user could then view that sentence and if interested, also read the surrounding sentences. Alternatively this task could be viewed as finding key sentences that could be useful as "hot spots" for collecting information to summarize an answer of length X to an information request.

Tasks

This year there will be four tasks which vary the kinds of data available to the systems and the kinds of results that need to be returned. There will be fifty topics, each with 25 relevant documents selected by the assessor who wrote the topic, as well as zero or more documents which were judged irrelevant. The documents are split into sentences.
The four tasks are, for each topic:

1. Given the full set of documents for the topic, identify all relevant and novel sentences. This is last year's task.
   (This task will be due first, on September 1, 2004.)

2. After the first due date, NIST will release the full set of relevant sentences for all documents. Given all relevant sentences, identify all novel sentences.

3. We will also release the novel sentences within the first 5 documents. Given the relevant and novel sentences in the first 5 documents ONLY, find the relevant and novel sentences in the remaining documents.
   (Tasks 2 and 3 will be due second, on September 15, 2004.)

4. Given all relevant sentences from all documents, and the novel sentences from the first 5 documents, find the novel sentences in the remaining documents.
   (Task 4 will be due last, on September 22, 2004.)

Participants are free to participate in any or all tasks. You may submit a maximum of five runs per task.

Topics and Documents

This year, the track will be using fifty new topics (numbered N51-N100) developed using the AQUAINT collection. AQUAINT contains newswire articles from three different wires: New York Times News Service, AP, and Xinhua News Service. All three sources have documents covering the period June 1998 through September 2000; additionally, the Xinhua collection goes back to January 1996.
The topics are evenly divided between two topic types:

- **Event** topics are about a particular event that occurred within the time period of the collection. Relevant sentences pertains specifically to the event.

- **Opinion** topics are about different opinions and points of view on an issue. Relevant sentences take the form of opinions on the issue reported or expressed in the articles.

The topics have traditional TREC topics statements with a title, description, and narrative. For each topic, the assessor has selected 25 relevant documents and some number (possibly zero) of irrelevant documents from the collection. They are probably not the only documents

for that topic, nor are they necessarily the best. You will be provided with those documents concatenated together in chronological order and separated into individual sentences. Each sentence is tagged with a source document ID and a sequence number.
The documents are on a protected web site located at http://trec.nist.gov/novelty/index.html. The web site is protected since it contains document text and we must be sure you have legitimate access to the document text before you can access it. To get the access sequence for the protected site, send an email message to Lori Buckland, lori.buckland@nist.gov requesting access. Lori will check our records to make sure we have signed data use forms for the AQUAINT data from your organization and respond with the access sequence. Please note that this is a manual process, and Lori will respond to requests during her normal mail-answering routine. Do not expect an instantaneous response. In particular, do not wait until the night before the deadline and expect to get access to the test data.

Task and training data restrictions

This task should be done completely automatically. Any fields in the topic can be used. It should be assumed that the set of relevant documents are available as an ordered set, i.e. the entire set may be used in deciding the sentence sets. However the topics must be processed independently. Both these restrictions reflect the reality of the application.
You are free to use any other TREC documents or training data you would like. Although there are probably other relevant documents in the collection, NIST will not be providing further qrels. You will be asked when runs are submitted to describe additional data used. Tasks 2 and 3 cannot be ordered such that all the test data is hidden from both tasks. Therefore, you are expected to keep the training and test sentences separate between your task 2 and 3 runs. Other training data may be kept in common, but do NOT (for example) submit a task 3 run which takes advantage of the relevant sentences released for task 2.
The topics and judgments for last year's Novelty Track data is available from the TREC web site (LINK). Keep in mind that last year, all documents were judged relevant, whereas this year there are irrelevant documents mixed in. Nevertheless, you may find the data useful for designing and/or training your system.

Format of results

Participants will return either one or two lists of doc id/sentence number pairs for each topic, one list corresponding to all the relevant sentences and the second list (a subset of the first) containing only those sentences that contain new information.
Only submit the sentences required for each task! For task 1, a run submission should have both relevant and novel sentences, but for task 2, a run should only contain novel sentences. Don't include any data given by NIST, only include the output your system is required to produce.
Results must be submitted in the following format. This format is a variation of the TREC ad hoc format, and is identical to last year's format without the sequence number field.

N1 relevant FT924-286 46 nist1

N1 relevant FT924-286 48 nist1

N1 relevant FT924-286 49 nist1

N1 relevant FT931-6554 7 nist1

N1 relevant LA122990-0029 14 nist1

N1 new FT924-286 46 nist1

N1 new FT924-286 48 nist1

N1 new FT924-286 49 nist1

N1 new FT931-6554 7 nist1

N1 new LA112190-0043 15 nist1

N2 relevant LA122490-0040 1 nist1

There should be one file per run, ordered by topic number, including both the relevant and new lists for each topic number.

Field 1 -- topic number, an N followed by a number

Field 2 -- "relevant" or "new"

Field 3 -- document id (the docid field exactly as it appears in the tag)

Field 4 -- sentence number (again exactly as it appears in the tag)

Field 5 -- the run tag; this should be a maximum of 12 characters, letters and digits only; it should be unique to the group, the type of run, and the year

Evaluation

The sentences selected manually by the NIST assessors will be considered the truth data. To avoid confusion, this set of sentences are called RELEVANT in the discussion below. Agreement between these sentences and those found by the systems will be used as input for recall and precision.

Recall = #RELEVANT matched/#RELEVANT

Precision = #RELEVANT matched/#sentences submitted

Recall = #new-RELEVANT matched/#new-RELEVANT

Precision = #new-RELEVANT matched/#sentences submitted

The official measure for the Novelty track will be the F measure (with beta=1, equal emphasis on recall and precision):

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

alternatively, this can be formulated

```
                  2 * (No. relevant sentences retrieved)
F  =  ---------------------------------------------------
```

(No. retrieved sentences) + (No. relevant sentences)

(for novel sentence selection tasks, substitute "new" for "relevant")

Definition for new and relevant

You are trying to create a list of sentences that are:

1. relevant to the question or request made in the description section of the topic,
2. their relevance is independent of any surrounding sentences,
3. they provide new information that has not been found in any previously picked sentences.

附录 2.　Penn Treebank Tagset

CC	Coordinating conjunction e.g. and,but,or...
CD	Cardinal Number
DT	Determiner
EX	Existential there
FW	Foreign Word
IN	Preposision or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List Item Marker
MD	Modal, e.g. can, could, might, may...
NN	Noun, singular or mass
NNP	Proper Noun, singular
NNPS	Proper Noun, plural
NNS	Noun, plural
PDT	Predeterminer, e.g. all, both ... when they precede an article
POS	Possessive Ending, e.g. Nouns ending in 's
PRP	Personal Pronoun, e.g. I, me, you, he...
PRP$	Possessive Pronoun, e.g. my, your, mine, yours...
RB	Adverb, Most words that end in -ly as well as degree words like quite, too and very
RBR	Adverb, comparative, Adverbs with the comparative ending -er, with a strictly comparative meaning.
RBS	Adverb, superlative
RP	Particle
SYM	Symbol, Should be used for mathematical, scientific or technical symbols
TO	to
UH	Interjection, e.g. uh, well, yes, my...
VB	Verb, base form, subsumes imperatives, infinitives and subjunctives
VBD	Verb, past tense, includes the conditional form of the verb to be
VBG	Verb, gerund or persent participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner, e.g. which, and that when it is used as a relative pronoun
WP	Wh-pronoun, e.g. what, who, whom...
WP$	Possessive wh-pronoun，e.g.
WRB	Wh-adverb，e.g. how, where why

参考文献

[Abdul-Jaleel 2004] N. Abdul-Jaleel, J. Allan, W.B. Croft, F. Diaz, L. Larkey, X. Li, M.D. Smucker, C. Wade. UMass at TREC 2004: Novelty and HARD, In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Baum 1972] Baum, L. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. Inequalities 3:1-8.

[Bikel 1999] D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. Machine Learning,34(1-3):211–231, 1999.

[Brill 1992] Brill, Eric. 1992. A simple rule-based part of speech tagger. Proceedings of the Third Annual Conference on Applied Natural Language Processing, ACL.

[Brill& Marcus 1994] Brill, Eric & Marcus, M. 1993. Tagging an unfamiliar text with minimal human supervision. ARPA Technical Report.

[Brill 1994] E. Brill. Some advances in rule-based part of speech tagging. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa., 1994.

[Brill 1995] Brill, Eric. 1995. Unsupervised learning of disambiguation rules for part of speech tagging.

[Carbonell 1998] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR, pages 335－336, 1998.

[CNNIC 2004] 中国互联网络信息中心,中国互联网络发展状况统计报告，2004.7. http://www.cnnic.net.cn/

[Conroy 2004] J.M. Conroy. A Hidden Markov Model for the TREC Novelty Task, In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Conroy 2003] J.M. Conroy, Daniel M., Dianne P.. From TREC to DUC to TREC Again In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.293

[Conroy 2001] J. Conroy, J. Schlesinger, D. O'Leary, and M. Okurowski. Using HMM and

Logistic Regression to Generate Extract Summaries for DUC". In DUC 01 Conference Proceedings, 2001.

[Dai 1999] Dai, Y., Khoo, C.S.G. and Loh, T.E. (1999). A new statistical formula for Chinese text segmentation incorporating contextual information. Proc ACM SIGIR99, pp. 82-89.

[Eichmann 2004] D. Eichmann, Y. Zhang, S. Bradshaw, X. Ying Qiu, L. Zhou, P. Srinivasan, A. Kumar Sehgal, H. Wong. Novelty, Question Answering and Genomics: The University of Iowa Response. In Proceedings of the 13th Text Retrieval Conference (TREC ’04) , Gaithersburg, MD USA, November 16-19, 2004, 71

[Erkan 2004] G. Erkan. The University of Michigan in Novelty 2004, In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Erkan&Radev 2004] Erkan, G., & Radev, D. R. (2004). Lexpagerank: Prestige in multi-document text summarization. In Lin,D., & Wu, D. (Eds.), Proceedings of EMNLP 2004, pp. 365.371 Barcelona, Spain. Association for Computational Linguistics.

[Fellbaum 1998] Fellbaum, C., ed. "WordNet: An Electronic Lexical Database" . MIT Press, Cambridge, MA. 1998

[Francis&Kucera 1982] W Francis and H Kucera. Frequency Analysis of English Usage. Houghton Mifflin, 1982.

[Ganesh 2003] Ganesh R., Kedar. B, Chirag Shah., Deepa P.. Generic Text Summarization Using WordNet for Novelty and Hard. In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.303

[Greene&Rubin 1971] Greene, B. B. & Rubin, G. M. 1971. Automatic grammatical tagging of English. Technical Report, Brown University. Providence, RI.

[Grefenstette 1994] Gregory Grefenstette and Pasi Tapanainen, `What is a word, what is a sentence? problems of tokenization', in The 3rd International Conference on Computational Lexicography, Budapest, 1994. 79—87

[Harman 2002] D. Harman. Overview of the TREC 2002 Novelty Track. In Proceedings of the 11th Text Retrieval Conference (TREC ‘02) , Gaithersburg, MD USA, November 19-22, 2002, 57-60.

[Helmut 1994] Helmut Schmid, Part-of-Speech Tagging with Neural Networks Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)

[Hideto 2002] Hideto Kazawa, Tsutomu Hirao, Hideki Isozaki and Eisaku Maeda. A machine learning approach for QA and Novelty Tracks: NTT system description.In Proceedings of the 11th Text Retrieval Conference (TREC '02) , Gaithersburg, MD USA, November 19-22, 2002, 472.

[Hockenmaier 1998] Hockenmaier, J. and Brew, C. (1998) Error-driven learning of Chinese word segmentation. In J. Guo, K. T. Lua, and J. Xu, editors, 12th Pacific Conference on Language and Information, pp. 218-229, Singapore. Chinese and Oriental Languages Processing Society.

[Jahna 2003] Jahna O.,Hong Q.. Ali H. and Dragomir R. The University of Michigan at TREC 2003. In Proceeding of the Eleventh Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp 732

[James 2003] James Allan, Courtney Wade, and Alvaro Bolivar. Retrieval and Novelty Detection at the Sentence Level. . In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003.

[Jin 2003] Qianli Jin, Jun Zhao, Bo Xu. NLPR at TREC 2003: Novelty and Robust . In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.126

[Joachims, 2002] Thorsten Joachims, Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.

[Joachims, 1999] T. Joachims, 11 in: Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 1999.

[Kim 2004] S.-M. Kim, D. Ravichandran, E. Hovy. ISI Novelty Track System for TREC 2004, In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Kit 2002] Chunyu Kit, Haihua Pan and Hongbiao Chen. Learning Case-based Knowledge for Disambiguating Chinese Word Segmentation: A preliminary study. First SIGHAN Workshop attached with the 19th COLING, 2002.8, pp.63-70

[Kullback 1951] S. Kullback and R. A. Leibler. On information and su-ciency. Annals of Mathematical Statistics, 22:79 86, 1951. 6

[Lawrence 1989] Lawrence. R.Rabiner.1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of IEEE 77(2): pp.257-286.

[Leah 2002] Leah S..James Allen, Magaret. E.,Alvaro B. and Courtey. W., UMass at TREC 2002: Cross Language and Novelty Tracks, In Proceeding of the Eleventh Text Retrieval Conference, Gaithersburg, Maryland, November 19-22, 2002, pp.721

[Lovins 1968] J.B. Lovins, 1968: "Development of a stemming algorithm," Mechanical Translation and Computational Linguistics 11, 22-31.

[Luo 2001]Luo H. and Ji Z. Inverse Name Frequency Model and Rules Based on Chinese Name Identifying. In "Natural Language Understanding and Machine Translation", C. N. Huang & P. Zhang, ed., Tsinghua Univ. Press, Beijing, 2001, pp. 123-128.

[Luo 2001(2)]Luo Z. and Song R. Integrated and Fast Recognition of Proper Noun in Modern Chinese Word Segmentation. Proceedings of International Conference on Chinese Computing 2001, Singapore, pp. 323-328.

[Luo 2002]Xiao Luo, Maosong Sun, Benjamin K Tsou. Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information. the 19th COLING, 2002.8, pp.598-604

[Lyman 2003] Peter Lyman, Hal R. Varian, et. How Much Information? 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm S., October, 2003

[Manning& Schütze 1998] Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

[Monz 2002] Christof Monz, Jaap Kamps, and Maarten de Rijke. The University of Amsterdam at TREC 2002. In Proceeding of the Eleventh Text Retrieval Conference, Gaithersburg, Maryland, November 19-22, 2002, pp.603

[Monz&Kuhn 1960] Maron M.E., Kuhn J.L. (1960). On relevance, probabilistic indexing and information retrieval.Journal of the ACM, 7, 216-244.

[O'Connor 2004] N. O'Connor, A.F. Smeaton, P. Wilkins,O. Boydell, B. Smyth. Experiments in Terabyte Searching, Genomic Retrieval and Novelty Detection for TREC 2004,In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Paice 1990] Paice, C.D., "Another Stemmer", SIGIR Forum 24 (3): 56-61 (1990).

[Palmer 1997] Palmer, D. (1997) A trainable rule-based algorithm for word segmentation Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL ’97), Madrid, 1997.

[Peng 2001] Peng, F. and Schuurmans, D. (2001). A hierarchical EM approach to word segmentation. In 6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001)

[Peng 2001(2)] Peng, F. and Schuurmans, D. 2001. Self-supervised Chinese Word Segmentation. In Proceedings of the Fourth International, Symposium on Intelligent Data Analysis (IDA-2001).

[Porter 1980] Porter, M.F., An Algorithm For Suffix Stripping, Program 14 (3), July 1980, pp. 130-137.

[Qi 2002] Hong Qi. Jahna O.,and Dragomir R. The University of Michigan at TREC 2002: Question Answering and Novelty Tracks. In Proceeding of the Eleventh Text Retrieval Conference, Gaithersburg, Maryland, November 19-22, 2002, pp 733

[Richard 2003] Richard Sproat, Thomas Emerson. The First International Chinese Word Segmentation Bakeoff, First SIGHAN Workshop attached with the ACL2003, 2003.7, pp.133-143

[Robertson 1994] Robertson S.E. and Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proceedings of the 17 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232-241,1994.

[Robertson 1996] Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M., Payne, A. Okapi at TREC-4, The Fourth Text REtrieval Conference (TREC-4), NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, pp. 73-86, October 1996.

[Robertson 1999] Robertson, S. E., and Walker, S., Okapi/Keenbow at TREC-8. In Proceeding of the Eighth Text Retrieval Conference, Gaithersburg, Maryland, November, 1999

[Rocchio 1971] Rocchio, J. J. Relevance Feedback in Information Retrieval. In The SMART Retrieval system, Prentice-Hall, Englewood NJ. 1971, 232-241.

[Ryosuke 2003] Ryosuke Ohgaya, Akiyoshi Shimmura and Tomohiro Takagi. Meiji University Web and Novelty Track Experiments at TREC 2003, In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.399

[Salton 1983] Salton, G., McGill, M. J., Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[Salton 1989] Salton, G., Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison Wesley Publishing, 1989.

[Salton 1990] Salton, G., Buckley, C. Improving retrieval performance by relevance feedback. JASIS 41, 1990, pp. 288-297.

[Salton& Buckley 1990] Salton, G., Buckley, C. A note on Term Weighting and Text Matching. TR 90-1166, Department of Computer Science, Cornell University, 1990.

[Schiffman 2004] B. Schiffman, K.R. McKeown. Columbia University, Columbia University in the Novelty Track at TREC 2004,In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Shai 1998] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical Hidden Markov Model: Analysis and applications. Machine Learning, 32:41, 1998.

[Singhal 1996] Singhal A., C. Buckley, and M. Mitra. Pivoted Document Length Normalization. In H. Frei, D.Harman, P. Schauble, and R. Wilkinson, editors, Proceedings of the Nineteenth Annual International A CM SIGIR Conference on Research and Development in Information Retrieval, 1996.

[Soboroff 2003] I. Soboroff. Overview of the TREC 2003 Novelty Track. In Proceedings of the 12th Text Retrieval Conference (TREC ’03) , Gaithersburg, MD USA, November 18-21, 2003, 57-60.

[Soboroff 2004] I. Soboroff. Overview of the TREC 2004 Novelty Track. In Proceedings of the 13th Text Retrieval Conference (TREC ’04) , Gaithersburg, MD USA, November 16-19, 2004, 57-60.

[Srikanth 2003] Srikanth K., Yongmei S., R. Scott Cost, Charles N., Akshay J., Christopher J., Sowjanya R., Vishal S., Sachin B., and Drew O.. UMBC at TREC 12. In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.699

[Stanley&Joshua 1996] Stanley F. Chen and Joshua T. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pages 310-318, 1996.

[Sun 2002] Sun J., Gao J. F., Zhang L., Zhou M Huang, C.N. Chinese Named Entity Identification Using Class-based Language Model, Proc. of the 19th International Conference on Computational Linguistics, Taipei, 2002, pp 967-973

[Sun 2003] Jian Sun, Wenfeng Pan, Hua-ping Zhang. TREC 2003 Novelty and Web Track at ICT, In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.138

[Sun 1993] Sun M.S. English Transliteration Automatic Recognition. In "Computational Language Research and Development", L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press, 1993

[Takagi 1995] T. Takagi, A. Imura, H. Ushida, and T. Yamaguchi, “Conceptual Fuzzy Sets as a Meaning Representation and their Inductive Construction”, International Journal of Intelligent Systems, Vol.10, pp.929-945, 1995.

[Tan 1999] Tan H. Y. Chinese Place Automatic Recognition Research. In "Proceedings of Computational Language ", C. N. Huang & Z.D. Dong, ed., Tsinghua Univ. Press, Beijing,, 1999.

[Taoufiq 2004] Taoufiq Dkaki, J. Mothe. TREC Novelty Track at IRIT-SIG, In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Taoufiq 2003] Taoufiq D., Josiane M..TREC NOVELTY TRACK AT IRIT – SIG. In Proceeding of the Twelfth Text Retrieval Conference (TREC '03), Gaithersburg, Maryland, November 18-21, 2003, pp.337

[Taoufiq 2002] Taoufiq Dkaki, J. Mothe. Novelty Track at IRIT-SIG,In Proceedings of the 11th Text Retrieval Conference (TREC '02) , Gaithersburg, MD USA, November 19-22, 2002, pp.332

[Teahan 2001] Teahan, W. J. and Wen, Y. and McNab, R. and Witten I. H. 2001, A Compression-based Algorithm for Chinese Word Segmentation. In Comput. Ling., 26(3):375-393.

[Thompson 2002] Kevyn Collins-Thompson, Paul Ogilvie, Yi Zhang, and Jamie Callan. Information Filtering, Novelty Detection, and Named-Page Finding.In Proceedings of the 11th Text Retrieval Conference (TREC '02) , Gaithersburg, MD USA, November 19-22, 2002

[Tomiyama 2004] T. Tomiyama, K. Karoji, T. Kondo, Y. Kakuta, T. Takagi. Meiji University Web, Novelty and Genomic Track Experiments,In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Tsai 2004] M.-F. Tsai, M.-H. Hus, H.-H. Chen. Similarity Computation in Novelty Detection and Biomedical Text Categorization,In Proceedings of the 13th Text Retrieval Conference (TREC '04) , Gaithersburg, MD USA, November 16-19, 2004

[Tsai 2003] Ming-Feng Tsai, Wen-Juan Hou, Chun-Yuan Teng, Ming-Hung Hsu, Chih Lee and Hsin-Hsi Chen, Similarity Computation in Novelty Detection and GeneRIF

Annotation. In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.474

[Tsai 2002] Ming-Feng Tsai and Hsin-Hsi Chen. Some Similarity Computation Methods in Novelty Detection. In Proceeding of the 11th Text Retrieval Conference, Gaithersburg, Maryland, November 19-22, 2002

[Vapnik 1995] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995

[Viterbi 1967] Viterbi, A. J. 1967. Error bounds for convolutional codes and asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory* 13: 260-269.

[Wu 1998] Andi Wu ,Zixin Jiang. Word Segmentation in Sentence Analysis. 1998 International Conference on Chinese Information Processing, Beijing, 1998. 169-180.

[Xue 2002] Nianwen Xue and Susan P. Converse. Combining Classifiers for Chinese Word Segmentation, First SIGHAN Workshop attached with the 19th COLING, 2002.8, pp.63-70.

[Ye 2002] Ye S.R, Chua T.S., Liu J. M., An Agent-based Approach to Chinese Named Entity Recognition, Proc. of the 19th International Conference on Computational Linguistics, Taipei, 2002, pp 1149-1155

[Yu 2001] Jiangsheng Yu, Shiwen Yu. Some Problems of Chinese Segmentation. The First International Workshop on MultiMedia Annotation (MMA2001), 2001

[Zhai 2001] Zhai C. and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of the 24th Annual International ACM SIGIRConference on pages 334--342, 2001.

[Zhai 2002] C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 49–56. ACM Press, 2002.

[Zhang 2002] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, L. Zhao and S. Ma. Expansion-Based Technologies in Finding Relevant and New Information: THU TREC2002 novelty track experiments.In Proceeding of the Eleventh Text Retrieval Conference, Gaithersburg, MD USA, November 19-22, 2002, pp.591

[ZHANG 2002(2)] ZHANG Hua-Ping, LIU Qun, Zhang Hao and Cheng Xue-Qi Automatic Recognition of Chinese Unknown Words Recognition. First SIGHAN Workshop attached with the 19th COLING, 2002.8, pp.71-77.

[Zhang 2003] M. Zhang, C. Lin, Y. Liu, L. Zhao, L. Ma, S. Ma. THUIR at TREC 2003: Novelty, Robust and Web.In Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, MD USA, November 18-21, 2003, pp.556

[ZHANG 2003 (2)] Hua-Ping ZHANG, Qun LIU, Xue-Qi CHENG, Hao Zhang, Hong-Kui Yu. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, Second SIGHAN workshop affiliated with 41st ACL; Sapporo Japan, July, 2003, pp. 63-70

[Zhang 2005] Hua-Ping ZHANG, Jian SUN, Bing WANG, Shuo BAI. Computation on Sentence Semantic Distance for Novelty Detection, Journal of Computer and Science Technology, No.3 2005 (to be published).

[Zheng 1999] Zheng J.H. and Wu F. F. Study on segmentation of ambiguous phrases with the combinatorial type. Collections of papers on Computational Lingustics, 1999. Tsinghua University Press, Beijing.129-134

[白 2001] 白硕，计算语言学教程，内部讲义，2001 年 6 月

[丁 2004] 丁国栋，文本检索中的统计语言建模及其相关研究，博士开题报告，中国科学院计算技术研究所，2004 年 3 月

[杜 1990] 杜淑敏 等编著《编译程序设计原理》，北京大学出版社 1990 年版，pp51-55

[冯 1995] 冯志伟，面向计算机的语言研究，《语文与信息》1995 年第 1 期

[高 2001] 高山,张艳等. 基于三元统计模型的汉语分词及标注一体化研究.自然语言理解与机器翻译.北京.清华大学出版社. 2001.8. p.116-p.122

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

[梁 2001] 梁焰,王海波. 信息检索首页题. http://lcc.ict.ac.cn/

[刘 2004] 刘挺，信息检索模型(演讲稿)，http://ir.hit.edu.cn/cgi-bin/download/weekforumcnt.cgi?186, 2004 年秋

[刘 2003] 刘群，计算语言学讲义，http://www.nlp.org.cn/categories/default.php?cat_id=26, 2003

[刘&张 2005] 刘群,张华平,骆卫华，孙健，自然语言理解(译著), 电子工业出版社, 2005 年 1 月. p.157-p.164

[王 2004] 王树西，刘群，白硕，自动问答研究综述，Proc. of 20th International Conference on Computer Processing of Oriental Languages, 2003 年 8 月, p 498-506, 中国，沈阳

[王 2005] 王树西，基于模式推理的自动问答研究，博士学位论文，中国科学院计算技术研究所, 2005 年 7 月

[杨 2003] 杨志峰，稳定的信息检索方法及其在分布式环境下的应用，博士学位论文，中国科学院计算技术研究所，2003 年 7 月.

[许 2005] 许洪波，程学旗，王斌，骆卫华，文本挖掘与机器学习，计算所信息快报（内部），2005 年第 3 期.

[许 2003] 许洪波，大规模信息过滤技术研究及其在 Web 问答系统中的应用，博士学位论文，中国科学院计算技术研究所，2003 年 7 月.

[张 1998] 张仕仁.利用语素词规则消除切分歧义.1998 年中文信息处理国际会议论文集. 1998. 157-162

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

[张 2004] 张华平，刘群. 基于角色标注的中国人名自动识别研究. 计算机学报, vol.27, No.1, 2004, pp.85-91

[朱 1982] 朱德熙.语法讲义.商务印书馆.1982.11-11

致　谢

拜读过几十上百篇博士论文，读得最明白、最感怀的往往就是最后的致谢。在我这篇百页的博士论文当中，这段文字是最不起眼的，在我心里却是最感性最沉重的，这是我五年来最想写也最害怕写的一段，因为这里承载着我五年来的学习、工作与生活，它沉浸着我曾经的感动与感触。我愿意敞开心扉，去掉矫情与伪饰，与诸君分享我的这段生活。

五年前，我本科毕业一年多，在理工科技大厦的一家公司做软件工程师。3 月 9 日，我知道自己在硕士研究生考试中得了 402 分。那天晚上，我很激动地憧憬着未来，盼望着自己能够成为计算机领域的科学家，盘算着种种美丽的梦想。

五年之后的今天，我终于在中国科学院计算技术研究所的大家庭里成长为一个有独立思考能力的科研工作者，不知不觉间潜入了中文语言分析与大规模内容处理的研究领域，在科学研究的道路上取得了一些小小的成绩，有上万的机构和研究者能从我的辛勤工作中受益，这是最值得我自豪的地方；在此同时，计算所和软件室的领导和老师们又给了我太多的荣誉和肯定，这是我自觉惭愧的地方。

今天，我就希望能够通过这个机会来表达我的感恩与愧意。

我要首先感谢我的恩师白硕研究员！白老师的博闻强记、朴实严谨一直是我们后辈望尘莫及的，和白老师的几次对话都给我了莫大的启发，他给了我相当大的自由空间，鼓励我的种种奇思妙想，并给出一针见血式的评论，往往让我最快地把握问题的要害。同时，他和蔼可亲与大度豁达的君子风范一直是我的楷模。在博士论文修改过程中，白老师提出了非常详尽的建议，从研究内容的严谨论证到细微的“的”与“地”混用，无一不倾注着老师对弟子的要求之严与栽培之美！我还记得 2004 年 12 月老师和师母带着小白龙开着车陪我兜风，逛上海外滩的种种情形。白老师不仅是我的学术领航者，更是我处事为人的精神导师。

软件室主任程学旗研究员是我的硕士导师，我有幸成为程老师的开山弟子，我也很有幸能够在 5 年间和程老师朝夕相处。他是我的良师益友，无论是在学术研究上，还是在我的工作生活上，程老师都给了我无尽的关照，倾注了大量的心血。程老师特别尊重我的工作方式，往往会在第一时间和我沟通做学问做事做人的想法。他常跟我说：“一会来找我谈一下”，至今犹在耳边。这份情谊让我没齿难忘！我真诚地希望程老师能做我一生的老师！

刘群副研究员是我在计算语言学尤其是中文信息处理方面的启蒙老师，刘老师近乎苛刻的高标准要求，让我做事不敢有一丝的懈怠，我也是从刘老师身上了解如何去认真做研究的。可以说，没有刘老师的耳提面授，我就不可能完成 ICTCLAS 这样接近完美的汉语词法分析系统。我们一起打拼的岁月让他不仅成为了我的老师，更让他成为了我

的诤友。

软件室副主任郭莉副研究员给了我不少机遇、肯定和鼓励，她常常在不知不觉间让你觉得热心与诚恳。王斌副研究员与许洪波博士是直接领导我的大小组长。他们事无巨细的关心和指导一直让我受益匪浅。王老师还是我一直敬重的鄱阳老乡，我为他的出色管理和广受年轻人爱戴而感到自豪。在博士论文的提纲编撰和具体写作的过程中，洪波给了我非常详尽的建议。平时的工作上，洪波也给了我直接而又无私的帮助！

计算所软件室的同事孙健博士、骆卫华、刘悦博士、张刚、张凯、郭岩博士、谭建龙博士以及我的学友谭松波博士、于满泉博士、吴丽辉博士、张丙奇博士、赵章界博士、姜吉发博士、王树西博士、周昭涛、米嘉、钟尚平博士、谢丰博士等给了我很多的建议和灵感；我的“哥们”丁国栋博士、赵红超博士、张浩、李继锋、朱海龙曾经和我一起打球嬉闹吹牛夜话，陪我度过了许多快乐时光；我的师弟师妹邹刚、俞鸿魁、熊德意、王小飞等在平时的工作中，曾经和我有过非常愉快的合作，在此，一并表示感谢！

同时，我还要感谢软件室秘书宋锢女士和孙萍女士。五年来，她们像大姐姐一样为我们提供热心而又细致的后勤保障，给我提供了诸多的便利。计算所研究生部的宋守礼老师、周世佳老师、张晓辉老师与靳晓明老师在学习与生活诸方面都给了我不少帮助，再次表示感谢。

北京大学的俞士汶教授、詹卫东博士、于江生博士，美国伊利诺斯大学的 Richard Sproat 教授，德国慕尼黑大学的刘乐中博士，北京市政府外事办的姜伟先生，首都信息之窗的单昌明先生、法国原子能实验室的李怡平博士等曾经给过我不同形式的支持和肯定，在此表示感谢！同时，我还要感谢上万名知名和不知名的 ICTCLAS 用户，他们给了我非常多的反馈和建议，并一直鼓励着我不断地创新与改进，他们是我科学研究的强大动力。

我的岳父岳母在学习、工作和生活上给了我无微不至的关怀，岳父敬业负责、刚正不阿与开明果敢的军人风采一直为我所敬佩，他是我人生道路上的楷模和良师！岳母给我准备的可口饭菜一直温暖着我贫瘠的胃口。同时，我也特别感谢远在农村老家的父母，感谢他们 27 年来对我的哺育之恩，感谢他们克服了重重困难，让我安心地完成了 20 年的学习生涯。他们通过自己的勤劳汗水给了我土地一样的朴实、高山一样的坚强、流水一样的善良！我的妹妹张燕、张丽以及妹夫和外甥女给了我不少慰籍，一并致以谢意。

最后，我要特别感谢我美丽大方、知书达礼的爱妻曾飞女士，她给了我世界上最美丽最神奇的爱情，她让我知道生活原来如此的丰富多彩，她激励着我奋发向上，她为我创造了宁静的港湾，她是我最大的收获！在求学期间，她以无畏的牺牲精神为我付出了许多，她细致地校对过我博士论文里包括标点符号在内的每一个错误。今天是我们相遇两周年的日子，我愿将这篇博士论文献给她，以纪念我们曾经走过的美好岁月！

日子在指尖悄悄流淌，不觉间却沉积出暗香阵阵……

张华平

2005-3-15 蓝靛厂软件研究室

作者简历

姓名：张华平　　性别：男　　出生日期：1978.2.1　　籍贯：江西鄱阳

教育经历：

2000.9 – 2005.7　　中国科学院计算技术研究所 计算机软件与理论 硕博连读

1995.9 –1999.7　　北方工业大学 计算机系计算机软件及应用专业 工学学士

【攻读博士学位期间出版的专（译）著】

[1]刘群,张华平,骆卫华,孙健，自然语言理解(译著), 电子工业出版社, 2005 年 1 月 (ISBN：7-121-00755-X)

【攻读博士学位期间获得的软件著作权】

[1]计算所汉语词法分析系统 ICTCLAS，软件登记号为 2003SR0087。

[2]汉语命名实体识别系统，软件登记号：2004SR00677

【攻读博士学位期间发表的论文】

[1] Hua-Ping ZHANG, Jian Sun, Bin WANG, Shuo BAI. Computation on Sentence Semantic Distance for Novelty Detection; Chinese Journal of Computer Science and Tech. vol.3, 2005

[2] Hua-Ping Zhang, Hong-Bo Xu, Shuo Bai, Bin Wang, Xue-Qi Cheng. Experiments in TREC 2004 Novelty Track at CAS-ICT. In Proc. of the 13th Text Retrieval Conference, Gaithersburg, Maryland, November, 2004, pp287

[3] 张华平,刘群. 基于角色标注的中国人名自动识别研究. 计算机学报，vol.27, No.1, 2004, pp.85-91

[4] Hua-Ping ZHANG, Qun LIU, Hong-Kui YU, Xue-Qi CHENG, Shuo BAI. Chinese Name 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. 29-602

[5] Hua-Ping ZHANG, Qun LIU, Xue-Qi CHENG, Hao Zhang, Hong-Kui Yu. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, Second SIGHAN workshop affiliated with 41st ACL; Sapporo Japan, July, 2003, pp. 63-70

[6] Hua-Ping ZHANG, Hong-Kui Yu, De-Yi Xiong, Qun LIU. HHMM-based Chinese Lexical Analyzer ICTCLAS, Second SIGHAN workshop affiliated with 41th ACL; Sapporo Japan, July, 2003, pp. 184-187

[7] Kevin Zhang (Hua-Ping Zhang), Qun Liu, Hao Zhang, Xueqi Cheng. Automatic

Recognition of Chinese Unknown Words Based on Role Tagging, First SIGHAN affiliated with 19th COLING, September 2002, pp71-77

[8] 张华平,刘群.基于 N-最短路径的中文词语粗分模型．中文信息学报．2002.9, Vol.16(5):pp.1-pp.7;

[9] 张华平,刘群. 基于角色标注的中国人名自动识别研究. 第七届中国科学院计算机专业研究生学术会议, 2002.7, 四川广元

[10] 刘群,张华平,俞鸿魁,程学旗. 基于层叠隐马模型的汉语词法分析; 计算机研究与发展, 41 卷, No.8, 2004, pp.1421-pp.1429

[11] Jian Sun, Wenfeng Pan, Hua-ping Zhang. TREC 2003 Novelty and Web Track at ICT, In Proc. of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, November 18-21, 2003, pp.138

[12] 俞鸿魁,张华平,刘群. 基于角色标注的中文机构名识别, Proc. of 20th International Conference on Computer Processing of Oriental Languages, 2003 年 8 月, pp79-87, 中国，沈阳

【近期投稿拟发表的论文】

[1] Hua-Ping Zhang, Xue-Qi Cheng, Hong-Bo Xu, Shuo Bai. Custmized Language Parsing for Novelty Detection. In Proc. of the 28th International ACM SIGIR Conference, Salvador , BRAZIL, October, 2005 (已投稿)

[2] 王小飞，张华平，王斌. 双数组 Trie 树的算法优化研究 (已投稿)

【攻读博士学位期间参加的科研项目】

[1] 国家重点基础研究项目:文本挖掘与知识检索(G1998030510),主要负责汉语的词法分析、命名实体识别、词性标注等，较好地完成任务，已经成功验收。

[2] 国家重点基础研究项目：基于 Internet 超大规模知识检索算法与应用(G1998030507-4),主要负责汉语的词法分析、命名实体识别、词性标注等，较好地完成任务，已经成功验收。

[3] 计算所领域前沿青年基金课题：基于多层隐马模型的中文词语一体化分析系统(20026180-23), 课题负责人, 2002 年 10 月~2003 年 10 月。

[4] 国家安全某话题发现与跟踪项目,参加了整体设计。主要负责汉语的词法分析,2003 年 1 月~2003 年 12 月。

[5] 国家安全某快速索引项目, 主要负责词典的管理与组织、快速的汉语词法分析,2003 年 1 月~2003 年 12 月。

[6] 参加第 13 届国际文本检索会议(Text Retrieval Text Retrieval Conference, TREC)。我独自一人全权负责 Novelty（新颖性检测）比赛全部的四个任务(2004 年 9 月)。综合排名在十四家国际知名研究机构中名列前茅。

[7] 文本检索会议(TREC12),参与了 Novelty 任务四的开发与研究。

[8] 参与设计并管理了中文自然语言开放平台(www.nlp.org.cn)，这是国际上最活跃、也是最受欢迎的中文自然语言处理学术与技术交流平台之一。

【攻读博士学位期间的获奖情况】

[1] 2004 年，获得中国科学院计算技术研究所所长特别奖

[2] 2004 年，获得中国科学院院长优秀奖

[3] 2003 年，获得中国科学院计算技术研究所软件研究室优秀学生奖

[4] 2002 年，获得中国科学院计算技术研究所所长三等奖

[5] 2001 年，获得中国科学院计算技术研究所软件研究室优秀新生奖

[6] 参加第 13 届国际文本检索会议(Text Retrieval Text Retrieval Conference, TREC)，该会议由美国国防部的高级研究发展署(Advanced Research and Development Activity，ARDA)与 NIST(国家标准技术研究所)共同主办。我独自一人全权负责 Novelty（新颖性检测）比赛全部的四个任务(2004 年 9 月)。综合排名在十四家国际知名研究机构中名列前茅。

[7] 2002 年，自主开发的计算所汉语词法分析系统 ICTCLAS 在国家 973 专家组评测中获得第一名，在 2003 年汉语特别兴趣研究组(the ACL Special Interest Group on Chinese Language Processing, SIGHAN)组织的第一届国际汉语分词大赛中综合得分获得两项第 1 名、一项第 2 名。同时，ICTCLAS 也是中文自然语言开放平台(www.nlp.org.cn)上最受欢迎的开放源代码项目，目前，我们已经向国内外的企业和学术机构颁发了 15,000 多份授权，我们提供的各种形式研究成果，在学术界和产业界得到了广泛的应用，其中包括：Yahoo、NEC 研究院、中华商务网、硅谷动力、云南日报等企业，新疆大学、清华大学、华南理工、麻省大学等研究机构。同时，ICTCLAS 广泛地被《科学时报》、《人民日报》海外版、《科技日报》等多家媒体报道。