中文分词研究综述

王琴瑶1），王柯云1），马冰1），张海波1）

（北京理工大学计算机学院 北京）

摘要：中文分词是中文信息处理的基础，是计算机中文信息处理中的难题，也是中文文献内容分析中必须解决的关键问题之一。对中文文本的分词处理目前已经应用到了中文自然语言理解、文献检索、搜索引擎以及文本挖掘系统等领域。本文对现有的中文分词技术进行了综述,1、基于词典的分词算法，2、基于统计的分词算法，3、基于神经网络的分词算法，分析了现有分词方法的技术特点，指出了部分分词方法存在的优缺点，并对分词系统评估方法进行了介绍及实例分析，最后介绍了中文分词在拼音输入法上的应用。

关键词：中文分词 分词方法 中文信息处理

A review of Chinese

word segmentation studies

WANG Qin-Yao1） WANG Ke-Yun1)  MA Bing1)  ZHANG Hai-Bo1)

(School of Computer Science, Beijing Institute of Technology Beijing)

**Abstract:** Chinese word segmentation is the basis of Chinese information processing. It is a difficult problem in computer Chinese information processing and one of the key issues that must be solved in the analysis of Chinese literature content. The word segmentation of Chinese text has been applied to Chinese natural language understanding, document retrieval, search engine and text mining system. This paper summarizes the existing Chinese word segmentation techniques, 1. Dictionary-based word segmentation algorithm, 2. Machine learning-based word segmentation algorithm, 3. Neural network-based word segmentation algorithm, analyzes the technical characteristics of the existing word segmentation method, and points out The advantages and disadvantages of partial word segmentation methods are introduced, and the word segmentation system evaluation methods are introduced and analyzed. Finally, the application of Chinese word segmentation in pinyin input method is introduced.

Keywords: Chinese word segmentation word segmentation method Chinese information processing

分工：王琴瑶（6 7章），王柯云（1 4 8章），马冰（5 章），张海波（2 3章）

**目录**

[1 绪论 3](#_Toc533533263)

[1.1 研究背景 3](#_Toc533533264)

[1.2 国内外研究现状 3](#_Toc533533265)

[2 分词方法的困难性与关键问题 5](#_Toc533533266)

[2.1分词的困难性 5](#_Toc533533267)

[2.2分词的关键问题 6](#_Toc533533268)

[3 基于机械匹配的分词方法 7](#_Toc533533269)

[3.1 正向最大匹配法 7](#_Toc533533270)

[3.2 逆向最大匹配法 8](#_Toc533533271)

[3.3 双向扫描 8](#_Toc533533272)

[3.4 N\_最短路径 9](#_Toc533533273)

[4 基于统计的分词方法 9](#_Toc533533274)

[4.1 基于统计的中文分词步骤 10](#_Toc533533275)

[4.2 基于统计的分词方法的语言模型 10](#_Toc533533276)

[4.3 切分歧义以及N-最短路径的切分排歧策略 14](#_Toc533533277)

[4.4 机械匹配与统计相结合的中文分词方法 15](#_Toc533533278)

[5 基于理解的中文分词 15](#_Toc533533279)

[5.1 关于LSTM 15](#_Toc533533280)

[5.1.1 什么是LSTM 15](#_Toc533533281)

[5.1.2 深入LSTM结构 16](#_Toc533533282)

[5.2 关于CRF 18](#_Toc533533283)

[5.2.1什么是CRF 19](#_Toc533533284)

[5.2.2 如何训练CRF 21](#_Toc533533285)

[5.3 LSTM + CRF 21](#_Toc533533286)

[5.4深度学习库keras及其在NLP中的应用 21](#_Toc533533287)

[6 分词系统评估方法及实例分析 22](#_Toc533533288)

[6.1 分词系统评估方法 22](#_Toc533533289)

[6.2 基于机械匹配和概率统计相结合的中文命名实体识别研究 22](#_Toc533533290)

[6.3 基于表示学习的中文分词算法探索 26](#_Toc533533291)

[6.4 统计与字典相结合的领域自适应中文分词 29](#_Toc533533292)

[7 汉字处理：拼音输入法 33](#_Toc533533293)

[7.1 拼音输入法（基于统计） 33](#_Toc533533294)

[7.2 单机输入法时代 33](#_Toc533533295)

[7.3 互联网时代的输入法 34](#_Toc533533296)

[7.4 云时代的输入法 35](#_Toc533533297)

[7.4分词引擎介绍 37](#_Toc533533298)

[8 总结与展望 39](#_Toc533533299)

[9 参考文献 40](#_Toc533533300)

# 1 绪论

## 1.1 研究背景

自然语言处理是利用计算机对人类特有的自然语言的信息进行各种类型的处理与加工的技术，是人工智能研究的重要组成部分。随着信息社会、信息技术的快速发展，中文信息处理的作用越来越大，其重要性不言而喻。经过众多学者在几十年间的研究，中文信息处理领域取得了许多优秀的成果，不过中文分词仍然是中文信息处理的瓶颈。“分词”即为将一个句子中的字符切分为词的过程，也是中文信息处理的最基础的研究工作。

中文与英语等语言不同，英语的单词是天然分隔开的，而汉语中的词并不是如此，并且汉语词语基本上没有形态变化。一个汉语句子由一串前后连续的汉字组成，词与词之间没有明显的分界标志。汉语的书面表达方式是以汉字为最小单位的，但在自然语言理解中，词是最小的、能独立活动的、有意义的语言成分。中文分词就是把没有分割标志的汉字串(没有词的边界)转换到符合语言实际的词串即在书面汉语中建立词的边界。中文分词是中文信息处理的第一道“工序”，中文分词在中文信息处理许多应用领域（机器翻译、文本分类、术语抽取、文本摘要、信息检索等）中扮演着极为重要的角色。随着中文信息处理的发展，中文分词也得到了很大的发展，出现了众多的算法。根据其特点，可以将现有的分词算法分为三大类：基于词典的分词算法、基于统计的分词算法、基于神经网络的分词算法。本文首先介绍了这三类中文分词方法，然后对分词系统评估方法进行了介绍及实例分析，最后介绍了中文分词在拼音输入法上的应用。

## 1.2 国内外研究现状

国外早在1963年就初步建成了自然语言处理系统，而国内对于这方面的研究起步较晚，中文自动分词的研究起始于1980年前后，1983年，北京航空航天大学的梁南元副教授实现了第一个汉语自动分词系统CDWS（Chinese Distinguishing Word System），完成了对2500万字的现代汉语词频统计工作。在该系统进行研发时，相关的研究人员在研究过程中首次论证了中文分词的可行性，并且初步建立了用于描述书面汉语的计算模型。此后又有数个系统问世，并提出了许多种分词方法。到今天为止，对于中文分词相关任务，国内外己经有很多学者做了大量的研究工作。主要使用的方法有传统的方法和近年来新提出的基于深度神经网络的方法。其中，传统方法主要包括基于词典的方法、基于统计的方法、统计与规则相结合的方法；基于深度神经网络的方法则是针对使用场景和各种神经网络模型的特点，选用各种不间的神经网络模型结构来完成分词任务。

早期的基于词典的方法主要是借助观察词语的语言学形态，编写规则模板进行过滤和抽取。在分词方面，Pak-kwong Wong等人，使用规则方法和基于词典的最大匹配方法进行分词。最大匹配算法实现简单，但处理歧义和新词的能力很弱。John S.Justeson等人则提出基于词典的方法进行术语抽取。其编写了几种种规则模板，利用词汇的前缀的词性作为判断一个词串是否为术语的依据。这样的方法的局限性是显而易见的：其编写的规则过于简单，无法对复杂的语言现象进行更为全面的覆盖。

在发现了基于词典的方法的局限性之后.人们开始考虑引入语料的统计信息。词语的统计特征是通过概率论和信息论的知识得到的，主要包括词频、互信息、上下文信息、词语度、领域度等。Patrick Hanks等人最早提出使用互信息的概念来进行术语抽取，即利用统计信息计算两个词语之间的结合程度，来对可能组合成术语的词串进行抽取。这项工作主要是用于英文语料的，由于汉语句子中没有天然的词汇界限，这种方法不能直接照搬到汉语语料中。Xue提出了采用基于字符在词内部位置标注的思想来解决分词问题，并采用手工标注数据基于最大熵模型进行了实验，实验表明，相比较最大匹配方法能够取得更高的准确率。刘群、张华平等人提出了一种基于层叠隐马模型的汉语词法分析方法，在分词方面采取基于类的隐马模型，未登录词和词典收录的普通词作同样处理，通过引入角色HMM来达到对未登录词的识别。沈勤中等奖中文分词视作一个对中文字符串进行分类的过程，采用条件随机场模型对各个汉字进行相应的标记，最后转换成分词结果，在现有的CRF模型的基础上从字的构词能力角度探究了字位置的概率特征，提出了基于字位置概率特征的条件随机场分词方法。

很快，学界提出了统计与规则相结合的方法。这种方法也是传统方法中最为常用的一种方法。Zhao Hai等便用条件随机场(Conditional Random Fields, CRFs)棋型，将分词任务转化为序列标注任务，在2006年的Backoff国际中文评测任务中取得了非常优异的成绩：Katerina Frantzi等人在其工作中提出了使用改进型C-value 算法进行术语抽取的方法。这种方法结合停用词表（Stop List)和若干词性规则模版，通过词频统计信息和上下文信息对语料中的潜在术语进行挖掘抽取：Dat Ba Nguyen等人使用将CRF应用到命名实体识别任务。该工作通过加入大大量的语言特征来进行词义消歧，在CoNLL-YAGO数据集上，相较于之前的最好结果，取得了1.8%的F1值的提升；刘辉，刘耀等人则针对专利文献的语言学特点，制定了一系列的特征模版，使用CRF对专利文献术语进行了抽取，其在通信相关的术语抽取实验达到了80.9%的准确率和75.6%的召回率。

尽管传统的中文分词方法现在己经发展成熟，并且在评测任务中取得过非常好的成绩，但是其始终存在一个无法避免的局限性，这种分词方法十分依赖于人工制定的规则和特征模版。近年的使用传统中文分词方法的工作，重点都集中在大量的特征工程上。规则和特征模版需要具有专业语言学只是的人进行编写。提高了人工成本，降低了工作效率，而且人工编写的模版有限，即便是专家所写也无法保证能够覆盖所有的语言现象，因此近年来学界开始研究不需要人工制定规则和特征模板的技术。

对着近年来深度神经网络技术的兴起，国内外学者也开始将深度神经网络应用在中文分词相关的任务上。这种方法的最大优势在于不需要人工制定规则和特征模版，模型可以自动从标注语料中学习到语言的内在规律。

国内外很多学者都在做相关方面的研究。Chen Xinchi等人在EMNLP2015上提出了使用长短时记忆网络（Long Short-Term Memory，LSTM）来进行中文分词，使用LSTM来对字向量进行编码，并使用线性层进行预测，在PKU数据集上取得了95.7%的F1值；Jason P.C.Chiu, Eric Nichols等在2015年首次提出了一种融合双向长短时记忆网络（Bi-directional Long Short-Term Memory，Bi-LSTM）和卷积神经网络（Convolutional Neural Network, CNN）的命名实体识别方法，这种方法使用Bi-LSTM网络进行编码，卷积神经网络作为输出层，最高取得了89.93%的F1值。

也有一些学者将神经网络与传统方法相结合。Guillaume Lample等人在ACL2016上提出了一种融合了Bi-LSTM网络和CRF的方法，同时结合神经网络的学习能力和CRF的分类能力，取得了非常好的效果，这种方法在CoNLL-2003测试集上达到了90.94%的F1值。Zhou Hao等人在2017年提出了一种利于词语窗口将上下文信息融入到分词模型中的方法，这种方法在PKU数据集上取得了96.0%的F1值。

Xu Mingbin等人在ACL2016上提出了一种使用FOFE编码（Fixed-size Ordinally Forgetting Encoding）和前馈神经网络的命名实体识别方法，解决了词串判断时出现的边长编码问题，该方法在CoNLL-2003测试集上取得了90.85%的F1值；Sun Zhiqing等人在2017年提出了将传统的基于字词的分词标注方法替换为基于词间隔的标注方法，在网络结构上使用深层卷积神经网络进行编码，该方法在PKU数据集上取得了96.0%的F1值。

# 2 分词方法的困难性与关键问题

## 2.1分词的困难性

分词涉及多个方面的问题，主要包括核心词表、词的变形问题、词缀问题、歧义字段问题和未登陆词问题。

首先是核心词表、词的变形结构和词缀问题。许多算法都要一个核心词表，在核心词表中的词都要切分出来，然而目前并没有一个统一的词表。一般情况下都是分领域的词表适用较好。其次是词的变形问题。汉语中许多动词和形容词都可以产生变形结构。而这些变形结构需要规范来指导分词，但是由于汉语的复杂性，没有合理又可操作的规范。汉语中有些词单独使用没有实际意义，也给分词造成困难。

其次是歧义字段问题。汉语分词中通常会出现具有多种切分可能分字段，称为歧义字段。歧义字段可分为交集型（OAS）、覆盖型（CAS）、和真歧义。对人来说，可以通过上下文理解，但是机器很难判断如何切分。如果在ABC子串中，A，AB，BC，C都是词表中的词，成为交集型字段。交集型字段占全部歧义的85-90%，是歧义中重点需要解决的问题。覆盖型字段，如AB中，A是词，B是词，AB也是词，例如“马上”。真正歧义是指没有上下文环境就不能理解的歧义，例如“乒乓球拍卖完了”，既可以理解成“乒乓球／拍卖／完了”又可以理解为“乒乓／球拍／卖完了”。歧义问题一般分为两个步骤，歧义发现与歧义消解，可以使用双向扫描法发现歧义，通过最短路径等方法进行粗分。

再者是未登录词的识别。未登录词指的是词典中没有的词。包括各种命名实体（人名、地名、机构名、译名、时间、货币等等）和网络新词（ｂ站、秀等）。其中人名、地名、机构名等有常用词并且规律性强，所以利用规则和统计信息可以识别。数字型的合成词个构成方式呈现出规律性，也容易识别。随着社会的发展，新词与常用词混用进而形成新的歧义。未登陆词分为五类：一是缩略词，如中共、非典；二是专有名词，主要包括人名、地名、机构名；三是派生词，如机械化、生产者；四是复合词。产生未登录词的直接原因主要是词典中词目选择和词目的数量，机器可读词典与待处理文本的匹配关系。目前处理未登录词的方法主要是基于分解和动态规划策略的方法。

最后汉语分词在不同的应用场景对于分词有不同的要求。例如在检索系统中，分词越小越有利于检索结果呈现；而在键盘输入系统中，分词的要求则是要忽略一些错误，例如输入“zhege”和“zege”都输出“这个”；在语音识别系统中，可能速度和最长匹配是关注的重点。

## 2.2分词的关键问题

中文分词处理的关键问题包含歧义字段的处理和未登录词语的识别。

从歧义字段的构成形式上来看，歧义字段可以分为两类：一类是交搭型歧义，这类歧义因词与词之间的交搭造成的，另一类是多义覆盖型歧义字段，这类歧义切分因词与词之间的串联而形成，还有一类是真正的歧义。真歧义是指没有上下文环境时，无法判别其切分形式的歧义现象。

现有的歧义处理方法中，一般基于两步的策略：首先对歧义进行发现，然后才是歧义消解。歧义发现一般使用双向最大匹配法、全切分发现算法等等。歧义消解的方法主要包括：基于语言知识的规则方法，主要使用词法、句法甚至语义等层次的语言知识作为歧义消解的启发式规则；基于统计的方法，例如基于 MM、HMM 和 n-gram 的分词词性标注来消解切分歧义，通过 Viterbi算法寻求最优路径；基于实例的方法，例如利用伪歧义消解不依赖于上下文的特点，将为歧义字段的正确切分方式预先纪录在一张表中，通过直接查表对分词歧义进行消解；基于词典+实例的方法，例如“正向最大匹配+回退一字”，回退后的处理以实例和规则为标准。

未登录词识别通常使用无监督学习的方式解决。无监督学习方式是一种独立于词典、从无标记语料中挖掘知识的分词方法。它的优点是不依赖与词典、训练语料，无需训练，可以用于无构词规律的未登录词发现。

# 3 基于机械匹配的分词方法

机械匹配法是自动分词中最基础的算法，其基本思想是：事先建立一个词库，其中包含所以可能分词；对于给定的字串s，按照某种原则切取s的子串，若该子串与某词条相匹配，则该子串是词，则继续分割剩余的部分，直到全部切完；否则，该子串不是词，转上重新切取s的子串进行匹配。

根据s切取的原则不同，机械匹配算法分为几类，一是按照方向分，分为正向匹配法和逆向匹配法，按照每次匹配时优先长词还是短词可分为最大匹配和最小匹配；三是根据匹配不成功时重新切取的策略，可分为增字法和减字法。机械匹配的一种优化算法是N—最短路径法。

## 3.1 正向最大匹配法

正向最大匹配法，也称为FMM法。根据经验设定切词的最大长度max\_len,每次扫描的时候寻找当前开始的这个长度的词与字典中的词匹配，如果没有找到就缩短长度继续寻找。其流程图如图所示。

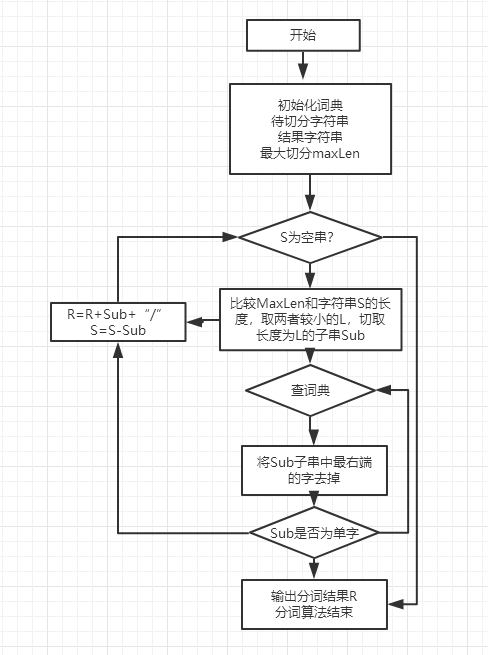


图3-1

## 3.2 逆向最大匹配法

逆向最大匹配法，也成为RMM方法。在切分汉字的时候，不是按照从左到右的顺利，而是按照从汉字尾端开始抽取。一般来说，逆向最大匹配算法要优于正向匹配。但是只依赖于最大匹配的精度还不能满足实际的需要，只是作为一种粗分的手段，还要与其他的方法进一步提高切分的准确率。

## 3.3 双向扫描

双向扫描法是将正向和逆向最大匹配得到的分词结果进行比较，从而决定正确的分词方法。据研究结果表明，只有9%的句子两种切分方法得到的结果不一样，且其中必有一个是正确的，只有1%的两种切分都是相同的，并且都是错误的。双向扫描法流程图见图2。

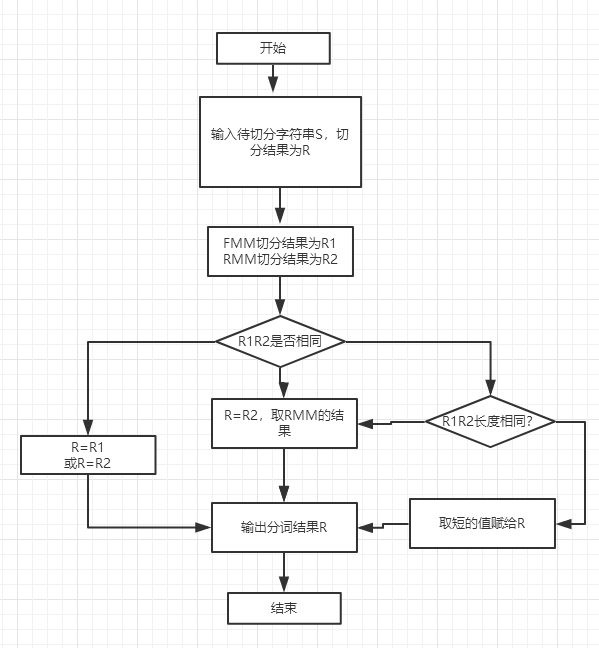


图3-2

双向扫描法的侧重点是检错和纠错，这种方法执行效率较慢，需要加载两个词典，占用较多的空间资源。

## 3.4 N\_最短路径

N-最短路径是对最短路径的一种改进，它的基本思想是基于现有的词典，将每个句子分解为一个带权的有向无环图，每个字代表图的表，边代表可能的分词，边的起点为词的第一个字，终点为词尾的下一个字，可以使用词的频率来表示边的权值。假定词与词之间是相互独立的，得到一个基于N-最短路径算法的一元模型，的搭配最终的边长公式为：

由公式知，词频越高边长越短，如此字串S和它所包含的词与有向无环图G的边一一对应，见下图。

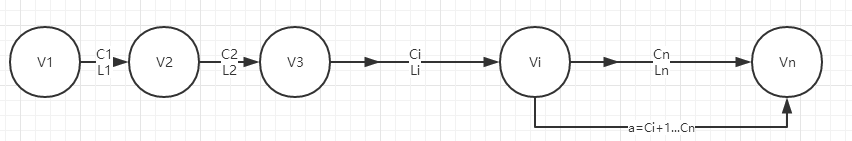


图3-3 有向无环图示意图

最终的结果即在上述带权的有向无环图的寻找N条权值和最大的路径。使用Dijkstra算法计算最短路径。假设N取5，则选取最短的前5条结果，将粗分结果进行停用词过滤。

机械的分词方法常被称为基于词表的分词方法。这种方法虽然实现简单、速度快，但处理分词歧义能力较差，严重依赖于词表，不能识别新词语。基于机械匹配的“演绎”推理方法，能较好的解决分词歧义和未登录词，具有一定的领域适应性，效率较高。但中文语言现象非常复杂，目前很难制定出涵盖所有语言现象的人工规则，近几年淡出人们的视线。

# 4 基于统计的分词方法

基于机械匹配的中文分词算法非常简单和方便，但是对于歧义词以及未登录词的分词效果不好，因为这些词语是词典里没有的。实际上，中文分词的主要瓶颈是切分排歧和未登录词识别。切分歧义和未登录词降低了自身正确切分的可能性，同时还干扰了其相邻词的正确处理。更糟糕的是，未登录词往往和切分歧义交织在一起，进一步增加了分词的难度。如：在 “克林顿对内塔尼亚胡说”中，“内塔尼亚胡”是一个词典中没有收录的译名，实际切分的时候，“对”与“内”，“胡” 与“说”往往会粘在一起，最终导致错误的切分结果：“克林顿/对内/塔尼亚/胡说/”。并且，基于机械匹配的中文分词算法没有考虑词语与词语之间的关系，也没有进行语法方面的考量。而基于统计的分词方法，优势就在于处理歧义，对于识别未登录词拥有较好的识别能力，分词精度比较大。现在基于统计的中文分词方法渐渐成为了主流方法。

基于机械匹配的方法是理性主义研究方法，基于统计的方法是经验主义研究方法。两种方法最明显的区别是：基于统计的分词方法摈弃了词典，在进行分词的时候不需要词典作为输入，而是需要输入各种各样的语料库，通过训练出的语言模型来分词。

基于统计的分词方法通过对大众语料进行统计、分析、提取和加工，得到用标识(tags)标记的标准化语料库供进一步的研究，如马尔可夫模型(Markov model)语法应用。这种基于统计分词方法有许多的应用：早期以真实语料为基础的语言学研究，有夸克和里奇等编写的《英语语法大全》。知名度高、影响力大的计算机语料库则有20世纪60年代由美国Brown大学整理的Brown语料库，约含100万个经过标注的词汇。第一本基于语料库语言学出版的词典是波士顿出版商霍顿一米福林(Houghton-Mifflin)出版的《美国传统词典》。继Brown美式英语言语料库产生之后，柯林斯(Collins)出版商出版了由英国挪威等国家的大学学者整理的COBUILD语料库，在此基础上研究的成果有诸如《柯林斯英语大辞典》等柯林斯英语学习系列词典300余种，可以说是语料库语言学在词典方面开发应用最广泛的一个例子。

## 4.1 基于统计的中文分词步骤

第一步，加载大规模中文语料集；这里需要输入中文语料集一般是包含大规模（十万数量级以上）的中文句子的文档。第二步，通过大规模中文语料集训练出相应的语言模型；这个“训练”的意思其实是，利用统计机器学习模型来学习词语切分的规律。第三步才是进行中文分词。

## 4.2 基于统计的分词方法的语言模型

（1）N元语言模型（N-Gram）

20世纪70年代，贾里尼克用一个简单的统计语言模型解决了判断文字序列是否合法的问题。他的思路很简单：一个句子是否合理取决于它的可能性大小。其中，这个“可能性”需要用概率来计算，如果第一个句子计算出来的概率最大，那么第一个句子就是结构最为合理的序列。也就是说，这个语言模型可以计算任何句子的概率。

N元语言模型的定义是：对于一个中文字符串S，可以将其看成一个连续的字符串序列，这个序列可以是一句话，一个词，也可以是一个单独的字。对于字符串S有一种切分方式为：w1 w2 w3 到 wn，N元语言模型实际上计算的是字符串S被这样切分为w1 w2 w3 到 wn的概率有多大，然后将这个概率记为P（s），则可以得到P(s) = P(w1 w2 … wn)；对于字符串中的每个词语wi，N元语言模型假定wi出现的概率与它前面的i-1个词语有关，然后由条件概率公式可以化简。

但这样做的话，当需要计算的句子很长时，那在计算末尾处的词语时，计算量就非常大。而且在一般情况下，第 i 个词的出现只是和相邻的几个词语的关联比较大。如果前面所有的词都计算进去的话，会产生干扰数据。因此在实际使用中，N会选择成1，2，3，4。选择更高的情况很少，因为训练它需要更庞大的语料集，而且使用更高元的语言模型，数据稀疏严重，时间复杂度高，精度却提高的不多。

例如，我们可以计算“我爱你”的概率：N元语言模型根据一个词语的前N−1个词语来计算这个词语的概率。如果我们有了一个二元模型，“我爱你”的概率就等于P（我）×P（爱 | 我）×P（你 | 爱）；以此类推，三元模型下“我爱你”的概率等于P（我）×P（爱 | 我）×P（你 | 我，爱）。

（2）互信息模型

前面有提到大规模的中文语料集，那么，应该怎样从语料集中找到“词语”呢？我们可以根据语料库中，字与字之间的紧密程度来判断这个组合是否为词语。如果两个字总是紧挨着出现，那么这两个字在很大程度上就可以被判定为是一个中文词语。如果两个字没有相邻出现，或只相邻出现了一次或者几次，而单独出现了很多次，那么这两个字的紧密度就不那么强，从而可以判定这两个字不能组成一个词语。

互信息描述的就是两个字之间的紧密程度，其公式为MI(x,y)=log2 (P(x,y)/P(x)P(y))。MI值越大，表示两个字之间的结合越紧密，两个字如果要么不出现，要么一起出现，则他们的亲密度就非常高，如“尴尬”“匍匐”等词；若MI值小于或约等于零，则两个字之间的关系很弱，基本不会成词。但互信息也会抽出一些紧密度高但并不是词语的常用字组，例如“这一”、“我的”、“许多的”之类，所以在实际应用时还需要与其它方法相结合。

（3）最大熵模型

熵的概念最早来源于物理学，熵是对一个随机事件不确定性的衡量。简单来说，东西变得无序化的过程，就是熵增大的过程。熵越大，系统越混乱无序。例如，房间变乱的过程，熵在增大。与此类似的还有，一滴墨水均匀地分布到一杯清水的这个过程中，熵也在增大。因为杯子里的墨水和水逐渐混合到了一起，当它们混合均匀时，熵最大，因为均匀分布时熵最大的模型。

最大熵原理是统计学的原理，也是概率模型学习的一个准则。最大熵原理认为，学习概率模型时，在所有可能的概率模型中，熵最大的模型是最好的模型。

举个例子，一个色子有6个面，问人掷到每个面的概率分别是多少，大家一定会说每面都是1/6。为什么大家会说1/6，而不是说什么1是1/3，2是1/2这样子的组合呢？因为大家潜意识里面觉得这样子最“保险”。“最保险”其实就是因为均匀分布刚好是熵最大的模型。最大熵模型认为，对于那些未知的事件，认为他们等可能，是最好的，也满足最大熵的情况。在已知条件下选择一个合适的分布来预测可能出现的事件，最主要的思想是在只掌握未知分布的部分知识时，应该选取符合这些知识，但熵值最大的概率分布。不确定性越大，熵值就越大；若随机变量退化成定值，熵为0。

由于最大熵模型在数学上十分完美，对科学家们有很大的诱惑力，因此不少研究者试图把自己的问题用一个类似最大熵的近似模型去套。但近似之后最大熵模型就变得不完美了，所以实验出来的结果并不怎么好。于是很多人又放弃了这种方法。而第一个在实际信息处理应用中验证了最大熵模型的优势的人是一个叫拉纳帕提(Adwait Ratnaparkhi)的人，原IBM现微软的研究员。拉纳帕提的聪明之处在于他没有对最大熵模型进行近似，而是找到了几个最适合用最大熵模型、而计算量相对不太大的自然语言处理问题，比如词性标注和句法分析。拉纳帕提成功地将上下文信息、词性（名词、动词和形容词等）、句子成分（主谓宾）通过最大熵模型结合起来，做出了当时世界上最好的词性标识系统和句法分析器。拉纳帕提的论文发表后让人们耳目一新。拉纳帕提的词性标注系统，至今仍然是使用单一方法最好的系统，这个系统也使人们看到了用最大熵模型解决复杂的文字信息处理的希望。

（4）隐马模型

隐马模型是一种非常经典的统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。下面用一个简单的例子来解释HMM模型：

假设这里有三个不同的骰子。第一个骰子是最常见的6面的骰子（称这个骰子为D6），每个面（1，2，3，4，5，6）出现的概率是1/6。第二个骰子是个4面的骰子（称这个骰子为D4），每个面（1，2，3，4）出现的概率是1/4。第三个骰子有8面的骰子（称这个骰子为D8），每个面（1，2，3，4，5，6，7，8）出现的概率是1/8。

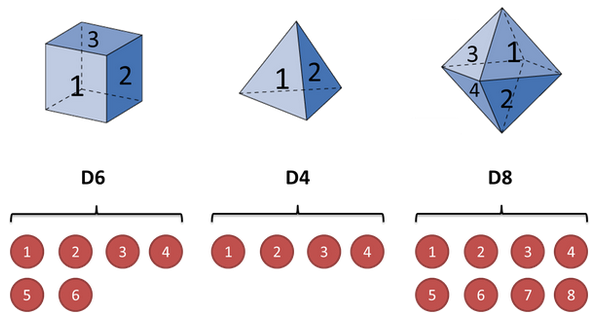


图4-1

假设我们开始掷骰子，我们先从三个骰子里挑一个，挑到每一个骰子的概率都是1/3。然后我们掷骰子，得到一个数字，1，2，3，4，5，6，7，8中的一个。然后不停的重复上述过程，我们就会得到一串数字，每个数字都是1，2，3，4，5，6，7，8中的一个。假设我们掷骰子10次，然后得到这么一串数字：1 6 3 5 2 7 3 5 2 4。那么这串数字就是可见状态链。但是在隐马尔可夫模型中，我们不仅仅有这么一串可见状态链，还有一串隐含状态链。在这个例子里，这串隐含状态链就是所用的骰子的序列。比如，隐含状态链有可能是：D6 D8 D8 D6 D4 D8 D6 D6 D4 D8。

一般来说，HMM中说到的马尔可夫链其实是指隐含状态链，因为隐含状态（骰子）之间存在转换概率（transition probability）。在我们这个例子里，D6的下一个状态是D4，D6，D8的概率都是1/3。D4，D8的下一个状态是D4，D6，D8的转换概率也都一样是1/3。这个例子里设定的概率是相等的，但其实转换概率是可以随意设定的，比如，我们可以这样定义，D6后面不能接D4，D6后面是D6的概率是0.9，是D8的概率是0.1。这样就能够构成一个新的HMM。同样的，尽管可见状态之间没有转换概率，但是隐含状态和可见状态之间有一个概率叫做输出概率（emission probability）。

（5）层叠隐马模型

隐马模型是一种非常经典的统计方法，但是相对于复杂的自然语言现象来说，传统的隐马模型显得略为简单。所以，有时候需要采用多个层次的隐马模型，对汉语词法分析中遇到的不同情况，进行分别处理，即为层次隐马模型HHMM。层次隐马模型是一种不同于隐马模型的数学模型，它更为复杂，并且比隐马模型具有更强的表达能力，但使用起来的时空开销也比较大。层次隐马模型求解的时间复杂度是O(NT3)，而隐马模型只有O(NT)。

而层叠隐马模型实际上是若干个简单隐马模型的组合，各层隐马模型形成一种紧密的耦合关系。整个系统的时间复杂度与隐马模型相同，为O(NT)。

基于层叠隐马模型的中文分词算法步骤：第一，在预处理的阶段，采取N-最短路径粗分方法，快速得到能覆盖歧义的最佳N个粗分结果；第二，在粗分结果集上，采用底层隐马模型识别出普通无嵌套的人名、地名，并依次采取高层隐马模型识别出嵌套了人名、地名的复杂地名和机构名；第三，未登录词与歧义词都不作为特例，而是与普通词一起参与各种候选结果的竞争；第四，在全局最优的分词结果上进行词性的隐马标注。

该方法已经应用到中科院计算所汉语词法分析系统ICTCLAS中，这个系统是目前最好的汉语词法分析系统之一。下图为应用的基于CHMM的汉语词法分析框架：

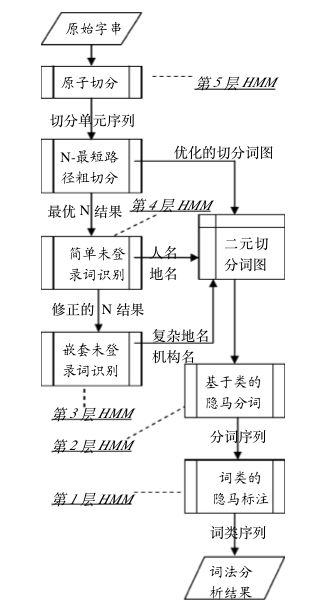


图4-2

## 4.3 切分歧义以及N-最短路径的切分排歧策略

前面有提过，从构成形态上划分，切分歧义一般分为交叉歧义和组合歧义；从排歧角度来看，可以分为全局歧义和局部歧义。“结合/成/分子/时”就是非常典型的交叉歧义，因为“合成”、“成分”、“子时”也都是常见的词语；组合歧义可以举例“这/个/人/手/上/有/痣 ”，在这句话里，“人”和“手”是两个词语，但它们组合起来的“人手”也可以作为一个单独的词语，由此形成歧义。

全局歧义指的是必须结合句子上下文才能准确排除的歧义，局部歧义则可以在句子内部进行排除。根据统计发现，局部歧义占大多数，全局歧义几乎可以忽略不计。例如“乒乓球拍卖完了”就属于全局歧义，必须要结合上下文才能确定这句话的意思是“乒乓球拍/卖/完/了”还是“乒乓球/拍卖/完/了”。

那么这些切分歧义的应该怎么处理呢？这里采取的是N-最短路径的切分排歧策略。它的基本思想是：在初始阶段保留切分概率最大的N个结果作为分词结果的候选集合。在未登录词识别、词性标注等词法分析之后，再通过最终的评价函数，计算出真正最优结果。

实际上，N-最短路径方法是最少切分方法和全切分的泛化和综合。该方法通过保留少量大概率的粗分结果，可以最大限度保留歧义字段和未登录词。最短路径法正确切分覆盖率率比较高，又解决了全切分搜索空间过大，运行效率低的弊端。下表给出了8-最短路径与常用算法在切分结果包容歧义方面的对比测试结果：



图4-3

说明：

1)切分最大数指的是句子可能的最大切分结果数。

2)切分平均数指的是单个句子平均的切分结果数。

3)正确切分覆盖率=正确切分被覆盖的句子数/句子总数

## 4.4 机械匹配与统计相结合的中文分词方法

基于统计的中文分词方法也存在着一些不足，例如，当输入的语料库只是有关某一领域时，那么其分词效果就会受到限制，对该领域之外的语言无法进行准确的处理。所以现在很多地方使用的是基于机械匹配与基于统计的分词方法相结合，例如前文提到的中科院计算所汉语词法分析系统ICTCLAS，就是首先使用机械匹配对语料进行初步的切分，然后通过统计以及其它简单规则来进行歧义消解和未登录词的识别，该系统框架如下：

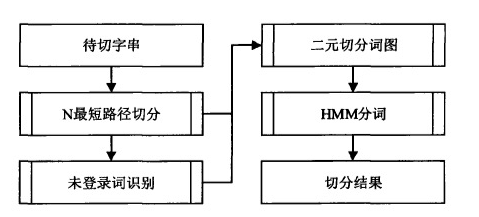


图4-4

# 5 基于理解的中文分词

接下来重点讲一下近几年的中文分词有方法，近些年很多学者都用深度网络模型来对中文进行分词，以下就是较新分词方法，它主要分为两类，一个是基于字符的中文分词，一类是基于词的中文分词。近些年很多学者致力于中文分词，对中文分词的方法进行了改进，其中运用到了RNN，CNN，GNN等深度神经网络来自动地获取特征，从而代替传统方法中手工定义的特征。从句子中获取简单的特征改为获取复杂的特征，从单一语料库单一标准的模型改进为可以使用多语料进行分词等等。接下来我将详细介绍基于双向循环神经网络和条件随机场的中文分词。

## 5.1 关于LSTM

LSTM 是循环神经网络的一种变种，是处理序列的小能手。

### 5.1.1 什么是LSTM

长短期记忆（Long short-term memory, LSTM）是一种特殊的RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。简单来说，就是相比普通的RNN，LSTM能够在更长的序列中有更好的表现。

LSTM结构和普通RNN的主要输入输出区别如下所示。

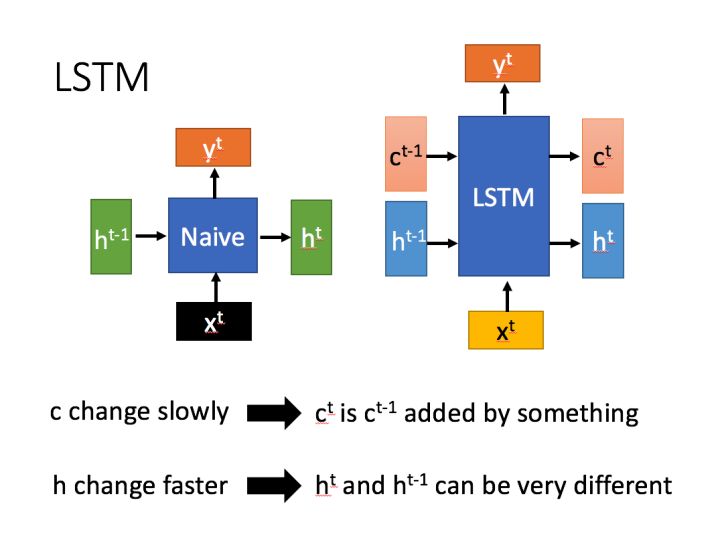


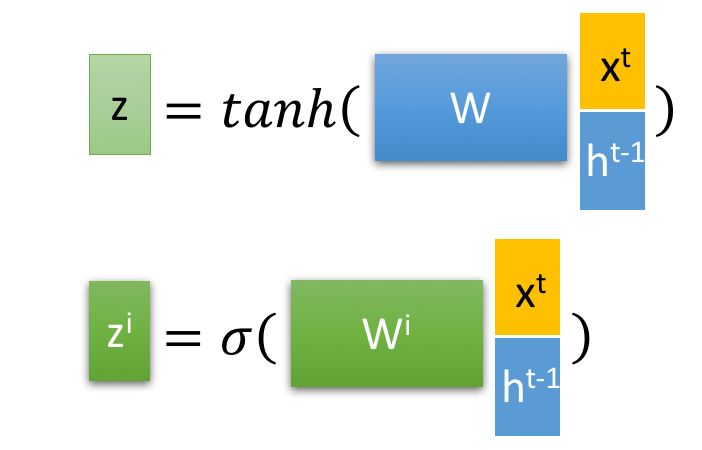
图5-1

相比RNN只有一个传递状态，LSTM有两个传输状态，一个（cell state），和一个（hidden state）。其中对于传递下去的改变得很慢，通常输出的是上一个状态传过来的加上一些数值。而则在不同节点下往往会有很大的区别。

### 5.1.2 深入LSTM结构

下面具体对LSTM的内部结构来进行剖析。

首先使用LSTM的当前输入和上一个状态传递下来的拼接训练得到四个状态。



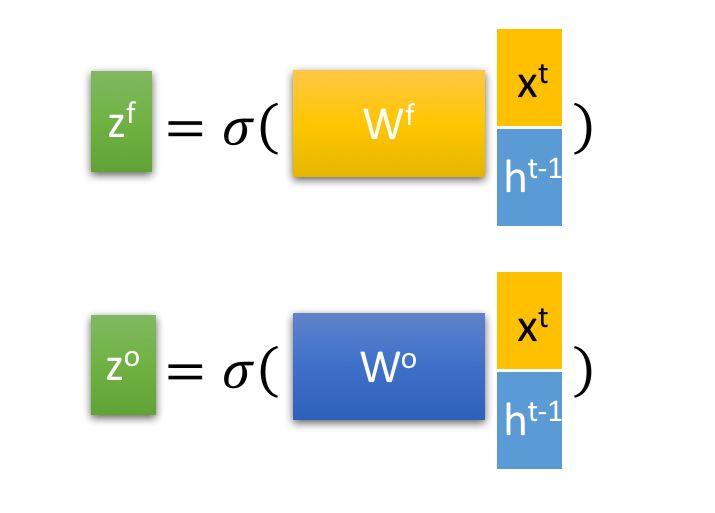


图5-2

其中，1543808334(1)，1543808352(1)，1543808378(1)是由拼接向量乘以权重矩阵之后，再通过一个sigmoid激活函数转换成0到1之间的数值，来作为一种门控状态。而z则是将结果通过一个tanh激活函数将转换成-1到1之间的值。

下面开始进一步介绍这四个状态在LSTM内部的使用。

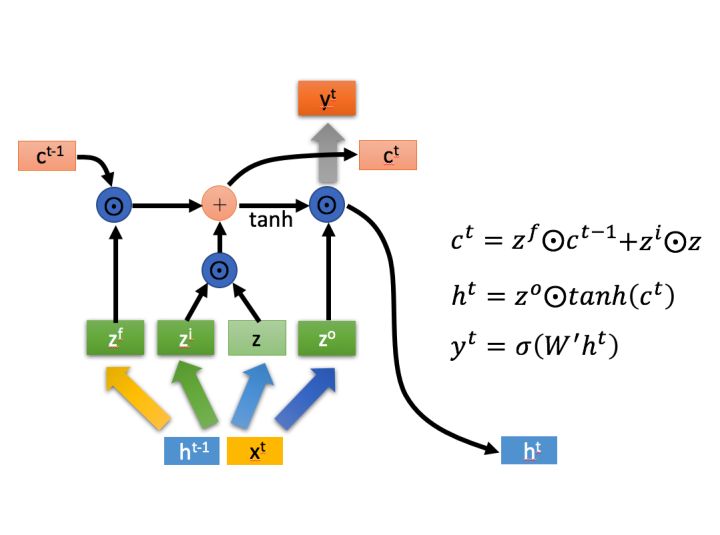


图5-3

1543807951(1) 是Hadamard Product，也就是操作矩阵中对应的元素相乘，因此要求两个相乘矩阵是同型的。 则代表进行矩阵加法。

LSTM内部主要有三个阶段：

1.忘记阶段:这个阶段主要是对上一个节点传进来的输入进行选择性忘记。简单来说就是会 “忘记不重要的，记住重要的”。

具体来说是通过计算得到的1543807996(1)（f表示forget）来作为忘记门控，来控制上一个状态的1543808034(1)哪些需要留哪些需要忘。

2.选择记忆阶段:这个阶段将这个阶段的输入有选择性地进行“记忆”。主要是会对输入1543808064(1)进行选择记忆。哪些重要则着重记录下来，哪些不重要，则少记一些。当前的输入内容由前面计算得到的1543808098(1)表示。而选择的门控信号则是由1543808125(1)（i代表information）来进行控制。

3.输出阶段:这个阶段将决定哪些将会被当成当前状态的输出。主要是通过1543808168(1)来进行控制的。并且还对上一阶段得到的1543808192(1)进行了放缩（通过一个tanh激活函数进行变化）。

与普通RNN类似，输出1543808216(1)往往最终也是通过1543808240(1)变化得到。

## 5.2 关于CRF

CRF是一种概率无向图模型，也是处理序列的小能手。

### 5.2.1什么是CRF

CRF即条件随机场（Conditional Random Fields），是在给定一组输入随机变量条件下另外一组输出随机变量的条件概率分布模型，它是一种判别式的概率无向图模型，既然是判别式，那就是对条件概率分布建模。

CRF较多用在自然语言处理和图像处理领域，在NLP中，它是用于标注和划分序列数据的概率化模型，根据CRF的定义，相对序列就是给定观测序列X和输出序列Y，然后通过定义条件概率P(Y|X)来描述模型。

CRF的输出随机变量假设是一个无向图模型或者马尔科夫随机场，而输入随机变量作为条件不假设为马尔科夫随机场，CRF的图模型结构理论上可以任意给定，但我们常见的是定义在线性链上的特殊的条件随机场，称为线性链条件随机场。

(1)概率无向图模型

前面说到CRF的输出随机变量是一个概率无向图模型，那么现在看看该模型。

概率无向图模型是由无向图表示的联合概率分布，假设联合概率分布P(Y)通过无向图来表示，则在图中节点表示随机变量，边表示随机变量之间的依赖关系，联合概率分布P(Y)满足马尔科夫性则称其为概率无向图模型，或者是马尔科夫随机场。

图是一个由节点和边组成的结构体，无向是指边没有方向，整个图记作G=(V,E)，其中V为节点的集合，E为边的集合。

每个节点v对应一个随机变量1543809156(1)，于是1543809180(1)，在观察序列X的条件下，每个随机变量1543809206(1)都满足马尔科夫特性，即

1543809225(1)

其中1543809242(1)表示w和v是图G中邻近的两个节点。

(2)线性链条件随机场

无向图的结构理论上可以是任意的，但在NLP中对于标记处理问题，对其建模主要用最简单最普通的链式结构，即线性链条件随机场。如下图，可以看到节点为线性链结构，节点对应了序列Y的元素，而观察序列X不做任何独立性假设，但X序列的结构也可以是线性链结构。

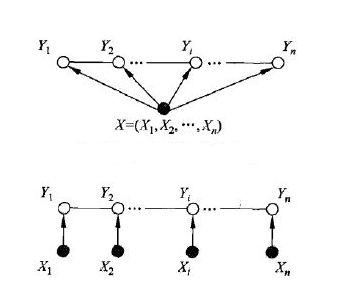


图5-4

综上所述，设有线性链结构的随机变量序X=(X1,X2,...,Xn),Y=(Y1,Y2,...,YN)，在给定观察序列X的条件下，随机变量序列Y的条件概率分布为P(Y|X)，若其满足马尔科夫特性，即

P(Yi|X,Y1,Y2...Yn)=P(Yi|X,Yi−1,Yi+1)

这时P(Y|X)则为线性链条件随机场。

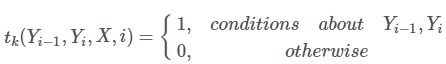
(3)概率的定义

在线性链条件随机场中，在给定的观察序列X情况下，某个特定序列Y的概率为P(Y|X)P(Y|X)，根据定义有，

P(Y|X)=exp(∑i,kλktk(Yi−1,Yi,X,i)+∑i,lμlsl(Yi,X,i))

其中，tk(Yi−1,Yi,X,i)表示转移函数，表示在序列X下序列Y在位置i-1及i对应的值转移概率，而sl(Yi,X,i)表示状态函数，表示在序列X下序列Y在位置i对应的值概率。另外λk,μl分别为两个函数的权重。

转移函数和状态函数都称为特征函数，特征函数一般取值0或1，满足特征函数的则为1，否则为0。比如下面的转移函数，只有当Yi−1,Yi满足一定的条件时才为1，否则为0。



如果我们令sl(Yi,X,i)=sl(Yi−1,Yi,X,i)，则转移函数和状态函数可以统一由特征函数表示，对特征在各个位置i求和，有

Fk(Y,X)=∑ni=1fk(Yi−1,Yi,X,i)

最后再加上归一化，最终条件随机场的条件概率为，

1543815488(1)

其中，

1543815510(1)

### 5.2.2 如何训练CRF

训练CRF主要就是要训练特征函数的权重，对于训练集(x1,y1),(x2,y2),...,(xn,yn)，采用极大似然估计法计算权重参数，条件概率的对数似然函数为：

1543812710(1)

其中p~(x,y)为训练样本集中xy的经验概率，它等于xy同时出现的次数除以样本空间容量；p~(x)为训练样本集中x的经验概率，它等于x出现的次数除以样本空间容量。

然后对λ求导，令其为0再求解出λ，即得到解。因为极大似然估计法不一定能得到一个近似解，所以需要利用一些迭代技术来确定参数，比如GIS或IIS算法，这里不再深入。

5.2.3 什么时候考虑CRF

如果信息是与时间或空间的前后有关联时要考虑到CRF。

## 5.3 LSTM + CRF

LSTM 和 CRF 我们都了解了，单独使用它们都挺好理解，但如何将它们结合起来是我们更关注的。其实如果没有 CRF 参与其实也是可以完成任务的，我们说单向 LSTM 网络因为没考虑上下文，所以引入了双向 LSTM 网络，此时每个词经过词嵌入层再进入前向和后向循环神经网络，这时输出能得到标签的概率。

在没有 CRF 参与的时候可能会存在一个小缺陷，它没办法约束标签的特征，比如某标签到另外一标签的转换概率。如果有标签的特征就能进一步提高学习能力。

## 5.4深度学习库keras及其在NLP中的应用

Keras是一个非常易用的深度学习框架，使用python语言编写，是一个高度模块化的神经网络库，后端同时支持Theano和TensorFlow，而Theano和TensorFlow支持GPU，因此使用keras可以使用GPU加速模型训练。Keras中包括了构建模型常用的模块，如Optimizers优化方法模块，Activations激活函数模块，Initializations初始化模块，Layers多种网络层模块等，可以非常方便快速的搭建一个网络模型，使得开发人员可以快速上手，并将精力放在模型设计而不是具体实现上。常见的神经网络模型如CNN，RNN等，使用keras都可以很快搭建出来，开发人员只需要将数据准备成keras需要的格式丢进网络训练即可。如果对keras中自带的layer有更多的需求，keras还可以自己定制所需的layer。

Keras项目中的example自带了多个示例，包括经典的mnist手写识别测试等，其中和NLP相关的示例有很多，比如基于imdb数据的情感分析、文本分类、序列标注等。其中lstm\_text\_generation.py示例可以用来参考设计序列标注问题，这个示例试图通过LSTM学习尼采的作品，通过序列标注的思想来训练一个文本生成器模型。

# 6 分词系统评估方法及实例分析

## 6.1 分词系统评估方法

1. 黄金标准

所谓的黄金标准是指评价一个分词器分词结果的好坏，必须有一个“公认正确”的分词结果数据作为参照。因此，要找权威的分词数据作为黄金标准

可以使用SIGHAN（国际计算语言学会（ACL）中文语言处理小组）举办的国际中文语言处理竞赛Second International Chinese Word Segmentation Bakeoff所提供的公开数据来评测，它包含了多个测试集以及对应的黄金标准分词结果。

1. 评价标准

精度(Precision)、召回率(Recall)、F值(F-measure)是用于评价一个信息检索系统的质量的3个主要指标，分别简记为P,R和F。同时，还可以把错误率(Error Rate)作为分词效果的评价标准之一。

精度表明了分词器分词的准确程度；召回率可认为是“查全率”，表明分词器切分正确的词有多么的全；F值综合反映整体的指标；错误率表明了分词器分词的错误程度。

P、R、F越大越好，ER越小越好。一个完美的分词器的P、R、F值均为1，ER值为0.通常，召回率和精度这两个指标会相互制约。P、R、F的计算公式如下：

精度(P)=×100%

召回率(R)=×100%

## 6.2 基于机械匹配和概率统计相结合的中文命名实体识别研究

命名实体(Named Entity, NE)是指一些具体或抽象的客观实体，例如人、组织、地点、时间等。文本中的命名实体大多是以特定的专有名称出现的，例如人名、组织名、机构名、地名等，也可以是时间、数量的表达式等形式。命名实体识别的任务实际上就是指从文本中发现出命名实体，并确定其类别的过程。

1995年9月举行的第六届MUC(Message Under⁃standing Conference)会议第一次引入了命名实体识别。当时给出的任务是识别出特定文本集中的专有名称和数量短语并对它们归类。1998年召开的第七届MUC会议已经将命名实体识别作为主要的研讨议题之一。此次会议将命名实体的类别确定为人名(Person)、地名(Location)、机构名(Organization)、日期 (Date)、时 间 (Time)、百 分 数 (Percentage)和 货 币(Monetary value)等七种类型。

20世纪90年代初期，国内外的一些学者开始对中文通用命名实体(如：人名、地名、组织机构名等)识别进行了一些研究。例如：清华大学的孙茂松是我国最早做中文姓名识别的，他采用统计的方法计算人名的用字概率。复旦大学的吴立德采用统计和规则相结合的方法对中文人名、组织机构名的识别进行研究，取得了较好的效果。因特尔中国研究中心的 ZHANG Yi-Min和 ZHOU Joe F等人采用基于记忆的学习(Memory-Based learning，MBL)算法开发了一个抽取中文命名实体及实体间关系的信息抽取系统，该系统在ACL2000上取得了较好的效果。

预处理：自然语言信息的特殊性决定了在进行命名实体识别前要做一些预处理工作，且中文文本的预处理对中文命名实体识别效果影响较大，在实际应用中，更应该引起重视。

中文分词。自然语言均是由字或词组成的，词是最小的有意义的语言单元。中文的词语之间没有明显的边界，因此，对于一个句子的准确切分将直接决定命名实体的正确识别。由此可见，中文词法分析中的分词是命名实体识别的第一步，也是中文信息处理的基础和关键。

词性标注。词性标注是指依据切分后的分词情况，标出每个词的类型，正确的词性标注是正确进行实体识别的前提。

信息抽取时采用的是中科院计算技术研究所的词法分析系统ICTCLAS，其主要功能包括中文分词、词性标注、实体识别、新词识别等。按照973专家组评测结果，ICTCLAS分词识别率97.58%，基于角色标注的未登录词识别能取得高于90%的召回率，其中中国人名的识别召回率接近98%。

中文命名实体识别的一般过程

(1)在分词的同时，标注出词表中已经收集的命名实体。

(2)在此基础上，调用构建好的命名实体识别模型，对文中的尚未标记出的命名实体进行识别。

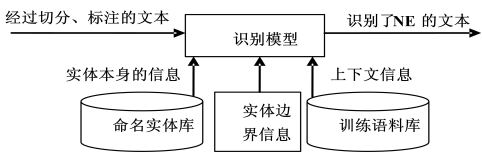


图6-1

基于机械匹配的中文命名实体识别方法：是利用中文命名实体的内部结构以及上下文的边界特征等信息来手工建立命名实体的识别规则。例如：中国人名“张三/nh” 可以利用词性信息结合中国人名姓氏用字表、名用字表来识别；复杂的人名可以利用头衔边界信息来识别，例如“王经理”、 “何主席”、 “老杨”、 “小王”等。但是，这些规则依赖于设计者的直觉，主观性很强，同时，由于不同类型文档的词语分布规则存在差异，所以将某一领域文档中提取的实体识别规则应用到另一个领域时通常就会出现问题。

人工构建规则的最大缺点是代价很大。基于机械匹配的识别系统的建立不仅需要有计算语言学背景的专家的参与，而且要付出长时间的代价。

基于统计的实体识别方法：基于统计的命名实体识别方法主要是利用大型标注语料库来训练，得出某个字作为命名实体组成部分的概率，并以此为基础来计算某个候选字段作为命名实体的概率，若大于某一阈值，则识别为命名实体。

（1）N-gram模型 该模型建立在统计概率理论的基础之上的，它的前提是假设文本中某一个词的出现概率取决于它前面的若干个词出现的概率。

显然，要计算出词的出现概率，必须知道它前面所有词的出现概率，这样导致计算非常复杂。

(2)基于HMM的模型（隐马尔科夫模型）

规则和统计相结合的实体识别方法: 自然语言处理领域常用的一种信息处理方法，它可以将信息的规则特征和统计特征结合起来，有利于提高系统的性能。在命名实体识别中，其优势体现在以下两个方面。

1. 通过统计概率的计算可以大幅度降低规则方法处理的复杂度，减少规则使用的盲目性
2. 加入的识别规则可以降低系统对大规模语料库的依赖。

规则和统计相结合的方法是自然语言处理领域常用的一种信息处理方法,它可以将信息的规则特征和统计特征结合起来，有利于提高系统的性能。在命名实体识别中，其优势体现在以下两个方面。

（1）通过统计概率的计算可以大幅度降低规则方法处理的复杂度，减少规则使用的盲目性。（2）加入的识别规则可以降低系统对大规模语料库的依赖。

在命名实体中人名、地名、机构名和专有名词构成规则相对于时间和日期来说更为复杂，是我们研究的重点对象。这里的规则包括用于确定实体边界的外部规划和描述实体内部结构的内部规则。

外部规则：处于具体语言环境中的命名实体，其上下文信息必然影响它的边界和类别。

在一个句子中，人名的前面或后面常有一些指示词，如“主席”、 “教授”、 “说”等，因此，这些标志性很强的指示词为我们确定实体的边界提高了很大的帮助。

内部规则：从汉语构词的角度来说，命名实体一般是由一个或多个词构成的，其内部的构成规则我们称之为内部规则。对于一些复杂的人名或人名指代，比如小王、小余、小林、王某。

基本流程：

(1) 从左向右扫描经分词和词性标注后的文本，若词性序列满足某一实体类别的构成规则，则产生一个该类别的候选实体。

(2) 利用该类别的外部规则进行确认，若满足规则，结束本次识别过程转向(1)；否则转(3)。

(3) 继续扫描紧跟其后的词串，按从语料库中提取的统计规则进行识别。结束本次识别过程后，若文本未结束，则转(1)。

本文采用1998年1月的人民日报语料前10000句作为测试语料，剩下的部分用于训练。

在测试集中，我们考察的几类命名实体构成情况，如表1所示。

表1测试集的构成

|  |  |
| --- | --- |
| 实体（NE）类别 | 实体（NE）个数 |
| 人名（Nh） | 3049 |
| 机构名（Ni） | 5276 |
| 地名（Ns） | 2491 |
| 日期（Nr） | 1531 |
| 时间（Nt） | 37 |
| 总计 | 12384 |

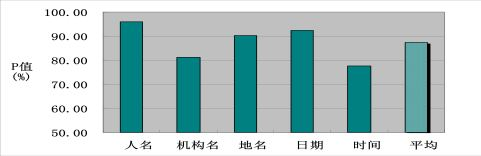


图6-2 不同类型NE识别结果的P值比较

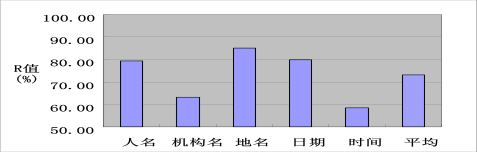


图6-3不同类型NE识别结果的R值比较

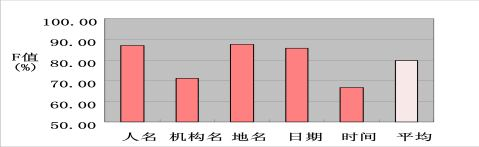


图6-4不同类型NE识别结果的F值比较

从图2和图3可以看出，系统对几个类型NE识别都取得了较高的精确度，但是召回率都相对较低，原因在于识别所采用的规则限制条件严格，针对性较强，因此它对实体的识别具有较高的准确率。但是规则并不适合所有的情况，所以召回率较低。从图4可以看出，系统获得的F值较高，但是这些规则受语言环境和领域的影响较大，一旦应用对象发生变化就很难再适用。

## 6.3 基于表示学习的中文分词算法探索

监督学习中的数据中是提前做好了分类的信息的，如垃圾邮件检测中，他的训练样本是提前存在分类的信息，也就是对垃圾邮件和非垃圾邮件的标记信息垃圾邮件筛选。

监督学习中，他的训练样本中是同时包含有特征和标签信息的，监督学习中，比较典型的问题就是像上面说的分类问题(Classfication)和回归问题(Regression)。

无监督学习是另一种常用的机器学习算法，与监督学习不同的是，无监督学习的样本是不包含标签信息的，只有一定的特征，所以由于没有标签信息，学习过程中并不知道分类结果是否正确，比较典型的是一些聚合新闻网站，利用爬虫爬取新闻后对新闻进行分类的问题。

例如 百度新闻。所有有关这个关键字的新闻都会出现，它们被作为一个集合，在这里我们称它为聚合(Clustering)问题。

随着基于神经网络的表示学习方法的兴起，使得自动学习特征成为可能。基于表示学习的中文分词方法首先从大规模语料中无监督地学习中文字的语义向量，然后将字的语义向量应用于基于神经网络的有监督中文分词。实验表明，表示学习算法是一种有效的中文分词方法，但是仍然发现，由于语料规模等的限制，表示学习方法尚不能完全取代传统基于人工设计特征的有监督机器学习方法。

基于字表示的有监督分词：

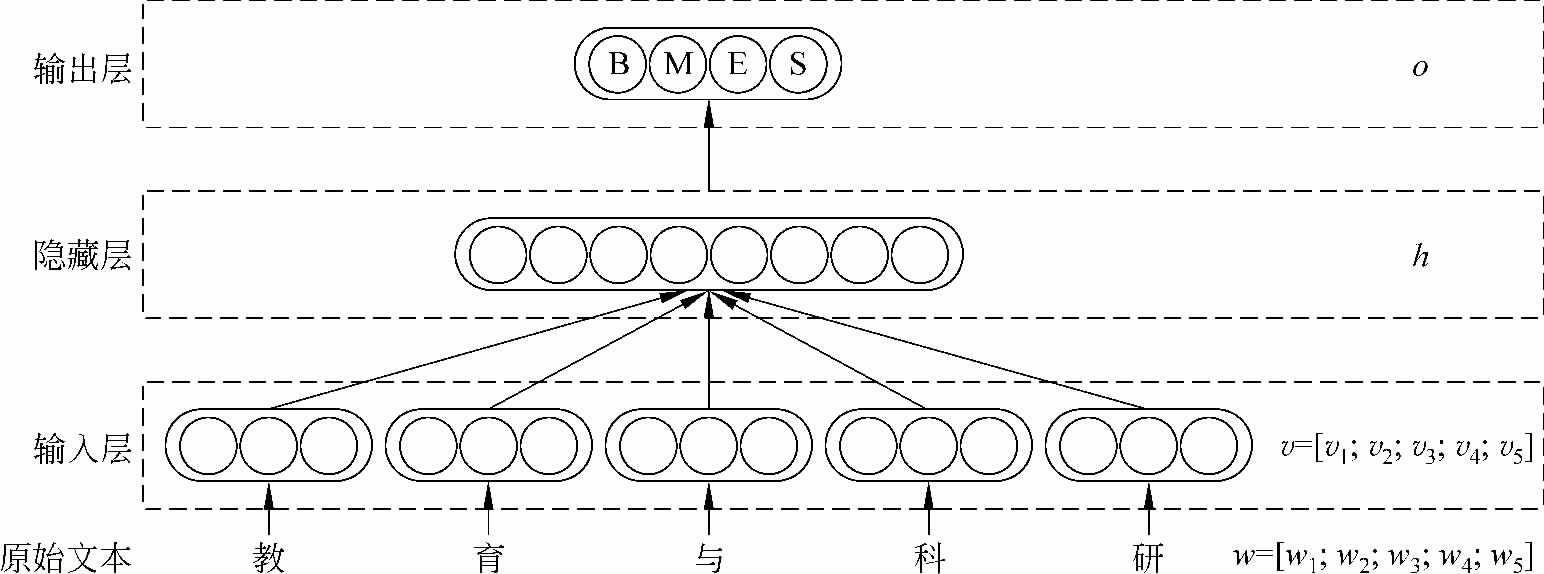


图6-5 算法基本结构图

采用BMES体系对汉字进行标注。对于单字词，其标签为S；对于多字词，词中的第一个汉字标签为B，最后一个汉字标签为E，中间字的标签为M。对训练数据的每个字进行标注后，本文采用一种3层神经网络结构对每个字进行训练，其结构如图所示。

对于句子中的每个字的标签分类任务，本文选取上下文以及当前字，共个字作为特征。其中上文和下文均为个字。图中最下方为这个字的原始文本，经过第一层，将每个字转换成其字向量表示，并把个字连接成一个维的向量V。该维的向量是神经网络的输入层。隐藏层h的设计与传统的BP神经网络一致，输入层的个节点与隐藏层的H个节点之间两两均有边连接。隐藏层选用tanh函数作为激活函数。输出层一共有4个节点，使用softmax归一化后，分别表示这个字被打上B、M、E、S标签的概率。

网络结构可以形式化的表示为：

其中U为输入层到隐藏层的权重， V为隐藏层到输出层的权重。这两层均可理解为简单的矩阵相乘。最后使用softmax函数可以将输出O转换成标签概率P。

网络的训练目标使用传统的最大似然估计法，即求出一组参数，并最大化：

其中参数包含各个字的字向量v以及两个网络中的参数矩阵U 和V。训练中这里使用了朴素的随机梯度下降法。

实验设置

在有监督学习部分，本文使用的语料为Sighan2005 bakeoff的分词语料。选取其中北京大学标注的数据用于训练、验证和测试。原始语料只包含了训练集与测试集，在实验前，我们将原始语料的训练集前90％ 当作我们自己的训练集，最后10%当作开发集。测试集保持不变。最后训练集共有1626187个字，验证集包含了160898个字，测试集有168973字。

在非监督实验中，我们使用了两个语料，第一个语料（实验中称“小语料”）直接采用了北京大学标注的数据中的训练集，共179万字。第二个语料（实验中称“大语料”）在第一个语料的基础上，加入了搜狗新闻语料的精简版，其中涉及教育、文化、军事等一共10个类型的新闻语料。删除其中有乱码的句子后，最后得到的语料一共有2723万字。

为了展示神经网络模型以及字表示对于实验的影响，本文设计了多组对比实验。



实验结果如表１所示。

表１中列举了本文所做的一共8组实验。其中#1和#2为上一节中描述的两个基准实验。 #3到#8依次为上一节中描述的各个实验。#1和#2为传统的最大熵分词方法得到的结果，与前人论文得到的结果相同，使用最大熵模型配合二元特征可以取得非常好的效果，该方法在Sighan2005的评测中，可以排到第三名。

#3和#6相比，有巨大的优势，这里主要有两点原因：第一，当特征数较少时，非线性的神经网络相比线性的最大熵模型有优势（与之相对的，如果特征数很多时，如#2中使用的二元特征，非线性模型无论是训练时间还是测试时间都会非常长）；第二，神经网络模型在反向传播时，可以修改词向量，这相比直接把词向量作为输入特征的最大熵模型更为灵活。

#3、#4、#5的比较中可以看出，无监督训练得到的字向量在作为有监督训练初始值时，可以显著地提升有监督学习的效果。其中#4虽然采用了无监督的数据进行训练，但实际上训练数据来自北京大学标注语料，因此可以看作是封闭训练的结果。#5只能看作开放训练的结果。

#2和#8中可以看出，将字向量作为附加特征辅助最大熵模型，效果几乎没有提升（只在小数点后第４位略有提升）。

#6和#7的对比实验。值得注意的是，即使使用随机数来描述一个字，也可以取得超过纯猜测的效果（不到0.25的准确率）

得出如下结论：字向量的表示是一种较好的特征，使用字向量配合神经网络实现的分词，相比一元特征有较大的优势。但是这种方法目前还不能取代人工设计特征，即使是简单的二元特征。随着数据量的增大，无监督学习得到的字向量也会越来越实用，相信使用更丰富的无监督训练语料，可以得到更有用的字向量。

## 6.4 统计与字典相结合的领域自适应中文分词

基于词典的方法利用词典作为主要的资源，这类方法不需要考虑领域自适应性的问题，它只需要有相关领域的高质量词典即可，但是这类方法不能很好的解决中文分词所面临的歧义性问题以及未登录词问题。

基于统计的方法是近年来主流的分词方法，它采用已经切分好的分词语料作为主要的资源，最终形成一个统计模型来进行分词解码。基于统计的方法在分词性能方面有了很大的提高，但是在跨领域方面都存在着很大的不足，它们需要针对不同的领域训练不同的统计分词模型。这样导致在领域变换后，必须为它们提供相应领域的分词训练语料，但是分词训练语料的获得是需要大量人工参与的，代价昂贵。而基于词典的方法却在领域自适应方面存在着一定优势，当目标分词领域改变时，只需要利用相应领域的词典即可。领域词典的获取相比训练语料而言要容易很多。如果把这两种方法结合起来，使得统计的方法能够合理应用词典，则可实现中文分词的领域自适应性。

在CRF模型中融入词典特征信息的方法来解决中文分词的领域自适应性问题。在训练CRF分词模型时，使用通用词典；而分词阶段通过额外再加入领域词典来实现领域自适应性。当分词领域改变时，只需要在原有词典的基础上再添加相应领域的词典，而且不需要改变原有已经训练得到的统计中文分词模型，就可以大大改善该领域的分词准确率。

本文中所使用的字符类别的定义以及相关例子如表1所示。

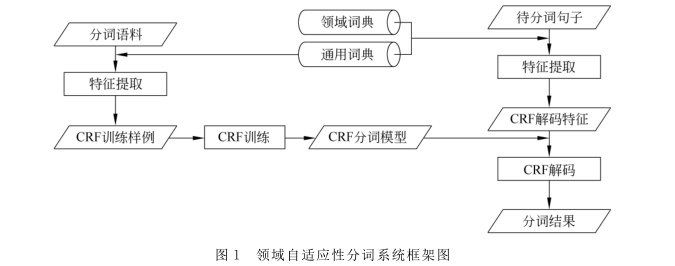


图6-6

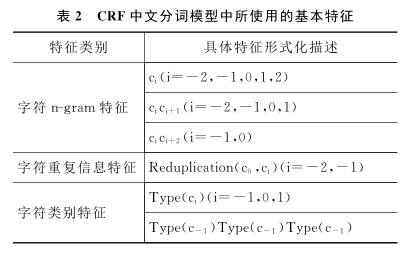
对于CRF模型，特征的选择尤为重要。本文首先使用了三类基本特征：字符n-gram特征，字符重复信息特征和字符类别特征。这里对字形态特征做了一 定 的 扩 展，将 输 入 字 符 分 为 九 类： Single,Prefix,Suffix,Long,Punc,Digit,Chinese-Digit,Letter以及Other。

给定一个中文字符，首先判断它是否属于标点符号、数字、中文数字或者字母；如果不属于其中的任何一类，则统计该字符在训练语料中出现的时候所处在的词的位置，仍用B、M、E、S表示，如果B位置出现的频次超过总次数的95％ ，则判定该字符属于Prefix类

别，如果 E位置出现的频次超过总次数95％ ，则为Suffix类别，同理S位置对应于Single类别，M位置对应于Long类别；如果还未找到该字符的类别，则标记为Other类别。



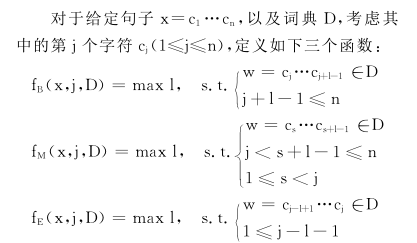
列举一下在CRF中文分词模型所使用的基本特征，如表２所示。



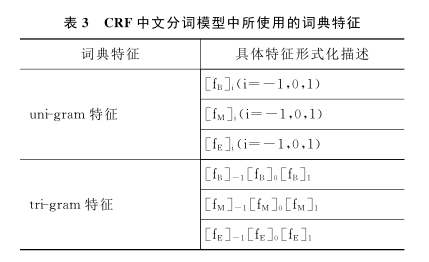
其中下标代表距离当前考察字符的相对位置，

例如，表示该字符的前一个字符。Reduplication()表示和是否为两个完全一样的字符，Type()表示字符的类别。

词典信息在统计模型中的特征表示



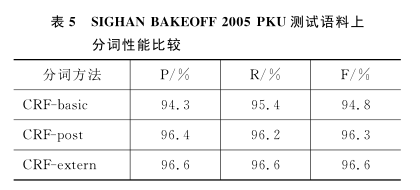
对CRF分词模型所引入的与词典D相关的扩展特征如表３所示



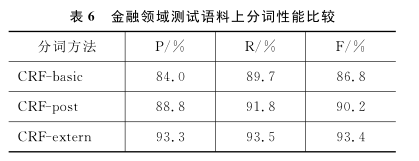
本文利用SIGHAN CWS BAKEOFF 2005中提供的PKU训练语料进行训练，训练过程中使用北京大学中国语言学研究中心公开的词典，该词典一共包含大约10万多个词。最后分别在相应的PKU测试语料和人工标注的金融领域语料上进行了评测，下表4给出了两个测试语料的统计信息。本文使用准确率（P）、召回率（R）和 F-measure值（F）来评价分词系统。本文采用CRFC＋＋工具包来进行训练和标注。



CRF-basic代表仅使用基本特征训练出来的模型；CRF-post表示使用拼接的后处理方法去纠正被CRF错误切分的词，这个方法假定词典中没有在训练语料中出现的词都应该是不可切分的；CRF-extern表示融入了词典信息特征之后所得到模型。在PKU的测试语料上，使用训练出来的模型，测试时所使用的词典和训练时所使用的词典一致。表５给出了PKU语料上测试的结果。从表中可以看出，CRF-extern与CRF-basic相比，F-measure提升了1.8% ；与CRF-post相比，提升了0.3%。



在金融领域的测试语料上，保持训练出来的CRF分词模型不变，使用的词典是在训练语料的词典基础上增加了1000个左右的金融领域专用词典。表6给出了金融领域测试语料上的结果。从表中可以看出，与 CRF-extern与CRF-basic相比，F-measure提升了7.6%；与CRF-post相比，提升了3.2%。



从上面的两个实验可以看出，

(a)无论测试语料的领域与训练语料领域是否相同，CRF-extern对比CRF-basic显著提高了分词的性能；

(b)当训练语料和测试语料领域相同时，CRF-extern 和CRF-post相比，分词性能有稍微的提高；但是当领域不同时， CRF-extern对比CRF-basic而言，有了非常显著的提高；

(c)测试领域和训练语料不同时，最终的分词 F-measure值达到了93.4% ，已经非常接

近于 CRF-basic在不考虑跨领域时的F-measure值94.8%。

综上所述，在统计模型中融入词典信息特征后，一方面分词性能有了一定的提高；另外一方面领域迁移后，分词性能依然能够保持在一定的水平。因此统计模型与词典结合后，使得中文分词具有良好的领域自适应性。

# 7 汉字处理：拼音输入法

## 7.1 拼音输入法（基于统计）

输入法的核心是词库，词库是记录的集合，每条记录包括词条、注音和同音词的词序（大多数体现为词频或概率）三个基本元素构成。早期的输入法的词条和注音通常来自专业机构和个人的人工整理。词序通常基于统计词频，拿较大规模的文本（例如几年的人民日报）作为标准训练库（通常称为语料）。但合理的词频统计一定是基于合理的分词方法。切分后每个词条的总出现次数称为词频，除以所有词条出现的总次数后得到出现的概率。当输入特定的拼音序列时，如果匹配上多个词条，将按照词频或概率的高低确定展现顺序。

理论上，如果词库集合无穷大，可以把所有文本片段放入词库中。例如“今天漫天大雪”所有字串全部放入词库一共有21个词条。所有这样的碎片加入词库，理论上可以匹配用户想输入的任何信息。但不幸的是，电脑的计算能力、存储能力、以及软件的安装包大小等，不允许过大的词库。早期的输入法词库的规模最大也只有十万条左右。

## 7.2 单机输入法时代

在单机时代，输入法的好坏往往取决于词库的大小、质量、合理性等。

组词和组句

当用户输入的拼音序列无法完整匹配上词库中的任何一个词条时，这时有两种做法，一种是早期输入法如全拼和智能ABC的做法，仅仅提供给用户最长匹配的词，例如输入“jin’tian’man’tian’da’xue”这个句子，给用户的候选是“今天”，当用户作出明确的选择后再给用户后续选择。另一种做法紫光、拼音加加等输入法，通过某种方法猜测用户想输入的完整句子或短语作为候选。

实现这一输入方式，首先需要对用户输入的整体拼音串进行切分，切分成不同的音节，比如用户输入“jin’tian’man’tian’da’xue”时，“jin’tian | man’tian | da’xue”，是一种切分方式，而“jin’tian | man | tian’da | xue”是另一种切分方式。之后把每一个拼音切分映射成词库中的词，得到不同的词与词的组合。最终可能的组合有“今天 | 漫天 | 大学”，“今天 | 漫天 | 大雪”，“今天 | 慢 | 天大 | 学”等。最终将哪种组合方式展现给用户呢？最普遍的做法是，计算每一种组合方式的组合概率。

对于组合 w1,w2,w3,w4，最简单的计算方式就是：

P(w1,w2,w3,w4) = P(w1)\*P(w2)\*P(w3)\*P(w4)

那么整体组合概率就依赖于每个词单独出现的概率（也即词频）。这种组合概率计算方法称为1元模型，统计信息少，计算复杂度低，也有一定的准确率，在早期有着很广泛的应用。但这种方法的错误率也是比较高的，例如，如果在全局信息中，P(大学) >P(大雪)，那么概率最大会是“今天漫天大学”，用户仍然需要进一步修改“大学”得到“大雪”才能输出选项。

## 7.3 互联网时代的输入法

互联网时代的输入法以搜狗输入法为代表，它诞生于互联网应用爆炸的时代，电脑已经不再是少数人手中的玩具，已经深入到千家万户中，聊天、论坛、写博、评论、电子邮件等网络应用已经取代专业输入成为主流需求。由于传播渠道有了质的改善，各个输入法产品之间的竞争也显得更为激烈。传统的词库制作方法已经不能满足网民的需求，体现在以下的三个方面：

1、词库中大都是比较权威的正式词条，而少有网民的口语词汇

2、互联网上的新词热词往往成为网民们热衷讨论的话题，需要非常快的词库更新速度

3、网民在进行某个特定领域输入的时候，严重缺乏该领域的专业词条

让我们看看搜狗输入法所代表的新一代输入法的进步所在：

互联网语料

互联网语料是解决上述问题的主要钥匙。网络上充斥着大量的信息，有的是权威发布的信息（如官方新闻等），有的是网民的交互活动产生的信息（如论坛，博客），当然也有大量的无效信息，例如广告传销、SEO作弊等。互联网语料的生成本身是一个复杂的问题，但依赖于搜狗搜索引擎的技术底蕴和海量数据处理能力，我们能够比较方便的获得最新的优质语料。这也是输入法能越来越“聪明”的基础。

新词和口语词

互联网具有这样两个特点：有大量网民产生的“非主流”的口语词汇；随时会产生新的词汇、词法或流行语。这就意味着，传统的基于人工构造的词典不会有很好的效果。口语词发现和新词发现成为互联网输入法的重要技术点。

专业细胞词库

专业词汇的发现技术和新词、口语词类似，只要我们能够获得特定专业领域的足够语料，后续的挖掘行为是雷同的。我们通过对文本分类和聚类算法，将某一个领域中被局部用户使用比较多的词汇聚合成一个类，形成我们的专业词库（也就是搜狗的细胞词库），并通过细胞词库推荐的形式，推荐给需要的用户使用来提高其输入的效率。

组词算法的革命

通过新词发现、口语词发现、专业细胞词库，输入法已经聪明了很多，能够做到“与时俱进”了。下面我们回到最开始的话题：输入法怎样变得更聪明，能更准确的理解用户的意图呢？

还是以“今天漫天大雪”为例。如果是一个人，光看到“daxue”时，第一印象很可能也是“大学”，但如果看到“mantiandaxue”，通常会反映过来是“漫天大雪”。这是因为人的思考结合了上下文和语境等因素。这也给我们一个启示，“漫天”和“大雪”之间的关系更强，一个改进的方法是，在分词统计的时候，不仅统计词条出现的次数，还统计“二元组”连续出现的概率，比如P(漫天，大雪)，那么在计算整体组合概率的时候，就可以简单的使用：

P(今天，漫天，大雪) = P(今天) \* P(漫天，大雪)

由于在统计语料中，（漫天，大雪）出现的次数会远远地大于（漫天，大学）的出现次数，因此最终“今天漫天大雪”的组合概率将更胜一筹。

显然的，这种基于“二元组”的概率计算方法更具有整体性。但相比基于词的概率计算方法，却要存储更多的信息，如果词典中有N个词，那么理论上需要存储的组合数目为N\*N，空间上是之前的平方级别。而且由于训练语料相对比较小而且不容易获取，因此实际统计出来的组合数目相比全局空间来讲要少很多。严格说来，二元模型并不是互联网时代的新鲜事物，微软拼音、黑马神拼等输入法早在单机时代也使用过类似的模型。但受限于训练语料的规模以及机器的计算存储能力，从计算效率和效果上，都不是非常令人满意，此外过大的安装包大小也影响了他们的传播。在互联网时代，用户桌面电脑的性能有了大幅提高，使较复杂的模型有了可能。

## 7.4 云时代的输入法

互联网时代的输入法仍然是传统意义上的桌面输入法，需要以安装包或者定期更新的形式把信息库存储到用户本地电脑上，而用户在输入的时候也必须使用本地的存储和计算资源。很显然，这种传播、更新、存储和计算方式，将是当前输入法发展的主要瓶颈。但是随着云时代的到来，网络延时不断降低，浏览器逐步取代本地操作系统，这就为输入法逐渐脱离桌面的限制，成为纯粹的网络输入法带来了契机。

直到“搜狗云输入法”的诞生，云时代输入法的雏形，才真正开始向世人崭露头角。

何谓云输入法

搜狗云输入法本质上是网络输入法，但为何称为“云输入法”？一方面，输入法本身的核心处理过程主要是信息存储、信息挖掘和概率计算，而非信息传输。搜狗云输入法的存储和计算都是在大规模服务器集群上完成的，这些服务器不仅存储能力巨大，而且多核处理器性能超强，能够同时支持成千上万的用户进行在线计算。用户端只需要通过网络把复杂的计算任务请求发送到服务器群上然后等待服务器群返回计算结果，这个大规模服务器的集群，正是云计算中的“云端”。另一方面，搜狗云输入法已经对外公开了计算接口，成为一个真正的输入法计算服务提供者，任何用户和第三方开发商都可以通过和服务器群直接交互来获得计算服务。可以看出，搜狗云输入法其本身已经具有了“云计算”的种种特质，因此冠名“云输入法”并不过分。

显而易见，相比于个人桌面电脑，云输入法使用的大规模服务器集群的存储计算能力已是不可同日而语；同时，由于服务器成了所有用户的“容器”，在这个容器中进行信息挖掘和信息更新就可以变得实时。这种利好是如何具体体现的呢？　　超大信息量

相比于桌面输入法的小而精，搜狗云输入法可以做到输入领域的大而全。通过搜索引擎抓取前所未有的超大训练语料（TB级别），可以做到无论是口语，古文，散文，现代文，诗词歌赋等不同文体，还是政治，经济，体育，娱乐等不同领域，都能面面俱到，应有尽有。权威词条，高频口语，方言俚语，专业词汇，网络热词等古今中外，都可以统统纳入词典而不受任何限制。

实时更新

通过对用户输入的实时统计和挖掘，可以根据用户对词条的输入情况，随时对词库进行补充和更新，进行基于用户词的新词发现；并且根据用户输入的分词统计，随时对词库的词频做出合理的调整，给用户最合理的词条排序；另外，还可以通过搜索引擎不断的抓取最新的网络资源，及时的分析网络新词并加入词库，以最快的速度使用户获利。

整句输入质的提升

由于“大脑”可以记住的文字信息量更大了，思考的速度也可以变得更快了，自然而然，此时的输入法就可以变更聪明了。通过扩大信息量并且采用更复杂的概率计算模型，整句输入的体验有了质的提升。

首先，以“缓解工作压力(huan’jie’gong’zuo’ya’li)”为例，如果采用基于“二元组”的概率计算模型，会得到什么样的结果呢？打开搜狗桌面输入法，经过拼音输入得到“换届工作压力”，令人失望。切出搜狗云输入法，输入结果则是“缓解工作压力”。为什么桌面输入法得不到最终的结果呢？

这种二元概率计算模型，会看到前后两个词之间的关系，但是却看不到“缓解”到“压力”之间联系。（换届，工作）是同音下最高频的二元组，因此对于二元概率计算模型，“换届工作压力”的整体概率强于“缓解工作压力”。但在搜狗云输入法中，我们对三元组（缓解，工作，压力）的概率也会进行存储并且在计算整体概率时使用。显然，这是一种更合理的整句概率计算方法，因为进一步加强了前后词之间的联系。有一个显而易见的结论：如果我们计算整体概率时采用N元组概率信息，那么N越大，我们对整体组合概率的评估也会越准确。当然，更大的“N”会导致存储空间成指数级的上升，这也是桌面输入法目前大都限制在二元模型上的最主要原因。但是云输入法却不受这样的限制，因此我们在构建模型库的时候，不仅存储了二元组信息，还存储了三元组信息已经更长多元组的信息。

当然，这个“N”不可能无限的扩大下去，计算复杂度问题和语料稀疏性问题终究不可避免。那搜狗云输入法又是如何解决这个问题的呢？在进行分词过程中，我们不仅统计词条和元组的频率，同时我们会统计远距离搭配的频率，这些远距离搭配，有的是实体意义上的搭配，有的是句式语法意义上的搭配，最终都能为整句输入的改善提供巨大的帮助。现在终于到了回答我们在前言中抛出的问题的时候了。“正当决策部门为弥合收入差距的鸿沟集思广益时”这句话，云输入法为何可以完美的输出？

通过观察可以发现，“当……时”，“弥合……的鸿沟”，都是比较固定的搭配，前者是句式方面的，而后者是实体意义上的搭配。而这些搭配都是我们可以通过分词过程中的搭配抽取得到的。有了这两个搭配参与到句子整体概率计算中，那么整个句子就可以正确无误的计算出来。

统计更长的元组频率和更远距离的词语搭配，并且海量存储任何可能出现的语言现象，做到见多识广，这就是“云输入”在理解用户输入意图方面如此之好的原因所在。

## 7.4分词引擎介绍

11款中文分词引擎大比拼

测试了11家中文分词引擎（各家分词系统链接地址），同时从分词准确度、歧义词切分、未登陆词（新涌现的通用词、专业术语、专有名词）三个方面进行论证。

BosonNLP：http://bosonnlp.com/dev/center

IKAnalyzer：http://www.oschina.net/p/ikanalyzer

NLPIR：http://ictclas.nlpir.org/docs

SCWS中文分词：http://www.xunsearch.com/scws/docs.php

结巴分词：https://github.com/fxsjy/jieba

盘古分词：http://pangusegment.codeplex.com/

庖丁解牛：https://code.google.com/p/paoding/

搜狗分词：http://www.sogou.com/labs/webservice/

新浪云：http://www.sinacloud.com/doc/sae/python/segment.html

语言云：http://www.ltp-cloud.com/document

最终的结果显示：



图7-1

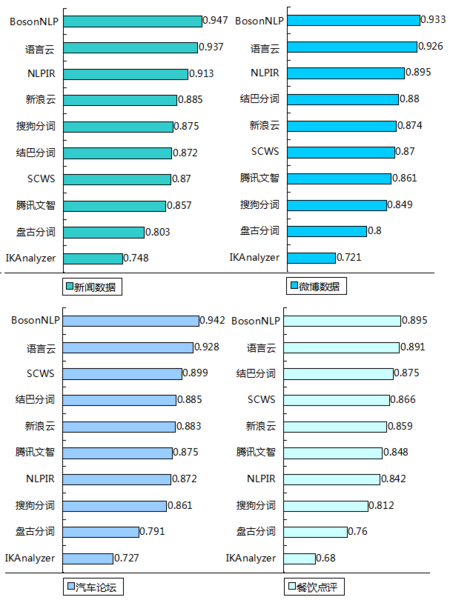


图7-2

从分词精度来看，哈工大的语言云表现的稳定一直在第二，BostonNLP分词更好，一直在这个领域保持第一。

当然在分词准确度可以接受的情况下，很多细节问题，包括是否有出错情况、是否支持各种字符、是否标注词性等都可能让我们望而却步。在分词颗粒度选择当中，BosonNLP、SCWS、盘古分词、结巴分词、庖丁解牛都提供了多种选择，可以根据需求来采用不同的分词粒度。与北大的分词标准对比来说，新浪云默认的分词粒度较大，而搜狗分词、腾讯文智分词粒度相对较小。除此之外，BosonNLP、新浪云、NLPIR、腾讯文智同时提供了实体识别、情感分析、新闻分类等其他扩展服务。下表给出了各家系统在应用方面的详细对比。



图7-3

# 8 总结与展望

中文分词是汉语自然语言处理的基石。现在，中文分词方法越来越成熟，但也依然存在能够改进之处。比如可以将现在分离开的术语抽取、命名实体识别等任务作为统一的任务来进行研究。并且，现在大部分中文信息处理工作都注重于通用领域，但其实专业领域也需要针对性的方法，现在已经有一些针对专业领域的中文信息处理方法，例如Maryam Habibi等人在2017年使用24个生物医学方面的语料库和Bi-LSTM-CRF框架，在生物医学领域的实体识别任务上取得了不错的效果。相信以后会有更多针对专业领域的中文信息处理方法，这些方法能够为学者们的研究提供非常多的帮助。除此之外，前文提到过全局歧义需要结合上下文才能判断正确分词方式，但现有的中文分词工具并没有很好的考虑到这一点，全局歧义虽然相对于局部歧义来说出现的频率更小，但这同样也是一个需要解决的问题，也许在不久的将来会出现新的算法能够将这个问题完美解决。

图灵奖获得者姚期智认为不同的学科、理论相互交叉结合，往往会出现理论上的突破和技术上的创新。中文分词正是计算机科学与语言学相互交叉融合的产物，使用更多的计算机技术理论解决更多的分词问题将是中文分词的研究目标。在众多学者的不断努力之下，中文分词的准确率一定会越来越高，应用的范围也会越来越广。

# 9 参考文献

[1]张华平.基于N\_最短路径方法的中文词语粗分模型.中国信息学报.2002.16(5):1-7;

[2] SUM M S, TSOU B K Y. Ambiguity resolution in Chinese word segmentation[C] T SOU B K, LAI T B Y.Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation, Hong Kong,27-28 December, 1995,City University of Hong Kong: 121-126.

[3]修驰.适用于不同领域的中文分词方法研究与实现:[D].北京：北京工业大学，2013.

[4]来斯惟,徐立恒,陈玉博,刘康,赵军.基于表示学习的中文分词算法探索[J].中文信息学报,2013,27(05):8-14.

[5]潘正高.基于机械匹配和统计相结合的中文命名实体识别研究[J].情报科学,2012,30(05):708-712+786.

[6]张梅山,邓知龙,车万翔,刘挺.统计与词典相结合的领域自适应中文分词[J].中文信息学报,2012,26(02):8-12.

[7]闫萍.基于机械匹配和概率统计相结合的中文命名实体识别研究[J].计算机与数字工程,2011,39(09):88-91.

[8]俞鸿魁,张华平,刘群,吕学强,施水才.基于层叠隐马尔可夫模型的中文命名实体识别[J].通信学报,2006(02):87-94.

[9]刘群,张华平,俞鸿魁,程学旗.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004(08):1421-1429.

[10]张华平,刘群.基于N-最短路径方法的中文词语粗分模型[J].中文信息学报,2002(05):1-7.

[11]张京楣. 基于统计方法的文本风格分析研究[D].山东大学,2012.

[12]孙铁利,刘延吉.中文分词技术的研究现状与困难[J].信息技术,2009,33(07):187-189+192.

[13]曹卫峰. 中文分词关键技术研究[D].南京理工大学,2009.

[14]张启宇,朱玲,张雅萍.中文分词算法研究综述[J].情报探索,2008(11):53-56.

[15]何莘,王琬芜.自然语言检索中的中文分词技术研究进展及应用[J].情报科学,2008(05):787-791.

[16]翟凤文,赫枫龄,左万利.字典与统计相结合的中文分词方法[J].小型微型计算机系统,2006(09):1766-1771.

[17]李质轩. 融合上下文信息的中文分词方法研究[D].北京交通大学,2018.

[18]杨世超. 古中文分词与词性标注方法研究[D].华北理工大学,2018.

[19]Wong P , Chan C . Chinese Word Segmentation based on Maximum Matching and Word Binding Force[C]// DBLP, 1996.

[20]Nguyen D B, Theobald M, Weikum G. J-NERD：Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features[J]. Transactions of the Association for Computational Linguistics,2016,4:215-229.

[21]Chiu J P C, Nichols E, Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics,2016(4): 356-370

[22]Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J].2016:260-270.

[23]Habibi M, Weber L, Neves M, et al. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics,2017,33(14):i37-i48.

[24]Zhou H, Yu Z, Zhang Y, et al. Word-Context Character Embeddings for Chinese Word Segmentation[C]. //Conference on Empirical Methods in Natural Language Processing.2017:760-766.

[25]Zhang S, Jiang H, Xu M, et al. A Fixed-Size Encoding Method for Variable-Length Sequences with its Application to Neural Netword Language Models[C]. //Meeting of the Association for Computational Linguistics. 2015:495-500.

[26]Xu M, Jiang H. A FOEE-based Local Detection Approach for Named Entity Recognition and Metion Detection[C]. //Conference on Empirical Methods in Natural Language Processing. 2017:760-766.

[27]Sun Z, Shen G, Deng Z. A Gao-Based Frameword for Chinese Word Segmentation via Very Deep Convolutional Networks[J]. arXiv preprint arXiv:1712.09509,2017.

[28]Xue N. Chinese Woed Segmentation as Character Tagging[J]. International Journal of Computational Linguistics & Chinese Language Processing, Volume 8,Number 1,February 2003:Special Issue on Word Formation and Chinese Language Processing,2003,8(1):29-48.

[29]Li S, Xue N. Towards Accurate Word Segmentation for Chinese Patents[J].2016.