分类号_____ 密级_____

UDC _____ 编号_____

中国科学院研究生院

博士学位论文

文本自动摘要方法研究

吴晓锋

指导教师 宗成庆 研究员 中国科学院自动化研究所

申请学位级别 工学博士 学科专业名称 模式识别与智能系统

论文提交日期 2010.10 论文答辩日期_____

培养单位 中国科学院自动化研究所

学位授予单位 中国科学院研究生院

答辩委员会主席_____

Research on

Automatic Document Summarization

Dissertation Submitted to

Institute of Automation, Chinese Academic of Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Engineering

by

Xiaofeng WU

(Pattern Recognition and Intelligent Systems)

Dissertation Supervisor: Professor Chengqing ZONG

独创性声明

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

签名：_____ 日　期：_____

关于论文使用授权的说明

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

（保密的论文在解密后应遵守此规定）

签名：_____导师签名：_____ 日　期：_____

摘要

自动文摘（Automatic Document Summarization, ADS）是自然语言处理领域的一个子领域。它是利用计算机系统自动生成文本摘要的应用技术，或者说是按读者（或用户）的要求以简洁的形式表达原文主要内容的技术。研究自动文摘技术的理论价值在于，一个完善的自动文摘系统几乎可以涵盖自然语言处理领域的方方面面，所以，该领域的研究对于整个自然语言处理的发展定能起到一定的推动作用。并且，这项研究也有着广泛的应用前景：在互联网技术高度发达的今天，自动文摘技术能够有效地帮助人们从检索到的文章中寻找自己感兴趣的内容，提高阅读速度和质量。

本论文主要工作和贡献归纳如下：

(1)在模型创建方面，本论文提出了一种基于序列分段模型（Sequence Segmentation Models，SSM）的有监督摘录型摘要提取方法。在这种方法里，摘要问题被看作“段标注”问题。与前人的工作相比较，SSM 方法的不同之处在于提取特征的单位不单来自句子，也可以来自于段。我们的 SSM 使用了可以对“段”建模并标注的半马尔可夫条件随机场（Semi-Markov Conditional Random Fields，SemiCRF）。实验表明，这种方法与单纯以句子为单位提取特征的摘要方法相比，有较明显的改善效果。

(2)在建模方面我们提出的另一种方法是采用排序学习方法（Learning to Rank，LTK）对通用型（generic）摘要问题建模。摘录型摘要的核心问题是给句子打分，打分的目的是为了后面的排序，并输出排名靠前的句子。而排序学习本质上就是为了解决排序问题，所以和摘录型摘要有很强的内在切合点。而且，采用排序学习建模更强调同一文本内的句子之间的相互比较，这和以往的建模方法有很大不同。我们将当前流行的几种排序学习算法在摘要问题上进行了比较，并第一次使用了逐列的排序学习方法。我们的实验证明，采用排序学习对通用型摘要建模是行之有效的，当采用 SVMMAP 这种逐列排序学习方法时，其总体效果还要优于以往建模方法。

(3)在特征提取方面，本论文提出了采用潜层狄利赫雷分配（Latent Dirichlet Allocation，LDA）来提取特征的方法。这种方法近年来被广泛应用于文本聚类、分类、段落切分等等，并且也有人将其应用于基于查询的无监督的多文档自动摘要。

该方法被认为能较好地对文本进行潜层语义建模。本论文在前人工作基础上，研究了 LDA 在有监督的自动文摘中的作用，提出了将 LDA 提取的主题（Topic）作为特征加入有监督模型中进行训练的方法，并分析研究了在不同 Topic 下 LDA 对摘要结果的影响。实验结果表明，加入 LDA 特征后，能够有效地提高以传统特征为输入的文摘系统的质量。

(4)在多文档摘要中，冗余句的识别和剔除是一个至关重要的问题。无论是采用摘录型摘要方法还是理解式摘要方法，这都是一个不可回避的问题。针对这个问题本论文着重研究了复述（Paraphrase）句的识别问题。传统的解决复述句识别方法是通过词频或句法上的相似度来判断的。可是哪怕用相同的文字书写的句子其含义也可能差别很大，而相同句法结构也不能保证意义一致。本文根据新闻语料的特点，提出了一种通过引入深层的语义角色标注来帮助识别新闻领域复述句的方法。该方法通过在语义角色这种结构化的含义表达形式中提取的特征来弥补传统方法的不足：先识别待判断的两个句子中所有谓词的语义角色，然后计算两个句子间对应语义角色的相似度，最后结合传统的句子相似度计算方法来进行相似性计算。实验证明，本文提出的方法能有效地提高复述语句的识别效果。

关键词：自动文摘，半条件随机场，排序学习，潜层狄利赫雷分配，复述句识别

Abstract

Automatic Document Summarization (ADS) is one of the subfield of Natural Language Processing (NLP). It can be defined as a technology which is to summarize documents with the help of computer, or to represent the original documents with short but comprehensive texts according to the demands of customers. The research of ADS is of both theoretical and applicational values: A complete ADS system can almost cover all of the NLP subfields, so, the research upon it must be a boost to the development of NLP; the applicational value is that with the advent of the information era and the development of the internet, ADS can efficiently speed up people"s reading.

In this thesis, therefore, we make an intensive study on the ADS" modeling and feature extraction. The main work and contributions are summarized as follows:

(1) We propose and implemented a new model which is based on Semi-Markov Conditional Random Fields (SemiCRF) and does ADS using Sequence Segmentation Models (SSM). Compared with existing approaches, SemiCRF can utilize features extracted from segments as well as sentences. According to our experiments, this new approach outperforms all exisiting approaches which only extract features from sentences.

(2) The other new approach we propose is based on Learnint to Rank (LTK). Because the core idea of extractive summarization methods is to rank sentences according some criteria, so it is quite natural to consider LTK in automatic summarization. We tested some of the most famous LTK approaches both pair-wised and list-wised. Our experiment results show that LTK could work well in generic extractive summarization, and one of the list-wised approaches SVMMAP could even outperform the best known result using the same feature space.

(3)We also extract a new feature for automatic summarization which based upon Latent Dirichlet Allocation (LDA). LDA has been a very hot point in recent years due to its sound theoretical foundation and good flexibility. Some of researches have shaped it for query-based summarization. Yet there has been no study of the influence of the num-

ber of topics to the summarization results and no one has used the original LDA to be a feature for doing automatic summarization. Our experiments are aiming to solve these two problems. Our results show that LDA is quite capable as an automatic summarization feature.

(4)We have also studied the problem of Paraphrase Recognition as an approach to solve the redundant sentence recognition problem. Our approach gives a thurough consideration of semantic roles of each word in sentence by utilizing UIUC"s SRL tools. This is quite different with former works upon paraphrase recognition that utilize only shallow informations of sentences. We extract features upon the labeled semantic roles of sentence and combined with other features which produced using former approaches. Our experiments show that this new approach could outperform all of the known results which performed on the same corpus.

Key words: automatic summarization, SemiCRF, Learning to Rank, Latent Dirichlet Allocation, paraphrase

目录

第一章 绪 论

1.1 引言

互联网技术的出现极大地丰富和便利了我们的生活，让我们可以方便地获取更多的信息，但是也带来了诸多问题。例如，如何在浩如烟海的信息里迅速地找到对特定用户有价值或感兴趣的信息？计算机擅长处理结构化的数据，而人类的语言本质上是开放的、发散的，而且在不断地演化，这也是人类富于创造性的一个重要表现。搜索引擎如今是我们主要借助的工具，可是目前的搜索引擎其实是用一种“偷懒”的方式答复用户提出的问题，用户依旧要在几十或者上百个搜索结果页面上去寻找自己需要的信息。

在这种情况下，对自动摘要技术的研究就显得尤为必要和迫切。

自动文摘（Automatic Summarization）是利用计算机自动编写文摘的应用技术[刘挺, 1999]，就是按读者（或用户）的要求以简洁的形式表达原文的主要内容。

一般认为，摘要应该具备的三个特点是：

1、文摘来源于单文本或多文本

2、文摘应该包含有重要信息

3、文摘应该比原文短小

进一步说，摘要技术是在文件检索的基础上给读者反馈一段或几段短小精炼的文本，该文本或者一般性地反映出原文章或几篇文章的主题意思，或者根据用户需要提供用户感兴趣的某些领域的段落大意。

摘要和信息抽取技术有一定的相似性，都是要帮助用户加速阅读文本检索后的内容，但是又有很大不同：信息抽取指的是从文本中抽取指定类型的实体、关系、时间等信息，并形成结构化数据输出[Grishman, 1997]，例如，针对恐怖事件抽取时间、地点、伤亡情况、使用武器、恐怖分子、受害对象等等；而摘要是提供文章主题信息，帮助用户判断一片文章是否具有价值。基于查询的摘要系统与自动问答系统[吴有政, 2006]的区别在于摘要并不要求寻找精确的答案，而是只给出和查询语句相关的句子，由用户来寻找答案，这可以看作是自动问答系统在可以预见的一段时

间里都很难做到精确、准确的一个折衷。自动文摘系统是搜索引擎有力的辅助工具，也是进行信息抽取和自动问答的有效准备。一个好的基于互联网的自动摘要系统的流程如图 1-1 所示。

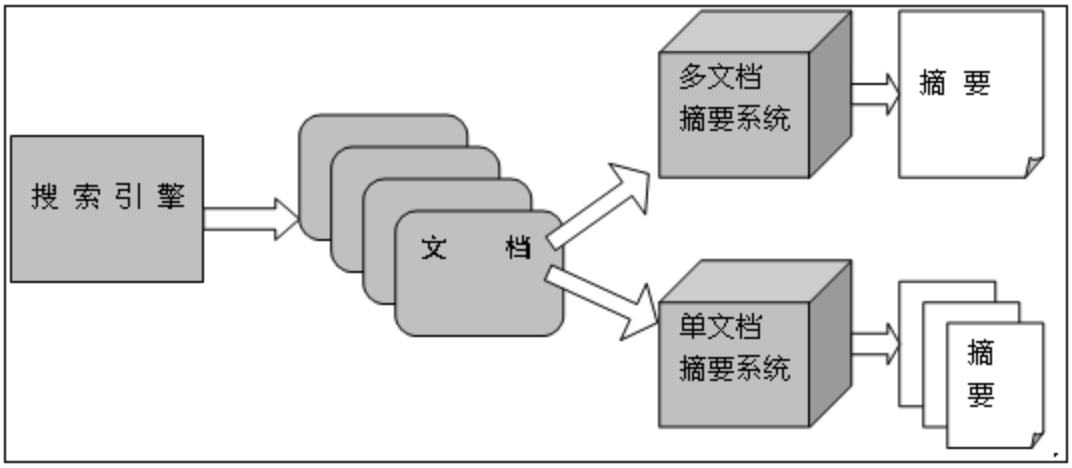


图1-1 自动摘要系统示意图

自然语言处理中的各个问题：从形态分析（汉语犹指分词）、命名实体识别、到句法、到语义角色标注、再到机器翻译、语言理解等，以及信息检索中的各项技术：从文本检索、指代消解、文本分类到问答系统、信息抽取以及自动摘要等，每项技术的进步与否都与相关领域和技术休戚相关。而一个好的摘要系统的搭建理论上应该涵盖从自然语言处理到文本检索中的几乎所有子领域的核心技术。目前来看，虽然全球许多知名研究机构都在自动文摘这一领域投入精力，但是，还没有真正实现能达到实用要求的摘要系统。这一领域的继续探索依旧吸引着众多科研人员，这一领域的发展也势必能进一步推动计算语言学的向前迈进。

1.2 本论文研究的主要问题和贡献

自动摘要是一个综合性很强的技术，合理地统计建模、有效地特征提取、句子压缩、篇章结构分析都非常重要。另外，对多文档摘要来说，句子聚类、复述句识别、句子生成、指代消解、排序方法等等，也都会对摘要系统的效果产生影响。上述每个问题的研究都有重要的意义，下面将自动摘要所涉及的各种技术以及本论文着重研究的问题及其作用列于表 1-1：

表1-1 与自动摘要相关的问题

重要技术	单文档		多文档	
	摘录型	理解式	摘录型	理解式
建模	✓			
特征提取	✓			
实体识别	✓			
篇章结构分析	✓			
指代消解	✓			
句子生成/压缩		✓		✓
句子聚类			✓	✓
复述句识别			✓	✓
句子排序			✓	✓

在表 1-1 所列出的摘要相关问题中，我们选择了单文档的建模、特征提取、以及复述识别问题。做出这种选择的原因如下：

- ✧ 建模和特征提取问题对摘要生成至关重要，而且针对性强，因此，本文对这一问题进行了深入研究。虽然本文只研究单文档摘要的建模和特征提取问题，但是，这些方法能够比较容易地扩展到多文档摘要中去。
- ✧ 复述句识别在多文档摘要中极为重要也比较困难，直接关系到对于冗余信息的处理。所以，我们选取这个问题加以研究，作为搭建多文档摘要系统的基础。

本论文的主要贡献在于：

（1）在模型创建方面，本论文提出了两种方法。一种是基于序列分段模型（Sequence Segmentation Models，SSM）的有监督摘录型摘要提取方法。在这种方法里，将摘要问题看作“段标注”问题。和前人的工作比较，SSM 方法的不同之处

在于提取特征的单位不单来自句子，也可以来自于段。我们的 SSM 使用了可以对“段”建模并标注的半马尔可夫条件随机场（semi-Markov Conditional Random Fields，SMCRF）。我们把结果和过去的方法做了对比，证明这种方法和比单纯针对句子为单位提取特征的方法有较明显的效果。

(2)在建模方面我们提出的另一种方法是采用排序学习方法（Learning to Rank，LTK）对通用型（generic）摘要问题建模。摘录型摘要的核心问题是给句子打分，打分的目的是为了后面的排序，并输出排名靠前的句子。而排序学习本质上就是为了解决排序，所以和摘录型摘要有很强的内在切合点。而且，采用排序学习建模更强调同一文本内的句子之间的相互比较，这和以往的建模方法有很大不同。我们将当前流行的几种排序学习算法在摘要问题上进行了比较，并第一次使用了逐列的排序学习方法。我们的实验证明，采用排序学习对通用型摘要建模是行之有效的，当采用 SVMMAP 这种逐列排序学习方法时，其总体效果还要优于以往建模方法。

(3)在特征提取方面，本论文提出了采用潜层狄利赫雷分配（Latent Dirichlet Allocation，LDA）来提取特征的方法。这种方法近年来被广泛应用于文本聚类、分类、段落切分等等，并且也有人将其应用于基于查询的无监督的多文档自动摘要。该方法被认为能较好地对文本进行潜层语义建模。本论文在前人工作基础上，研究了 LDA 在有监督的自动文摘中的作用，提出了将 LDA 提取的主题（Topic）作为特征加入有监督模型中进行训练的方法，并分析研究了在不同 Topic 下 LDA 对摘要结果的影响。实验结果表明，加入 LDA 特征后，能够有效地提高以传统特征为输入的文摘系统的质量。

（4）在多文档摘要中，冗余句的识别和剔除是一个至关重要的工作。无论是采用摘录型摘要方法还是理解式摘要方法，这都是一个不可回避的问题。来自不同新闻机构的两句话，是否是说的同一件事情，同一个问题，如何判断？ 针对这个问题本论文着重研究了复述（Paraphrase）句的识别问题。传统的解决复述句识别方法是通过词频或句法上的相似度来判断的。可是哪怕用相同的文字书写的句子其含义也可能差别很大，而相同句法结构也不能保证意义一致。本文根据新闻语料的特点，提出了一种通过引入深层的语义角色标注来帮助识别新闻领域复述句的方法。该方法通过在语义角色这种结构化的含义表达形式中提取的特征来弥补传统方法的不足：先识别待判断的两个句子中所有谓词的语义角色，然后计算两个句子间对应语义角色的相似度，最后结合传统的句子相似度计算方法来进行相似性计算。实验证明，本文提出的方法能有效地提高复述语句的识别效果。

1.3 论文其余章节的组织

论文第二章为自动文摘技术的综述，主要介绍单文档和多文档摘要的发展历程、自动摘要技术分类方法、评测方法、典型系统介绍以及机器学习在各种方法中的应用。

第三章简要介绍本论文依托的几个主要模型和算法，包括隐马尔可夫模型、条件随机场模型和支持向量机。

第四章至第七章是本论文的核心部分，详细介绍本人的研究工作和各创新点。其中，第四章介绍采用半条件随机场对摘要问题建模的方法；第五章介绍采用排序学习对摘要的建模方法；第六章介绍新特征的获取方法，研究采用潜层狄利赫雷分配提取潜层主题作为特征对摘要问题的影响；第七章介绍在复述句识别方面的研究工作。

最后一章为本论文的结束语，总结并展望下一步的工作。

第二章 自动文摘技术综述

2.1 引言

对自动摘要技术的研究工作最早见于H. P. Luhn在1958年发表的一篇题为“The Automatic Creation of Literature Abstracts (Auto-abustracts)”的论文[Luhn, 1959]。此后的四十多年中，Baxendale、Oswald、Edmundson、Wyllys、Earl、IBM 公司以及俄亥俄州立大学等相继进行了自动文摘技术的研究，该技术得到了不断的发展与完善。1993年12月，在德国Wadern召开了历史上第一次以自动文摘为主题的国际研讨会。1995年，Information Processing & Management组织了一期标题为Summaring Text的专刊，标志着自动文摘时代的到来，自动文摘技术的研究进入了空前的繁荣期。在ARDA(Advanced Reaserch and Development Activity)资助下，NIST于2000开始举办的文本理解会议(DUC, Document Understanding Conference)，是对文本摘要技术最有力的推动，该会议每年举办一次，吸引了大批研究者和研究机构。该会议于2008年并入文本分析会议(TAC, Text Analysis Conference)。另外，(TREC, Text Retrieval Conference)和(MUC, Message Understanding Conference)也举办了一些公开评测。

2.2 自动文摘技术的分类

自动文摘技术可按不同的标准分为不同的类型。

根据文摘的功能可以分为指示型文摘（indicative）、报道型文摘（informative）和评论型文摘（evaluative）[Mani, 1999b]。根据输入文本的数量，自动文摘技术可以分为单文档文摘（sigle-document summarization）和多文档文摘(multi-document summarization)。根据原文语言种类自动文摘可以分单语言文摘（monolingual）和跨语言文摘（cross-lingual）。根据文摘和原文的关系又可以分为摘录型文摘（extract）和理解型文摘（abstract），前者是由从原文中抽取的片段组成，后者则是对原文主要内容重新组织形成的。根据文摘的不同的应用，可以分为通用型文摘（generic）和面向用户查询的文摘（query-oriented），前者提供原文作者的主要观点，后者则反映用户感兴趣的内容[Hovy, 1999]。

另一个有趣的分类法来自[Mani, 1999b]，它将摘要方法分为表层（surface）、实体（entity）以及篇章（discourse）三个层次。表层方法主要指利用词频、句子位置、以及线索词等信息的方法。实体方法主要指通过实体以及其关系——包括共现（co-occurrence）、共指（co-reference）等，并通过定义在实体上的模型来寻找重要信息的方法。篇章方法则是对文本格式、修辞结构等等进行建模。

表 2-1 列出了各种分类方法的分类依据以及相应类别。

表2-1 摘要分类法

分类依据	类别 1	类别 2	类别 3
输入文本数量	单文档	多文档	/
语言	单语言	跨语言	/
文摘和原文关系	摘录型	理解式	/
功能	指示型	报道型	/
应用	通用型	面向查询	/
依赖信息	表层	实体层	篇章结构

下面分别从单文档、多文档以及统计学习算法在摘要中的应用这条主线介绍文摘技术的发展历程。

2.3 单文档摘要的发展

虽然摘录型文摘存在呆板生硬的问题，但是由于大部分自然语言处理技术还处在雏形阶段，所以大多数研究者还是对理解型文档摘要持保留态度。简言之，摘录型文摘就是直接从文章中抽取句子，因为基本不牵扯对冗余句子的判断，它在单文档文摘中的应用更是顺理成章。既然文档摘要的主要任务是找出并抽出文本中涵盖重要意义的部分，那么如何界定这个“重要”，或者说如何找到反映这个“重要”的特征，就基本主导了自动文摘的发展方向。另外，机器学习的兴起也大大带动了文档摘要技术的发展，人们在机器学习方法的推动下，致力于寻找更有效更能反映摘

要问题本质的方法。下面从特征选取和算法这两个方面简要介绍单文档摘要技术的发展，这些方法在多文档摘要中也能得到应用。

2.3.1 特征

Luhn 在上世纪 50 年代就开始了自动文摘技术的研究工作，可谓文档摘要技术的鼻祖。早期的单文档摘要主要面向技术文献。他的工作基本可以归纳为：取词干、去停用词，然后计算词频，并用词频来作为“重要性”的度量。Luhn 的工作虽然简单，但是他所采用的方法为摘要研究奠定了基调，后面的研究者在摘录型摘要上基本都遵循这个思维方式。

同一时期，与 Luhn 同在 IBM 公司的 Baxendale 对 200 段文字进行了研究，他发现 85%的段落其第一句话为重要信息，而 7%的段落的重要信息在最后一句。这样，人的书写习惯，这个看似简单却容易被忽视的特征也被引入到自动文档摘中来，并且其重要性在自动文摘后来的发展中一直无法被忽视。

另外两个经典的特征：线索词（cue words，如文档中出现“强调”、“突出”等字眼）和骨架（skeleton，也就是一句话是否是标题）被 Edmundson 于 1969 年引入。作者用这两个特征结合句子位置和词频来为句子打分。

词频可以在一定程度上反映句子的重要性，但是从信息论的角度看，越是小概率事件，其信息量或者说重要性越大，那么，单纯用词频的方法来反映句子的重要程度就显得不合适了。广泛应用于信息检索的 TF×IDF[Frakes, 1992]是这两种观点的折中，所以它一出现很快就被引入自动文摘中。

Gong[Gong, 2001]引入了一种计算相似性的办法。作者认为，基于潜层语义分析（LSA, Latent Semantic Analysis）可以得到文章的隐含主题（topic），这样句子的相似性以及句子的重要性可以通过在这个主题上的投影计算得到。Gong 的这种通过奇异值分解来获得隐含主题的方法具有很大的开创意义，不少人在他的影响下开始寻找有趣的数学模型来探索文本中词与词、句与句以及文本间的关系。

还有一些其它特征如对数似然相似度、指示词、大小写和句子长度也被引入到摘要里。同一个特征在不同作者的方法中经常也会有一些差别。这些林林总总的特征一般会结合各种各样的机器学习算法来帮助摘要系统提升性能。

2.3.2 算法

本节将自动摘要算法主要分为有监督和无监督两大类，另外，有些也部分地采用启发式的方法。我们主要就这两类算法在摘要中的应用进行详细介绍。

1）有监督方法

（1）朴素贝叶斯分类器

用朴素贝叶斯分类器来进行自动文摘的方法是由[Kupiec, 1995]提出的。作者假设各个特征相互独立，用贝叶斯公式（2-1）来得到一个句子是否为摘要的概率：

$$P(s \in S \mid f_1, f_2, ..., f_N) = \frac{\prod_{i=1}^{N} P(f_i \mid s \in S) \cdot P(s \in S)}{\prod_{i=1}^{N} P(f_i)} \tag{2-1}$$

其中，s 代表句子，S 代表摘要集合，f 为特征（共有 N 个），句子的概率作为分值，排序后，输出需要的摘要长度。

Kupiec 采用的特征包括：句子长度、固定词语搭配（作者给出了 26 个，如 in summary, in conclusion 等固定词语搭配，这些搭配更倾向于包含进摘要）、段落特征（作者将文章分为开头、结尾、中间三部分）、专题词（thematic words，一篇文章中使用较为频繁的词被视为专题词，句子中专题词的多少也被看为一个特征）以及大写词语特征（因为专有名词一般首字母大写，且一般来说含有专有名词的句子会包含比较重要的信息）。

最后，Kupiec 在他的实验里用人工打分的方法进行评测。

[Aone, 1999]的系统 DimSum 也采用了贝叶斯模型，Aone 引入了 TF×IDF 作为特征。该系统做了命名实体识别、词语搭配识别、简单的指称统一（如将 U.S. 统一表示成 United States）以及用 WordNet 来识别不同的词形变换和同义词。

贝叶斯模型简单有效，但其模型固有的缺陷：特征独立性假设，限制了它在复杂问题上的应用。

（2）决策树方法

针对贝叶斯模型的独立性假设缺陷，[Lin, 1999]采用了决策树来尝试解决基于查询语句的摘要生成。Lin 采用了更多的特征，并研究了它们对于抽取句子效果的影响。比较新颖的特征有：IR signature（语料中排名靠前的最重要的词）、Query signature（一个句子里含有多少个查询语句中的词汇）、数字日期以及是否含有代词

和形容词。

Lin 选用的语料来自 TIPPSTER-SUMMAC[Sundheim, 1998]，系统为人工评测和匹配评测。实验结果表明，决策树在很多语料里都比贝叶斯模型效果好。这也从一方面说明了决策树方法可以一定程度的克服贝叶斯模型过强的独立性假设，并在一定程度上揭示了特征间潜在的联系。

（3）最大熵模型

最大熵模型也能在一定程度上克服贝叶斯模型独立性约束过强的问题，[Osborne, 2002]将这种模型引入摘要中，并只规定了两种标注：摘要或非摘要。其公式如（2-2）

$$P(c \mid s) = \frac{1}{Z(s)} \exp\left(\sum_i \lambda_i f_i(c, s) \right) \tag{2-2}$$

其中，c 代表标注类型，s 为句子，$Z(s)$ 是归一化因子，f 为特征，λ 为特征权重。作者在以贝叶斯作为基线的模型和最大熵模型中都加入一个先验值，这种方法克服了返回句子过多的问题，

Osborne 在文中将最大熵模型和贝叶斯模型在同样的特征下做了比较，肯定了最大熵分类器的效果。

但是，这三种方法都只是在单独句子上进行建模、分类。然而，任何句子都不是孤立的，句子间必然存在某种联系，如能将这种联系加入到模型中去，直觉上看，对摘要问题是非常合适的。

（4）隐马尔可夫模型

[Jing, 1999]采用隐马尔可夫模型来分解人工写成的摘要，也就是把人工撰写的摘要的细节成分投射回原文。在 Jing 的启发下[Conroy, 2001]尝试采用隐马模型来自动生成摘要，第一次将摘要问题作为序列标注问题进行研究。其原理简言之，就是认为当前一句话是否为摘要的概率不但与该句的特征有关，还与上一句话是否是摘要有关。

Conroy[Conroy, 2001]假设了 s 种摘要状态和 $s+1$ 种非摘要状态，仅在摘要状态允许状态跳跃，且仅在非摘要状态允许状态保持。用这个模型计算每句话属于各个状态 i 的概率，如果 i 是偶数，则说明这个是第 $i/2$ 句摘要，反之则说明这句话不是摘要。该模型仅仅用了三个特征：句子位置、句子长度、句子及文本的相似度。作

者用 TREC 语料对模型进行训练，得到转移概率以及每个状态下产生不同特征的概率。Conroy 的实验采用人工评测。

隐马尔可夫模型把句子和句子建立起了简单的联系，但它和贝叶斯模型一样，也存在特征独立性假设这个缺陷。

（5）条件随机场模型

能够既考虑序列模型问题，又能降低独立性假设，并且没有最大熵马尔可夫（MEMM, Maximum Entropy Markov Model）的标注偏置问题的条件随机场模型[Lafferty, 2001]近年来在 POS 标注、命名实体识别、语块识别、句法分析等许多自然语言处理领域获得成功。[Shen, 2007]参照最大熵模型的方法，也将摘要考虑成 0-1 标注问题，并采用条件随机场模型进行标注。Shen 的方法是在整篇文章上考虑句子序列，也就是序列标注问题，使得标注序列在整个文章上的概率最大。其公式如下（2-3）。

$$P(Y \mid X) = \frac{1}{Z_X} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, X) + \sum_{i,l} \mu_l g_l(y_i, X) \right) \qquad (2\text{-}3)$$

其中，$X = (x_1, x_2, ..., x_3)$ 为句子序列，y 为标注，f 是考虑当前句子和上一句子的标注情况下产生的特征，g 是仅仅考虑当前句子情况下产生的特征，μ 和 λ 为特征权重。作者认为用条件随机场来处理摘要问题的优势是可以充分考虑前面句子对当前句子的影响，并且没有特征的独立性假设的约束。作者的实验使用的是 DUC01 单文档摘要语料，在与 SVM、最大熵、HMM 及一些非监督方法的比较上，条件随机场取得了最佳的结果。

条件随机场模型很好地体现了序列标注模型的各种优点，并且当特征数及类别数不多的情况下速度较快。但序列标注模型都是以句子为单位，对有些以词汇为基础的特征来说，这样的粒度明显偏小。

（6）遗传算法

[Yeh, 2005]介绍了一种利用遗传算法的可训练摘要算法，采用的特征有：句子位置、句子和文章题目的相似度的特征等，并且在他的方法里对句子位置这个特征做了进一步细分：句子位置特征根据在文章中的不同位置给予不同的权重。Yeh 用台湾新闻周刊作为语料测试了该方法。[Svore, 2007] 也采用了遗传算法 RankNet[Burges, 2005]对句子进行打分，该方法还加入了一些从外部新闻网站获取

的知识来帮助提高性能。

还有一些有监督方法，如[Goldstein, 1999]单纯采用线性组合特征训练参数的方法，[Ishikawa, 2002; Fuentes, 2007]采用 SVM 算法，[Fisher, 2006]采用感知机算法等。

以上介绍了摘要问题中使用的一些有监督的机器学习方法。当文本的写作风格、使用词汇等变化不太大的时候，一般来说有监督方法能够取得比无监督方法好的效果。但是有监督方法也迫切面临着语料匮乏的问题。语料匮乏在摘要问题上非常突出，上述很多实验都是在规模有限的语料上实现的。

2）无监督方法

无监督方法更倾向于寻找单词分布的隐含规律和文章的结构信息，如修辞结构（Rhetorical Structure）、词链（Lexical Chain）、基于图的句子相似性计算、文本的隐含主题（Latent Topic）等。尽管不借助于训练语料，很多无监督方法也取得了不错的成绩。下面介绍几种经典的无监督摘要方法。

（1）基于词链

词链[Morris, 1991]代表的是词汇黏着（Lexical Cohesion）[Hoey, 1991]，可以看成是文本中一连串语义相关的词，并且这些词在文本中可以不连贯。词汇黏着可以分为重复（reiteration）和共现（collocation）两类：重复包括单词的简单重复、同义词重复和下义词；共现指经常出现在同一句子中的词，比如“老师”和“学校”。

[Barzilay, 1997]采用重复黏着的词链做为判断一个句子重要性的标准，文中使用了 WordNet 作为计算词链的辅助工具。摘要的生成分为三步：首先对文本按主题做段落切分；然后创建词链；最后通过词链判断句子的重要性并抽取。Barzilay 的工作特点是仅仅采用了词链作为特征，摒弃了其它的特征，并且取得了不错的结果。[Silber, 2002]采用了线性时间复杂度的计算词链的方法用于摘要提取。[Bergler, 2003]对文章做了共指消解，用名词短语的共指消解链作为计算句子重要性的尺度。

采用词链作为摘要问题的主要特征的合理性在于：词链是分析篇章语义的一个重要替代工具，它通过潜层词汇间的上下文关系一定程度上反映出文本，乃至语义的发展趋势。这在篇章分析远处于雏形的现阶段，词链的意义固然非常重大。然而采用词链的方式面临的主要困难是词链计算的正确性很难得到保证。

（2）基于潜层语义分析

潜层语义分析（LSA，Latent Semantic Analysis）[Deerwester, 1990]是信息检索

中使用的一种经典方法，这种方法巧妙地用 SVD 分解这种常用的数学工具来将词汇形成某种程度的聚类，其所说的“语义”其实只是用奇异值来描述词与词之间潜在的共现关系。[Gong, 2001]将这种方法应用于自动摘要中，其原理如下式（2-4）：

$$A = U\Sigma V^T \tag{2-4}$$

在 $A=[A_1, A_2, ..., A_N]$ 中，每个列向量 A_i 代表第 i 句话中每个词的词频，对其按(2-4)式做奇异值分解后 $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_N)$，并满足 $\sigma_1 \geq \sigma_2 ... \geq \sigma_r = \sigma_{r+1} ... = \sigma_N = 0$。因为 A 的稀疏性，这里的 r 远远小于 N。通过奇异值分解，将 A 中的每个代表句子词频的列向量 A_i 投影到 V^T 的列向量 $\psi_i = [v_{i1}, v_{i2}, ..., v_{ir}]^T$。从语义学的角度[Deerwester, 1990]讲，$v_{i1}, v_{i2}, ..., v_{ir}$ 代表的就是各个潜层 topic：$\sigma_1, ..., \sigma_r$ 在句子 i 里的权重，这二者的内积就反映了该句子在这篇文章中的重要性。

Gong 的摘要生成系统流程如下：

a) 将文本切分成句；

b) 创建矩阵 A；

c) 对 A 做奇异值分解；

d) 选出 V^T 中的前 k 个索引最大的列向量生成摘要；

潜层语义分析的方法在数学上看非常简单直观，但是其所定义的“语义”非常缺乏合理的物理解释。

（3）基于潜层狄利赫雷分布（LDA, Latent Direchlet Allocation）

LDA 是近些年来比较流行的一种分析潜层主题的方法[Blei, 2003]，它本质上是一种贝叶斯图模型。LDA 在图像检索、文本段落切分等很多领域都有很重要的应用。它的思想脱胎于 LSA 以及其后续的发展，pLSA[Hofmann, 1999]，也是通过无监督的方法获取词汇间的共现规律，但它的数学模型更加完备，具体的 LDA 算法见第三章。

[Daume III, 2006]将 LDA 模型针对基于查询的摘要做了变型。传统的 LDA 模型应该以文章为单位，但是 Daume 在文章中将其缩小到句子级。其产生过程如下图 2-1 所示，其算法在图 2-2 给出。

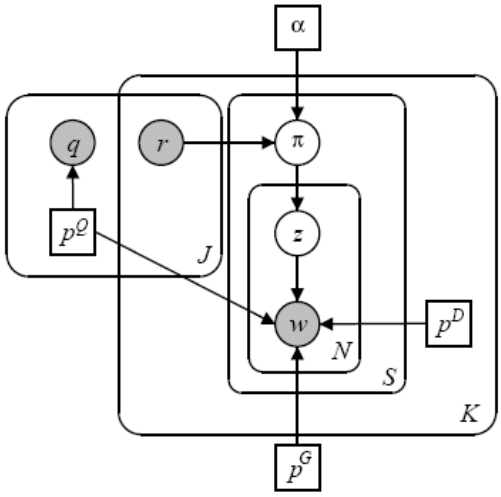


图2-1 Blei 的基于查询的贝叶斯图模型

算法：

1、每个查询语句 $q_{j=1...J}$ 中的单词 q_{jn} 由模型 $p^{q_j}(q_{jn})$ 生成

2、对每个文档 $k=1...K$，以及文档中的句子 s

a) 通过 $Dir(\pi_{ks}|\alpha)$ 得到这个句子被产生的概率 π_{ks}

b) 对于句子 s 中的每个词 w_{ksn}

用 $Mult(z|\pi_{ks})$ 为其选择一个类别 z_{ksn}

最后根据不同的类别 z_{ksn} 由下式产生出单词 w_{ksn}

$$\begin{cases} P^G(w_{ksn}) \ \ if \ \ z_{ksn}=0 \\ P^{d_k}(w_{ksn}) \ \ if \ \ z_{ksn}=k+1 \\ P^{q_j}(w_{ksn}) \ \ if \ \ z_{ksn}=j+K+1 \end{cases}$$

图2-2 Blei 的算法

算法中的 $Dir(\pi_{ks}|\alpha)$ 表示以 α 为参数的狄利赫雷（见第六章）分布。$Mult(z|\pi_{ks})$ 表示以 π_{ks} 为参数的多项式分布。

图 2-1 中的实心圆代表可见变量；空心圆为隐含变量；圆角方框代表概率独立的一组变量，变量数目标注在右下角；正方形方框代表 4 个已知参数。这个模型的联合概率需要用马尔可夫蒙特卡罗方法或者 EM 方法来计算[Bishop, 2006]。

Daume 对查询语句做了词汇扩展，选用 Duc02 的语料用来验证这种方法。Daume 声称该系统的结果在 Duc02 的参赛系统中名列前茅，但是他对抽取到的句子采用了

压缩处理，这样的处理是否会有助于系统的性能提高，提高了多少？这些并没有在他的文章里反映出来。

LDA 模型有着明确的物理含义，其定义的“语义”概念比 LSA 模型清晰直观。它的主要缺点是计算复杂，不论通过马尔可夫蒙特卡洛方法，或者通过求近似解的期望传播方法，它们的计算量都是非常庞大的。

（4） 基于图的句子相似性

受谷歌（Google）的 PageRank 算法[Brin, 1998]以及[Kleinberg, 1999]的 HIST 算法的启发，[Mihalcea, 2005]提出了在基于图的框架下用非监督的方式计算句子的权值的方法：将文档中的句子看作节点，句子间的相似度为节点间的边，然后采用 PageRank 和 HIST 的方法计算每个节点的权重。PageRank 的计算公式见公式（2-5）。

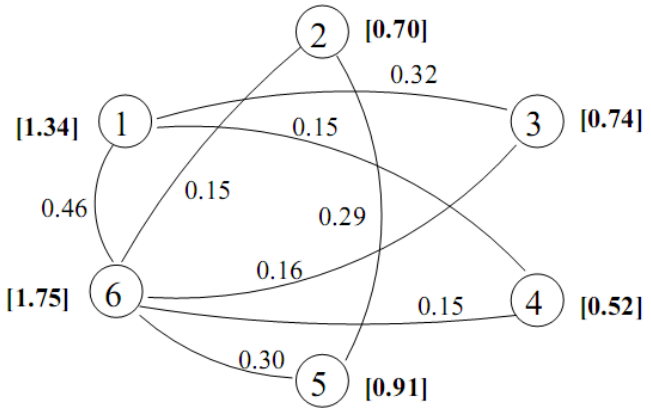
$$PR(V_i) = (1-d) + d \bullet \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \tag{2-5}$$

公式中的 $PR(V_i)$ 表示从外部节点上根据 PageRank 方法给当前节点 V_i 计算得到的分值。$Out(V_j)$ 表示节点 V_j 所拥有的外部节点数量。

PageRank 采用的是有向图，当用作句子的权值计算时则采用无向图。Mihalcea 的方法里句子相似度采用单纯的句子间的词汇交迭来计算。图 2-3 给出了一个简单文本的示例。

[1] Watching the new movie, “Imagine: John Lennon,” was very painful for the late Beatle’s wife, Yoko Ono.
[2] “The only reason why I did watch it to the end is because I’m responsible for it, even though somebody else made it,” she said.
[3] Cassettes, film footage and other elements of the acclaimed movie were collected by Ono.
[4] She also took cassettes of interviews by Lennon, which were edited in such a way that he narrates the picture.
[5] Andrew Solt (“This Is Elvis”) directed, Solt and David L. Wolper produced and Solt and Sam Egan wrote it.
[6] “I think this is really the definitive documentary of John Lennon’s life,” Ono said in an interview.

(a)



(b)

图2-3 Mihalcea 的基于图算法的摘要生成方法

图 2-3(a)中给出了 6 个句子，图 2-3(b)中给出句子间的相似度，每个节点旁边的分值为最终由 PageRank 算法计算得到的句子分值。评测实验的语料采用的是 DUC02，评测方法为 ROUGE。

Mihalcea 的方法的巧妙之处是通过有效的 PageRank 算法找到所有句子相互间的联系（这里是通过相似度反映出的一种联系），并通过这种联系定义出了句子的重要性。但是摘要问题所需要的“重要性”，很多时候不能完全由句子间的相似性反映出来，而且句子级别的“相似”也存在比较严重的稀疏问题。

归纳起来说，采用无监督方法的优点有：①不需要语料标注；②能够方便地对不同领域的文本进行处理，比如不同的写作风格，可能导致句子位置这个在有监督方法中的重要特征失效，必须要重新训练模型才行，而这一点在无监督方法中是不需要考虑的；③无监督方法往往能找到文章中所有句子之间的相互关系，并借此推导出句子的重要性，这在有监督方法里是非常难做到的。但无监督方法的缺点也很明显：比如，要么过分依赖词与词间的关系而忽略句子间的空间关系，要么呆板地依赖事先制定的启发式规则而没有变通等等，而这恰恰是有监督摘要方法的优势所在。

句子抽取是单文档文摘所采用的主要方式，虽然也有少数方法采用了基于理解的摘要方式，但基于理解的文摘系统大多见于多文档摘要中。我们将在多文档摘要方法中介绍基于理解的方法。

2.4 多文档摘要的发展

在使用搜索引擎的时候，同一主题往往会返回成千上万的网页，将这些网页形成一个统一的、精练的、能反映主要信息的摘要是非常有意义的。另外，互联网上某一新闻单位针对同一事件的系列报道，或者对某一事件数家新闻单位同一时间的报道，若能从这些有很强相关性的文章中提炼出一个覆盖性强并且简洁的摘要也是很有意义的。这两种情况是当前多文档文摘的两种典型应用。所谓多文档文摘就是将同一主题下的多个文本描述的主要信息按压缩比提炼出一个文本的自然语言处理技术[Radev, 2002]。

最早的多文档摘要方法是哥伦比亚大学的 Mckeown 和 Redev 开发的 SUMMONS[McKeown, 1995]，这是一个基于模板的特定领域的信息理解系统，它更像是信息抽取而不是文摘。

单文档文摘所采用的方法在多文档文摘中都能找到应用，所以多文档也可以按照单文档的分类方法进行分类，比如基于理解的摘要和抽取的摘要、面向查询的摘要和通用型的摘要等等。但和单文档摘要比起来，多文档摘要面临的主要问题是：

(1) 几乎全部的单文档文摘都是基于抽取的，抽取对于单文档也的确行之有效。而对于多文档，在同一主题中的不同文档中不可避免地会有信息交迭，同样不可避免地会有信息差异，如何能最大程度地既避免冗余又能反映出这些差异是多文档文摘中的首要目标。

(2) 单文档的输出句子一般来说一成不变的是按在原文中出现顺序排列的，而在多文档情况里，这种空间关系不存在了，所以当前的多文档文摘大都采用时间顺序，而如何准确地得到每个句子的时间信息，也是多文档文摘要解决的一个重要问题。

下面我们从这两个方面介绍多文档摘要的工作。

2.4.1 冗余的识别和处理

人们采取了许多办法来识别摘要中的冗余句子。一个普通的做法是聚类法，该方法将相似的信息聚为一类。另一种做法是候选法，系统首先测量候选文段与已选文段的相似度，仅当候选段有足够的新信息才将其入选。这种方法中以 MMR 最为流行。

下面详细地叙述这两种方法中有代表性的工作。

（1）聚类法

[McKeown, 1999; Barzilay, 1999a]的工作属于第一种做法，其主题识别所采用的方法实际上就是通过词频等特征信息以及机器学习算法来判断两个段落之间是否相似，然后将最相似的段落聚为一类。

与传统的基于 TF×IDF 的方法不同，[Radev, 2000b]开发的 MEAD 系统采用质心（Centroid）用来生成同一事件的新闻报道的聚类。质心定义如下式（2-6）

$$c_j = \frac{\sum_{d \in C_j} \tilde{d}}{|C_j|} \tag{2-6}$$

其中，c_j 代表第 j 个聚类的质心，C_j 代表属于该类的文章，其数量由分母给出。d 代表该类中的某篇文章，而 $\tilde{d}$ 代表文章的 TF×IDF 向量，该向量中的值只有大于某个阈值的才被保留下来。质心可以看成代表了这个聚类中的重要信息的向量。

[Marcu, 2001]使用 C-Link[Defays, 1977]聚类算法将相似的句子聚类，每个类别给出其大小和相似度，该相似度为每个类中最不相似的两句话的余弦相似度。

[Mani, 1997]在信息抽取的框架下采用基于图的扩散激活（spreading activation）方法来识别一组文本里的相似和不同。Mani 使用由图的节点和边代表的实体和关系来形成摘要。[Zha, 2002]对文章采用加权无向图和加权二分图（bipartite graph）建模，然后用图谱聚类法对相似主题的句子进行聚类。基于图的方法还有[Mihalcea, 2004]的 TextRank 以及[Erkan, 2004]的 LexPageRank，这些方法都是先构建基于句子相似度的图，然后借助 PageRank 等方法寻找重要的句子。Erkan 认为基于质心的方法会有将聚类中的信息过于泛化的缺点，而创建句子相似度的图有利于更好地描述句子的重要性。在 Zha 的方法的基础上，[Wang, 2008]提出了一种基于句子语义分析和对称非负矩阵分解的聚类方法：首先对所有句子进行语义分析，然后创建句子到句子的语义相似度矩阵，最后将这个对称矩阵分解，这就构成了相似句的句群，达到聚类的目的。

聚类方法考虑的是先将相同类别的句子聚集在一起，然后再从每一个类里找出代表性的信息。如果类别数适当，聚类理想，这种做法理论上能够保证所取得的信息之间既有尽可能少的交迭，又能保证覆盖面最广。然而如何合理的定义类别数，以及聚类所依赖的相似度信息是否准确是这种方法能否成功运用的瓶颈。候选法则从另一个角度考虑这个问题。

（2）候选法

候选法中以最大边缘相关（Maximum Marginal Relevance, MMR）[Carbonell, 1998]方法最有代表性。面对高冗余的海量文本，传统基于用户查询的最大相关度（maximizing relevance）的信息检索方法，显得无能为力。 MMR 就是致力于解决这种问题而提出的。

简言之，有着高 MMR 值的文本应该既与所做的查询相关，又与已入选的文本相似度最小，用公式表示如下（2-7）：

$$\mathrm{MMR} \overset{def}{=} Arg \max_{D_i \in R \setminus S} \left[\lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right] \qquad (2\text{-}7)$$

上式中，当λ=1时，MMR 算得的是某文本和查询语句的相关度的排序（也就是只有Sim_1的值）；λ=0时，算得的是某文本和已入选文本最大的不同（diversity）的排序；而当λ介于这两值之间时，MMR 为上述两值的线性组合。

MMR 有着既能保证相关又能减小冗余的特性，如果以句子为单位而不是以文本为单位，可以很方便的将之应用于多文档文摘中去[Goldstein, 2000; Lin, 2002a]。MMR 的另一个特点是它的主题或查询相关性，这可特性可以很方便地根据用户需求生成相应的摘要。

[Daume III, 2005]是针对查询语句的多文档摘要方法。Daume 采用基于潜层狄利赫累分布（Latent Dirichlet Allocation）的变体模型来计算相似度，求得每个句子由查询语句产生的概率，并以这个概率作为和查询语句的相关度，最后用类似 MMR 的方法选取句子。

一般来说，基于查询的多文档或单文档摘要都要考虑采用 MMR 或其某些变体来进行摘要抽取。但 MMR 在通用型的多文档摘要里也有应用。

候选法与聚类法最大的不同在于它往往需要先对文本进行初步的摘要句提取，然后再进行 MMR 判别。这种思路下，如果一个信息未出现在摘要句中则其将不会再在 MMR 中进行考虑。而且，基于候选的方法也非常依赖语句间的相似度计算方法。

2.4.2 提取重要信息

当找到输入文本中的相似部分后，可以按照单文档摘要的方法直接抽取相关句子。如在 MEAD 中，只要根据压缩率选出分值最高的句子就行了（如压缩率为 v，

句子数为 N，则选出排序在最前面的 N×v 个句子）。因为多文档冗余性太大，虽然这样的做法可行，很多学者还是愿意从理解的角度，采用句子压缩、融合等生成策略来解决这个问题。

（1）融合法

融合法[Barzilay, 1999a]比较复杂。信息融合（Information Fusion）的目的是要生成一个简洁、通顺并能反映一组相似句子，如图 2-4（也就是主题），之间的共同信息的句子。为达成这个目标，要识别出对所有入选的主题句都共有的短语，然后将之合并起来。显然词袋（bag of words）的方法在这里肯定无法使用，为实现融合需要更深层的自然语言处理技术。

On Friday, a U.S. F-16 fighter jet was shot down by Bosnian Serb missile while policing the no-ly zone over the region.
A Bosnian Serb missile shot down a U.S. F-16 fighter over northern Bosnia on Friday
On the eve of the meeting, a U.S. F-16 fighter was shot down while on a routine patrol over northern Bosnia.
O"Grady"s F-16 fithter jet, based in Aviano, Italy, was shot down by a Bosnian Serb SA-6 anti-aircraft missile last Friday and hopes had diminished for finding him alive despite intermittent electronic signals from the area which later turned out to be a navigational beacon.

图2-4 Barzilay 的融合法

Barziley 认为，理想状态下，句子融合方法应该不需要完备的语义表达，而仅依靠输入文本和可以从语料中自动获得的潜层的语言知识（比如一个句法树）就能完成句子融合。如何由主题得到应该入选摘要的短语呢？对于图 2-4，也就是如何能确定"On Friday, U.S. F-16 fighter jet was shot down by a Bosnian Serb missile"为摘要想得到的唯一的表达呢？

为得到谓词论元结构并识别功能角色，Barziley 首先用句法分析器得到每句话的依存句法树，如图 2-5 所示，然后遍历每棵树，把相同的节点输出，一旦找到一个完整的谓词论元结构，就将其标记为摘要输出。如果出现根节点相同但实际上不相同的两个短语，则判断它们是否互为复述（paraphrase）（通过规则方式）。当得到了需要的谓词论元结构后则将其输入到一个语言生成系统生成摘要句。

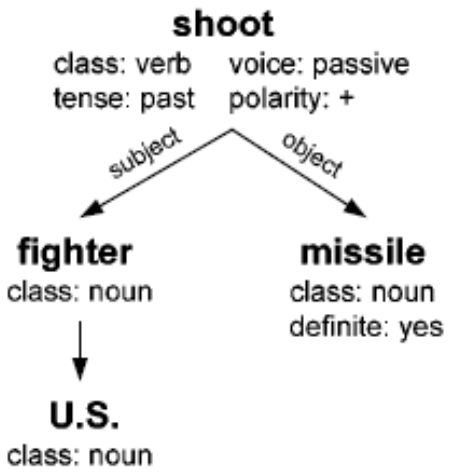


图2-5 Barziley 融合法的依存句法树示例

[Marsi, 2005]采用基于树对齐算法[Meyers, 1996; Barzilay, 2003]来达到句子融合的目的——每对节点的相似度只和这两个节点的单词相似度以及这两个节点的子孙的最佳匹配对有关。

（2） 压缩法

压缩法也就是在句子内部做修改，剔除一些不重要的成分，一般用于摘要生成后的后处理。这种方法属于理解型摘要方法，它并不局限在多文档文摘中使用——在单文档文摘里也能使用，只是因为多文档文摘中更倾向于使用理解型的摘要方法，所以在这里介绍。

压缩法一般是人工制定规则或采用统计学习获得规则。

[Mani, 1999a]所说的改写（revision）方法实际上就是采用一组人工给出的句子压缩和合并规则。首先从文本中根据权重抽出候选的摘要句，然后根据指定的规则对单句进行压缩或对两个句子进行合并。[Hovy, 2005]针对不同实体的重要性从保证信息量的角度采用内容和句法两个方式决定取舍。[Daume III, 2006]在生成摘要的最后阶段也采用简单的规则方法对句子进行压缩。

[Knight, 2000; Knight, 2002]将翻译中经典的噪声信道模型用于句子压缩——将待压句子看作压缩句经过噪声信道后的结果，这样用统计翻译的方法就可以得到一个句子的压缩。Knight 手工制作了 1035 个句子进行训练，并把结果和基于决策树的方法进行了比较，实验证明噪声信道模型有明显的优势。[Jing, 2000; Jing, 2002]认为人写摘要的一个主要习惯是对原文采用裁剪和粘贴（cut-and-paste）——写摘要的人从抽得的句子里去除不重要的成分，然后合并剩余的部分得到连贯的摘要。Jing

定义了去除（reduction）、合并（combination）、句法变形（syntax transformation）、词汇改写（lexical paraphrase）、一般化和特殊化(generalization and specification)以及重排序（reordering）6 种操作，然后用隐马尔可夫模型生成最终的摘要句。

其它基于统计的方法还有：[Riezler, 2003]采用词汇功能文法（Lexical Function Grammar）为句子做句法分析、生成来得到压缩句；针对压缩句语料不足，[Turner, 2005]使用了非监督和半监督的方法。[McDonald, 2006]使用基于最大边缘学习框架的判别式模型，模型所用的特征来自压缩的二元文法以及深层句法分析。

采用理解的方法来做摘要是自动摘要的一个重要发展方向，然而其对自然语言处理技术有严重的依赖性。当前自然语言处理技术的发展程度使得这种方式只能停留在理论探索阶段，完全无法做到实际应用。

2.4.3 确保一致性

对多文档摘要来说，确保一致性（coherence）是非常困难但也非常重要的一步工作，因为既要考虑事件发生的时间顺序，又要考虑内容的连贯性。从理论上说，如果要做到一致性，需要有对整个篇幅的内容的了解以及对叙述结构的清楚认识，可是篇章结构的识别对于现在的自然语言处理来说本身也很困难。针对一致性问题，许多学者做了研究，其方法可以大致分为两类，一种是依赖时间的，另一种是不依赖时间的。

（1）时间依赖的方法

仅按照时间顺序对句子进行排序[Lin, 2001; McKeown, 1999]也不失为一个好办法。因为直观的讲，新闻的系列追踪报道都是以时间作为基本脉络的。这样，在同一篇文章里的句子按照空间顺序输出，不同文章里的句子按照时间顺序输出就能在一定程度满足一致性的要求。

这样做要考虑的一个问题是时间消歧：做多文档摘要，同一个诸如“周一”的时间，在不同的文摘里很可能意味着不同的时间。如果没有绝对时间作为参考，生成的摘要会让读者以为事情是发生在一个星期里。为避免读者对来自不同日期但都提到诸如“昨天、明天”等词的段落产生混淆，系统应将其换为准确的时间戳（time stamp）。

[Lin, 2002a]使用出版日期作为参考点，并为下列时间表达式计算绝对时间

- Weekdays（Sunday,Monday,etc）
- (past|next|coming)+weekdays
- Today, yesterday, last night

最后将摘要句子按时间顺序排序。另外,[Lin, 2002a]还采用了加入引导句的方法来提高连贯性。就是给每个入选的摘要句配一个合适的引导句（introductory sentence），除非摘要句本身就是个引导句了。Lin 在 DUC－2001 中使用每篇文摘的第一句作为引导句。

只按时间来输出带来的一个问题是内容的重要性在摘要里得不到体现。[Barzilay, 2002]又给出了一种数量占优排序（Majority Ordering，MO）的方法，并对排序方法做了个总结，他认为：使用时间排序（Chronological Ordering，CO）——如果话题（topic）在文本中很不集中，也就是散布在整个文本中的时候，使用 CO 算法容易破坏连贯性；而 MO 也不会取得好的效果。Barzilay 最后在 CO 的基础上提出了一种扩张算法：其实质是在时间连贯性和内容重要性之间做折中。[Okazaki, 2004]在 Barzilay 方法的基础上，先将待输出摘要句按时间排序，然后对每个摘要用其在原文中的前几个句子来估计其预设信息（presupposition information），最后对排序做改进。[Bollegala, 2010]采用了自底向上的排序策略：通过制定时间、主题相关度、前驱和后继四个标准并以这四个标准逐对合并摘要句并排序，最终得到排好序的摘要。

依赖时间的排序方法方便有效，当文章的时间信息能够准确的到的情况下，该方法往往能取得比较好的效果。但这种方式对于时间信息不完整的文本则会完全失效，而且其对文本内部的时间消歧往往依赖一些启发式的规则，缺乏适应性。

（2）不依赖时间的方法

这些方法往往需要更深的篇章结构理解，如在修辞结构理论（Rhetorical Structure Theory，RST）关系基础上发展的针对多文档分析的交叉文本结构理论（Cross-Document Structure Theory，CST）[Radev, 2000a]，但是 CST 过于复杂，很难得到好的结果。

[Lapata, 2003]提出了句子交替的概率模型：用动词、名词、依存关系等特征代替的句子间的笛卡尔积作为特征，然后从语料里训练并计算两个句子交替的概率。[Barzilay, 2004]提出了一种基于内容的 HMM 概率模型对特定领域文本的主题转换进行建模的方法。Barzilay 的方法里状态的迁移，也就是主题的变换，能够一定程

度上反映句子的排序信息。[Ji, 2006]提出了一种基于流形（manifold）结构的半监督句子分类，以及历史参照排序策略为句子排序。Ji 的排序方法能一定程度保证主题连贯。

不依赖时间的排序方法往往需要借助于篇章理解。在基于句子的语义理解尚属于雏形的今天，篇章理解更是尚处于理论探索阶段。但这毕竟是摘要问题最终解决需要跨过的一道阶梯。

从上面的介绍可以看出，虽然多文档摘要相对较新，这里介绍的方法都不甚成熟，但却引入了很多很有趣的问题分支。多文档摘要对句子排序、指代消解、句子生成、时间消歧、篇章结构等领域都提出了更高的要求。也有很多研究者致力于如：如何让检索到的文档生成的摘要能给出人物传记、一系列同一类事物的描述、一系列社论的综述、以及事件的起因等等的方向，另外，如何给单个或多文档生成题目也很具有挑战性。

2.5 评测方法

很明显，摘要的评测也是一项个非常棘手的工作。可以想象，人工的摘要评测和人工摘要生成同样也不会有很好的一致性，因为人和人对文章的理解本身就可能有很大偏差。有数据说：即便是相对平铺直叙的新闻，人工摘要也只能做到最多 60％的吻合（测量句子内容的交迭）[Radev, 2002]。可是如果不用自动评测，评测问题也会成为一个大灾难。因此，建立自动评测的出发点是不要求完美，但要尽量和人工评测一致。自动评测对于整个自动文摘领域技术的进一步发展和完善起着至关重要的作用。

[Radev, 2002]将评测方法分为两类，内部评测（intrinsic evaluation）和外部评测（extrinsic evaluation）。

内部评测侧重于从生成的摘要的连贯程度及内容完整性的角度来评价摘要系统，它又可大致分为形式度量（form metrics），又叫一致性(coherence)，主要由人工参与评测；及内容度量（content metrics），又叫信息量（informativeness），主要通过计算系统生成摘要与原文或参考摘要内容的覆盖度来评测。外部评价方法就是通过使用文摘系统而达到对某个任务效率的提高与否来衡量文摘系统的性能。不同的应用，比如文本分类、自动问答、传统（ad-hoc）的信息检索等，都有各自相应的外部评价方法。

目前国际公认的几个摘要评测组织都是综合内部评测及外部评测来共同对摘要系统进行评估的。因为外部评测依赖于其它的任务，而人工评测依赖于人为因素，所以这里主要介绍内部评测中的内容度量方法的发展。

Lin对摘要自动评测做了大量工作，他开发了SEE （Summary Evaluation Enviroment）[Lin, 2002b]，以及参考了机器翻译中BLEU方法的ROUGE（Recall-Oriented Understudy for Gisting Evaluation）[Lin, 2004]。[Lin, 2006]从信息论的角度做自动评测，其实质是计算人工摘要和自动摘要概率分布的JS距离（Jensen-Shannon divergence）。Lin的评测方法里使用最广的是ROUGE的各种变体如：ROUGE-N，ROUGE-L，以及ROUGE-S（N取1-5，代表语言模型；L代表最大公共匹配；S和N类似，引入了类似F1值的方式）。在DUC01和DUC02的评测中ROUGE-N当N取2，以及L、S都取得了很好的效果。但在和人工摘要的互相关性上多文档摘要比单文档摘要的要差。

2.6 典型系统介绍

表2-2 经典摘要系统

系统名称	多/单文档	查询/通用	主要方式：A（理解）E（抽取）	排序方式	开发时间
Summarist	单	查询	A、E、概念融合、句子生成	空间	1997
NeATS	多	查询	E、MMR	时间	2001
MEAD	多	通用	E、质心、聚类	时间	2001-2004
Columbia's Newsblaster（multiGen）	多	领域	A、句子融合	时间	2001

表2-2给出了几个经典的摘要系统，以及各个系统的技术特色和发布时间。

下面简要介绍其中有代表性的几个多文档摘要系统，其中NeATS和Mead是基于抽取的，是当前的主流，而MultiGen的特点是采用了信息融合，虽然当前效果并不甚理想，但应该具有更大的潜力。

2.6.2 NeATS 系统

NeATS [Lin, 2001]：是一个基于抽取的多文档摘要系统，在DUC2001&2002中排名前二。NeATS大都用的是单文档摘要的技术，但是将这些技术用于多文档摘要还是首次。将引导句放入摘要是该系统的创新。该系统包括以下三个主要部分：

1、主题选择：它的主要任务是去识别一个文档集中的重要概念。NeATS 通过计算似然率来识别一、二以及三元文法代表的主要概念，然后将这些概念聚类以得到一个大主题下的若干重要子主题。然后，文档集中的句子根据重要概念结构（key concept structure）进行重排。

2、内容过滤：NeATS 采用了句子位置、关键词（stigma words）及 MMR 来选取句子。

3、内容表现(Content Presentation)：为确保文摘的连贯性(choherence)NeATS 给每个入选句额外配一个引导句(一般为入选句的前一句),然后按每句的时间顺序输出。

NeATS的设计是以实用为出发点的。其中的句子抽取、加入引导句、考虑句子位置等特性无一不是这种观点的体现。它的成功也给出了一种在当前自然语言处理技术发展水平下如何处理摘要这种综合性很强的复杂问题的思路。

2.6.3 MEAD 系统

MEAD[Radev, 2000b]：是个基于抽取的摘要方法。它的第一版是在2001年发布的，最新的版本为3.08。NewsInEssence网站的核心算法就是MEAD。NewsInEssence是个能自动做相关主题文本聚类，并形成摘要的多文档文摘系统，由密西根大学的CLAIR小组开发完成。

MEAD 的系统原理如下：

1、 它的输入是聚类后的文本集（通过一种变形的 TF×IDF 方法来做 TDT，这种方法称为 CIDR[Radev, 2000b]），将其切分成句子，并且给出一个压缩率。

2、 根据质心值（Centroid value）、位置值（段首句最高分等等）以及与首句交迭三个值加权求和，得到同一个聚类中的句子的分值。

3、 根据压缩率，得到入选句，并对其按照时间先后顺序输出。

MEAD系统的创新之处在于提出了采用质心算法来计算句子与句子，句子与篇章间的相似性的方法。除了质心、位置、以及首句交迭，它没有考虑更多的特征，这使得该系统简单实用，但也很难生成高质量的摘要。

2.6.4 MultiGen 系统

MultiGen[McKeown, 1999]：句子融合是MultiGen的核心技术。MultiGen的输入为已聚类的同一主题的新闻报道，输出各个报道的共同信息。 与句子抽取相比，句子融合是多文档摘要的一个重要尝试，它可以真正地将来自一个文本集中的所有句子，综合成一个可以很大程度上体现这个集的信息的句子。这种方法也避免了句子抽取的偏向性（biased）。

该系统分为以下部分：

1、主题创建，Simfinder (与 mead 的 TDT 相似)将输入的句子进行聚类成主题（theme），同一主题中的句子表达的信息相似；

2、主题选择（theme selection）：为按所要求的压缩率输出摘要，要给各个主题打分，从中选出若干最高分。打分按照：①该主题中含句子多少；②该主题中句子的相似度；③重要性分数（salience score）。前两个由 simfinder 计算，后一个用词汇链（lexical chain）计算[Morris, 1991; Barzilay, 1997]；

3、主题排序（theme Ordering）：通过捕捉各个主题间的时间关系（chronological order）把选出的主题排列成一致（coherent）的文本；

4、句子融合（sentence fusion），用一个主题中的所有的相似句来创建一个紧凑通顺的，并且能反映该主题主要信息的句子。这主要通过识别这些句子中共同的短语（phrases），然后将它们组合成一个新句子。

MultiGen是基于理解的摘要方法的代表性的尝试，其中的句子融合部分基本给出了用理解的方式解决摘要问题的主要框架。然而当前计算机理解的水平大大限制了这可系统的实用性。

2.7 本章小结

本章介绍了单文档和多文档摘要近年来的主要发展，以及因为摘要问题而引申出的，在其它 IR 领域也广泛讨论的一些子问题。早期自动文摘问题的提出，是在自然语言处理各个分支都处在雏形甚至尚未形成的阶段，所以早期的自动摘要，都倾向于使用表层信息及一些简单特征。随着自然语言处理技术不断进步，人们在摘要问题上的研究也一步步深入，对于深层的信息的挖掘也不断被尝试。可以谨慎地预言，随着时间的推移，现阶段主流的摘录型方法将会被理解式的方法所取代。

第三章　统计机器学习方法

3.1　引言

“自然语言处理要研制表示语言能力和语言应用的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并讨论这些实用系统的评测技术”[宗成庆, 2008]。

自然语言处理的发展从一开始就伴随着计算机的成长，从计算机的出现开始，人们就不断关注机器语言和人类语言的关系。进入上世纪 90 年代，随着计算能力的飞跃性提升，基于统计方法的各种自然语言处理技术也获得了长足发展，各种统计模型被应用到自然语言处理的各个领域。

人们围绕语言问题建模的热情一直推动着自然语言处理的发展。合理的模型能够帮助人们捕捉优秀的、更反映问题本质的特征。而每个模型有其自身的优点和固有的缺点，应该说，没有完美的模型，只有合适的模型。针对不同的问题选用合适的模型是极富创造力的工作。

本章介绍论文中要用到几个统计模型，包括隐马尔可夫模型（Hidden Markov Model，HMM）、条件随机场模型（Conditional Random Field，CRF）、支持向量机模型（Supportive Vector Machine，SVM）。

下面分别讨论上述模型的特点。

3.2　隐马尔可夫模型

HMM 模型最早在语音识别领域获得了成功的应用，并起到了关键的作用,近十几年在自动分词、词性标注、机器翻译等多个方面都得到了广泛的应用。

我们经常说本质决定现象，在观察到的现象背后隐藏着不可见的，但是决定性的本质。隐马尔可夫模型模拟的正是现象和本质的关系：假设在观测物（现象）的背后有一系列隐藏的状态（本质），现象是可见的，而本质是不可见的，观测序列是由这些隐藏状态序列所生成。HMM模型可以定义成一个5元组：

$$\eta = (\Omega_x, \Omega_o, A, B, \pi) \tag{3-1}$$

其中 $\Omega_x = \{q_1, q_2, ..., q_N\}$，为隐藏状态的有限集合；

$\Omega_o = \{o_1, o_2, ..., o_M\}$，为观测物体的有限集合；

A 是隐藏状态转移概率 $N \times N$ 矩阵；

B 是从某个隐藏状态产生某个观测物体的 $N \times M$ 输出概率矩阵；

π 是初始状态概率。

HMM模型建立在三个假设上：

- 马尔可夫假设：某一状态序列出现的概率只和它前面的一个状态相关，这也就是一阶马尔可夫性；
- 时间无关假设：状态的转移和时间无关；
- 独立性假设：某时刻输出的可见状态只和当前状态相关。

HMM的三个基本问题是：

（1） 估计问题：也就是给定了模型和一个观察序列，如何得到这个序列由这个模型产生的概率。

（2） 序列问题：给定了模型和一个观察序列，如何找到最优的状态序列，使该模型在这个状态序列下能最好的“解释”这个观察序列，也就是概率最大。

（3） 参数估计问题：如何由观察序列得到模型的各个参数。

下面给出解决办法。

（1）估计问题

给定了模型 η 和观察序列 $O = (o_1, o_2, ...o_T)$，现在的任务是找到概率 $P(O \mid \eta)$。在通用型的情况下，任一时刻，可能的状态都有N个，那么能输出观察序列 O 的可能的状态序列总共有 N^T 个。为简单起见，我们先给出观察序列由某一个状态序列 $Q = (q_1, q_2, ..., q_T)$ 产生出来的概率（3-2）（这里用到了独立性假设）。

$$P(O \mid Q, \eta) = \prod_{i=1..T} P(o_i \mid q_i, \eta) = \prod_{i=1...T} b_{q_i}(o_i) \tag{3-2}$$

接着，我们再给出模型 η 产生出状态序列 Q 的概率（这里用到了马尔可夫性假设）：

$$P(Q \mid \eta) = P(q_1, q_2, ..., q_T \mid \eta) = \pi_{q_1} \times \prod_{i=1...T-1} a_{q_i q_{i+1}} \tag{3-3}$$

最后，我们回到原来的问题，就是指给定模型的情况下观察序列的概率：实际

上就是先让模型产生所有可能的状态序列，然后从这些状态序列中产生出我们给定的观察序列的概率合，其公式：

$$\begin{aligned} P(O|\eta) &= \sum_{q_{1\ldots T}} P(Q,O|\eta) = \sum_{q_{1\ldots T}} P(Q,O|\eta)P(Q|\eta) \\ &= \sum_{q_{1\ldots T}} \pi_{q_1} b_{q_1}(o_1) \prod_{i=1\ldots T-1} a_{q_i q_{i+1}} b_{q_i+1}(o_{i+1}) \end{aligned} \tag{3-4}$$

上式的计算复杂度成指数级增长，也就是“指数灾难”，实际计算是不可能的。为此人们提出前向算法（forward algorithm），利用动态规划的方法来解决这个问题。为了实现前向算法，需要定义一个前向变量$\alpha_t(i)$，其定义如下

$$\alpha_t(i) = P(o_1 o_2 \ldots o_T, q_t = s_i | \eta) \tag{3-5}$$

前向变量的含义是给定时刻t，HMM已经输出了观察序列$o_1 o_2 \ldots o_t$，且当前隐含状态为s_i的概率。有了前向变量定义，在t+1时刻的前向变量就可以很方便的给出前向算法如下：

算法：前向算法

第一步 初始化

$$\alpha_1(i) = \pi_i b_i(o_1), i = 1\ldots N \tag{3-6}$$

第二步 递推

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right) b_j(o_{i+1}), \quad t = 1\ldots T-1 \tag{3-7}$$

第三步 求和

$$P(O|\eta) = \sum_{i=1\ldots N} \alpha_T(i) \tag{3-8}$$

直观上看，上式的计算在每一个观测值要计算所有N 个状态，对于每个状态需要考虑其前面N 个状态，那么总的时间复杂度为$O(N \times N \times T) = O(N^2 T)$，远远小于先前的$O(N^T)$，这个复杂度是很容易实现的。

有了前向算法后，很容易得到它的逆过程，后向算法——实际上就是从后向前看观测序列。先定义后向变量

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \ldots, o_T | q_i = s_i, \eta) \tag{3-9}$$

有了后向变量，很容易得到后向算法：

算法：后向算法

第一步 初始化

$$\beta_T(i)=1,i=1...N \tag{3-10}$$

第二步 递推

$$\beta_t(i)=\sum_{j=1...N} a_{ij}b_j(o_{t+1})\beta_{t+1}(j),t=1...T-1;i=1...N \tag{3-11}$$

第三步 求和

$$P(o\mid\eta)=\sum_{i=1...N}\pi_i\beta_1(i) \tag{3-12}$$

这样有了前后向算法，很容易得到观察序列某个时刻t取某个状态i的概率：

$$P(O,q_t=s_i\mid\eta)=\alpha_t(i)\beta_t(i) \tag{3-13}$$

（2）序列问题（维特比算法）

维特比（Viterbi）算法适用于解决HMM的第二个问题：如何最好的“解释”看到的观察序列。这个问题是HMM最常用的特性，在序列标注问题里，把标注看作隐含状态，序列作为观察值，这样解决第二个问题实际上就是找到最合理的标注。这个问题要取决于对“最合理”的理解。

如果把序列割裂开看，每个观测值最可能的状态的取值应该是：

$$\hat{q}_t=\arg\max_{i=1...N}\left(\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1...N}\alpha_t(i)\beta_t(i)}\right) \tag{3-14}$$

这种理解明显忽略了状态之间的联系，失去了序列的马尔可夫性。另一种常用的理解是“最合理的序列”：

$$\hat{Q}=\arg\max_{Q}P(Q\mid O,\eta) \tag{3-15}$$

这样，优化的不是单个状态，而是整个状态序列。

算法：维特比算法

第一步：初始化

$$\begin{aligned}&\delta_1(i)=\pi_i b_i(o_1),\ i=1...N\\&\psi_1(i)=0\end{aligned}\tag{3-16}$$

第二步 迭代

$$\begin{aligned}&\delta_t(j)=\max_{i=1...N}\left(\delta_{t-1}(i)\times a_{ij}\right)\times b_j(o_t),\ t=2...T;\ j=1...N\\&\psi_t(j)=\arg\max_{i=1...N}\left(\delta_{t-1}(i)\times a_{ij}\right)\times b_j(o_t),\ t=2...T;\ j=1...N\end{aligned}\tag{3-17}$$

第三步 终结

$$\begin{aligned}&\hat{Q}_T=\arg\max_{i=1...N}\left(\delta_T(i)\right)\\&\hat{P}(\hat{Q}_T)=\max_{i=1...N}\left(\delta_T(i)\right)\end{aligned}\tag{3-18}$$

第四步 回溯（寻找最优路径）

$$\hat{q}_t=\psi_{t+1}(\hat{q}_{t+1}),\quad t=T-1,T-2,...,1\tag{3-19}$$

和评估问题一样，维特比算法也采用了动态规划算法。维特比的复杂性和前向、后向算法复杂性一样。往往在实际应用时需要搜索*n*个最优结果，这样在计算过程中要记录多个结果。

（3） HMM的参数估计（Baum-Welch算法）

这是HMM的第三个问题，找到使$P(O|\eta)$最大的η，也就是给定训练数据后，模型如何训练的问题。由于含有隐含变量，相当于在所以采用最大似然估计法是不可行的（intractable）。期望最大化（Expectation Maximization，EM）算法可以用于解决这种含有隐含变量的模型参数估计。EM采用一种迭代爬山算法，可以局部使$P(O|\eta)$最大化。其基本思想是先给模型参数赋初始值，然后在E步，用已有的模型参数求得隐含变量的期望，也就是得到每个时间点可能状态的概率；在M步，根据E步得到的状态序列的期望值用最大似然法重新估计模型参数；重复E和M直至收敛。

下面给出EM算法在求解HMM参数估计时的形式化算法，又叫前向后向算法（forward-backward algorithm）：

算法：前向后向算法

第一步 初始化，随机的给模型参数赋初始值η^0，但要满足概率和为1的约束。

$$\xi^n{}_t(i,j)=\frac{P(q_t=s_i,q_{t+1}=s_j,O\mid\eta^n)}{P(O\mid\eta^n)}=\frac{\alpha^n{}_t(i)a^n{}_{ij}b^n{}_j(o_{t+1})\beta^n{}_{t+1}(j)}{\sum_{i=1\ldots N}\sum_{j=1\ldots N}\alpha^n{}_t(i)a^n{}_{ij}b^n{}_j(o_{t+1})\beta^n{}_{t+1}(j)} \tag{3-20}$$

第二步 由模型η^n计算（第n次）隐含变量（状态）的期望（ E步）：

$$\gamma^n{}_t(i)=\sum_{j=1\ldots N}\xi^n{}_t(i,j) \tag{3-21}$$

由上面求得的期望值重新估计模型参数得到η^{n+1}（M步）

$$\begin{aligned}\pi_i^{n+1}&=\bar{\pi}_i=\gamma_t^n(i)\\ a_{ij}^{n+1}&=\bar{a}_{ij}=\frac{\sum_{t=1\ldots T}\xi_t^n(i,j)}{\sum_{t=1\ldots T}\gamma_t^n(i)}\\ b_j^{n+1}(k)&=\bar{b}_j(k)=\frac{\sum_{i=1\ldots T}\gamma_t^n(j)\times\delta(o_t,v_k)}{\sum_{i=1\ldots T}\gamma_t^n(j)}\end{aligned} \tag{3-22}$$

这里的δ函数是克罗奈克函数（Kronecker），即示性函数

第三步 循环第二步直到模型收敛。

前向后向算法是在训练集不包含状态序列的情况下计算的，因为隐含变量的“卷绕”，使得无法求出概率分母。EM算法只能保证给出局部最大值，还可以采用蒙特卡罗马尔可夫模型(Markov Chain Monte Carlo, MCMC) [Bishop, 2006]来计算参数，这样做的话耗时很长，但可以保证得到全局最优值。另外，如果训练集中包含有状态序列，则可以直接用极大似然估计方法求得模型参数。

3.3 条件随机场模型

条件随机场（Conditional Random Fields, CRF）[Lafferty, 2001]是一个在给定输入节点条件下计算输出节点条件概率的无向图模型（Undirected Graphical Model）。CRF和HMM、最大熵马尔可夫MEMM一样也是用来做统计序列标注、切分。CRF近年来广泛应用在自然语言处理的各个领域，如潜层句法分析[Sha, 2003]、命名实体识别[McCallum, 2003]、自动文摘[Shen, 2007]、信息抽取[Pinto, 2003]等等都获得

了比较大的成功。

给定的输出序列标识Y和观测序列X，CRF是通过定义条件概率$P(Y|X)$，而不像HMM是通过定义联合概率$P(X,Y)$，来刻画序列的，这也就是为什么CRF以及MEMM通常归类为判别式模型（discriminative model），而HMM通常叫产生式模型（generative model）。定义CRF的特点是： 第一，放松了HMM的特征（状态）的独立性假设——该模型可以很好地拟合真实世界的数据，在这些数据中，标记序列的条件概率依赖于观察序列中非独立的、相互作用的特征，并通过赋予这些特征不同的权值来表示该特征的重要程度；第二，纠正了MEMM中的标注偏置（Label bias）问题——CRF公式中的用于归一化的分母是对整个序列取归一化，而MEMM是针对某个节点。CRF可以说囊括了MEMM的优点。

CRF也可以看作无向图模型(undirected graphical model)或者马尔可夫随机场（Markov random field），简单讲，就是任何一个节点的条件概率都只和与其连接、或邻近的节点有关。其定义如下：

定义：$G=(V,E)$为一个无向图，V是节点集，E是无向边集。V中的每个节点对应着一个随机变量Y_v，也就是$Y=\{Y_v|v\in V\}$，Y的可能取值为$\{y\}$。对每个随机变量Y_v，以观测序列X为条件，则都满足无向图上的马尔可夫特性：

$$p(Y_v|X,Y_w,w\neq v)=p(Y_v|X,Y_w,w\sim v) \tag{3-23}$$

其中$w\sim v$标识两个节点在图G中是邻近节点。那么，(X,Y)为一个条件随机场。应该说，图模型的结构是随意的，比如给出如下最简单、最普通的链式结构图3-1

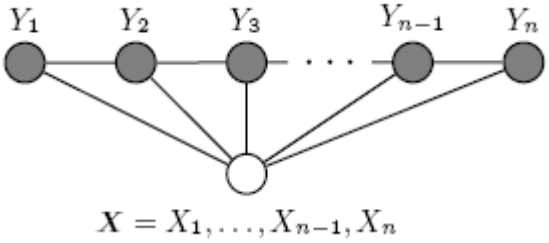


图3-1 链式 CRF 示意图

因为观察序列X在上式只是条件，并不考虑其独立性，所以X之间并不存在图结构。CRF的形式化描述如下：

对于第j种特征，定义局部特征函数$f_j(y_{i-1},y_i,X,i)$和对应该特征的权重λ，每个$f_j(y_{i-1},y_i,X,i)$表示状态特征$s(y_i,X,i)$和转移函数$t(y_{i-1},y_i,X,i)$，其中y_{i-1},y_i为标注序列，X为输入观测序列，i为输入序列X的第i个位置。则对于输入序列X和

标注序列Y，定义CRF的特征函数为

$$F(Y, X) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, X) \tag{3-24}$$

所以，CRF的条件概率可以由下式给出：

$$p(Y \mid X, \lambda) = \frac{\exp(\lambda \times F(Y, X))}{Z(X)} \tag{3-25}$$

其中，分母$Z(X)$为归一化因子：

$$Z(X) = \sum_{Y} \exp(\lambda \times F(Y, X)) \tag{3-26}$$

条件随机场模型需要解决的三个主要问题是：如何选特征、参数训练和解码。其中，参数训练过程可以在训练数据集上基于对数似然函数的最大化进行。具体算法请参见[Lafferty, 2001]。

3.4 支持向量机

支持向量机[Vapnik, 2000]近年来在模式识别的各个领域都取得广泛应用，在自然语言处理上，经典的应用有文本分类[Joachims, 1998]、语块分析[Kudoh, 2000]等。SVM建立在统计学习基础上，是一种基于两类模式分类问题的有监督机器学习算法，它在函数表达、泛化推广、和学习效率上都有极其优越的性能。

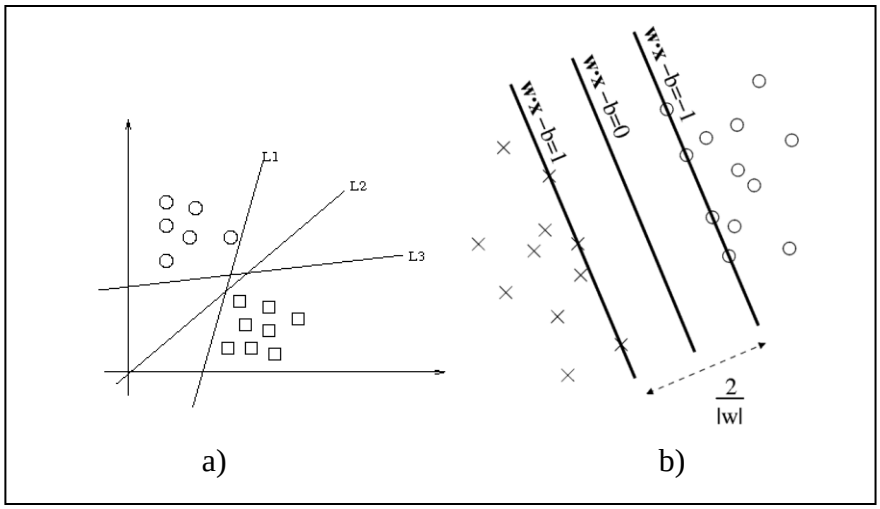


图3-2 SVM 原理图

SVM的主导思想是“升维”：将在普通维数下线性不可分的特征点，通过某种核函数变换，将其投影到一个高维空间，并在这个高维空间寻找最优分类面——使

两个平行超平面的距离最大化，这里假定平行超平面间的距离或差距越大，分类器的总误差越小。图3-2a中，L1和L3明显分类效果不如L2。

具体步骤是：

1) 通过预先选定的一些非线性映射（核函数）将输入空间映射到高维特征空间。
2) 在高维特征空间中对给定数据寻找最优超平面（Optimal Hyperplane, OHP），并实现分类。

这样做避免了在原特征向量空间中进行非线性曲面分割计算。

形式化定义

对于一个两类问题，假设有一组样本点 x_i，以及该样本点的类别 $y_i = \pm 1$，+1和-1分别代表不同的类，这样就构成了一组训练数据 $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$。现在要求支持向量机将这些点正确的用超平面分隔开。超平面可以写成下式，如图3-2b：

$$\mathbf{wx} - b = 0 \tag{3-27}$$

而支持向量如下式指的是，如果数据点是线性可分的，则这两个平面之间没有任何样本点且两个超平面间距离最大（见图3-2）

$$\begin{aligned} \mathbf{wx} - b &= -1 \\ \mathbf{wx} - b &= 1 \end{aligned} \tag{3-28}$$

由几何知识知道这个最大距离为 $2/\|\mathbf{w}\|$，所以最大化平面间距离，等价于最小化 $\|\mathbf{w}\|$，这是一个二次规划（Quadratic Programming）最优化问题。

而所有的样本点 x_i 应该满足

$$y_i(\mathbf{wx} - \mathbf{b}) \geq 1, \quad i = 1...n \tag{3-29}$$

采用拉格朗日求极值的方法，可以得到

$$L(\mathbf{w}, b, \alpha) = \frac{\|\mathbf{w}\|^2}{2} - \sum_{i=1...n} \alpha_i \left[y_i(\mathbf{wx} + b) - 1 \right] \tag{3-30}$$

最优化问题的解就是上式的极值点。上式的 α_i 为非负系数。

SVM采用的核函数具有这样的性质：它是一组以支持向量为参数的非线性函数的线性组合。因此分类函数的表达式仅和支持向量的数量有关，而独立于空间的维度。在处理高维输入空间的分类时，这种方法尤其有效。

理论上支持向量机是一个二类分类问题，在实际应用中处理多分类问题的时候有两种解决方式：一个是一对多模式，另一个是一对一模式；例如：采用一对多组合模式，对于n 类分类问题，则需要构建n 个二类分类器；而采用一对一组合模式则需要n(n-1)/2 个分类器。

3.5 本章小结

本章介绍了几个对本论文至关重要的模型，包括HMM、CRF和SVM。 其中，HMM和CRF模型对于第4、6、7章的内容比较重要，SVM对于第5章的内容比较重要。

第四章　基于半条件随机场的摘录型自动摘要建模方法

4.1　引言

根据摘要生成方式的不同，摘要方法主要可分为摘录型（Extractive）和理解式（Abstractive）。[Radev, 2002]认为，所有不是基于句子抽取的摘要方式，都可以归结为理解式——包括句子压缩、信息融合、篇章结构、以及基于本体论等等。现今，自然语言处理中的很多子领域，尤其是语义理解方面，还都处在实验性阶段，单纯对句子的语义理解都还谈不上实际应用，勿论针对篇章的理解了。所以笔者认为，虽然现在很多学者致力于理解式的方法，但摘录型的摘要方式还是，而且还将在很长一段时间里是实用性自动摘要方法的主流。

摘录型摘要的主要方法是从文章里提取特征然后采用有监督或者无监督的机器学习算法，对句子进行分类、打分，然后抽取并排序。因为是句子抽取，所以特征提取的基本单位也都是句子。

本章介绍了一种基于序列分段模型（Sequence Segmentation Models，SSM）的有监督摘录型摘要提取方法。

在这种方法里，我们将摘要问题看作“段标注”问题。和前人的工作比较，SSM方法的不同之处在于提取特征的单位不单来自句子，也可以来自于段。我们的 SSM方法使用了可以对“段”建模并标注的半马尔可夫条件随机场（Semi-Markov Conditional Random Fields，SemiCRF）。我们把结果和传统方法进行了对比，证明这种方法和比单纯针对句子为单位提取特征的方法有较明显的优势。

4.2　相关工作

早期的摘录型摘要方法主要集中在根据某些基本特征启发式地给句子打分上，如词频、词的分布特性、一些专有名词[Luhn, 1959]，句子在段落中的位置[Baxendale., 1958]等。

[Paice, 1990] 采用启发式方法确定句子重要性，并将句子打分的特征分为如下：

- ✧ 频度关键词（Frequency-Keywords）特征：使用频度较高的词被认为和主题关系密切，而含有这种词数量较多的句子也应该比较重要，可以用这种词的数量及频度来给句子打分。
- ✧ 标题关键词（Title-Keywords）特征：如果一句话含有标题中的词，则应该给予这个句子较高的重要性。
- ✧ 位置（Location）特征：重要的信息应该在文章的首和尾，以及段落的首和尾。
- ✧ 线索词（Cue Phrase）特征：含有如“极大的”、“极重要的”以及“几乎不可能”等等词汇的句子应提高重要性。
- ✧ 指示词（Indicator Words）：含有例如“In summary”、“This report”等词的句子重要性会高。

这些特征被证明是有效的。近年来，基于有监督的摘录型方法的工作主要都集中在寻找更有效的机器学习算法来利用这些特征上，当然，有些算法也会提出一些新的特征。下面详细介绍有监督摘录型摘要方法的发展历程。

[Kupiec, 1995]提出了用贝叶斯分类器来提取摘要句的方法，该方法采用的特征有：

- ✧ 句长特征：句子是否长于某个阈值（比如 5）。
- ✧ 是否有某些固定搭配：是否有“this letter”、“In conclusion”等 26 个线索词。
- ✧ 位置信息：句子是否在文章开始的 10 段和结尾的 5 段中的段首、段尾、段中。
- ✧ 主题词特征：频度比较高的词被定义成主题词。根据每个句子含有高频主题词的多少。
- ✧ 大写词特征：大写专有名词一般比较重要，如果一个句子出现了一些频度较高的专有名词则这句话也比较重要。

Kupiec 采用的贝叶斯模型如公式(4-1)：并假设条件独立。

$$\begin{aligned} P(s \in S \mid F_1, F_2, ..., F_k) &= \frac{P(F_1, F_2, ...F_k \mid s \in S)P(s \in S)}{P(F_1, F_2, ...F_k)} \\ &= \frac{\prod_{i=1...k} P(F_i \mid s \in S)P(s \in S)}{\prod_{i=1...k} P(F_i)} \end{aligned} \tag{4-1}$$

Kupiec 将实验结果和基于主观确定权重的方法[Edmundson, 1969]做了比较，

认为基于语料训练的方法和 Edmundson 的结果一致。[Aone, 1999]也采用了基于贝叶斯分类器的方法。

贝叶斯模型需要很强的独立性假设。首先这些特征很难说是相互独立的；其次该模型对每个句子分别取特征并识别，这样做实际上也割裂了句子和句子之间的关系。[Yeh, 2005]采用了遗传算法来处理类似的特征，遗传算法是一种最优化的方法，可以不做特征间的独立性假设，但他的方法也是逐句孤立计算的。

针对以上缺点，[Conroy, 2001]采用了隐马尔可夫模型来抽取摘要，第一次将文本看作句子序列来建模，采用 HMM 处理特征间的依存。其核心思想就是认为下一句话是否为摘要的概率不但与该句的特征有关，还与上一句话是否是摘要有关。

Conroy 假设了 s 种摘要状态和 $s+1$ 种非摘要状态，仅在摘要状态允许状态跳跃，且仅在非摘要状态允许状态保持。用于抽取前两个摘要句的 HMM 摘要模型见图 4-1，图中最后两个状态允许有任意多的摘要或非摘要句。用这个模型计算每句话属于各个状态 i 的概率，如果 i 是偶数，则说明这个是第 $i/2$ 句摘要，反之则说明这句话不是摘要。因为任何一条从始至终的路径都要经过开始的 $s-1$ 该模型用于抽取前 $s-1$ 个摘要句。该模型仅仅用了三个特征：句子位置、句子长度、句子及文本的相似度。

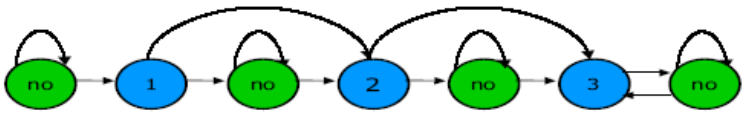


图4-1　Conroy 的 HMM 摘要模型

相对于朴素贝叶斯，HMM 放松了特征间的独立性假设，并以序列的角度看待摘要的问题。但是 HMM 处理复杂特征的能力非常有限，特别是难以应用大量的关联特征以及难以获得远距离特征。

针对这个问题，[Shen, 2007]采用了 CRF 模型来做序列标注，CRF 属于区分式模型，它能有效的利用远距离特征。Shen 的模型如下图 4-2，将摘要问题看作典型的序列标注问题，0 代表非摘要句，1 代表摘要句。与[Conroy, 2001]的方法相似，Shen 的模型里观察序列也用一系列特征表示。Shen 在自己的实验里将 CRF 的摘要结果和 SVM、HMM、朴素贝叶斯等等做了比较，证实 CRF 能得到最好结果。

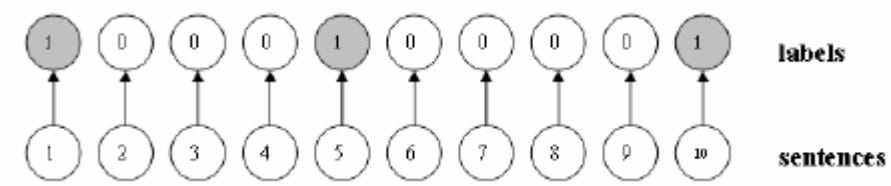


图4-2 Shen 的 CRF 摘要模型

以上这些方法共同的一个观点是，句子应该作为摘要的最基本单位，所有的特征都应该在句子层上得到。可是很容易看出，摘要句，尤其是非摘要句很容易形成连续的“段”，这些段上应该也包含有丰富的特征，而以上方法对这些特征是无能为力的。

近年来，SemiCRF 为代表的半马尔可夫模型（Seim-Markov Model，SM）[Ge, 2002]，我们称之为段序列模型（Sequence Segmental Model，SSM）吸引了不少学者的兴趣。SSM 将传统的序列标注问题改进为段序列问题，以结合更多的特征：以命名实体识别为例：实体长度以及其它实体间的相似度等等。SemiCRF 在一些经典的序列标注问题上性能优于 CRF。

4.3 本项研究工作的动因

为改进传统方法不能利用“段”特征的局限，本文基于 SemiCRF 提出了新的摘要问题建模思路：将有监督摘要问题的特征提取单元从句子放大到“段”，采用段序列模型。在这种方法里，特征不单来自句子，也可以来自“段”。

这里的“段”的意思是指一句或几句有共同标注的句子，而文本则看作是“段”的序列。SemiCRF 是适合做段序列标注的模型，理论上它能包含 CRF 的所有优点。使用 SemiCRF 可以囊括传统基于句子级的特征以及基于段的特征——比如对数似然度或余弦相似度，如果单纯从句子的角度度量，因为粒度过小，其数据稀疏的程度肯定要大于以段为单位的度量。进一步说，SemiCRF 是 CRF 的更一般化的模型，所有在 CRF 中可以利用的特征在 SemiCRF 中也得找到应用。

在下面的介绍中将先给出 SemiCRF 模型，然后给出我们针对这个模型所选用的特征。

4.3.1 半马尔可夫条件随机场

CRF 是由[Lafferty, 2001]提出的区分式模型 $P(Y \mid X)$，这里的 Y 和 X 都可能有复

杂的结构。作为一个序列标注模型，CRF 的优点一是放松了 HMM 过强的独立性假设，二是克服了最大熵马尔可夫模型的标注偏置（label bias）。SemiCRF 是在 CRF 基础上的一般化。之所以叫“半”，是因为它的特点是可以允许系统在每一个状态保持一段时间，当这段时间结束后，系统再选择是否跳出当前状态。而进入下一个状态要遵从马尔可夫性，也就是要考虑当前状态。当系统处于保持状态的时候，是没有马尔可夫性的。

（1） SemiCRF vs. CRF

给定一个观察序列 $X=(x_1,x_2,...,x_M)$，以及其标注（状态）序列 $Y=(y_1,y_2,...,y_M)$，其中的 y_i 在一个固定的集合 Ψ 中取值。对于摘要问题，本论文的 SemiCRF 仿照 Shen 的 CRF 方法，令 $\Psi=\{0,1\}$，0 代表非摘要，1 代表摘要。对于 CRF 来说，应该是寻找可以使下式最大化的标注序列。

$$P(Y\mid X,W)=\frac{1}{Z(X)}\exp(W\cdot F(X,Y)) \tag{4-2}$$

这里的 $F(X,Y)=\sum_{i=1}^{M}\mathrm{f}(i,X,Y)$ 是一个大小为 T 的列向量，$\mathrm{f}=(f_1,f_2,...,f_T)'$ 代表共有 T 个特征函数。每个特征函数可以被写成 $f_t(i,X,Y)\in R,t\in(1,\cdots,T),i\in(1,\cdots,M)$。举例来说，如果有如下假设：

第 10 个特征是：

[*当前句子的长度是否大于 25*]&[*当前句子是否为一个摘要句*]

第 11 个特征是：

[*当前句子是否位于段首*]&[*当前句子是否为一个摘要句*]

那么，当这两个特征函数作用于某篇文本，如 *text*_1 ，并且该文本的标注序列为 *label_ sequence*_1，那么特征函数 f_{10}(3,*text*_1, *label_sequence*_1)和 f_{11}(3,*text*_1, *label_sequence*_1)的意义分别是：

在带有标注 *label_ sequence*_1 的文本 *text*_1 中，

[*第 3 句的长度是否大于 25*]&[*第 3 句是否为一个摘要句*]

[*第 3 句是否位于段首*]&[*第 3 句是否为一个摘要句*]。

W 是一个大小为 T 的列向量，表示每个特征函数的权重。

上式中的 $Z(X)$ 为归一化因子：

$$Z(X)=\sum_{Y'}\exp(W\cdot F(X,Y')) \tag{4-3}$$

本论文方法的核心思想是：将序列 X 切分成段 $S=<s_1,s_2,\ldots,s_N>$，也就是将 X 按某种方式分割成 N 段。三元组 $s_j=<t_j,u_j,y_j>$代表在这种分割方式下的第 j 段，t_j 代表该段的开始位置，u_j,代表终止位置，y_j 代表该段的输出标注（在前面提到，每一段只有一个标注）。这种条件下，段和段之间不能存在交迭现象，也就是：

$$t_1=1,u_N=|X|,1\le t_j\le u_j\le|X|,t_{j+1}=u_j+1 \tag{4-4}$$

$$\sum_{j=1}^{N}|s_j|=|X| \tag{4-5}$$

上式中的$|\bullet|$代表取长度。要完成这一步，就可以引入 SemiCRF 了。

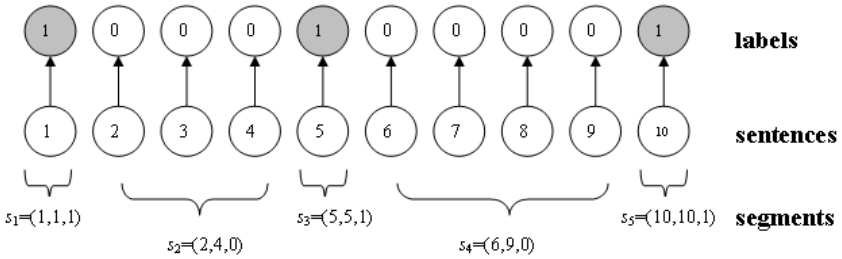


图4-3 SemiCRF 摘要模型

图 4-3 表示了一个有 10 句话的文本的 SemiCRF 模型，其中第三排（segments）为其中的一种分段方法 $S=<(1,1,1),(2,4,0),(5,5,1),(6,9,0),(10,10,1)>$，第二排（sentences）中的句子为该句子的属性（特征）。

原 CRF 原理公式中的特征函数 f 转变成了段特征函数 $g=(g_1,g_2,\ldots,g_{T'})$。和 CRF 中的 f 一样，$g_t(i,x,s)\in R$ 也把三元组(i,x,s)映射为一个实数。同样，可以定义 $G(X,S)=\sum_{i=1}^{N}g(i,X,S)$，这样最终得到了用于估计段序列 S 的概率表达式，为方便对比，将 CRF 的表达式一起写出，列于表 4-1。

（2） 推断（Inference）

SemiCRF 中的推断问题就是给出文本和模型，求出最可能的标注序列（摘要序列）。为了搜索最优段序列，SemiCRF 采用了一个类似 Viterbi[Viterbi, 1967; Sarawagi, 2005]的动态规划搜索算法：

表4-1 从 CRF 到 SemiCRF

CRF	**SemiCRF**
$F(X,Y)=\sum_{i=1}^{M}\mathrm{f}(i,X,Y)$	$G(X,S)=\sum_{i=1}^{N}\mathrm{g}(i,X,S)$
$P(Y\mid X,W)=\frac{1}{Z(X)}\exp(W\cdot F(X,Y))$	$P(S\mid X,W)=\frac{1}{Z(X)}\exp(W\cdot G(X,S))$
$Z(X)=\sum_{Y'}\exp(W\cdot F(X,Y'))$	$Z(X)=\sum_{S'\in\Delta}\exp(W\cdot G(X,S'))$
Y' 为任意 可能序列	Δ 为任意可能的段分割方法

假设：

- ✓ 语料中段的最长的值是 K，
- ✓ $S_{1:i,y}$ 表示所有可能的从 1 到 i 的段序列且最后的段输出为 y。
- ✓ V(i,y) 表示 P(S'|X,W)中的最大值，根据表 4-1 知，它也是 W·G(X,S'), S'∈$S_{1:i,y}$.中的最大值。

有了以上假设，可以给出类 Viterbi 最优段序列搜索算法如下：

算法：类 Viterbi 最优段序列搜索算法：

第一步 初始化:

令 $V(i,y)=0,\ for\ i=0$

第二步 迭代:

$for\ i>0$

$$V(i,y)=\max_{y',k=1,\ldots,K}V(i-k,y')+W\cdot\mathrm{g}(y,y',x,i-k+1,i)$$

$for\ i=0$

$$V(i,y)=0$$

$for\ i<0$

$$V(i,y)=-\infty$$

第三步 停止，回溯:

$$bestSeg = \max_{y} (|V|$$

很明显，这个类 Viterbi 算法相对于普通 Viterbi 要复杂，分析算法可知，其复杂度是普通 Viterbi 的 K 倍，这是线性增长的[Sarawagi, 2005]。

（3）参数估计（训练）

对于给定的训练数据 $T=\{(x_l,s_l)\}_{l=1}^{N}$，训练的目标是找到使 $P(S\mid X,W)$ 最大的 W，也就是使下面的对数似然函数最大：

$$L_W=\sum_l \log P(S_l\mid X_l,W)=\sum_l (W\cdot G(X_l,S_l)-\log Z(X_l)) \tag{4-6}$$

上式是凸的（convex），这样的函数可以用很多方法求得最大值，如梯度下降法等。对 L 求导后得：

$$\begin{aligned}\nabla L(W)&=\sum_l G(x_l,s_l)-\frac{\sum_{s'} G(x_l,s')e^{W\cdot G(x_l,s')}}{Z_W(x_l)}\\&=\sum_l G(x_l,s_l)-E_{P_{rs}(W\mid)}G(x_l,s')\end{aligned} \tag{4-7}$$

求解这个方程第一项比较容易，但求第二项归一化因子时需要借助于类似于前向、后向算法的动态规划算法，并利用 G 的马尔可夫性。首先定义：

$$\alpha(i,y)=\sum_{\mathbf{s}'\in\mathbf{s}_{1:i,y}} e^{\mathbf{WG}(\mathbf{s}',\mathbf{x})} \tag{4-8}$$

则对于 $i>0$，并给出 $\alpha(0,y)=1$ 且 $\alpha(i,y)=0,\ for\ i<0$ 有如下迭代式

$$\alpha(i,y)=\sum_{d=1}^{L}\sum_{y'\in Y}\alpha(i-d,y')e^{\mathbf{Wg}(y,y',\mathbf{x},i-d+1,i)} \tag{4-9}$$

这样，就有：

$$Z_{\mathbf{w}}(\mathbf{x})=\sum_y \alpha(|\mathbf{x}|,y) \tag{4-10}$$

同理，我们给出分子的计算方法，对 G 中第 k 个成份，设：

$$\eta^k(i,y)=\sum_{\mathbf{s}'\in\mathbf{s}_{1:i,y}} G^k(\mathbf{s}',\mathbf{x}_l)e^{\mathbf{WG}(\mathbf{x}_l,\mathbf{s}')} \tag{4-11}$$

则有如下迭代公式：

$$\eta^k(i,y)=$$

$$\sum_{d=1}^{L}\sum_{y'\in Y}(\eta^k(i-d,y')+\alpha(i-d,y')g^k(y,y',\mathbf{x},i-d+1,i))e^{\mathbf{Wg}(y,y',\mathbf{x},i-d+1,i)} \quad (4\text{-}12)$$

这样最终有：

$$E_{\Pr(\mathbf{s}'|\mathbf{W})}G^k(s',\mathbf{x})=\frac{1}{Z_{\mathbf{W}}(\mathbf{x})}\sum_{y}\eta^k(|\mathbf{x}|,y) \quad (4\text{-}13)$$

4.3.2 特征空间

文献[Shen, 2007]采用了 CRF 模型进行摘要生成，为了验证 SemiCRF 模型在摘要问题上的性能，并方便结果比较，本文的特征空间主要在 Shen 的框架下构建。

表4-2 SemiCRF 特征

编号 \ 方法	**semi-CRF**	**CRF**
1	**Ex_Position**	Position
2	**Ex_Length**	Length
3	**Ex_Log_Likelihood**	Log Likelihood
4	**Ex_Similarity_to_ Neighboring_Segments**	Similarity to Neighboring Sentences
5	**Ex_Segment_ Length**	*
6	*Thematic*	Thematic
7	*Indicator*	Indicator
8	*Upper Case*	Upper Case

4-2 表中间一列给出了本方法所采用的所有特征，为了方便比较，将[Shen, 2007]中所采用的 CRF 的特征也列出，并命名为 Regular Features。中间一列中以 Ex 开头的黑体字特征是相应的为了应用于 SemiCRF 而扩展过的特征，Extended Features。本方法中也同样采用了与 CRF 一样的特征，即中间一列中没有以 Ex 开头的特征。列表中的星号表示在普通 CRF 方法中没有对应的特征。

下面给出这些特征的详细介绍：

Extended Features（扩展特征）

1) Ex_Position：扩展的位置特征，表示当前段（Segment）处在段落（paragraph）中的位置。如果当前段包含有段落的起始句，则该特征为 1；如果包含有结尾句，则为 2；其他情况为 3;
2) Ex_Length：去除停用词后该段中的词数。
3) Ex_Log_Likelihood：当前文本的概率模型里生成当前段的对数似然度。其定义如下式：

$$\log p(s_i \mid D) = \sum_{w_j} N(w_j, s_i) \log p(w_j \mid D) \tag{4-14}$$

其中，$N(w_j, s_i)$表示词 w_j 在段 s_i 中出现的次数，我们使用

$$p(w_j \mid D) = N(w_j, D) / \sum_{w_k} N(w_k, D) \tag{4-15}$$

来估计一个词从文本文本里产生的概率。

4) Ex_Similarity_to_Neighboring_Segments：计算一个段何其临近的段之间的基于 TF×IDF[Frakes, 1992]的余弦相似度。与 Shen 的方法不同的是，这里只考虑了和该段紧邻的段。
5) Ex_Segment_Length：这个特征在 CRF 中没有对应的，它给出当前段中句子的数量。

以上的特征，除了第 5)个，都是 Shen 在 CRF 模型中采用特征的扩展。可以看出，如果将段长度设为 1，则这些扩展特征就和原特征一样了。

有些在 Shen 的方法中使用的特征，对其扩展后对于结果并没有帮助，所以本方法还是用其原型。这些特征称为一般特征(Regular Features)。

Regular Features（一般特征）：

6) 专题词（Thematic）：去除停用词后，文本中出现的高频词。这个特征给出了一个句子中包含专题词的个数。
7) 提示词（Indicator）：诸如“conclusion”、“briefly speaking”等词对于是否是摘要句有提示作用。用这个特征来表示一个句子中是否含有这种类型的词。
8) 大写词（Uper Case）：句子中的大写词语很可能是专有名词，包含有这种词的句子很可能是作者希望强调的，这样的句子有可能是摘要句。所以，我们采用这个特征表示一句话中是否（除了开头）含有大写词语。

4.4 实验

4.4.1 实验准备

为了验证本章的方法，这里采用广泛使用的 DUC2001 语料。DUC（Document Understanding Conference）评测由 DARPA 资助，NIST 举办，从 2001-2007 年每年举行一次。DUC 的主要宗旨是让学者们能够参与到大规模的实验中去，并进一步推动摘要的发展。我们使用的 DUC2001 语料包含 147 篇新闻文本，每一篇都由人工标注出了是否为摘要句。使用这一语料的另一个原因是本实验的基线实验为 Shen 的 CRF 摘要方法[Shen, 2007]，而他的实验也是在这个语料上实现的。这样我们基于段序列标注的模型能更好的和基于序列标注的模型进行比较。

本次实验采用的评测标准为 F1 值，其定义如下式：

$$\mathrm{F1}=\frac{2\times \mathrm{Precesion}\times \mathrm{Recall}}{\mathrm{Precesion}+\mathrm{Recall}} \tag{4-16}$$

我们还采用了十折交叉验证来降低模型训练的不确定性。

以上这些步骤均严格按照 Shen 的方法，以方便比对。

4.4.2 实验结果

我们的实验目的是测试 SemiCRF 模型用于摘要问题的效果。正如 4.3.2 节提到的只有当段长度大于 1 的时候扩展特征才起作用，所以，扩展特征可以和表 4-2 右栏的一般特征一起组合。

为了测试那些扩展特征各种组合后提取摘要的效果，我们设扩展特征集合为 A={1,2,3,4,5}，其子集为 A_i（i=1,…,32)；一般特征集合 B={1,2,3,4,5,6,7,8}。则我们的第 i（i=1,…,32）次实验所采用的特征为： A_i+B/ A_i。举例来说，取扩展特征集合{1,2,3,4,5}的子集{1,3}，则取一般特征 B 中的{2,4,5,6,7,8}。

表 4-3 给出了实验结果，最左边一列为行号。其中为斜体字的第 1，3，5，7，9，11，13 行为所采用的 A_i（如“*1+2*”表示采用第一个和第二个扩展特征，其余特征采用一般特征）。第 2，4，6，8，10，12，14 行为实验 F1 值结果，“*All*”对应的单元为采用所有扩展特征的结果。而“*CRF*”对应的单元为全采用一般特征对应的结果。

表4-3 SemiCRF 实验结果

1		*1*	*2*	*3*	*4*	*5*
2	F1	0.395	0.391	0.398	0.394	0.392
3		*1+2*	*1+3*	*1+4*	*1+5*	*2+3*
4	F1	0.395	0.396	0.396	0.395	0.382
5		*2+4*	*2+5*	*3+4*	*3+5*	*4+5*
6	F1	0.389	0.384	0.398	0.399	0.380
7		*1+2+3*	*1+2+4*	*1+2+5*	*1+3+4*	*1+3+5*
8	F1	0.398	0.397	0.393	0.403	0.402
9		*1+4+5*	*2+3+4*	*2+3+5*	*2+4+5*	*3+4+5*
10	F1	0.402	0.403	0.401	0.403	0.404
11		*1+2 +3+4*	*1+2 +3+5*	*1+2 +4+5*	*1+3 +4+5*	*2+3 +4+5*
12	F1	**0.407**	0.404	0.406	0.402	0.404
13		*All*	*CRF*			
14	F1	0.406	0.389			

表4-4 与非监督方法的比较

	LSA	*HITS*	*Seim-CRF*
F1	0.324	0.368	0.407

表 4-4 给出了 SemiCRF 最好结果与一些经典的非监督方法的比较，包括 LSA[Frakes, 1992]和 HITS[Mihalcea, 2005]

从表 4-3 可以看出，如果单个的加入扩展特征，则实验结果较不加入扩展特征的（CRF 的结果）略有提高，而提升最大的是特征 3（第二行，第四列）。最好的结果是结合扩展特征 1，2，3，4，这时的绝对提高为 1.8%，相对提升为 4.6%。另一

个可以观察到的现象是，虽然有少数几个扩展特征组合会降低系统表现，但绝大多数都是有益的。这样，我们的实验结果充分说明了采用段序列标注模型，SemiCRF，来对摘要问题建模的有效性；以及扩展特征在该模型下的有效性。

从表 4-4 可以看出，我们的 SemiCRF 模型效果远超过了另外两个经典的非监督学习的摘要方法。相对于 LSA 和 HITS 分别有 8.3%和 4.9%的提高。

SemiCRF 虽然提供了一种可计算（tractable）的高阶 CRF 扩展，但是它的复杂度随段长度的增加也在急剧增加，而且随着段的增加特征也会变的稀疏。我们的实验里，SemiCRF 的训练时间为 CRF 的 20 倍左右，测试时间约为 10 倍。

4.4.3 实验结论

我们的实验框架以及具体细节均是完全比照着 Shen[Shen, 2007]的 CRF 模型来做的，Shen 的方法也作为我们的一个重要的基线实验。这里我们虽然采用相同的特征名称，但是 SemiCRF 具有可以缩放粒度的特性，这使得同样的特征在 SemiCRF 下比 CRF 等模型更为丰富。实验中最好的扩展特征组合取得了比 CRF 方法 4%的相对提高。并且其它的组合也绝大多数会提升系统的性能。从结果看，采用段序列标注方法的 SemiCRF 模型的确可以通过整合扩展特征将 CRF 的标注能力进一步放大，提高摘要效果。这也就证明了我们的假设：以句子为粒度获取特征的稀疏性可以通过从以段为粒度获取特征的段序列标注模型来获得一定的弥补。

SemiCRF 的主要缺陷是它的速度。使用扩展的特征空间，则训练和测试的搜索范围都将大大增加。文献[Sarawagi, 2005]从理论上证明，SemiCRF 的搜索速度随着粒度的增大呈线性增长，即便对于大型语料来说，这个复杂度还是可以接受的。

进一步的工作，主要应该从继续寻找可扩展的特征入手，以及寻找更合适的语料进行更加细致的实验。

4.5 本章小结

在本章的工作里，我们首先针对摘录型摘要，分析了传统方法中以句子为单位抽取特征时存在的稀疏性问题，这种问题对于摘录型摘要普遍存在，而且采用传统的点标注或序列标注模型都是无法克服的。以往人们只能通过对句子中的词语做语义相似度扩展来弱化这种稀疏。

针对这个问题，我们提出了采用段序列标注模型 SemiCRF 来对摘要进行建模的

思想。SemiCRF 已经在命名实体识别上取得了不错的效果，我们借鉴这种思想，并应用到摘要中。采用 SemiCRF 模型，可以将特征粒度从一个句子扩大到多个句子，从而在一定程度上减弱稀疏对特征抽取的影响。

通过以 CRF 为基线的实验，我们证明了采用 SemiCRF 模型的有效性。从而为提取摘要特征，以及摘要建模提供了一条新的途径。

第五章　基于排序学习的摘录型自动摘要建模方法

5.1　引言

基于句子抽取的摘要方法可以采取很多种策略。传统上，一般将其作为分类或者是回归问题来建模，然后对句子进行打分、排序。如前面章节所述，在这一框架下，人们已经尝试了很多的机器学习方法来解决这一问题。

将这一问题作为分类或者回归问题来研究有其合理性，但是这种框架割裂了同一篇文本中句子与句子之间的比较关系。换句话说，模型训练好后，这种传统框架强调的是当前待分析的各个句子在已有模型之下的表现，根据这个表现来得到句子之间的孰优孰劣。这个“已有模型”，是通过训练集中的文章中提取的特征训练出来的。可是摘要问题同样需要关注的一点是当前句子在当前文本下，与该文本中其它句子之间的关系。对于和“已有模型”差别较大的文章，这种相对关系显得更为重要。

排序学习（Learning to Rank，LTR）[Joachims, 2003; Liu, 2009]的思想正好弥补了传统模型这方面所欠缺的考虑。近几年来，LTR 成为信息检索领域的一个新的热点，已经有一些学者开始尝试用这个方法来做基于查询的摘要。

本章所要介绍的工作的创新性在于：首先我们探索了基于排序学习的通用型摘要生成方法，并证明其行之有效；另外，我们对于一些排序学习中经典的逐对（Pair-Wise）方法，如 RankingSVM[Joachims, 2003]、RankingBoost[Freund, 2003b]，以及较新的逐列（List-Wise）方法，如 SVMMAP[Yue, 2007]、ListNet[Cao, 2007]进行了比较，得出 SVMMAP 等逐列算法在摘要问题上效果更加优越的结论。

5.2　相关工作

我们先简要回顾摘录型摘要的主要方法，主要介绍传统的摘要建模方式和以 LTR 为摘要建模的方式。然后，我们介绍 LTR 在近些年的发展，包括 Pair-Wise 方法和 List-Wise 方法的发展历程。

5.2.1 摘录型摘要回顾

本章引言中指出，传统的摘录型摘要建模的方法将其看作普通的分类或者回归问题。在这个框架下寻找合适的机器学习算法来使用各种摘要问题的特征。诸如贝叶斯分类器[Kupiec, 1995]，决策树[Lin, 1999]，最大熵[Osborne, 2002]，条件随机场[Shen, 2007]，半条件随机场[Wu, 2008]，遗传算法[Yeh, 2005]等等。

文献[Wang, 2007]指出，这种传统框架下，所有来自不同文章的句子都应在所提取的分类特征信息下是可比较的。这就是说，句子是在来自训练集文档里训练出来的模型下进行比较的。

具体到摘要问题上，假设将之看作二分类问题，也就是“是”或“不是”摘要。传统的方法就是去计算“是”的可能性大，还是“不是”的可能性大。而所依据的标准，就是用特征表示的当前句子，在现有模型下的打分。这样的模型在处理相似（比如类似的风格、类似的词汇等等）的文章的时候比较合理，但是如果待处理的文章和模型训练所用的文章有较大差异的时候，其合理性就要大打折扣了。

针对传统建模框架的不足，文献[Wang, 2007]提出了采用排序学习的思想来处理基于查询的摘要问题。采用这种方法的出发点是让同一篇文章里的句子更侧重于相互间的比较，从而对于文本间的不一致性有比传统的分类方式得到的模型更强的鲁棒作用。文献[Wang, 2007]采用的排序学习方法是 RankingSVM[Joachims, 2003]，Wang 认为，RankingSVM 在摘要问题上能起到很好的作用。因为在这个方法中，句子的分值是通过同一篇文本中句子的相互比较得到的。另一个采用类似方法的工作来自[Metzler, 2008]，Metzler 采用的方法是梯度自举决策树（Gradient Boosted Decision Tree, GBDT）[Li, 2007]。

Wang 和 Metzler 的方法都仅仅研究了基于查询的摘要问题，而没有考虑通用型的摘要问题，并且，他们对于众多排序学习方法在摘要问题中的表现也缺乏细致的研究结果。

5.2.2 排序学习方法回顾

近几年来，在信息检索和机器学习领域，如何从训练数据中得到模型，然后根据相关度、倾向性以及重要性排序的问题，也就是排序学习引起了越来越多的关注。排序学习除了在信息检索领域有着重要的作用，在其他领域也有广泛的应用，如在机器翻译领域里，如何给可能的翻译结果排序；在计算生物学的蛋白质结构预测里，如何给可能的 3 维结构排序等。

传统的排序方法，如 BM25 或者信息检索里的语言模型，都在灵活性上有所欠缺。随着机器学习算法的发展，排序问题更多地侧重于如何从已有的标注数据中得到模型的方法。

现有的排序学习方法大致可以分为三种：逐点的（Point-Wise）、逐对的（Pair-Wise）以及逐列的（List-Wise）。逐对的方法是用针对单个对象采用回归或者分类的方法来解决排序问题。逐对的方法是将排序问题转换成对象对的分类问题，也就是先将对象划分成对，然后逐对采用分类算法[Joachims, 2003]。逐对和逐点方法都采用现有的分类学习算法，这样得到的排序模型是间接的。逐列的方法采用的是全新的思路，直接对排序问题建模，它得到的是直接的排序模型[Cao, 2007; Xia, 2008]。因为逐点法和传统的排序方法没有太大区别，本章的工作只考虑了逐对和逐列方法。本节剩余部分给出这两种方法的简要介绍。

（1）逐对方法

通过将优化目标设置成归一化减值求和增益（Normalized Discount Cumulative Gain，NDCG）以及 K 位精度（Precision at K）等，人们采用各种机器学习算法如：罗杰斯特回归（Logistic Regression）、支持向量机、神经网络以及感知器等得到的排序模型可以划归为逐对方法。比较知名的方法有采用支持向量机作为分类器的 RankingSVM[Herbrich, 2000]；采用自举（Boosting）算法分类的 RankBoost[Freund, 2003a]；采用基于交叉熵作为损失函数，并用梯度下降算法训练获得的神经网络模型的 RankNet[Burges, 2005]。

逐对方法框架的核心思想是：将待分类的目标对分为“正确排序”或者“错误排序”两类。通过这种分类最终获得正确的排序。在信息检索领域，这种思想已经有了不少的应用[Joachims, 2003; Burges, 2005]。

采用逐对的方法有一定的优势。比如，已知的机器学习模型可以方便的服务于这个框架；并且，用于训练的实例（也就是目标对）很容易获得。但它的缺点也非常明显：其一，最小化目标对的分类错误和最小化所有目标的排序错误有着本质的区别；再者，在逐对方法的框架下，要求目标对独立同分布，而这个假设对于很多问题是个过强的约束。

（2）逐列方法

与逐对方法不同，逐列方法的思想是将整个目标序列作为训练对象，并在该序列上定义损失函数（用真实排序和现有排序作为输入，定义损失）。

[Cao, 2007]是最早提出这种思想的文献之一，作者将其方法命名为 ListNet。ListNet 采用文章列表而不是文章对作为学习实例。损失函数定义为真实排列和预测排列的参数化的概率分布之间的交叉熵。Cao 通过实验验证，认为 ListNet 比传统的逐对算法性能更为优越。另一个算法 RankCosin[Qin, 2007]和 Cao 的方法比较类似。

SVMMAP[Yue, 2007]，采用中间平均精度（Mean Average Precision，MAP）作为排序问题的优化目标。因为 MAP 本质上是针对整个文章列表计算的（不是在文章对或者单个文章上），所以也将该方法划入逐列方法中。Yue 在文献中指出，SVMMAP 在排序问题上有较好的效果，并且有较快的学习速度。

因为排序问题本质上就是应该以序列为考察对象的，而逐列方法因为是针对整个目标序列的优化过程，所以从理论上就胜过了逐对方法一筹。在实际应用中也有很多逐列方法声称取得了比逐对方法更好的结果，但是这种方法的优化算法很不容易获得，其算法一般比逐对方法复杂的多。

5.3 本项研究工作的动因

在这一部分，我们先给出在摘录型摘要问题中采用排序学习的思路；然后，简要介绍在本章中将要考察的一些经典的逐对以及逐列的排序学习算法的定义；最后给出本章用于模型计算的特征列表。本章的工作主要侧重于对于排序算法在摘要问题中的优劣研究，所以采用的都是常用的摘要特征。

5.3.1 排序学习和摘要

有监督的摘录型摘要可以被看作句子排序以及选择问题。训练数据是标注了摘要信息的文本，也就是一句话是否是摘要，测试部分是对于新文本中的句子给出合适的标注。

本章 5.2 小节提到，采用排序学习算法来对句子进行排序并生成摘要，从直观上讲，对于不同类型的文本更具备合理性。图 5-1 给出了传统的摘录型摘要方法和采用排序学习方法的不同思路的示意图。从图中可以看出，传统方式下模型更侧重横向信息的比较，即让各种分类信息在该模型下相互竞争；而排序学习方式下，更侧重的是同一文章间不同的句子之间的竞争关系。

排序学习一般使用在基于查询的检索中。举例来说，排序学习的使用大致如下：给出一个文章集和一个查询语句，排序学习模型输出相应每篇文章的分值。这个分值一般指该文章和查询语句的相关度，相关度越强的分值越高。在训练过程中，给

出一定数量的查询语句和与查询语句相关的文本，以及相应的相关度，最终得到一个排序函数。训练所用的特征来自查询语句和文本组合成的单元。

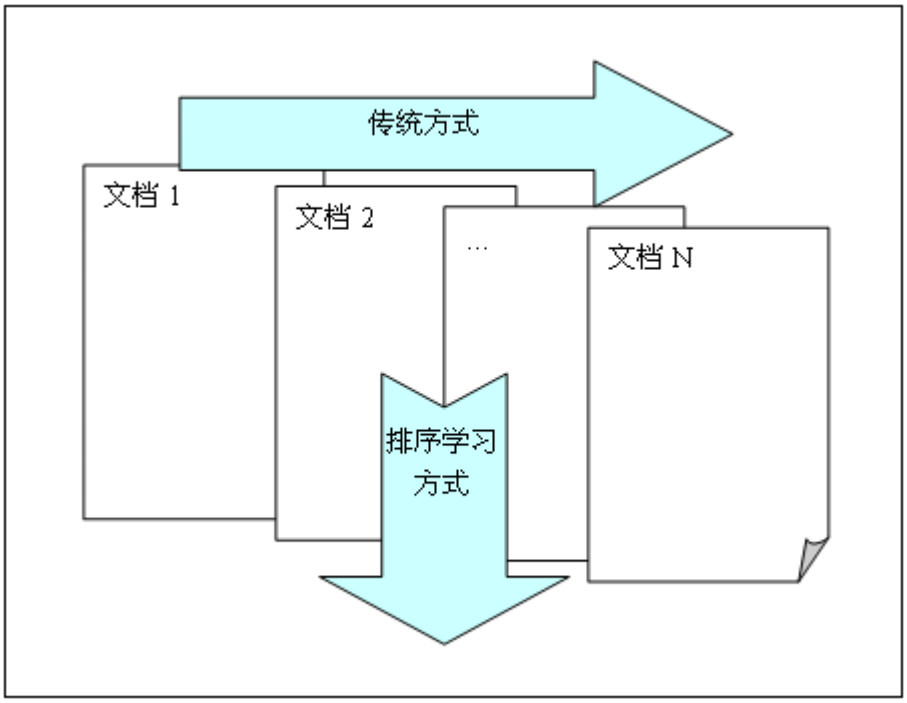


图5-1 基于排序学习的摘要方法和传统摘要方法区别示意图

这个模式可以方便地应用到基于查询语句的摘要生成问题中去，所需的变换是：将检索中的以文章为单位换为以句子为单位；将检索中的相关变成摘要中的是或不是摘要（“是”为相关，“不是”为不相关）。文献[Wang, 2007; Metzler, 2008]所采用的方法属于这个思路。本文考虑的是通用型摘要问题，我们将查询部分省去，直接从句子上提取特征。

5.3.2 排序学习算法

在这一小节中我们介绍本章工作所用到的排序学习算法，包括逐对的 RankingSVM、RankBoost 以及逐列的 ListNet、SVMMAP。我们使用这些算法针对通用型摘要问题进行句子排序并抽取。因为所介绍的排序学习算法大都采用支持向量机（SVM）[Vapnik, 2000]作为框架，在第三章里我们已经简要介绍了 SVM，这里我们做一个简单简要回顾 SVM 算法，并采用 SVM 作为基线。

（1） 支持向量机（SVM）

SVM 以其卓越的分类效果，优秀的泛化能力而有着广泛的应用。其核心思想是

寻找距离最大的超平面将训练数据分成两类。

最优分类面也可以看成下式的解

$$\begin{aligned} &to\ \ Minimize: 1/2\times\|\mathbf{w}\|+C\sum\nolimits_{i=1}^{n}\xi_i\ (for\ \mathbf{w},\xi\geq 0)\\ &subject\ to: y_i(\mathbf{w}\bullet x_i)\geq 1-\xi_i(for\ \forall i\in n)\end{aligned} \tag{5-1}$$

ξ_i 是松弛变量，当超平面分类错误的时候，其值大于 1。C 为惩罚因子，控制正规化的总量值。SVM 是一种泛化的线性分类器，有着同时最小化经验分类错误和最大化几何边距的能力。

后面介绍的算法是本章工作所采用的排序学习算法，这里我们先给出排序学习的一般性的定义：

排序学习是一种监督或者非监督的机器学习方法，其目标是从训练数据中自动创建一个排序模型。训练数据包括数个部分有序（partial order）的元素列表。这种“有序”一般是指对于一个特定元素的数值分值或序数号，还有一种是给定一个元素的二元判断（比如相关或不相关）。排序模型的用途是对新的未知的元素列表生成“类似于”训练数据的一种排序。

下面介绍具体的一些排序学习算法。

（2）逐对排序学习方法

1） RankingSVM

RankingSVM 可以看成 SVM 的一种一般化的表示，其模型的训练数据不是标注成二元类别，而是给出逐对的倾向性（preference）比较。一般认为，RankingSVM 在排序问题上隐含着对数据的结构上的考虑，而这个是 SVM 所欠缺的。

RankingSVM 的简单的形式化定义如下：

$$\begin{aligned} &to\ \ Minimize: 1/2\times\|\mathbf{w}\|+C\sum\nolimits_{i,j}\xi_{i,j}\ (for\ \mathbf{w},\xi\geq 0)\\ &subject\ to: (for\ \forall i,j\in n)\ \ (\mathbf{w}\bullet x_i-\mathbf{w}\bullet x_j)\geq 1-\xi_{i,j}\end{aligned} \tag{5-2}$$

这里的 n 代表逐对的倾向性集合的大小，权重向量 $\mathbf{w}$ 是要从训练数据中得到的模型。得到训练的模型后，新的句子 x 的分值可以通过 $\mathbf{w}\bullet x$ 来计算，这个分值就可以用于排序的计算。RankingSVM 如今已近成为逐对排序算法中的经典算法并应用于信息检索领域。在排序问题上，其效果明显优于 SVM 算法。

2） RankBoost

文献[Freund, 2003b]给出了另一种逐对排序学习算法，采用 boosting 算法结合倾向性排序。和普通的 boosting 算法一样，RankBoost 反复循环。每次循环，从称为弱分类器的单独过程里得到一个弱的排序。其算法如下图 5-2

RankBoost 算法:

初始化: 设 D 为 $X \times X$ 的一个分布矩阵

第一步: $D_1 = D$

第二步: 令 t=1...T

用 D_t 训练弱分类器

得到弱分类器 $h_t : X \to \mathbf{R}$

选取 $\alpha_t \in \mathbf{R}$

更新

$$D_{t+1}(x_0, x_1) = \frac{D_t(x_0, x_1)\exp(\alpha_t(h_t(x_0) - h_t(x_1)))}{Z_t}$$

这里 Z_t 是归一化因子

第三步 3: 输出结果

图5-2 RankBoost 算法

$D_t(x_0,x_1)$表示的是排序单元 x_0 排在单元 x_1 前面的概率分布，这个概率在第 t 轮传递给弱分类器。对于每一对实例，其值越大，说明弱分类器将该对实例正确排序的可能性越大。每轮迭代，产生一个新的弱分类器更新分布矩阵。每次更新过后，如果弱分类器对于实例正确分类，则 $D_t(x_0,x_1)$变小，否则变大。最终结果，$D_t(x_0,x_1)$会集中在排序最难确定的那些实例对上。最后的排序 H 是所有弱排序 h_t 之和。对于 α_t 的选择，请参阅[Freund, 2003b]。

（3）逐列排序学习方法

在逐对方法里，我们可以看到每个特征向量对形成了一个新的实例。而排序问题的实现，实际上是由对这些新实例的分类问题转化而来。而在逐列方法中，损失

函数反映的是训练数据上的整体损失。

1）ListNet

文献[Cao, 2007]给出了一种称为 ListNet 的排序学习算法。ListNet 的损失函数是定义在数据列上的。其模型采用的是神经网络，并用梯度下降方法求解。我们在图 5-3 给出了算法。

ListNet 算法:

初始化: 输入训练数据

$$\{(x^{(1)},y^{(1)}), (x^{(2)},y^{(2)}),..., (x^{(m)},y^{(m)})\}$$

第一步: 设置神经网络参数ω

第二步: 令 t= 1 to T

令 i = 1 to m

使用当前ω 计算新的分值列表 z_{ω}^{i}

使用梯度下降法计算损失函数和ω相关的梯度，然后得到 $\Delta\omega$.

更新 $\omega = \omega - \eta\Delta\omega$

Step 3: 输出最终的 ω

图5-3 ListNet 算法

算法中输入的 $x^{(i)}$代表从列上获得的特征，相应的 $y^{(i)}$ 代表真实的排序。

2）SVMMAP

中间平均精度（Mean Average Precision, MAP）是一种信息检索里广泛采用的评估指标。在 SVMMAP[Yue, 2007]之前已经有了一些以 MAP 为优化目标的排序学习算法，但是这些方法均不能达到全局最优解，并且复杂度较高。文献[Yue, 2007]给出的这种 SVMMAP 算法以 SVM 为基础，针对 MAP 做优化。它可以找到最优解，并且速度较快。

MAP 可以形式化的定义成：

$$MAP(r1,r2)=\frac{1}{R}\sum_{i:r1_i=1}\Pr ec@i \tag{5-3}$$

这里的 *r*1 和 *r*2 代表给定的文章集（或者别的排序对象）的两种不同的排序。(5-5)式中，*r*1 代表真实排序，*r*2 代表需要被评估的系统输出。R 代表被认为是相关文章的总数，这些相关文章被标注为 1。Prec@*i* 表示 *r*2 在 *i* 位置以前被正确标注为 1，也就是相关，的百分数。

与给予不同位置上的错误相同的惩罚的 ROCArea 方法不同的是，从(5-5)中可以看出，MAP 对于高位置上的错误给予更高的惩罚。

以 MAP 为损失目标的优化函数的主要困难是，它无法分解成独立计算出来的，描述正反例对的相互关系的数值之和。Yue 在文献[Yue, 2007]里巧妙地利用了 MAP 损失函数有交换两个相同标注且不改变数值的特点。而这也赋予了 SVMMAP 算法逐列的属性。那些以 F1 值或 ROCArea 等为优化目标的算法不具备这个性质。

Yue 指出，SVMMAP 的复杂度是迭代次数的多项式时间，而每次迭代复杂度在 *O*(n*log*n)。

5.3.3 特征空间

对于摘录型摘要问题，学者们已经尝试了许多特征。本章的工作旨在测试新模型的性能。并且，为了与以前的模型，如[Shen, 2007; Yeh, 2005]的比对更有说服力，我们采用了一些最常用的摘要特征。这些特征只需要很少量的计算就可以得到。

位置特征（Position）：如果该句是在当前文章的开始，则其值置为 1；如果在文章结尾则置为 2；其它位置置为 3。

长度（Length）：移除停用词后句子的长度。

对数似然（log_likelihood）：当前句子由所在文本产生出来的对数似然值。我们使用公式(5−6)来计算该值。$N(w_j, x_i)$ 代表第 w_j 个词在第 x_i 词句里出现的次数。我们采用 $N(w_j,D)/\sum_{w_k}N(w_k,D)$ 来估计一个词从文章里产生的概率。

$$\log P(x_i \mid D)=\sum_{w_j}N(w_j,x_i)\log p(w_j \mid D) \tag{5-4}$$

与邻近句的相似度（Similarity to neighboring sentence）：我们定义句子间的相似度为基于 TF×IDF[Frakes, 1992]上的余弦相似度。所考虑的邻近句子为上下个三个。

语义词（Thematic）：移除停用词后，我们定义出现频率最高的词为语义词。这个特征给出了每句话中语义词的数量。

指示词（Indicator)：我们认为语句中如果出现诸如“conclusion”、“briefly speaking”等词，那么该语句成为摘要句的可能性会增大。我们用这个特征表示一句话中是否含有这种词。

大写词（Uppercase）：语句中的大写词很可能是一些名词，另外作者希望强调的词语也可能会采用大写形式。所以我们认为含有大写词的语句成为摘要句的可能性会增大。我们用这个特征表示一句话中是否含有大写词。

5.4 实验

为方便与以前的工作进行比较，本次实验我们依然采用 DUC2001 语料。我们将 DUC2001 语料中标注的摘要句看作基于查询的文本检索中的相关文本，非摘要句看作不相关文本。

我们采用（5-7）式定义的 F1 值和 ROUGE-2 作为评估指标。对于 F1 值，将手工标注的摘要句看作正确值 C；系统输出作为 S。

$$P=\frac{|C\cap S|}{S}\quad R=\frac{|C\cap S|}{C}\quad \mathrm{F1}=\frac{2PR}{P+R} \tag{5-5}$$

ROUGE-2[Lin, 2004]是另一个较广范采用的自动文摘评测标准。文献[Lin, 2004]认为 ROUGE 和人的评估结果具有较高的相关性。因为 ROUGE-2 具有良好的性能以及简单的形式，我们采用它作为评测标准。

实验采用十折交叉验证来减小模型训练的不确定性。最终的 F1 值和 ROUGE-2 值均为十折后的平均值。

5.4.1 基线实验

我们的实验采用两个基线作为参照，一个是逐对排序学习方法：包括 RankingSVM 和 RankBoost；另一个是广泛使用的传统的有监督摘要模型：包括 SVM 方法和条件随机场（CRF）方法。在相同的特征条件下，CRF 方法是已知的传统模型中效果最好的，我们也将其结果列于表 5-1 中。

5.4.2 实验结果

表 5-1 中给出了本章实验的主要结果。无色区域中的是传统模型，包括 CRF 和 SVM；浅色区域中的是逐对排序学习模型，包括 RankingSVM 和 RankBoost；深色区域为逐列方法，包括 ListNet 和 SVMMAP。从表中可以清楚的看到，我们的基于逐列的排序学习算法在两个评测指标上均取得了不错的成绩。

表5-1 实验结果

	CRF	***SVM***	***Rank SVM***	***Rank Boost***	***ListNet***	***SVM MAP***
ROUGE-2	0.455	0.417	0.434	0.431	0.454	**0.460**
F1	0.389	0.343	0.372	0.375	0.386	**0.394**

与 SVM 相比，逐对方法中的 RankSVM 在 F1 值上有大约 8.5%的提高，在 ROUGE-2 上有 4%的提高。另一个逐对方法 RankBoost 分别提高了 9.3%和 3.5%。逐列的方法中，ListNet 分别提高了 9%和 12.5%；SVMMAP 在 ROUGE-2 和 F1 上分别提高了 10%和 14%。

使用相同的特征，并使用相同的语料，已知最好的方法是 CRF[Shen, 2007]。在与 CRF 的比较中，可以看出 SVMMAP 也取得了微小的提升。

在表 5-2 我们给出了逐列方法与逐对方法之间的比较。从表中可以看出，我们选取的这两种经典的逐列方法在两个指标上的性能均优于另两种经典的逐对方法。

5.4.3 结果分析与展望

从实验结果可以看出，我们采用的逐对和逐列的排序学习算法比单纯的采用 SVM 的传统模型要优越。这也证明了强调同一篇文章内部之间句子的竞争在摘要问题上的重要性，以及采用排序学习模型的合理性。

表5-2 逐对方法和逐列方法的比较

ROUGE-2	***SVMRank***	***RankBoost***
ListNet	+4.6%	+5.3%
SVMMAP	**+6.0%**	**+6.7%**
F1	***SVMRank***	***RankBoost***
ListNet	+3.8%	+3.0%
SVMMAP	**+5.9%**	**+5.1%**

我们的方法除了 SVMMAP 超过了采用相同特征的 CRF 方法外，其余都略微低于 CRF 序列标注方法。分析其原因，我们认为：第一，CRF 模型提取非独立特征的能力较强；另外，CRF 作为一种序列标注模型，其最有序列搜索的过程一定程度上也体现了“内部竞争”的这种概念。

下一步更深入的工作，主要有两点：第一，采用更大规模的、写作风格差异也更大的语料，进行更深入细致的比对试验；第二，现有的排序学习算法有近 20 种，而且不断地在推陈出新，如何针对摘要问题建模，获得更合理的排序算法也是非常有益的尝试。

5.5 本章小结

在本章的工作中，我们采用了逐对和逐列的排序学习算法来进行通用型摘录型的摘要工作。

首先我们提出采取排序学习处理摘要问题的合理性分析，认为摘要问题的模型训练除了要考虑从历史数据训练出的模型外，也应该考虑当前待测试文本内部句子间的竞争关系。而排序学习恰恰能针对这个问题建模，并且从句对（逐对方法）以及序列（逐列方法）上抽取特征，进行排序。然后我们介绍了摘要的传统建模方法以及排序学习的发展历程。最后我们给出实验，证明了排序学习，尤其是逐列排序学习算法在通用型摘要问题上也有较优越的性能。

第六章　潜层主题特征提取

6.1　引言

自动文摘技术中特征的重要性不言而喻。除了句子位置，线索词等类似启发式的规则等重要特征，对词的分布规律特征的捕捉也是人们研究的重点方向。从词汇集（Vocabulary）到 TF×IDF，从 TF×IDF 再到潜层语义索引等等，无一不体现了人们希望从表层信息上挖掘出较深层语义的愿望。

潜层狄利赫雷分配（Latent Dirichlet Allocation，LDA）方法近年来被广泛应用于文本聚类、分类、段落切分等等，并且也有人将其应用于基于查询的无监督的多文档自动摘要。该方法因其完善的数学模型，清晰的语义层的定义，非常为学者们推崇，并且为人们提供了一个发挥想象的广阔的空间。

LDA 被普遍认为能较好地对文本进行潜层语义建模，但其在自动摘要中的作用尚少有研究。

本章的工作旨在研究 LDA 作为摘要特征时的性能。同时，我们还实现了基于 LDA 的有监督自动文摘方法。我们研究了 LDA 在单文档自动文摘中的作用，并分析研究了在不同 Topic 数量下 LDA 对摘要结果的影响。实验结果表明，加入 LDA 特征后，能够有效地提高以传统特征为输入的文摘系统的质量。

6.2　相关工作

在信息检索领域，词频特征及词的分布规律特征都是非常重要的。鉴于词汇集规模一般非常巨大，如何能有效的避免稀疏，如何能有效的对文本进行建模，并从模型中发掘潜在的词汇分布规律一直是学者们关心的问题。而进行降维一直是这个问题的主导解决思路。

通常使用的 TF×IDF 方法[Frakes, 1992]，是单纯地把基本词汇的频度与该词的稀疏度相结合作为该词的分值，它的降维贡献在于把任意长度的整个语料中的数据规模减小到定长的词汇集的级别。这种缩减虽然非常可观，但即使是词汇集的级别

仍是非常巨大的，并且文本内和文本间的词分布关系在这个降维中也没有反映出来。

潜层语义索引（Latent Semantic Indexing，LSI）[Deerwester, 1990]是从事信息检索研究的学者们进行的一个有意义的尝试，LSI 对 TF×IDF 的矩阵进行歧义值分解，由此构筑的线性空间比原空间的维度大大下降，并且被认为能捕捉到一些诸如同义词和多义词等基本的语言学特征。LSI 方法认为，每一篇文本都是一个主题的产生物，反映了这个主题（topic）。但显然这个假设有点过于硬性。

概率潜层语义索引（pLSI）[Hofmann, 1999]通过引入概率放松了这个限制。它认为文本中可以有多个主题（topic），每个词从主题中产生出来，而用主题的分布来表征这篇文本。这样表征文本的维数就从词汇集的量级降到了主题数的量级，这是一个重大的进步。pLSI 的缺陷主要在于，它不是个完备定义的文本生成模型，它的参数只和训练文本有关，理论上不能将其直接用在未见过的文本上。

近年来，继 TF×IDF、潜层语义索引（LSI）、pLSI 模型之后，潜层狄利赫雷分配（Latent Dirichlet Allocation, LDA）[Blei, 2003]引起了人们的重视。LDA 是一种包含三层的层级贝叶斯生成概率模型，它把文本语料看作离散数据，数据中的每一个元素看作是由底层的有限个混杂在一起的主题产生出来的，而每一个主题又被看作是从一个更底层的概率模型中产生出来的。LDA 克服了 pLSI 的理论缺陷，并且继承了 pLSI 的降维优势。

6.3 本项研究工作的动因

在 6.2 节中我们介绍了在信息检索中人们为了降维所作出的不同努力。而作为信息检索的重要分支之一的自动文摘往往借鉴信息检索中的成熟技术，所以上述各种技术在摘要问题上都有不同的应用（参见第二章）。

文献[Daume III, 2005]提出了用一种基于 LDA 的概率图模型的变型体而得到的主题作为特征，并将之应用于基于查询的多文档文本抽取摘要的方法，取得了较为明显的效果。但是 Daume 采用的是一种无监督的方法，而且对于主题数目对于摘要生成的影响也没有加以分析。

目前来看，LDA 模型在摘要问题上应用的还比较少，它的各种潜力还有待进一步挖掘、验证。这也是本章工作希望解决的问题。

我们下面首先给出 LDA 模型的介绍。

6.3.1 LDA 模型

向量空间模型、潜层语义检索模型，都是信息检索领域（IR）对于文本语料或其它离散的数据集的建模。这种建模一般都基于词袋（Bag-of-Words）的理念，从概率论的角度也就是同一篇文本里的单词或同一个数据集里的数据元素是可交换（exchangeability）的。

LDA[Blei, 2003]也是一个建立在词袋理念上的、三层的贝叶斯概率模型——一个对如文本语料等建模的生成概率模型。在这个模型中，语料中的每篇文本都认为是来自于一个有限的服从某种分布的主题（topic），而这个topic的分布能反映这篇文本的内容信息。

LDA模型是建立在狄利赫雷分布以及多项式分布的基础上的模型。多项式分布可以简单地理解为二项式分布的推广，这里不再介绍。狄利赫雷分布是伽玛分布的推广，而伽玛分布是二项分布的共轭分布，所以狄利赫雷分布是多项分布的共轭分布。在介绍LDA模型时，我们也会简单介绍狄利赫雷分布。

LDA用到的变量定义如下：

- 一个单词（最小的数据元素）用 w 表示， 词汇集（Vocabulary）是单词的总集，一个单词可以根据其在词汇里的位置唯一确定，其索引属于集合 $\{1,2,...,V\}$，这里的 V 也就是词汇的大小。
- 一篇文章是含有 N 个词的向量，用 $W=(w_1,w_2,...,w_N)$ 表示， w_n 代表文章中第 i 个单词。
- 一个语料是文章的集合，用 $D=\{W_1,W_2,...,W_M\}$， M 代表文章的数量。

有了定义好的变量， LDA对语料里的每篇文章 W 按如下方式建模：

1） 选择 $N \sim poison(\xi)$： N 服从参数为 ξ 的泊松分布
2） 选择 $\theta \sim Dir(\alpha)$： θ 服从参数为 α 的狄利赫累分布
3） 对于文章中的每个词 w_n
   a） 选择一个主题 $z_n \sim Multinomial(\theta)$： z_n 服从参数为 θ 的多项式分布
   b） 选择一个词 $w_n \sim p(w_n \mid z_n,\beta)$

这个模型有以下简化假设：主题的数目 k 是已知的；每个单词 w^i 在不同的主题 z^j 下的概率为 $\beta_{ij}=p(w^i \mid z^j)$，这个 β 就成为一个 $k \times V$ 的矩阵，这个矩阵需要用训练数据估计出来；最后，文本长度 N 和别的系数都没有关系，所以忽略其随机性。

一个 k 维的狄利赫累随机变量 θ 其取值在 $k-1$ 维的单纯形（simplex）上（满足 $\sum_{i=1...k}\theta_i=1$ 并且 $\theta_i\geq 0$），并且有如下的概率密度：

$$p(\theta|\alpha)=\frac{\Gamma(\sum_{i-1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}...\theta_k^{\alpha_k-1} \tag{6-1}$$

这里的 α 是一个 k 维向量且 $\alpha_i>0$，Γ 是伽玛函数（Gamma Function）。选用狄利赫累分布的原因是：它是指数族函数，并且有有限维充分统计量（sufficient statistics），再有就是它和多项式分布是共轭的（conjugate）。这些特性都有助于后面的推断和参数估计。当给定了模型参数 α 和 β，则代表着一篇文章里主题的分布的 k 维向量 θ，代表该文章中每个单词来源的主题集 N 维向量 Z，以及代表文章中的每个单词的 N 维向量 W 的联合分布如下面的(6-2)式：

$$p(\theta,Z,W|\alpha,\beta)=p(\theta|\alpha)\prod_{n=1...N}p(z_n|\theta)p(w_n|z_n,\beta) \tag{6-2}$$

可以用图6-1代表上式：

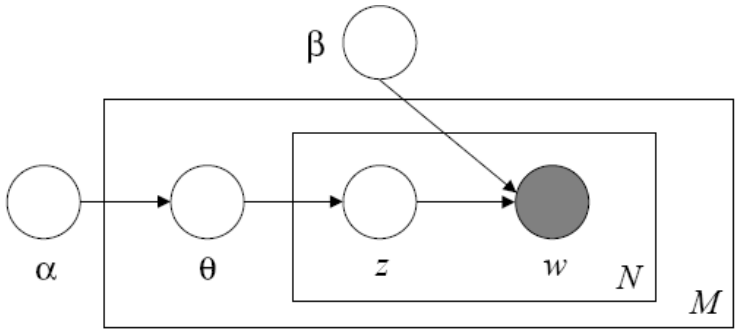


图6-1 LDA 概率图模型

图6-1中，矩形框代表“重复”，也就是可交换性，即Bag-of-Words的理念；实心圆代表可见变量，空心圆代表隐含变量。最外层的矩形代表语料，也就是 M 篇文本，里面的矩形代表每篇文本（这里假设都为 N 个单词）。

上式6-2中 z_n 所代表的那个主题为 i，则对于 k 维向量 θ，上式中 $p(z_n|\theta)=\theta_i$。如果对上式 θ 积分、Z 求和可以得到：

$$p(W|\alpha,\beta)=\int p(\theta|\alpha)\left(\prod_{n=1...N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta \tag{6-3}$$

这是一篇文章产生的概率表达式，也就是图中的内层矩形，如果要求的整个语料的概率表达式，则要再对语料中所有的文本联乘：

$$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d \tag{6-4}$$

从图中也可以清楚的看到LDA共分为三层，语料级参数α、β——整个语料不变，文本级参数θ——每篇文本不同，以及单词级参数w，z——每个单词不同。

推断（Inference）问题

推断问题类似HMM的解码问题，给定了模型，如何求出观测数据最佳的“解释”。如下式，

$$p(\theta,Z|W,\alpha,\beta)=\frac{p(\theta,Z,W|\alpha,\beta)}{p(W|\alpha,\beta)} \tag{6-5}$$

上式(6-5)分母为：

$$p(W|\alpha,\beta)=\frac{\Gamma(\sum_i\alpha_i)}{\prod_i\Gamma(\alpha_i)}\int\left(\prod_{i=1\ldots k}\theta_i^{\alpha_i}\right)\left(\prod_{n=1\ldots N}\sum_{i=1\ldots k}\prod_{j=1\ldots V}\left(\theta_i\beta_{ij}\right)^{w_n^j}\right)d\theta \tag{6-6}$$

这里遇到了因为隐含变量的存在而复杂度过高导致无法计算（intractable）的情况。对于上式的求解一般采用马尔可夫蒙特卡罗方法（MCMC）、变分推断（Variational Inference）、以及期望传播（Expectation propagation）等近似方法。

变分推断的基本思想是：使用Jesen不等式来得到对数似然的一个可变下界。本质上说，就是用带索引的变分参数来得到一个下界函数族，然后通过最优化过程来确定最终的变分参数，最后得到一个最紧的下界。通常逼近两个函数的方法是使两者的KL距离最小：

$$(\gamma^*,\phi^*)=\arg\min_{(\gamma,\phi)}\mathrm{D}(\mathrm{q}(\theta,\mathrm{z}|\gamma,\phi)\,||\,\mathrm{p}(\theta,\mathrm{z}|\mathrm{w},\alpha,\beta)) \tag{6-7}$$

可以通过类似EM的迭代方法来搜索参数，所以总的运算约在N2K的量级。

参数估计

LDA模型的模型参数是指语料级参数α和β，其参数估计公式如下

$$l(\alpha,\beta)=\sum_{d=1}^{M}\log p(W_d|\alpha,\beta) \tag{6-8}$$

也就是给定语料$D=\{W_1,W_2,\ldots,W_D\}$寻找使上式最大的参数α和β。同样因为隐含变量的缘故，一般采用变分EM算法。

LDA的理念指明了一条对于有层级概念的物理结构的概率模型建模方法。LDA的主要优点在于它的模块化（modularity）和可交换性（exchangeability）。作为一个概率模型，LDA比同样有很强的维数约减能力的LSI更好的适应能力和物理解释。LDA有很多进一步的发展，比如用到时间序列上等等。

6.3.2 基于 LDA 的摘要特征

为什么采用LDA作为摘要特征呢？

在第6.2节相关工作中我们介绍，自动文摘方法中和词频有关的特征有着从词汇（Vocabulary）级别发展到潜层语义索引或概率潜层语义索引的主题上的不断降维的趋势。这样做的一个很重要的目的是避免稀疏问题。

采用潜层语义索引或者概率潜层语义索引所得到的主题的物理意义并不明确，所谓的主题的含义比较模糊。而相反，LDA的模型不但在数学上比较完善，而且其中主题的含义非常清晰。这样，从直观上讲，采用LDA得到的主题为计算句子概率模型的单位既可以有效避免稀疏，又能比较好的反映句子和文章主题的关系。

这样，我们定义了如下两个基于LDA的特征：

相邻句子的主题相似度特征（FA）：我们用余弦相似度来度量两个句子主题的相似度。这个特征用来记录一句话与其前 3 个和后 3 个句子的主题相似度。其计算方法和句子余弦相似度的计算方法相同。

句子和所处文本的主题相似度（FB）：用余弦相似度度量每个句子和其所处文本的 LDA 主题相似度。

本节我们通过介绍LDA模型的原理，引出了本章工作的动因，并给出了基于LDA模型提取到的简单特征。下一节将介绍我们的实验。

6.4 实验及分析

因为本章的工作和前两章的工作有很大相关性，所以我们实验所采用的语料（DUC2001）和评测方法和前两章一致。本节首先给出摘要问题中常用的其它特征；然后给出实验用的系统框架，我们将在有监督和无监督两种情况下测试 LDA 特征对于摘要问题的影响；最后我们给出结论以及工作展望。

6.4.1 基本特征

这里我们采用有监督文本摘要中常用的一些特征。这里我们称之为基本特征：

长度特征：去除停用词后的句子长度。

位置特征：句子处在文章中的位置。在文章的开始为 1，结尾为 2，其余位置为 3。

对数似然特征：句子 xi 由其所在文章 D 生成的对数似然值 logP(x_i|D)。 其定义为：$\sum_{w_k} N(w_k, x_i) \log p(w_k \mid D)$。这里的 $N(w_{ki},x_i)$是词 w_k 在句子 x_i 中出现的次数；p(wk|D)可以用 $N(w_k, x_i) / \sum_{w_j} N(w_j, D)$ 来估计。

主题词特征：指去除停用词后的高频词。句子包含的这种词越多，被标定为摘要句的可能性也越大。我们用这个特征来记录句子中的主题词个数。

指示词特征：一些含有诸如“conclusion”、“briefly speaking”这种词的句子很可能是摘要句。用这个特征来标识一个句子是否含有这种词。

大写词特征：一些专有名词和作者要强调的词语一般会被大写，含有这种词语的句子是摘要的可能性也较大。用这个特征来标识一个句子是否含有被大写的词。

相邻句子相似度特征：我们用余弦相似度来度量两个句子的相似度。这个特征用来记录一句话与其前三个和后三个句子的相似度。

6.4.2 系统设计

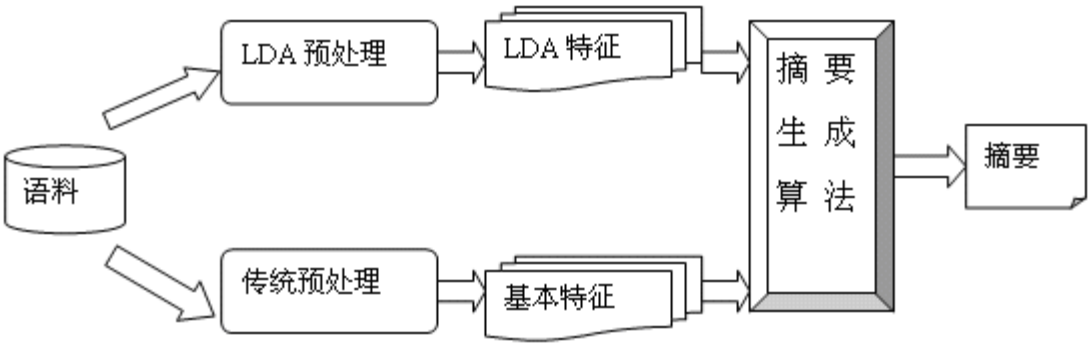


图6-2 基于 LDA 的摘要系统

我们的摘要生成系统框图见 6-2。首先语料经过两种预处理模块：传统预处理

和 LDA 预处理，分别得到基本特侦和 LDA 特征；然后将这两种特征加入我们的摘要生成算法，包括 CRF、SemiCRF（见第四章）、排序学习中的 SVMMAP（见第五章）。

6.4.3 实验 1：LDA 特征获取

我们在 DUC2001 年的语料上用 LDA 模型求得其 topic 分布，也就是多项式概率分布 p(w|z)，所得的结果部分展示在表 6-1。

表6-1 部分主题

Topic1	Topic2	Topic3	Topic4
eruption	kong	oil	ice
ash	hong	million	fish
vocanic	patten	valdez	antifreeze
vocanologist	chinese	ship	antarctica
bloom	people	tanker	sheet
information	politic	vessel	stream
acitivity	system	captain	scientist
life-threatenning	council	alaska	earth
scientist	public	mile	cold
warning	concession	set	water
dollar	government	crew	seal
torist	legislative	official	university

表 6-1 按概率从高到低的顺序给出了一些 topic 的分布情况。如 topic1 从高到

低依次为 eruption，ash， vocanic，vocanologist，bloom，information，activity，life-threatenning 等等，可以认为这个 topic 应该主要是和火山活动的新闻有关的；topic2 为 kong，hong，patten，Chinese，people，politic，system，council，public，concession 可以认为这是一个反映香港与大陆政治关系的一个主题；依次 topic3 和 topic4 分别为海上石油和北极问题。

这个实验中我们采用的是基于变分近似的算法，采用 EM 迭代，因为语料规模较小，算法收敛的很快。

6.4.4 实验 2：CRF 摘要系统实验

我们采用 CRF 算法对摘要系统进行了实验，实验结果见表 6-2。

表6-2 CRF 算法实验结果

Topic	**FA**	**FB**	**FA+FB**
20	0.372	0.38	0.386
30	0.386	0.392	0.391
40	0.361	0.405	0.404
50	0.379	0.402	0.401
60	0.376	0.393	0.406
70	0.387	0.399	0.405
80	0.36	0.371	0.37
90	0.376	0.372	0.381
100	0.377	0.38	0.379

表 6-2 给出的数据是分别采用新特征 FA 和 FB 以及一起采用这两个特征 FA+FB，并结合基本特征得到的 F1 结果。纵坐标为采用不同的主题数量，我们取从 20 个主题开始，每 10 个主题为单位逐渐递增，一直到 100 个主题。图 6-3 中给出了线条图形表示。

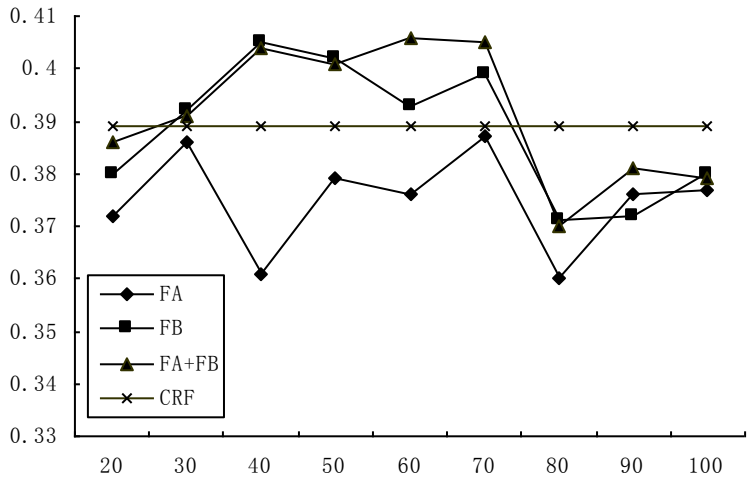


图6-3 CRF 算法系统实验结果

6.4.5 实验 3：SemiCRF&SVMMAP

在这个实验中，我们测试 LDA 特征在前两章的算法下的对性能的影响。

在针对 SemiCRF 的算法中，我们仿照第四章的扩展特征的方法对 LDA 特征进行了扩展，并加入所有的扩展特征。我们针对不同主题数量做了实验，其分布规律与 CRF 模型下大致相同，故表 6-3、6-4 仅列出了 40、50、60 的结果。

6.4.6 结果分析与展望

表 6-2 中的黑体数字为 topic 数量，当 topic 取 40，取特征 FB，以及当 topic 取 70，取 FA 特征和 FB 特征时，我们的系统达到最优值，0.405。它比[Shen, 2007]中的结果 0.389 提高了约 4.1%。另外，从实验结果上看，FB 特征比 FA 效果要好，它的曲线比较接近于 FA+FB 的曲线。本实验除了研究这两个特征的有效性，还对 topic 的数量对实验结果的影响作了测试。从结果上看 topic 取 30 到 70 之间的时候 FB 以及 FA+FB 的效果大都好于基线实验，而且 FA+FB 对系统的提升对于 topic 的变化更不明显，应该说 FA+FB 的效果更具有鲁棒性。

表 6-3 列出了 SemiCRF 算法下的结果，表 6-4 列出了 SVMMAP 下的结果。从结果上我们可以明显得出和 CRF 算法下相似的结论：FB 特征要普遍优于 FA 特征，并且最好的结果出现在 FB+FA 中。另外，和不采用 LDA 特征的原始方法比较，加入 LDA 后的结果要略优于原始结果，但涨幅并不大。

表6-3 SemiCRF 实验结果

Topic	FA	FB	FA+FB
40	0.390	0.407	0.407
50	0.397	0.404	**0.410**
60	0.394	0.404	0.407
原始 SemiCRF		0.407	

表6-4 SVMMAP 实验结果

Topic	FA	FB	FA+FB
40	0.394	0.398	**0.401**
50	0.392	0.390	0.398
60	0.392	0.395	0.395
原始 SVMMAP		0.394	

而从图 6-3 中我们可以清楚地看到，当主题数量由少到多增加时，实验结果无论是 FA、FB 还是 FB+FA，都有明显的改善。而当主题数量过 60-70 时，结果又开始变差。

分析其原因，我们认为这是由于语料规模偏小，而主题数目设置过多从而导致

了数据稀疏，使特征的作用下降。因此，在下一步研究中，我们将考虑语料的规模和 topic 的数量之间的关系。

另外，从表 6-1 中我们还可以看到，虽然 LDA 展现了一定的获得潜在主题的能力，但因词袋（bag-of-words）的前提理论假设，使得同一个词可以被不同的主题产生出来，如表中的“people”，“scientist”等词在不同的主题中都占有了很高的频率。[Blei, 2003]中也提到了这种情况，并提出了通过采用部分可交换性，或用马尔可夫性来描述词序列的方法来约束词袋理论假设。

下一步工作，我们希望继续考察 LDA 模型对文本切割的作用，以及在文本初步切割的基础上进一步取摘要的效果。另外，我们还希望研究在不同的语料规模下主题的数量对系统的影响。

6.5 本章小结

本章提出了一种新的基于 LDA 建模后抽取特征的自动摘要方法，我们在一些有监督摘要算法上进行了实验。结果证明，采用这种 LDA 特征，比单纯采用传统的特征有较大的提高。我们还进一步分析了不同的主题对系统性能的影响并初步指出了可能的原因。

第七章　基于语义的新闻领域复述句识别

7.1　引言

复述（Paraphrase）和蕴含（Text Entailment）可以看作为一个问题。对于两个语言片段（短语、句子、或篇章）A 和 B，如果能从 A 的语义中推理出 B，那我们说 A 蕴含 B，反之说 B 蕴含 A，而复述则可以看作 A 蕴含 B 且 B 也蕴含 A。

最早研究复述的文献见于[McKeown, 1979]，复述在自然语言处理中有许多应用：如机器翻译中[Callison-Burch, 2006; Zong, 2001; 宗成庆, 2002]，可以借鉴复述识别中的技术处理实时处理遇到的未登录短语；自动问答中[Harabagiu, 2006]，可以用于识别多种问句形式来提高系统性能；以及多文档自动摘要[Barzilay, 1999b; 秦兵等, 2005]的句子生成、句子压缩、相似句子的识别中等等。

作为一个独立的问题提出，句子级的复述识别（Paraphrase Recognition）一般指的是对于给定两个句子，判断其是否在语义上一致，如图 7-1。

Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence. *Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.* a)复述句
Armstrong, 31, beat testicular cancer that had spread to his lungs and brain. *Armstrong, 31, battled testicular cancer that spread to his brain.* b)非复述句

图7-1　复述句示例

图 7-1 中 a)部分的两个句子是互为复述句，直观的看，虽然单从用词的重合度上看这两个句子很类似，但从句法角度讲它们差别很大；而 b)部分的两个句子用词重合度也比较大，并且句法也类似，但是第一句包含有一些第二句没有的信息。

本文的工作主要致力于句子级别的复述识别，采用经过语义角色标注后的信息

为特征，然后通过机器学习算法来识别复述。虽然通过语义角色标注来识别复述的工作前人也有涉及[Qiu, 2006]，但本文将从一个新的角度来获取特征,并考虑了新闻语句本身的特点，实验证明本文的方法能够取得满意的结果。

7.2 相关工作

我们将复述识别的方法大致分为三类，一种是基于词袋信息的，第二种是基于句法信息的，第三种是基于深层语义的。下面从这三个方面简要介绍前人的相关工作。

7.2.1 基于词袋信息的方法

虽然复述问题的提出是鼓励学者从语义角度寻找判断两个句子是否可以相互替代的方法，但因为语义角色标注正确率有待提高，所以从实用的角度，还是有许多学者尝试用表层信息来解决这个问题。基于词袋的方法中向量模型是最主要的方法，这种方法广泛应用于文本分类，信息检索，以及句子相似度计算，在此方法上的改进一般包括：去停用词、词干化（stemming）、POS 标注以及词义扩展等。

文献[Corley, 2005]在词袋的基本思想下，引入基于 WordNet[Fellbaum, 1998]的世界知识，并且将传统的词到词的相似性计算方法扩展到文本到文本。其方法是将两段文本的相似性定义成了文本中名词和动词的加权函数，以及从别的语料中得到的倒排文档频率。作者在有监督和无监督条件下分别测试了该扩展方法。

在机器翻译中的自动评测启发下（以 BLEU[宗成庆, 2008]为例，其核心思想是计算 n 元文法匹配的对数几何平均），文献[Finch, 2005]采用机器翻译中常用的自动评测机制 BLEU、NIST、WER 以及 PER 来提取特征对分类器进行训练，并进行句子的语义相似性计算。

还有些基于潜层语义索引（LSI）的计算相似度方法，但 LSI 缺乏合理的物理解释，并且计算复杂度较高。

基于词袋信息的方法往往简单实用，但是很明显这样做违背了复述这个问题提出的初衷。单纯采用词袋信息，不能也可以肯定不会达到很好的效果。

7.2.2 基于句法信息的方法

两个意思相近的句子虽然有可能在句法上有差异，但是其内部子结构往往能找

到很多相似之处，而且，依存句法和语义表示有一定的相似性，所以有不少学者试图在句法层面上寻找复述问题的解决办法。

文献[Wu, 2005]提出了使用基于反向转换文法（ITG）为复述问题建模的方法。简单讲 ITG 实际上是面向双语的上下文无关文法，其目的是让双语平行语料分析的鲁棒性最大化，而不是验证语料的合法性，从而应用于平行语料的标注，包括划界、对齐、切分等[宗成庆, 2008]。借助 ITG 模型，文献[Wu, 2005]没有采用任何外部知识，而是仅仅凭借句法层面上的相似就取得了不错的效果。

文献[Bar-Haim, 2005]提出了一种基于词汇—句法的方法，该方法要求判断蕴含时，不但要比较两句话词汇上的一致性，还要比较句法上的一致性。这种词汇—句法关系包括由词形变化引起的句法改变、动词的被动到主动的变化、共指等等。通过实验，[Bar-Haim, 2005]证明词汇—句法方法优于仅仅基于词汇的方法。

文献[Wan, 2006]系统归纳了 19 个特征用于训练分类器来识别复述句。这 17 个特征中前 9 个为词汇特征，10-15 为依存句法特征，16-17 为句长特征。

文献[Das, 2009]的理论假设是如果两个句子是依存句，则它们的句法树虽然可以允许有不同，但总体对齐应该比不是依存句的好。Das 结合词汇、句法信息，采用准同步依存文法（quasi-synchronous dependency grammar）建模。

文献[Zhang, 2005]采用了依存句法分析将被动句式变为主动句式，然后采用如编辑距离、词汇相似度、以及类似于 BLEU 的 n 元文法作为特征进行训练。

从句法的角度尝试复述问题比单纯从词汇的角度有说服力，而且效果也普遍比后者好。可是句法层毕竟离语义层还有很大差距，使用句法特征也都只能集中在句法的某个片段上，无法从整体句法树上寻找句子间的相似性。

7.2.3 基于深层语义的方法

文献[Qiu, 2006]是比较经典的基于语义角色标注的复述识别方法，文中不再单纯用词片段或句法树片段作为信息单位，而是用基于 PropBank[Kingsbury, 2002]的谓词论元结构（Predicate Argument Structure, 结构化的表示一个动词谓词及其参量——也就是论元，见图 7-2）。这种结构可以很好的表示动作、概念、及其相互关系。而单纯用动词、名词来表示这个含义有可能面临相同的词袋但却不同意义的情况（如：他打我 和 我打他；意义正好相反）。

predicate: 打	predicate: 打
arg0: 我	arg0: 他
arg1: 他	arg1: 我

图7-2 谓词论元结构

文献[Qiu, 2006]的系统流程图见图 7-3。首先将句子对用 Charniak 句法分析器[Charniak, 2000]进行句法分析；其次采用语义角色标注工具 ASSERT[Pradhan, 2004]进行语义标注，找到句子里的谓词论元结构；然后通过贪婪算法寻找匹配的谓词论元结构，在这个过程中用到了外部词库；最后没有匹配上的其它成分被送进一个有监督分类器，来判断这些成分是否为重要成分，并作出是否这两个句子互为复述的判断。

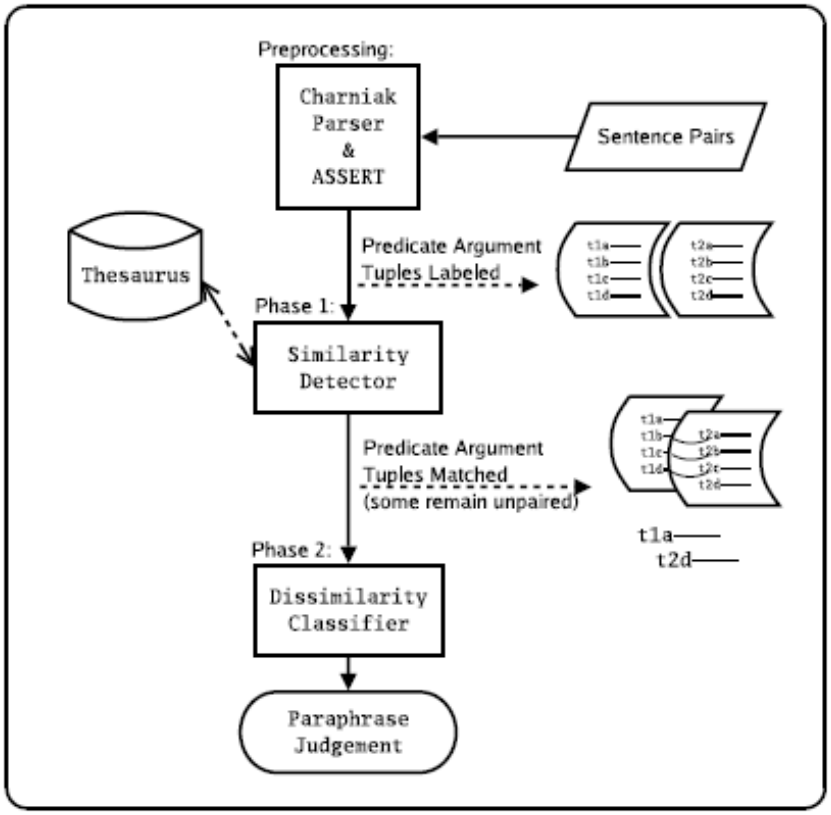


图7-3 文献[Qiu, 2006]方法流程

送入分类器的特征有两类：一类是句法树路径特征，指的是在句法树上的未匹配的单元与最近的有公共父节点的匹配单元之间的距离，作者认为这个特征可以在一定程度上表征这个未匹配部分的重要性；另一类是谓词论元结构中谓词，也就是

动词之间的相似性。

基于深层语义的方法相对较少，主要原因是语义角色标注本身正确率有待提高。虽然如此，从语义层考虑复述问题，毕竟最合乎逻辑。我们认为，要想做到真正实用的句子复述识别，必须要从这个角度着手加以研究。

7.3 本项研究的动因及思路

7.3.1 动因

我们分析了文献[Qiu, 2006]中采用的方法，认为其主要针对的矛盾是类似于图 7-2 的论元互换，或者主动语态和被动语态等这些从句法分析以及词汇分析上无法得出正确语义的情况。因为 ASSERT 语义角色工具包只能保证谓词的 arg0 和 arg1 的一致性，所以其方法仅仅能识别最基本的谓词以及这个谓词的发起者和接收者（arg0,arg1），并没有利用更多的语义角色标注信息识别句子里的其它成分。然而现实中很多的复述问题要面临的情况复杂的多，比如一个单纯的补语或状语的不同，就有可能导致两个极为相似的句子不称其为复述：如“我在家吃饭”和“我在学校吃饭”。最后，文献[Qiu, 2006]给出的方法过于一般化，并没有考虑新闻语料自身的特点。

为了得到更完备的语义角色，我们选用伊利诺伊斯大学认知计算组（Cognitive Computational Group，UIUC）的语义角色标注工具包 SRL。SRL 工具包有比较好的语义角色标注效果，对时间、地点、数字、指代等都有很好的识别，其标注方式见于表 7-1。

图 7-4 给出了采用 SRL 标注好的两个复述句。这两句话均来自微软的复述语料库（Microsoft Paraphrase Corpus，MPC[Dolan, 2004]）。以图 7-4 a)为例，SRL 针对每一个动词（包括 publish，offer，add）给出了谓词论元结构，也就是着色的三列，其中不同的灰度代表不同的论元。

如动词 publish，其主语 A0 也就是 They，宾语 A1 为 an advertisement on the internet。另外 SRL 还清楚的找到了时间状语 on July 4，以及状语 offering cargo for sale。

表7-1 SRL 工具标注信息

A0	**主语**	**A1**	**宾语**
A2	间接宾语		
AM-DIR	方向	**AM-DIS**	篇章标记
AM-EXT	范围	**AM-LOC**	位置
AM-MOD	一般修饰	**AM-NEG**	否定
AM-MNR	举止	**AM-PRD**	第二谓词
AM-PRP	提议	**AM-REC**	反义
AM-TMP	时间	**AM-ADV**	副词修饰

图 7-4 b)给出了这句话的一个复述句的分析。

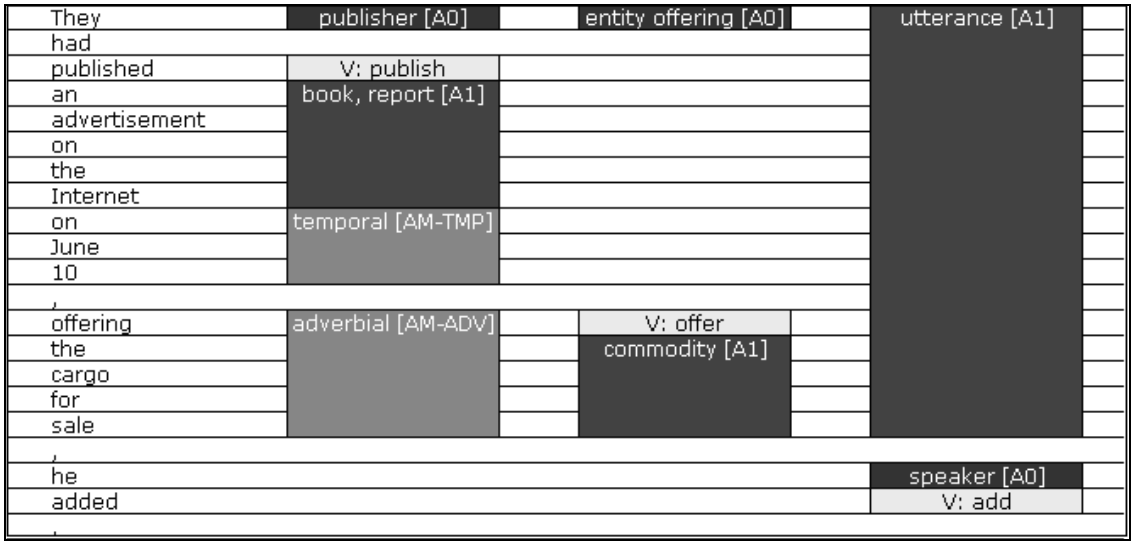
可以看到，虽然这两句话句法结构有较大差异，但是 publish 的论元中，但除了 A0 不一致外，其它语义角色都一致。

MPC 语料均来自于不同的新闻文稿，所以，虽然在 a) 句中出现了无法消解的代词 They，以及后面的代词 he，但是因为新闻背景关系，在其它语义角色相同或大致相同的情况下，还是认为这两句话是复述——这在 MPC 语料中是很常见的现象。

从新闻语句自身特点看，很多新闻句中可能主、谓、宾都一致，但是不同的时间、不同的地点以及不同的宾补成分，都可能使得这两句看似相似的句子实际上在诉说不同的事情。分析这种情况，我们认为，单纯从由谓词、A0、A1 论元组成的元组为单位寻找匹配并不能很好的解决新闻语句的复述识别问题，句子的各种修饰成分在复述句识别中同样占有很重要的作用。

7.3.2 设计思路

复述句识别方法以有监督训练为主，事实上几乎所有人的工作都集中在如何找到最能反映这个问题本质的特征上。在文献[Qiu, 2006]的基础上，我们针对新闻复述句的特点，提出了一种基于语义角色标注的有监督新闻复述句识别方法，其流程见图 7-5，下面介绍每个模块的功能，并给出选用的特征。



a） SRL 例句

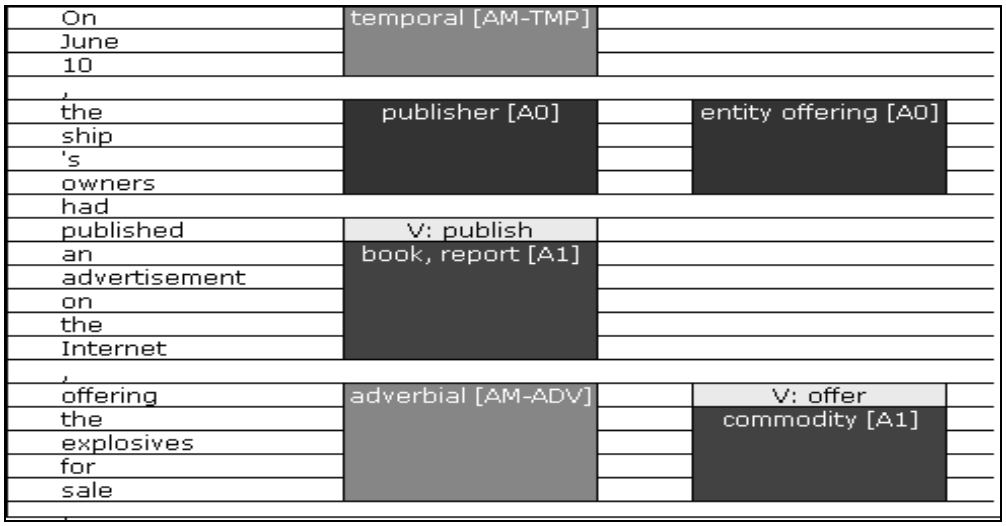


图7-4 b) SRL 例句

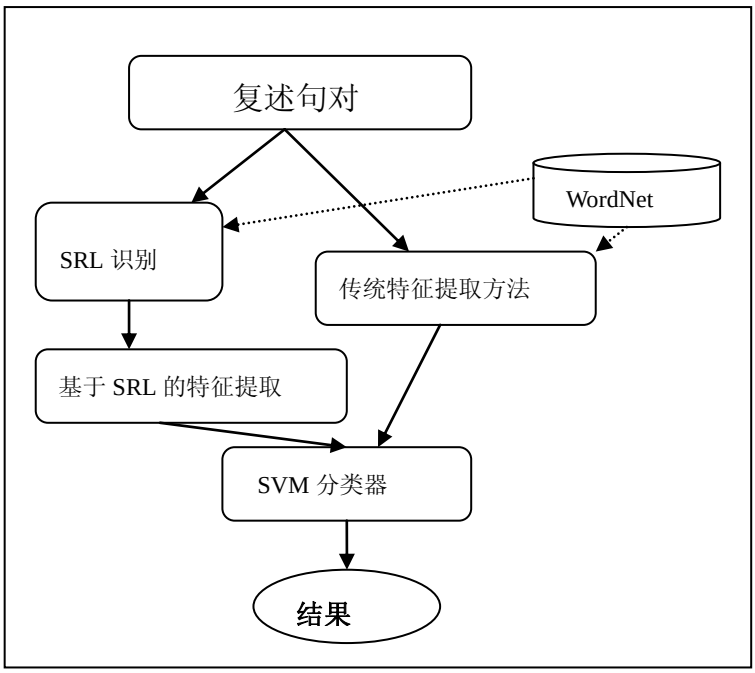


图7-5 基于语义角色标注的有监督新闻复述句识别方法流程

首先将复述句输入传统特征提取模块。在传统特征提取模块之中，我们先对句子进行了去停词、词干化工作，然后我们考虑了如下几个特征，我们称之为基本特征：

基本特征（BF）：

句长差特征：两个句子长度相减所得的差，可以为正值或负值。

绝对句长差特征：对句长差特征取绝对值。

这两个特征参考了文献[Wan, 2006]。

句子相似度特征：我们用如下公式(7−1)计算相似度

$$Sim_{S_1S_2} = \frac{2\bullet(S_1 \cap S_2)}{|S_1|+|S_2|} \tag{7-1}$$

基于 WordNet 的相似度特征：对公式（7-1）中的分子，我们对词干化以前的词采用 WordNet3.0 计算相似度，为句子中的每个词寻找其对应复述句中相似度分值最高的词，然后将这些相似度叠加并乘以 2。

我们同时也将句子输入 UIUC 的 SRL 标注工具模块，进行语义角色标注。得到

标注好语义角色的句子后，将其输入 SRL 特征提取模块，我们提取了如下一些特征，我们称之为语义特征。

语义特征（SF）：

谓词数差特征： 两句话的谓词数相减所得，可以为正值或负值。

匹配的谓词数量特征： 两句话中有多少谓词可以匹配，我们经验的规定两个谓词在 WordNet 上的相似度超过 0.5 即为匹配成功，我们最多寻找三对最相似的谓词。

未匹配的词是否被其它谓词的论元覆盖： 我们经验的判定一句话中越是重要的部分被各个谓词论元覆盖的次数越多。以图 7-4a 为例，“They”这个主语被三个论元所覆盖：publish 的 A0，offer 的 A0，以及 add 的 A1，而对于谓词 add 以及其论元 A0 都只被覆盖了一次。

对于匹配的谓词，我们按照表 7-1 寻找是否有匹配的论元，并给出如下一组特征。

句子 1 中的论元特征：第一句话中是否包含某个论元成分。这是个 0、1 特征。

句子 2 中的论元特征：第二句话中是否包含某个论元成分，同上。

这两个特征反应了论元的匹配情况。

匹配的论元成分的相似度：我们用公式(7-2)计算相似度。

$$P_1 = \frac{A_1 \cap A_2}{|A_1|}$$
$$P_2 = \frac{A_1 \cap A_2}{|A_2|} \tag{7-2}$$
$$Sim = \frac{2P_1 \cdot P_2}{P_1 + P_2}$$

其中 $A_{i=1,2}$ 表示句子 1 和 2 在某个谓词下相对应的论元，和计算相似度特征的方法一样，我们这里也采用了 WordNet 外部知识。对于不存在的论元成分，这个特征取零。这个特征会产生 3×15 个子特征，它们的设立是考虑到不同的论元成分，哪怕是一些修饰成分，在新闻语料中都可能扮演很重要的角色这个特点。

代词特征：我们从 SRL 中得出的结果并没有考虑指代消解，我们用这个特征给出某个论元成分中是否包含代词。给出这个特征，是因为考虑到新闻语料中的复述句往往来自不同的新闻机构或者不同的日期，单纯从孤立的句子上看，会有一些无

法消解的指代，比如图 7-4a 中的 They 和 he。图 7-4 a)和 b)这两句话在新闻背景下是复述句，但是严格意义上说，它们不应该是互为复述句的。

扩展的句子相似度特征：在语义角色标注后，我们引入了一个新的计算句子相似度方法，其公式如（7-3）

$$extSim = \frac{\sum_{w\in\{S_1\cap S_2\}} (N_1^w + N_2^w)}{\sum_{w\in\{S_1\}} N_1^w + \sum_{w\in\{S_2\}} N_2^w} \tag{7-3}$$

公式(7-3)中的 N_i^w 表示有多少个谓词论元成分覆盖了这个词，如图 7-4 a)中的 offering 为 3，added 为 1。选用这个特征也是基于越是重要的成分被覆盖的越多这个直观假设。

这些特征的设计都是充分考虑了新闻语句中识别复述与普通意义上的复述的不同。得到的特征我们直接送入 SVM 分类器进行训练和测试，并由 SVM 输出最终结果。

7.4 实验

这一部分介绍我们采用的语料和实验结果。

7.4.1 MPC 语料介绍

我们的实验里采用的是 MCP[Dolan, 2004]语料。MPC 是一个用于一般用途的复述语料库，它来源于在线新闻网站。其创建分两步，第一步是由仅仅依靠编辑距离过滤词汇上不相似的句对，这样共得到了 5801 个句子对；第二步为人工标注，其中 3900 个句子对标为正例，其余为反例。这样这个语料中的句子单从词汇上看句对间相似度很大，这就给复述识别任务提出了更高的要求。图 7-1 中给出的两组句子来自 MPC，其中第二个例子稍微做了一点改动。

为便于比较，我们的评测采用 F1 值作为评测标准，和文献[Qiu, 2006]的评测方法一致，见公式（7-4）。

$$\mathrm{F1} = \frac{2 * \mathrm{Precesion}}{\mathrm{Precesion+Recall}} \tag{7-4}$$

7.4.2 实验及结果

在 MCP 语料中，训练集有 2753 个正例，1323 个反例；测试结中有 1147 个正例，578 个反例。从训练集抽取特征后我们用 SVM 分类器进行训练，然后再测试集上做测试。

我们选用基于语义角色标注的方法 Qiu[Qiu, 2006]和基于句法的方法 Zhang[Zhang, 2005]以及 Wu[Wu, 2005]作为基线实验。在实验中，我们先测试了传统的基于基本特征（BF）的结果，然后给出基于语义特征（SF）的结果，最后将这二者结合给出 SF+BF 的结果。试验结果及基线实验在表 7-2 给出。

7.4.3 结果分析与展望

从表 2 我们看到，当取依赖于词频等表层信息的基本特征 BF 时结果为最低 0.789，而加入语义特征 SF 后，系统性能得到很大改善，取得了 3.7%的相对提高，达到 0.818。这证明我们采用的语义特征效果比较明显。

表7-2 试验结果

Me-	Preci-	Re-	F1
Zhang	0.743	0.88	0.8
Wu	0.723	0.92	0.8
Qiu	0.725	0.93	0.8
BF	0.725	0.86	0.7
SF+BF	0.739	0.91	0.8

我们的结果比依赖句法和词袋特征的 Wu 和 Zhang 的方法都有较明显提高。前面提到 Zhang[Zhang, 2005]的方法依赖依存句法分析对句式做的调整，如将被动句式变为主动句式等。这种方法效果并不理想，仅取得了 0.807 的 F1 值。这个结果正反映出新闻语料自身的特点，也就是是否是复述句很大程度上也依赖句子中某些状语、补语等修饰成分。我们的方法和 Qiu[Qiu, 2006]相比提高虽然并不明显，但是我们的方法是在完全没有依赖句法特征做出的，这也是我们下一步将要进行的工作。

7.5 本章小结

本章介绍了一种新的基于语义角色标注的新闻领域复述语句识别方法。针对新闻语句不容易从词频、句法等信息做出复述判断的特点，对经过语义角色标注的新闻语句，我们提取了能反映句子不同成分匹配情况的更为丰富的特征。从实验结果看，我们的结果比依赖词袋信息，以及句法信息的方法都有明显的提高。

第八章 结束语

自动摘要是一个综合性很强的问题，几乎可以囊括信息检索和自然语言处理的方方面面。根据现有的技术发展，本论文主要选择了更具代表性的建模以及特征提取问题来做了重点研究。我们的主要思想是通过篇章的表层信息来寻找能反映摘要问题本质的方法。

本论文的方法虽然都是在单文档上做的。从目前看，多文档摘要与单文档摘要的不同主要集中在后处理上。所以我们的方法也可以很容易的扩展到多文档中去。

我们在摘要建模的工作上提出了两种方法。

一种是基于 Semi-CRF 的序列分段模型（SSM）的有监督摘录型摘要提取方法。我们将摘要问题看作“段标注”问题。我们采用基于 SeimCRF 的方法来实现 SSM。我们方法与前人的不同之处在于提取特征的单位不单来自句子，也可以来自于段。我们的实验证明这种方法比单纯针对句子为单位提取特征的方法有较明显的效果。

另一种方法是采用排序学习（LTR）来对通用型摘要问题建模。摘录型摘要的核心问题是给句子打分，打分的目的是为了后面的排序，并输出排名靠前的句子。而排序学习本质上就是为了解决排序，所以和摘录型摘要有很强的内在的切合点。而且，采用排序学习建模更强调同一文本内的句子之间的相互比较。除了建模，我们还将当前流行的几种排序学习算法在摘要问题上进行了比较。

在特征提取方面，本论文考虑了潜层狄利赫雷分配（LDA）建模方法。我们研究了 LDA 特征在几种有监督的自动文摘中的作用，并分析研究了在不同 Topic 数量下 LDA 特征对摘要结果的影响。实验结果表明，加入 LDA 特征后，能够有效地提高以传统特征为输入的文摘系统的质量。

最后，作为对多文档摘要工作的一个铺垫，我们研究了复述句识别的问题。传统的解决复述句识别方法是通过词频或句法上的相似度来判断的。但即使相同的文字书写的句子其含义也可能差别很大，而相同句法结构也不能保证意义一致。本文根据新闻语料的特点，提出了一种通过引入深层的语义角色标注来帮助识别新闻领域复述句的方法。该方法通过在语义角色这种结构化的含义表达形式中提取的特征

来弥补传统方法的不足：先识别待判断的两个句子中所有谓词的语义角色，然后计算两个句子间对应语义角色的相似度，最后结合传统的句子相似度计算方法来进行相似性计算。实验证明，本文提出的方法能有效地提高改写语句的识别效果。

下一步工作，我们希望能找到几种模型相互融合的方式为摘要问题建模。这种考虑是希望各种模型能够在摘要问题上取长补短，以求达到更好的效果。但这也需要更大的语料的支持。另外，更多考虑无监督的建模方法对于提高摘要系统的适应性也很有益处。

在特征提取上，我们的 LDA 建模只是初步的探索。概率图模型数学完备并且富于变化，已经有很多学者利用它在各种不同的问题上进行演化、变形，形成新的模型。下一步，如何能够针对摘要问题利用概率图模型搭建合理的模型，集特征提取和摘要建模于一身，将是一个非常值得尝试的工作。

在绪论中我们还提到，自动摘要技术牵扯自然语言处理的方方面面，是一个综合性很强的技术。一个完善的系统除了本论文讨论的建模以及特征选择，还应该包括词性标注、句法分析等相关模块，将这些模块合理的搭建成一个完整的摘要系统，也是下一步工作的重点。

我们认为，摘要问题面临的困境主要是：如果自然语言处理技术用得太深，则囿于当前技术水平，难以形成通顺的摘要；如果用得太浅，比如只停留在词袋（bag of words）的概念，或加入一些启发式规则，则只能永远作为一个简单的辅助工具，也很难真正具有实用意义。综上所说，我们认为摘要技术未来的发展方向应该着重于在上述两个方面齐头并进，寻找最佳的平衡点，并为推进自然语言处理技术的发展做出贡献。

致谢

值此博士学位论文完成之际，回想攻读博士学位这段既艰苦又充实的岁月，我感慨万端，不能平静。曾经得到难以计数的老师的关心、亲人的鼎助和朋友的支持，心中油然生出无限的感激之情。

首先感谢我的导师宗成庆研究员!感谢他在学业上对我的悉心指导和不倦教诲，他严谨的工作态度都深深地感染了我，无论是过去和现在，还是将来，他都是我学习和工作中应当效仿的楷模。对于我的论文，他更是倾注了大量的心血来具体指导。同时，也深深地感激他对我生活上的给予的不尽的帮助以及对我的家人的尽可能的照顾。

感谢模式识别国家重点实验室语音语言技术研究组的赵军老师、刘文举老师、陶建华老师以及实验室其他老师！他们以不言之教昭示出科研工作者宝贵的敬业精神，通过他们我学到了研究课题以外的丰富知识，开阔了研究视界，提高了科研素质。

我衷心地感谢刘非凡、吴友政、段湘煜、曹文洁、周玉、柴春光、徐昉、王根、蔡勋梁、李寿山、刘鹏、张家俊、陈钰枫、鉴萍、周可艳、方李成、何彦青、李茂西、夏睿、汪昆、庄涛、刘康、杨帆、韩先培、翟飞飞、涂眉、齐振宇诸同学，以及陆征助理。感谢他们长期给予的无私帮助，他们的友情使我的学习生活丰富多彩。他们让我的博士生涯增添了难以忘怀的乐趣和温暖。

感谢研究生部的邸凌、卜树云、李磊等老师以及模识实验室综合办连国臻、赵微等老师的关心和支持。

特别感谢我的父亲和母亲！感谢他们多年的养育之恩，当我遇到困难的时候，他们总是给我无私的关爱和支持。感谢我的妻子，没有她，我无法完成博士学业；感谢我的女儿，她的到来让我的生活充满了美好的期望。

最后，对参加论文评审、答辩的各位老师表示衷心的感谢！

个人简历

吴晓锋，男，1976 年 4 月 23 日出生于山东省东营市。

2005.9 — 2010.9　　中科院自动化所，模式识别与智能系统专业，博士生

2000.9 — 2003.7　　中国人民解放军防化学院，核技术及应用专业，研究生

1993.9 — 1996.7　　西北大学，电子学与信息系统专业

攻读博士学位期间参加的项目

[1] 国家自然科学基金项目“基于语言理解的机器翻译方法研究”，项目编号：60975053

[2] 国家支撑计划项目“多语言信息服务环境关键技术研究及应用”，项目编号：2006BAH03B02

攻读博士学位期间发表的论文

1. Xiaofeng Wu and Chengqing Zong. A New Approach to Automatic Document Summarization. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, India, January 8-10, 2008. Pages 126-132
2. Xiaofeng Wu and Chengqing Zong. An MAP Based Sentence Ranking Approache to Automatic Summarization. *In Proceedings of 6th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE),Beijing*, China Aug. 21-23,2010,Pages 237-241
3. Xiaofeng Wu and Chengqing Zong, Automatic Summarization using Learning to Rank, *Topics in Multilingual Language Technology (in honour of Hans Uszkoreit)*.（Accepted）
4. 吴晓锋，宗成庆. 一种基于 LDA 的 CRF 自动摘要方法. 中文信息学报，No.6, Vol.23, 2009. Pages 39-45.
5. 吴晓锋，宗成庆, 一种基于语义角色的复述句识别方法. 中文信息学报，No.5，Vol.24，2010. Pages 3-9.

参考文献

[Aone, 1999] C Aone, M Okurowski, J Gorlinsky, et al. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. Advances in automatic text summarization: 71-80.

[Bar-Haim, 2005] R Bar-Haim, I Szpektor, O Glickman. 2005. Definition and analysis of intermediate entailment levels, Empirical Modeling of Semantic Equivalence and Entailment, pp. 55-60.

[Barzilay, 2003] R Barzilay. 2003. Information fusion for multidocument summarization: paraphrasing and generation. Columbia University New York, NY, USA.

[Barzilay, 1997] R Barzilay, M Elhadad. 1997. Using lexical chains for text summarization, Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization.

[Barzilay, 2002] R Barzilay, N Elhadad, K Mckeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 17(35-55).

[Barzilay, 2004] R Barzilay, L Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization, In HLT-NAACL 2004: Proceedings of the Main Conference, pp. 113–120.

[Barzilay, 1999a] R Barzilay, Kr Mckeown, M Elhadad. 1999a. Information fusion in the context of multi-document summarization, Association for Computational Linguistics, pp. 550-557.

[Barzilay, 1999b] R Barzilay, Kr Mckeown, M Elhadad. 1999b. Information fusion in the context of multi-document summarization 550-557.

[Baxendale., 1958] P.B. Baxendale. 1958. Man-made Index for Tech-nical Literature -An Experiment. IBM Journal of Research and Development, 2(4): 354-361.

[Bergler, 2003] S Bergler, R Witte, M Khalife, et al. 2003. Using knowledge-poor coreference resolution for text summarization, Proceedings of the HLT/NAACL Workshop on Text Summarization (DUC 2003). Document Understanding Conference.

[Bishop, 2006] Cm Bishop. 2006. Pattern recognition and machine learning, Springer New York.

[Blei, 2003] Dm Blei, Ay Ng, Mi Jordan. 2003. Latent dirichlet allocation. The Journal of Machine Learning Research, 3(993-1022).

[Bollegala, 2010] D Bollegala, N Okazaki, M Ishizuka. 2010. A bottom-up approach to sentence ordering for multi-document summarization. Information Processing and Management, 46(89-109).

[Brin, 1998] S Brin, L Page. 1998. The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems, 30(1-7): 107-117.

[Burges, 2005] C.J.C. Burges, T. Shaked, E. Renshaw, et al. 2005. Learning to Rank using Gradient Descent, Proceedings of the 22nd international conference on Machine learning, pp. 96-108.

[Callison-Burch, 2006] C. Callison-Burch, P. Koehn, M. Osborne. 2006. Improved statistical machine translation using paraphrases, Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 17-24.

[Cao, 2007] Z Cao, T Qin, Ty Liu, et al. 2007. Learning to rank: from pairwise approach to listwise approach, ICML, pp. 136-144.

[Carbonell, 1998] J Carbonell, J Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, In Proceedings of ACM SIGIR 1998, pp. 335-336.

[Charniak, 2000] E Charniak. 2000. A maximum-entropy-inspired parser, In Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'2000). pp. 132-139.

[Conroy, 2001] Jm Conroy, Dp O'leary. 2001. Text summarization via hidden markov models, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 406-407.

[Corley, 2005] C Corley, R Mihalcea. 2005. Measuring the semantic similarity of texts, Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp. 13-18.

[Das, 2009] D Das, Na Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 468-476.

[Daume III, 2005] H Daume Iii, D Marcu. 2005. Bayesian multidocument summarization at MSE, In ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures.

[Daume III, 2006] H Daume Iii, D Marcu. 2006. Bayesian query-focused summarization, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 305-312.

[Deerwester, 1990] S Deerwester, St Dumais, Gw Furnas, et al. 1990. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6): 391-407.

[Defays, 1977] D Defays. 1977. An efficient algorithm for a complete link method. The Computer Journal, 20(4): 364-366.

[Dolan, 2004] B Dolan, C Quirk, C Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources, Proceedings of the 20th international conference on Computational Linguistics, pp. 350.

[Edmundson, 1969] Hp Edmundson. 1969. New methods in automatic extracting. Journal of the ACM (JACM), 16(2): 285.

[Erkan, 2004] G Erkan, Dr Radev. 2004. Lexpagerank: Prestige in multi-document text summarization, Proceedings of EMNLP.

[Fellbaum, 1998] C Fellbaum. 1998. WordNet: An electronic lexical database, MIT press Cambridge, MA.

[Finch, 2005] A Finch, Ys Hwang, E Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence, Proceedings of the Third International Workshop on Paraphrasing (IWP2005), pp. 17-24.

[Fisher, 2006] S Fisher, B Roark. 2006. Query-focused summarization by supervised sentence ranking and skewed word distributions. Proc. DUC-2006.

[Frakes, 1992] Wb Frakes, R Baeza-Yates. 1992. Information retrieval: data structures

and algorithms, Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

[Freund, 2003a] Y Freund, R Iyer, Re Schapire, et al. 2003a. An efficient boosting algorithm for combining preferences. The Journal of Machine Learning Research, 4(933-969).

[Freund, 2003b] Yoav Freund, Raj Iyer, Robert E Schapire, et al. 2003b. An efficient boosting algorithm for combining preferences. In Machine Learning: Proceedings of the Fifteenth International Conference.

[Fuentes, 2007] M Fuentes, E Alfonseca, H Rodr ¨Guez. 2007. Support Vector Machines for Query-focused Summarization trained and evaluated on Pyramid data, In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pp. 57–60.

[Ge, 2002] Xianping Ge. 2002. Segmental semi-markov models and applications to sequence analysis. California; Irvine: pp. 182.

[Goldstein, 1999] J Goldstein, M Kantrowitz, V Mittal, et al. 1999. Summarizing text documents: sentence selection and evaluation metrics, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 121-128.

[Goldstein, 2000] J Goldstein, V Mittal, J Carbonell, et al. 2000. Multi-document summarization by sentence extraction, NAACL-ANLP 2000 Workshop on Automatic summarization, pp. 40-48.

[Gong, 2001] Y Gong, X Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieva, pp. 19-25.

[Grishman, 1997] R Grishman. 1997. Information extraction: Techniques and challenges. Lecture Notes in Computer Science, 1299(10-27).

[Harabagiu, 2006] S Harabagiu, A Hickl. 2006. Methods for using textual entailment in open-domain question answering, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 905-912.

[Herbrich, 2000] R. Herbrich, T. Graepel, K. Obermayer. 2000. Large margin rank boundaries for ordinal regression In Advances in Large Margin Classifiers.

[Hoey, 1991] M Hoey. 1991. Patterns of lexis in text, Oxford University Press Oxford.

[Hofmann, 1999] T Hofmann. 1999. Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50-57.

[Hovy, 1999] E Hovy, Cy Lin. 1999. Automated text summarization in SUMMARIST, Advances in automatic text summarization. The MIT Press: pp. 94.

[Hovy, 2005] E Hovy, Cy Lin, L Zhou. 2005. A BE-based multi-document summarizer with sentence compression, Proceedings of Multilingual Summarization Evaluation.

[Ishikawa, 2002] K Ishikawa, S Ando, S Doi, et al. 2002. Trainable automatic text summarization using segmentation of sentence, In Proc. 2002 NTCIR 3 TSC workshop.

[Ji, 2006] Pd Ji, S Pulman. 2006. Sentence ordering with manifold-based classification in multi-document summarization, In Proceedings of Empherical Methods in Natural Language Processing, EMNLP, pp. 526-533.

[Jing, 2002] H Jing. 2002. Using hidden Markov modeling to decompose human-written summaries. Computational linguistics, 28(4): 527-543.

[Jing, 1999] H Jing, Kr Mckeown. 1999. The decomposition of human-written summary sentences, Proceedings of the 22nd Conference on Research and Development in Information Retrieval (SIGIR-99), pp. 129-136.

[Jing, 2000] H Jing, Kr Mckeown. 2000. Cut and paste based text summarization, Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pp. 178-185.

[Joachims, 1998] T Joachims, C Nedellec, C Rouveirol. 1998. Text categorization with support vector machines: learning with many relevant, In Proceddings of the 10th European Conference on Machine Learning, pp. 137-142.

[Joachims, 2003] T. Joachims. 2003. Optimizing Search Engines using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data

Mining, pp. 133-142.

[Kingsbury, 2002] Paul Kingsbury, Martha Palmer, Mitch Marcus. 2002. Adding semantic annotation to the penn treebank. In Proceedings of the Human Language Technology Conference: 252-256.

[Kleinberg, 1999] Jm Kleinberg. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5): 604-632.

[Knight, 2000] K Knight, D Marcu. 2000. Statistics-based summarization-step one: Sentence compression, AAAI, pp. 703-710.

[Knight, 2002] K Knight, D Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence, 139(1): 91-107.

[Kudoh, 2000] T Kudoh, Y Matsumoto. 2000. Use of support vector learning for chunk identification, In Proceedings of the Fourth Conference on Computational Natural Language Learning(CoNLL-2000), pp. 142-144.

[Kupiec, 1995] J Kupiec, J Pedersen, F Chen. 1995. A trainable document summarizer, Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 68-73.

[Lafferty, 2001] J Lafferty, A Mccallum, F Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Intl. Conf. on Machine Learning, pp. 282-289.

[Lapata, 2003] M Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pp. 545-552.

[Li, 2007] P. Li, C. J. Burges, Q. Wu. Mcrank. 2007. Learning to rank using multiple classification and gradient boosting. Advances in Neural Information Processing Systems.

[Lin, 1999] Cy Lin. 1999. Training a selection function for extraction, Proceedings of the eighth international conference on Information and knowledge management, pp. 55-62.

[Lin, 2004] Cy Lin. 2004. Rouge: A package for automatic evaluation of summaries, In

Proceedings of the Workshop on Text Summarization Branches Out, post-conference workshop of ACL 2004, pp. 25-26.

[Lin, 2006] Cy Lin, G Cao, J Gao, et al. 2006. An information-theoretic approach to automatic evaluation of summaries, In Proceedings of HLT-NAACL, pp. 463-470.

[Lin, 2001] Cy Lin, E Hovy. 2001. Neats: A multidocument summarizer, Proceedings of the Document Understanding Workshop(DUC).

[Lin, 2002a] Cy Lin, E Hovy. 2002a. From single to multi-document summarization: A prototype system and its evaluation, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 07-12.

[Lin, 2002b] Cy Lin, E Hovy. 2002b. Manual and automatic evaluation of summaries, In Proceedings of the ACL-02 Workshop on Automatic Summarization, pp. 52.

[Liu, 2009] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval, Foundations and Trends in Information Retrieval: pp. 225-331.

[Luhn, 1959] Hp Luhn. 1959. The automatic creation of literature abstracts, . IBM J Res. Develop, 2(2): 159-165.

[Mani, 1997] I Mani, E Bloedorn. 1997. Multi-document summarization by graph search and matching. In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97), pp. 622-628.

[Mani, 1999a] I Mani, B Gates, E Bloedorn. 1999a. Improving summaries by revising them, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 558-565.

[Mani, 1999b] I Mani, Mt Maybury. 1999b. Advances in Automatic Text Summarization. The MIT Press; Cambridge, MA.

[Marcu, 2001] D Marcu, L Gerber. 2001. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation, In Proceedings of the NAACL-2001 Workshop on Automatic Summarization, pp. 1-8.

[Marsi, 2005] E Marsi, E Krahmer. 2005. Explorations in sentence fusion, 10th European Workshop on Natural Language Generation pp. 109-117.

[McCallum, 2003] A Mccallum, W Li. 2003. Early results for named entity recognition

with conditional random fields, feature induction and web-enhanced lexicons, In Proceedings of Seventh Conference on Natural Language Learning (CoNLL), pp. 191.

[McDonald, 2006] R Mcdonald. 2006. Discriminative sentence compression with soft syntactic constraints, Proceedings of the 11th EACL, pp. 297-304.

[McKeown, 1995] K Mckeown, Dr Radev. 1995. Generating summaries of multiple news articles, Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 74-82.

[McKeown, 1979] Kr Mckeown. 1979. Paraphrasing using given and new information in a question-answer system, Proceedings of the 17th annual meeting on Association for Computational Linguistics, pp. 67-72.

[McKeown, 1999] Kr Mckeown, Jl Klavans, V Hatzivassiloglou, et al. 1999. Towards multidocument summarization by reformulation: Progress and prospects, Proceedings of The National Conference On Artificial Intelligence, pp. 453-460.

[Metzler, 2008] D. Metzler, T. Kanungo. 2008. Machine Learned Sentence Selection Strategies for Query-Biased Summarization, SIGIR Learning to Rank Workshop.

[Meyers, 1996] A Meyers, R Yangarber, R Grishman. 1996. Alignment of shared forests for bilingual corpora, In Proceedings of 16th International Conference on Computational Linguistics (COLING-96), pp. 460-465.

[Mihalcea, 2005] R Mihalcea. 2005. Language independent extractive summarization, Proceedings of the ACL 2005 on Interactive poster and demonstration sessions, pp. 52.

[Mihalcea, 2004] R Mihalcea, P Tarau. 2004. TextRank: Bringing order into texts, In Proceedings of EMNLP2004, pp. 404-411.

[Morris, 1991] J Morris, G Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational linguistics, 17(1): 21-48.

[Okazaki, 2004] N Okazaki, Y Matsuo, M Ishizuka. 2004. Improving chronological sentence ordering by precedence relation, In Proceedings of 20th International Conference on Computational Linguistics (COLING 04), pp. 750–756.

[Osborne, 2002] M Osborne. 2002. Using maximum entropy for sentence extraction,

Proceedings of the ACL-02 Workshop on Automatic Summarization, pp. 8.

[Paice, 1990] Cd Paice. 1990. Constructing literature abstracts by computer: Techniques and prospects. Information Processing and Management, 26(1): 171-186.

[Pinto, 2003] D Pinto, A Mccallum, X Wei, et al. 2003. Table extraction using conditional random fields, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 235-242.

[Pradhan, 2004] S Pradhan, W Ward, K Hacioglu, et al. 2004. Shallow semantic parsing using support vector machines, In Proceedings of HLT/NAACL.

[Qin, 2007] T. Qin, X.D Zhang, M.F. Tsai, et al. 2007. Query-level loss functions for information retrieva. Information Processing & Management, 44(2): 838-855.

[Qiu, 2006] L Qiu, My Kan, Ts Chua. 2006. Paraphrase recognition via dissimilarity significance classification, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 18-26.

[Radev, 2000a] Dr Radev. 2000a. A common theory of information fusion from multiple text sources step one: cross-document structure, In Proceedings of the 1st Workshop on Discourse and Dialogue of the Association for Computational Linguistics.

[Radev, 2002] Dr Radev, E Hovy, K Mckeown. 2002. Introduction to the special issue on summarization. Computational linguistics, 28(4): 399-408.

[Radev, 2000b] Dr Radev, H Jing, M Sty , et al. 2000b. Centroid-based summarization of multiple documents. Information Processing and Management, 40(6): 919-938.

[Riezler, 2003] S Riezler, Th King, R Crouch, et al. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 118-125.

[Sarawagi, 2005] S Sarawagi, Ww Cohen. 2005. Semi-markov conditional random fields for information extraction. Advances in Neural Information Processing Systems, 17(1185-1192).

[Sha, 2003] Fei Sha, Fernando Pereira. 2003. Shallow Parsing with Conditional Random Fields, In Proceedings of the 2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-03), pp. 134-141.

[Shen, 2007] D Shen, Jt Sun, H Li, et al. 2007. Document summarization using conditional random fields, proceedings of IJCAI.

[Silber, 2002] Hg Silber, Kf Mccoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational linguistics, 28(4): 487-496.

[Sundheim, 1998] B Sundheim. 1998. The TIPSTER Question-and-Answer (Q&A) Summarization Task: Test Design and Test Advances in Automated Text Summarization. MIT; Cambridge.

[Svore, 2007] Km Svore, L Vanderwende, Cjc Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. EMNLP 2007: Empirical Methods in Natural Language Processing, Prague, Czech Republic, pp. 448-457.

[Turner, 2005] J Turner, E Charniak. 2005. Supervised and unsupervised learning for sentence compression. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 297.

[Vapnik, 2000] Vn Vapnik. 2000. The nature of statistical learning theory, Springer Verlag.

[Viterbi, 1967] A Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory, 13(2): 260-269.

[Wan, 2006] S Wan, M Dras, R Dale, et al. 2006. Using dependency-based features to take the "para-farce" out of paraphrase. Proc. of the Australasian Language Technology Workshop, pp. 131-138.

[Wang, 2007] C. Wang, F. Jing, L. Zhang, et al. 2007. Learning query-biased web page summarization, Conf. on Information and Knowledge Management, pp. 555-562.

[Wang, 2008] D Wang, T Li, S Zhu, et al. 2008. Multi-document summarization via

sentence-level semantic analysis and symmetric matrix factorization, In SIGIR ’08: Proceedings of the31st annual international ACM SIGIR conference on Research and development in information retrieval,, pp. 307-314.

[Wu, 2005] D Wu. 2005. Recognizing paraphrases and textual entailment using inversion transduction grammars, Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp. 25-30.

[Wu, 2008] Xiaofeng Wu, Chengqing Zong. 2008. A New Approach to Automatic Document Summarization, International Joint Conference on Natural Language Processing (IJCNLP), pp. 126-132.

[Xia, 2008] F. Xia, T.-Y. Liu, J. Wang, et al. 2008. Listwise approach to learning to rank - theory and algorithm, In ICML ’08: Proceedings of the 25th international conference on Machine learning, pp. 1192-1199.

[Yeh, 2005] Jy Yeh, Hr Ke, Wp Yang, et al. 2005. Text summarization using a trainable summarizer and latent semantic analysis. Information Processing and Management, 41(1): 75-95.

[Yue, 2007] Y. Yue, T. Finley, F. Radlinski, et al. 2007. A support vector method for optimizing average precision., Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 271-278.

[Zha, 2002] H Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering, In Proceedings of the 25th Annual ACM SIGIR Conference, pp. 113-120.

[Zhang, 2005] Y Zhang, J Patrick. 2005. Paraphrase identification by text canonicalization, Proceedings of the Australasian Language Technology Workshop, pp. 160-166.

[Zong, 2001] Chengqing Zong, Yujie Zhang, Kazuhide Yamamoto, et al. 2001. Approach to Spoken Chinese Paraphrasing Based on Feature Extraction. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS), Tokyo, Japan, pp. 551-556.

[刘挺, 1999] 刘挺, 王开铸. 1999. 基于篇章多级依存结构的自动文摘研究. 计算机研究与发展, 36(4).

[秦兵等, 2005] 秦兵等. 2005. 多文档自动文摘综述. 中文信息学报, 19(6): 14-20.

[吴有政, 2006] 吴有政. 2006. 汉语问答系统关键技术研究. 博士论文.

[宗成庆, 2002] 宗成庆, 张玉洁, 山本和英, et al. 2002. 面向口语翻译的汉语语句改写方法. 汉语语言与计算学报（Journal of Chinese Language and Computing）, 12(1): 63-77.

[宗成庆, 2008] 宗成庆. 2008. 统计自然语言处理. 清华大学出版社.