# 摘 要

自然语言是人类文明的载体，随着计算机科学的发展，人工智能技术的研究方向自然而然聚焦在计算机与人类之间通信的可行性。对于命名实体的识别成为了自然语言处理领域的重要课题之一、至今为止，其研究愈发受到重视。而在其概念之上，知识推理的概念被提出，知识推理是机器学习、深度学习研究又是最重要、最核心的问题。因此基于知识推理的863课题“类人求解系统”被提出。

以CRF（条件随机场）模型识别为基础，实现了对于初等数学文本中数学命名实体信息的标注。选择了合适于数学实体标注的标志特征作为系统训练的的特征集合。结合前人的经验，选择了较好的特征模板，验证并分析了在CRF模型中使用的特征的有效性。

本文结合理论和实践两方面重点研究了如何正确高效抽取初等数学概率与统计题中关键信息的方法，主要进行了以下几个方面的研究：

1、对于可用于命名实体标注的相关算法研究

对于中文命名实体识别，世界上已经有很多成功的实践，其应用通常基于概率图模型。本文对于已经应用于中文命名实体识别的一些概率模型进行了研究，并作出一些对比。

2、初等数学概率统计题自动求解场合下的命名实体标注研究：

首先分析了初等数学概率统计题的语言特点，依据实际的解题过程确定了初等数学概率统计题需要的命名实体标注集合，有实际意义。

3、基于CRF的初等数学命名实体识别算法：

CRF算法在应用时应规定特征函数，本文对特征函数进行了分析，并由于没有相关的应用于解题的数学实体研究，本文从词性、词形出发设计了原子特征与组合。最终构建了一个基于CRF的初等数学概率与统计问题命名实体识别系统。

**关键词：**自然语言处理，中文命名实体识别，条件随机场，初等数学

# ABSTRACT

Natural language is the carrier of human civilization. With the development of computer science, the research of artificial intelligence technology focuses on the possibility of communication between computer and human. The recognition of named entities has become one of the most important subjects in the field of Natural Language Processing. So far, the research has become more and more valued. On the basis of named entities, the concept of knowledge reasoning has been put forward, and knowledge reasoning is the most important and the core problem of machine learning and deep-learning. Based on CRF (Conditional Random Fields) model identification, we reached the goal of tagging mathematical named entities in elementary mathematics texts. Learning from previous experiences, we selected the feature of tagging which is suitable for labeling mathematical entities as the characteristic set for system training, and chose a proper feature template. The effectiveness of the features used in the CRF model has been verified and analyzed. In this paper, by combining theory and practice, we study the method of how to select the key information from the probability and statistic questions of elementary mathematics . So we research from the fellow aspects.

1. Algorithms available for named entity tagging

For Chinese named entity recognition, there have been many successful practices at home and abroad, applications of which are usually based on probabilistic graph models. In this paper, some probabilistic models had been applied to Chinese named entity recognition are researched and compared.

2.The named entity tagging in the case of elementary probability and statistic questions.

First the language characteristics of elementary mathematical probability and statistic questions are analyzed. Then according to the actual problem-solving process, a set of named entity labels with practical significance in elementary mathematical probability and statistical questions is determined.

3.Algorithm available for named entity tagging in elementary mathematics based on CRF.

Characteristic function must be provided while using CRF algorithm. In this paper we analyzes the characteristics function, and design the combination characteristics and atom characteristics as well, inspiring by the from and the part-of-speech of words. Finally, a named entity recognition system for probabilistic and statistic problems of elementary mathematics based on CRF is constructed.

**Keywords:** Natural language processing, named entities, Conditional random field, elementary mathematic

# 目 录

# 第一章 绪论

## 1.1 研究工作的背景与意义

纵观人类历史，语言是人类文明发展的载体。语言以及与其对应的文字为人类特有的，用来传递信息的符号系统，是人类与动物在智能方面最具决定性的差异。这里所描述的语言，是指人类发展自然形成的语言，如日语、汉语等。和人类为某种目的创造的语言（比如编程语言）不同，人们通常称其为自然语言。

自然语言处理是计算机科学领域的一个重要方向，它所研究的就是计算机与人类之间使用自然语言交流的理论和方法。众所周知，科技、文学和宗教，人类能够理解的信息几乎都通过自然语言传承和传播。在现代社会，甚至人类的逻辑与思考都是以语言为基础的。由此可见此学科方向的实际意义十分显著，如果计算机能够理解人类自然语言的特殊含义，那么就能让计算机帮助人类完成一些需要大量人力的工作，比如翻译工作，对互联网上的大量自然语言文本进行分析等。而人类操作计算机工作的方式也不再以学习繁杂的、不符合人类习惯的计算机语言为前提。而在研究计算机理解人类语言的过程中，也能增进我们对人类是如何理解语言这里问题的认识。

为能够实现人类与计算机的通信，研究既要完成计算机理解自然语言文本意义的目标，也要完成计算机使用自然语言表达特定的含义的目标。前者称为自然语言理解，后者称为自然语言生成。

中文文本形式上可以看做是由汉字及标点符号组成的字符序列。含义通常由完整的句子表达，而句子是由词组组成的，词组又由单个汉字组成。但无论是任何层次都存在着歧义和多义的情况，举个例子，在某个句子中的词组，可能在另一个句子中含义大相径庭，同时，两个完全不同的句子中两个不同的单词也有可能有相同的含义。这给自然语言理解带来了最直接的困难。但实际上，自然语言通常是不存在歧义的，这时由于我们在交流中，大脑通常会结合特定的上下文和经验，得到准确的含义。如何让计算机获得理解上下文的能力和人类特有的经验便是自然语言理解的主要工作。

现代的自然语言处理工作通常是基于机器学习，通常是统计机器学习。

## 1.2 命名实体的概念以及其地位

命名实体指的是文本中具有特殊意义的词语、短语。比如专有名词、人名、地名、机构名等，有时也根据需求包括时间、数量等。根据不同的应用场景，命名实体的确切含义则会有所不同。由此可见命名实体通常是一串文本中最基本、最重要的信息元素，是理解文本的基础。

至今，命名实体的研究越来越受到重视。1995年9月举行的MUC-6会议首次定义了“命名实体”这一术语，同时提出了一个新的领域，此领域旨在对英文命名实体的结果进行评测。目光转回国内，863计划中文信息处理与智能人机交互技术评测会议将中文命名实体识别作为分词和词性标注的子任务引入。中文与英文的差距较大，具有特殊性，便与命名实体的开放性和发展性产生了矛盾，导致中文命名实体研究进展较为缓慢。

## 1.3 初等数学自动类人求解场景下命名实体的概念

初等数学中，概率和统计在我国数学高考试卷中占据着重要地位，其贴近生活而情景多变，以应用题为主。对于此类问题的计算机自动类人解题充满了挑战性。若将解题过程分为多个过程，最开始的过程便是题目的形式化表征。通俗来讲，是如何才能让计算机理解自然语言编写的题干含义的问题。自然语言和数学语言是存在差距的，而连接他们的桥梁便是问题的表征。问题表征即是挖掘问题中元素和元素间的关系，而这些元素以及元素间的关系将成为计算机自动类人解题过程中用于求解模型的参数。也就是说，我们创造的问题形式化表征生成系统，首先应拥有搜索题目中的元素以及元素之间约束关系的能力。结合上面提到的实体的定义，在这个应用场景下，题目中的元素以及可能的元素约束关系，将是文本中最具有特殊意义也最首要的信息，也就是我们需要的命名实体。

## 1.4 国内外研究现状

世界命名实体识别领域已经取得很多进展，尤其是英文命名实体的识别技术已经达到了较高的水平，并且已经可以应用于实际生产。和英文相比，中文有很多特殊性：

英文单词单词独立，且专有名词首字母大写，而命名实体多为专有名词；中文词汇没有明显分界，在命名实体识别之前，通常要进行分词，专有名词没有明

显特征。显然，最终处理结果受预先分词的效果影响极大。这些都提高了中文命名实体识别工作的难度。

中文单词在不同句子中，通常有更多的歧义。例如，“一切为了人民”和“人民路”，“人民”一词在前一短语中仅为普通名词，而在后者短语中则可看做为地理命名实体。这一特点说明我们无法仅仅利用词库来进行实体的识别。同时，由于中文名词组成极度丰富，任何一个文字都可能出现在某一个名词中，大量的命名实体无法登陆词库。

综上，中文命名实体研究工作在进展落后的情况中同时面临了更多的困难。

国内，中文命名实体的研究中，结合规则和统计方法是常用的方案，即在使用中文本身的构成规则后，再结合一定的统计算法，后者减轻了前者研究的巨大代价，并显著提升了识别的效果。比如潘正高结合N-gram模型和基于隐马尔科夫模型（HMM）[1]的统计算法制定了特定规则，对人名的识别准确率达到了97%。何炎祥结合基于条件随机场模型的算法和特定规则[2]，针对地名和组织名的识别效果较好，F值分别达到了91.61%和85.74%。相对于人名、地名的识别来说，中文机构名的识别存在较大的困难，并且研究较少，因此周俊生团队提出了一种新的基于层叠条件随机场模型的中文机构名识别算法[3]，其思想是底层模型进行人名与地名的识别并为高层模型识别复杂机构名工作提供决策支持。

## 1.5 本论文的主要研究内容

本文首先构想了一定的应用场景，即初等数学中概率与统计题的自动求解引擎场合。并在此应用场景的需要出发，研究了初等数学题的题干结构以及自动求解的可能算法，从此得出本文需要自然语言处理完成的任务，即初等数学的命名实体标注。并迎合此需求需要设计了较为科学的命名实体标注集合。本文使用了概率图模型中条件随机场作为中文命名实体识别任务的预测模型，因此本文自宏观至具体的研究了机器学习，自然语言处理需要完成的任务，并研究了多种概率图模型的区别。在研究了条件随机场的基础上，研究了特征函数的原理，并根据原理和需求制定了特征模板。整篇文章通过训练，测试和统计分析，证明了条件随机场对于初等数学的命名实体标注有较好的效果。

## 1.6 本论文的结构安排

本文的章节结构安排如下：

第一章，绪论：简单描述了研究工作的背景与意义，通过阐述自然语言在人类社会中的地位说明自然语言处理任务的重要性。同时交代了初等数学自动类人求解场景下命名实体的概念。并交代了国内外相关研究现状。

第二章，相关理论基础：宏观说明了概率模型的概念。并比较了判别式和生成式模型的区别。并由此引出概率图模型的概念。为了准确描述概率图模型，本文用语言描述加数学公式的方法，定义并解释了几种知名的概率图模型：隐马尔科夫模型、马尔科夫随机场和条件随机场。

第三章，初等数学的自动求解场合下命名实体标注研究：首先交代了初等数学中概率与统计题的地位。并提出一个有趣的设想，即：如何构建初等数学概率与统计题自动求解引擎。在此设想出发，深入研究了问题表征相关内容，并在此研究之上设计了标注集合。

第四章，基于CRF的初等数学命名实体识别算法：首先交代了整个本文实验的总体设计。然后详细描述了本文实验的过程，即如何以CRF++工具为核心，通过编译安装CRF++工具，语料的预处理，数据集切分，等步骤，最终得到最终机器识别命名实体的结果。

第五章，实验结果与评测分析：首先对命名实体标注任务中一些常见的概念进行描述，并在这些概念上定义了评测中文命名实体识别系统性能的指标。然后根据实验已经获得的机器标注数据，通过脚本统计出这些指标，并根据指标进行简单的分析。

第六章，全文总结与展望：对整篇论文进行总结，并对今后的研究工作做一定的设想与计划。

# 第二章 相关理论基础

## 2.1 概率模型

在机器学习最首要的任务，是根据已观察到的证据，来对感兴趣的未知变量进行估计和推测[4]。前者就是所谓的训练样本，而后者通常为类别标记。概率模型（probabilistic model）提供了一种描述框架，将整个机器学习要完成的任务抽象为计算随机变量的概率分布。

概率模型通常分为产生式模型和生成式模型，假设我们关心的变量集合为$Y$，可观测的变量集合为$O$，其他变量的集合为$R$。生成式模型考虑联合分布$P(Y,R,O)$，而判别式模型考虑的则是条件分布$P(Y,R|O)$。在给定观测变量值的条件下，概率模型所做的推断工作，即是从考虑的分布中的得到条件概率分布$P(Y|O)$。

当我们仅考虑各种变量集合为离散状态的时候，由于为联合分布，考虑所有情况下求和消除变量$R$即可解出条件概率分布$P(Y|O)$，但由于算法复杂度极高并不现实。同样，由于特征向量之间往往相关，为了能够针对不同的情况研究高效的学习算法，人们发明了概率图模型（probabilistic graphical model）。

## 2.2 概率图模型

为了能够简洁紧凑地研究变量之间的概率关系，使用了图的形式作为表示工具。一般用一个结点表示一组随机变量，结点之间的边表示变量之间的概率相关关系。概率图模型通常分为两类：使用有向无环图表示变量之间依赖关系，成为有向图模型；使用无向图表示变量之间的相关关系，称为无向图模型。这里举三个模型为例进行详细描述和比较。

### 2.2.1 隐马尔科夫模型

隐马尔科夫模型（Hidden Markov Model，简称HMM），是非常著名的生成式有向图模型之一。主要用于对时序数据的建模，在自然语言处理、语音识别等领域有非常广泛的应用。

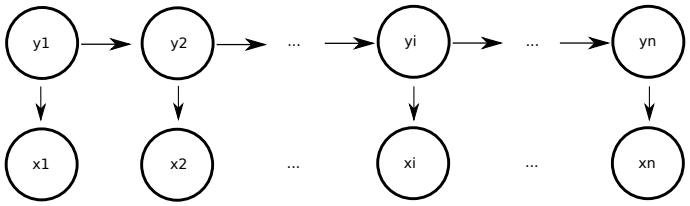


图 2-1 隐马尔科夫模型的图结构

如图2-1，位于上方的结点表示状态变量$\{y_1, y_2, ..., y_n\}$，其中$y_i \epsilon Y$表示第$i$时刻的系统状态，通常状态变量不可被观测。下方的结点则表示观测变量$\{x_1, x_2, ..., x_n\}$，其中$x_i \epsilon X$表示同一$i$时刻的观测值。图中的箭头表示变量间的依赖关系，可见，某一时刻观测变量值$x_t$仅依赖于状态变量$y_t$，与其他的状态变量取值无关系。同时某时刻状态变量$y_t$的取值，也仅仅相关于前一时刻的状态变量取值，与其他任何状态变量取值无关。人们在此模型上定义了马尔科夫性质，此性质描述为：某一系统在下一状态的概率仅与当前状态有关，而与之前的状态无关。基于这种性质，所有变量的联合分布为：

$$P(x_1, y_1, ..., x_n, y_n) = P(y_1)P(x_1|y_1)\prod_{i=2}^{n} P(y_i|y_{i-1})P(x_i|y_i) \tag{2-1}$$

假设系统在多个状态$\{s_1, s_2, ..., s_n\}$之间转换，那么状态变量的$y_i$的取值范围$Y$即是有N个可能取值的离散空间。这个空间我们称为状态空间。假如观测变量取值范围$X$为$\{o_1, o_2, ..., o_n\}$，此取值范围被称作观测空间。

初始时刻时，系统应该处于某一个状态，这个状态是根据初始状态概率决定的。由于状态只有N个可能取值，所以可以记为$\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_N)$。其中

$$\pi_i = P(y_1 = s_i), 1 \leq i \leq N \tag{2-2}$$

表示模型的初始状态为$s_i$的概率。

下一时刻，模型将转化为下一个状态，将会转化为哪一个状态，则需要参考一个N行N列的矩阵。通常记为$\boldsymbol{A} = [a_{ij}]_{N \times N}$，其中

$$a_{ij} = P(y_{t+1} = s_j | y_t = s_i), 1 \leq i, j \leq N \tag{2-3}$$

表示任意时刻$t$，若状态为$s_i$，则在下一时刻为$s_j$的概率。

每一时刻，我们希望得到对应当前状态的观测值的概率，由于观测值可能有M种取值，这里记为矩阵$\boldsymbol{B} = [b_{ij}]_{N \times M}$，其中

$$b_{ij} = P(x_t = o_j | y_t = s_i), 1 \leq i \leq N, 1 \leq j \leq M \tag{2-4}$$

表示任意时刻$t$若状态为$s_i$，观测值为$o_j$的概率。

至此，我们有了状态空间$Y$，观测空间$X$和三个概率矩阵，便能构建一个隐马尔科夫模型。

根据上文，我们可以将一个马尔科夫模型用其参数表示，即$\lambda = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}]$[4]。结合应用来讨论我们利用隐马尔科夫解决的三类问题，以便于与其他模型构成类比。

（1）模型与观测序列之间的匹配程度：即给定模型$\lambda = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}]$，计算其产生$\boldsymbol{x}$的概率$P(\boldsymbol{x}|\lambda)$。比如一个预测任务，在已知以往的观测序列，推断当前最有可能的观测值。

（2）根据观测序列推断状态变化：即给定$\lambda = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}]$和观测序列，找到与其依次对应的状态序列。比如说在语音识别的任务中，观测序列为语音信号序列，而状态为字符序列，目标就是根据信号序列推断最有可能的字符序列。

（3）根据观测序列，如何调整$\lambda = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}]$参数，以求该序列出现的概率最大？这便是一个使用一定量训练数据训练模型参数的过程。

基于变量联合分布的条件独立性，这三个问题都能高效求解[4]。

### 2.2.2 马尔科夫随机场

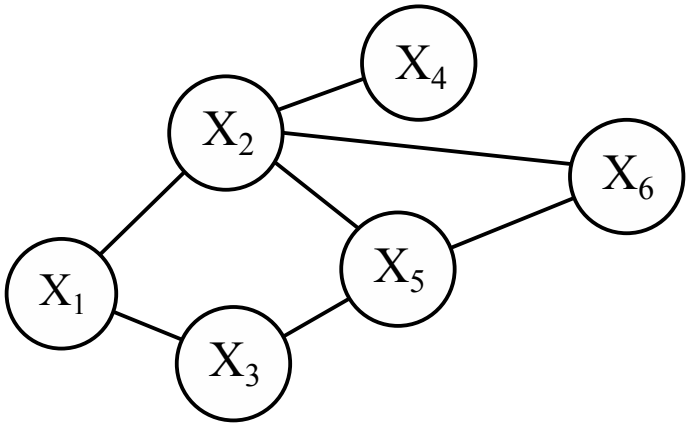


图 2-2 一个马尔科夫随机场的图结构

与隐马尔科夫相同的是，马尔科夫随机场同样为生成式模型，即对联合分布进行建模。但却为一种无向图模型（又称为马尔科夫网）。在一个表示马尔科夫随机场的无向图中，每一个结点都表示一个或一组变量。结点之间的边表示两个变量之间的依赖。为了能够准确地描述马尔科夫随机场的联合概率分布，在此阐明一些概念。

#### 2.2.2.1 团与最大团

对于给定无向图$G=(V,E)$，一个两两之间有变的结点集合便是此图的团（clique）。若一个团中加入任何一个结点都不再形成团，那么这个团就称作“极大团”（maximal clique）。也就是说，极大团不能被任何团包含，即不是其他任何一个团的真子集。顶点最多的极大团，称为图$G$的“最大团”（maximum clique）。

#### 2.2.2.2 势函数

势函数是一个表示其对应团状态的非负实值函数，即一个定义在团上的测度（Measure）。测度是一个函数，把给定集合的某个子集指定一个数，这个数可以代表大小、体积、概率等。在马尔科夫随机场中，一个势函数可以看做无向图中某个团映射到代表概率的某个数。

#### 2.2.2.3 马尔科夫随机场的概率分布函数

有了团和势函数的定义，多个变量之间的联合概率分布便可以基于团分解为多个势函数的乘积，每个势函数仅与一个团相关。具体来说，对于n个变量$\boldsymbol{x}=\{x_1,x_2,...,x_n\}$，所有团构成的集合为$C$，与团$Q\in C$对应的变量集合为$\boldsymbol{x}_Q$，则联合概率$P(\boldsymbol{x})$定义为

$$P(\boldsymbol{x})=\frac{1}{Z}\prod_{Q\in C}\Psi_Q(\boldsymbol{x}_Q) \tag{2-5}$$

其中$\Psi_Q$为与团$Q$对应的势函数，用于对团$Q$中的变量模型进行建模，因为概率应该在0和1之间，所以需要一个规范化因子$Z$，定义为

$$Z=\sum_{\boldsymbol{x}}\prod_{Q\in C}\Psi(\boldsymbol{x}_Q) \tag{2-6}$$

用于确保$P(\boldsymbol{x})$是被正确定义的概率。

### 2.2.3 条件随机场模型

与隐马尔科夫模型不同，条件随机场（Conditional Random Field，简称CRF）则是一种判别式无向图模型。

在本章第一节介绍过，生成式模型对条件分布进行建模。在已经获得一个观测序列的即多个变量的观测值的情况下，CRF在此观测序列的基础上建立条件概率模型：若令$\boldsymbol{x}=\{x_1,x_2,...,x_n\}$为观测序列，$\boldsymbol{y}=\{y_1,y_2,...,y_n\}$为与之相对应的标记序列，CRF希望建立一个条件概率模型$P(\boldsymbol{y}|\boldsymbol{x})$。CRF在自然语言处理领域应用

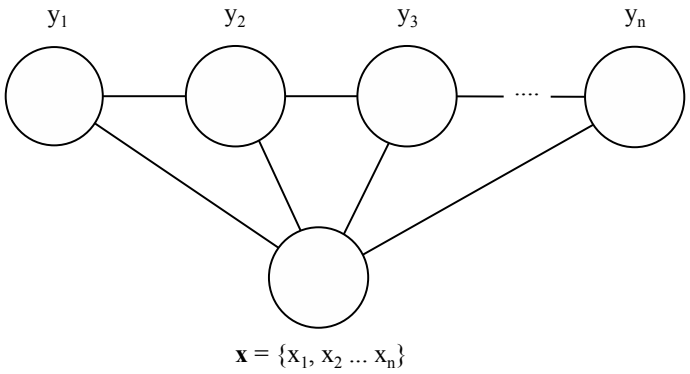


图 2-3 链式条件随机场的图结构

非常广泛，比如词性标注，观测数据为单词构成的短语，标记为每个短语中单词的词性序列。

假设有一无向图$G =< V, E >$，其中每一结点$v$对应一个标记变量$y_v$，$n(v)$表示结点$v$的临近结点。若图$G$的每一个变量都具有马尔科夫性，即：

$$P(y_v|\boldsymbol{x}, \boldsymbol{y}_{V\setminus\{v\}}) = P(y_v|\boldsymbol{x}, \boldsymbol{y}_{n(v)}) \tag{2-7}$$

，$(\boldsymbol{y}, \boldsymbol{x})$构成一个条件随机场。

由于的工作为命名实体识别，实际上是对标记序列建模，其使用的如图2-3所示的链式结构。我们主要讨论这种条件随机场。

条件随机场也可以使用势函数和团来定义条件概率$P(\boldsymbol{y}|\boldsymbol{x})$。给定观测序列$\boldsymbol{x}$，选择指数势函数并引入特征函数，就可以这样定义条件概率

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z} exp(\sum_{j}\sum_{i=1}^{n-1}\lambda_j t_j(y_{i+1}, y_i, \boldsymbol{x}, i) + \sum_{k}\sum_{i=1}^{n}\mu_k s_k(y_i, \boldsymbol{x}, i)) \tag{2-8}$$

其中，$t_j$是定义在观测序列两个相邻标记位置上的转移特征函数，可以看做是定义在图中两个表示相邻标记变量的结点之间的边上的转移特征函数（transition feature function），用来刻画相邻标记变量之间的相关关系和观测序列对其影响，$s_k$是定义在观测序列的标记位置上的状态特征函数（state feature function），用于刻画观测序列对标记变量的影响，$\lambda_j$和$\mu_k$为参数。同时也拥有规范化因子$Z$用于保证概率正确定义。

所谓特征函数是一组实值函数，观察公式可以看出，每个函数与一个特征权重相对应。在模型训练前，特征函数实际上是确定的，训练过程就是得到特征函数权重的过程。下面结合一个标注任务例子来解释转移特征函数和状态特征函数。

假如有一个词性标注任务，如表2-1 则某个转移特征函数如下

表 2-1 一个词性标注任务示例

| 观测序列 | 标记序列 |
|---|---|
| The | [D] |
| boy | [N] |
| knocked | [V] |
| at | [P] |
| the | [D] |
| watermelon | [N] |

$$t_j(y_{i+1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1 & \text{if } y_{i+1} = [P], y_i = [V] \text{ and } x_i = "knocked", \\ 0 & \text{otherwise,} \end{cases}$$

表示第$i$个观测值$x_i$为单词knocked时，其标记很可能是[V]，并且后一个标记很可能是[P]。而状态特征函数如下

$$s_k(y_i, \boldsymbol{x}, i) = \begin{cases} 1 & \text{if } y_i = [V] \text{ and } x_i = "knocked", \\ 0 & \text{otherwise,} \end{cases}$$

表示第$i$个观测值$x_i$为单词knocked时，其标记很可能是[V]。

由此可见，在使用条件随机场的时候，首先要确定某种特征函数，并确定此特征函数集合。

CRF是给定观察序列的情况下，计算整个观察序列对应的标记序列的联合概率分布，而不是给定当前状态标记下一个状态。标记序列的条件分布属性，让CRF可以很好地拟合标记序列的条件概率依赖于观察序列中非独立的特征数据。CRF模型近年来广泛应用于序列标注问题中[2, 3, 5]，并且在多个研究领域都取得了良好的效果。因为条件随机场模型综合了隐马尔科夫模型和最大熵模型的有点，并克服了HMM模型那样严格的独立性假设，并解决了标记偏置缺点。

## 2.3 本章小结

本章首先宏观地介绍了概率模型是通过概率值排序或者通过全局及局部方式对概率空间进行最优搜索来判别目标的数学模型。并由此引出了研究概率模型的

重要工具，概率图模型。并从隐马尔科夫模型开始，介绍了概率图模型通常解决的问题及解决方法中条件独立性的重要性，再通过引入势函数和团的概念，介绍了马尔科夫随机场和条件随机场，循序渐进地解释了生成式模型与判别式模型、无向图模型与有向图模型之间的区别。最终，通过模型本身的特点及相关研究，确定了CRF模型在解决序列标注问题上具有优势。

# 第三章 初等数学自动求解场合下命名实体标注研究

## 3.1 背景描述

概率和统计是我国数学高考试卷的重要组成部分之一，其特点主要是贴近生活、情境多变、题型灵活等，题型以应用题为主，因此概率和统计部分的类人自动解题充满挑战性。总体上看，概率和统计部分的类人自动解题的第一步是进行初等数学概率与统计问题的题意自动分析，将标准化试卷形式的概率与统计类问题转化为问题求解工具可以接受的语义形式，即解决做什么的问题，第二步即是解决怎么做的问题，并生成类人的答题过程，具体的研究方案如图3-1所示。

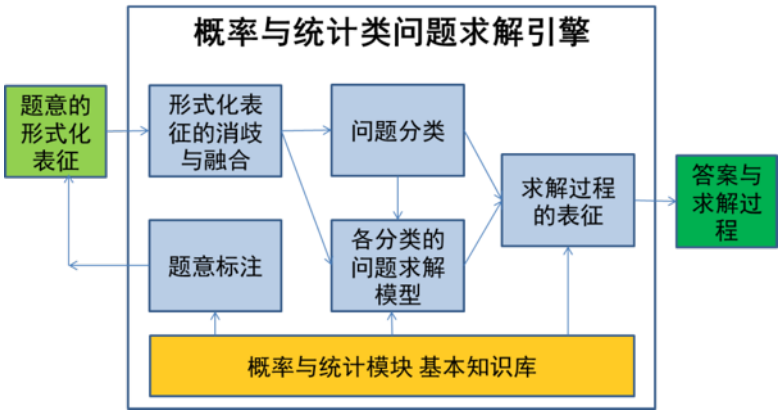


图 3-1 初等数学中概率统计题的自动求解引擎

可以看到，此引擎完成任务的第一步为题意的形式化表征。形式化表征对于人类解题来说，相当于对题干中关键条件理解内化的过程，也就是“我们是如何理解题目中的已知条件的？”这个问题。转化为自然语言处理任务，也就是对于计算机的“如何让计算机理解题目中已知条件？”的问题。

## 3.2 问题表征

自然语言与数学语言是存在差距的，为了能够让计算机理解问题中元素，即将对于计算机本身毫无意义的自然语言字符组成的字符串通过一定的方法转化为用计算机可以识别的数据内容表示的数学语言，我们应设计一种翻译工具完成

这一工作。数学语言通常表示的是问题中已知条件元素之间的相关关系和约束条件。不难想象，仅拥有了数学语言表示的题意，便可以让自动求解的工作成为可能。那么为了实现自然语言到数学语言的转换，首先我们的翻译工具应拥有识别题干中最重要的、以自然语言描述的关键条件，即应拥有搜索题目中的元素以及元素之间约束关系的能力。结合上面提到的实体的定义，在这个应用场景下，题目中的元素以及可能的元素约束关系，将是文本中最具有特殊意义也最首要的信息，也就是所谓的命名实体。

以一个初等数学中概率与统计题目的分层抽样题目为例，其通常为具有一个总体，而总体则有多个总体层组成，并会提供一个样本，而样本则有多个样本层构成，其中样本层中，各个层数量的比例和总体层中各个层数量的比例相同。而题目则会在提供其他数量的情况下，要求解题者根据分层抽样的原理，求出未知的数量。这些命名实体，我们称其为实体组件。让我们以一个题目为例子：

甲、乙两套设备生产同类型产品共4800件，采用分层抽象的方法从中抽取一个容量为80的样本进行质量检测。若样本中有50件产品由甲设备生产，则乙设备生产的产品总数为多少件？

表 3-1 分层抽样问题的实体组件表

| - | 实体名称 | 实体数量 | 单位 |
|---|---|---|---|
| 总体 | 产品 | 4800 | 件 |
| 样本 | 样本 | 80 | 件 |
| 总体层1 | 甲设备 | ? | 件 |
| 总体层2 | 乙设备 | ? | 件 |
| 样本层1 | 甲设备 | 50 | 件 |
| 样本层2 | 乙设备 | ? | 件 |

首先我们人力对此题目进行分析得到的实体组件列表如表3-1所示。考虑解答这道题，于样本实体数量已经给出，再结合样本层1实体数量，可以得到样本层2的实体数量。同时总体和样本之间实体数量比，暗示了总体层各层总量和样本层各层总量的比例。此实体组件表的所有未知量已经解出。此问题实际上也解答完成。

由此我们可以看出，对于一道初等数学概率与统计题分层抽样题，我们可以将其解答过程看成这样的过程：（1）抽取题干中所有的实体组件，并形成表格；（2）抽取题干中明确表示的实体组件之间约束关系；（3）识别题目类型，并

根据题目类型补充隐藏的实体组件约束关系；（4）利用前三部的条件对实体组件表进行补充。可以看出，此解答过程的第一步与第二步，都可以看成自然语言处理的命名实体识别的过程，而第三部的工作实际上也依赖于前两步提供的决策支持。至此，我们得知问题表征的重要工作为命名实体识别标注。

## 3.3 标注集合设计

通过上节的描述，可以看出，对于初等数学的概率统计中分层抽样问题的命名实体，可以分成实体名称、实体数量和单位三种。为了能够实现训练模型、测试模型以及可靠地统计分析识别模型的工作效果。我们应该拥有大量的已经识别完成并人工标记好的语料。为获取这些语料，首先我们要确定我们的命名实体标注方法。

本文使用了B（单词的开头），I（单词的中间），O（单词的结尾）标注方法。结合实体名称，实体数量，和单位三种。最终设计的标注集合如下表3-2和表3-3所示。

表 3-2 分层抽样问题的实体组件表

| 标注 | 含义 |
|---|---|
| B | 当前词为实体的首部 |
| I | 当前词为实体的内部 |
| E | 当前词为实体的尾部 |
| S | 当前词是实体 |
| O | 当前词不是实体 |

表 3-3 分层抽样问题的实体组件表

| 标注 | 含义 |
|---|---|
| NAM | 名字实体 |
| NUM | 数字实体 |
| UNI | 单位实体 |

根据此种标注集合设计，可以对初等数学中概率统计题进行人工标注，示例如表3-4所示。

表 3-4 分层抽样问题的实体组件表

| 字符序列 | 标注 |
|---|---|
| 在 | o |
| 工 | o |
| 厂 | o |
| 中 | o |
| 有 | o |
| A | S-NAM |
| 100 | S-NUM |
| 件 | S-UNI |
| 打 | B-NAM |
| 印 | I-NAM |
| 机 | E-NAM |
| 200 | S-NUM |
| 件 | S-UNI |
| C | S-NAM |
| 300 | S-NUM |
| 件 | S-UNI |
| 三 | o |
| 种 | o |
| 商 | o |
| 品 | o |

## 3.4 本章小结

本章首先分析得出初等数学概率与统计题的自动求解引擎的第一步应该是计算机对于题目的理解，也就是抽取其中首要信息，并转化为形式化表征。然后在尝试解题的过程中得到了题目中究竟哪些文字才是真正需要的实体组件，由此得出了此应用场景下命名实体识别的具体要求。根据此要求，再结合领域内常见的[4]标注方法，最终确定了中文命名实体标注集合。

# 第四章 基于CRF的初等数学命名实体识别算法

通过之前的讨论，我们能够看出基于CRF的初等命名实体识别效果将是较好的。因此我们在此基础上设计了基于CRF的初等数学命名实体识别算法。

## 4.1 总体设计

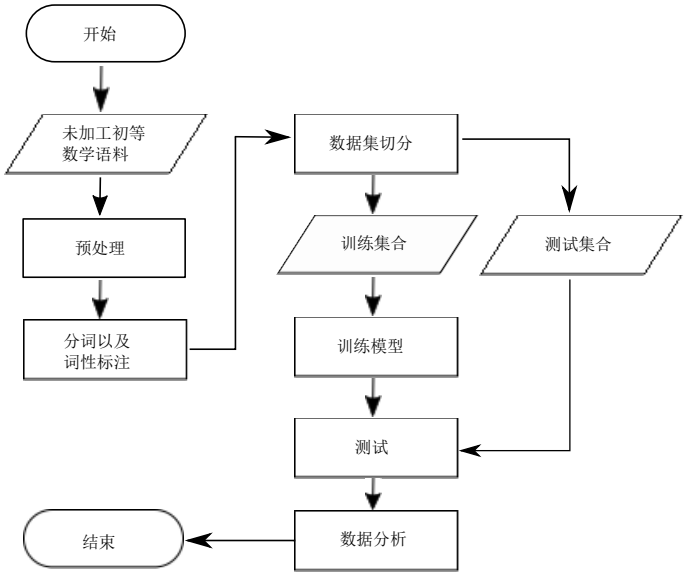


图 4-1 总体设计流程示意图

### 4.1.1 总体实验流程设计

想要完成命名实体识别的识别实验工作，最首要的工作应该是收集大量未标注处理过得相关领域语料。然后对语料进行预处理：（1）除错，由于语料来源不稳定，除错步骤需要尽量去除多余字、错字。（2）修改格式和数据合并。接着通过研究当前环境的语言特点以及应用场合，合理安排标注集合，并对数据进行人工标注。第四步，对数据集合进行切分，分为训练集与测试集合。第五步，构建测试平台并训练测试，统计结果并分析。整体设计流程图如4-1所示。

### 4.1.2 实验环境搭建

由于CRF模型已经有非常多的实现，并不需要自己编码完成。考虑到参考资料和稳定性等因素，选择了CRF++作为实验软件。由于有相当多的预处理工作，应选择一种易用方便的语言作为处理工具使用，在此我设计使用Python作为处理语言。Python是一种面对对象的解释型语言，目前世界范围内广泛使用的版本是2.0和3.0系列，两种语言版本兼容性不完美，因此通常我们使用其中某一个版本。出于方便考虑，选择了Python3.5作为处理工具的编程语言，为了能够良好的管理代码，使用了Git作为版本管理工具。为了防止Python多个版本产生冲突，使用了Virtualenv作为Python运行的虚拟环境。综上所述，最终使用的实验环境如表4-1

表 4-1 实验环境

| 项目 | 名称 | 版本 |
|---|---|---|
| 操作系统 | Ubuntu | 16.04 |
| 编程语言 | Python | 3.5.2 |
| 虚拟环境 | Virtualenv | - |
| 算法软件 | CRF++ | 0.58 |

## 4.2 详细实验流程

### 4.2.1 CRF++软件的安装与使用方法分析

为了能够尽量减少无用工作，首先安装CRF++软件并对其使用过程，数据格式要求有明确的认识是必要的。

#### 4.2.1.1 CRF++的编译与安装

在安装了GCC等工具，在Github得到了软件的源代码。根据参考源代码中说明网页，使用make工具进行第一次编译，发生了缺少链接的库的错误。检查错误后，参考Github页面上issue分页，执行命令：

尝试编译得到了crflearn和crftest软件。

表 4-2 训练数据集数据格式

| 观察序列 | 词性 | 语法树 |
|---|---|---|
| Confidence | NN | B-NP |
| in | IN | B-PP |
| the | DT | B-NP |
| pound | NN | I-NP |
| is | VBZ | B-VP |
| widely | RB | I-VP |
| expected | VBN | I-VP |
| to | TO | I-VP |
| take | VB | I-VP |
| another | DT | B-NP |

#### 4.2.1.2 训练及测试数据集合格式

参考源代码包中example文件夹，可以看见其中有四份例子。结合文档，可以掌握训练和测试的方法。CRF++使用的训练数据集和测试数据集的格式如下。

如表格4-2所示，示例中训练数据集由三列组成，第一列为观察序列，也就是单词或单字；第二列为第一列单词所对应的词性；第三列表示了此句子的语法结构。可以轻而易举地看出，此数据集训练出的模型将会解决英文句子的语法分析任务。在机器学习领域，有监督学习输入数据为一个特征向量和一个标签，在此例子中即为第一列的观察序列和第二列的词性，标签为语法树。考虑我们的初等数学命名实体标注的任务，特征向量的第一个分量也将是观察序列，也就是输入的初等数学题的字符串，而第二个分量可以是单字对应的词性，而标签则应该是命名实体标签。

#### 4.2.1.3 特征模板文件

本文在第二章介绍过，条件随机场模型在使用前应该制定好特征函数，在CRF++中，提供了特征模板作为一种方便的创建特征函数的方法。本文先以CRF++提供的日文分词示例训练集以及特征模板文件为例子，详细解释分析特征模板的使用。

如表4-3所示，在特征模板文件中，每一行代表一个特征，而其中每一个形如%[-2,0]的内容，代表着一个宏，宏的结构为“%[行偏移,列位置]”，代表了相对于当前指向的行的位置和列的绝对位置。比如说，扫描到第三行时，%[-2,0]代表

表 4-3 CRF++中提供的一个特征模板

| 特征模板文件内容 |
| --- |
| # Unigram |
| U00:%x[-2,0] |
| U01:%x[-1,0] |
| U02:%x[0,0] |
| U03:%x[1,0] |
| U04:%x[2,0] |
| U05:%x[-2,0]/%x[-1,0]/%x[0,0] |
| U06:%x[-1,0]/%x[0,0]/%x[1,0] |
| U07:%x[0,0]/%x[1,0]/%x[2,0] |
| U08:%x[-1,0]/%x[0,0] |
| U09:%x[0,0]/%x[1,0] |
| # Bigram |
| B |

表 4-4 日文分词任务训练数据集

| 训练数据集内容 |
| --- |
| 每 k B |
| 日 k I |
| 新 k I |
| 聞 k I |
| 社 k I |
| 特 k B |
| 別 k I |
| 顧 k B |
| 問 k I |
| ４ n B |

着“每”这个日语汉字单个字特征。很显然，根据特征模板，从第一行扫描到第最后一行，生成许多特征，姑且设这个特征数目为$L$。我们又知道输出类型是第三列也就是标记，在此例子中只有B和I。假如说某个训练集合中，输出类型为N种。使用Unigram设计的特征模板将会生成$L \times N$个特征函数。其形式如表4-5

表 4-5 特征函数集合示例

| 特征函数集合示例 |
| --- |
| func1 = if (output = B and feature=”U01:日”) return 1 else return 0 |
| func2 = if (output = I and feature=”U01:日”) return 1 else return 0 |
| .... |
| funcXX = if (output = B and feature=”U01:問”) return 1 else return 0 |
| funcXY = if (output = I and feature=”U01:問”) return 1 else return 0 |

### 4.2.2 未加工初等数学语料的获取及预处理

互联网上有大量的中学数学题库可以使用，但手工下载收集工作量过大，因此使用了Python爬虫获得了大量的免费初等数学题，并进行人工初步筛选去除明显无法使用的文件，比如空文件；不包含答案的文件；乱码文件等。

由于获得的题库文件格式通常为Word，而训练CRF++的文件通常为UTF-8编码纯文本文档，因此选择使用python-doc库进行处理，去除了其中空文件和并仅仅抽取出题干。

利用工业上常用的分词及词性标注软件包jieba，对所有的数据进行了合并分词并标注了词性。效果如表4-6

随后，本文根据第三章讨论得出的标注集合，对所有数据进行人工标注处理。处理后，由于人工标注难免出现错误，使用python正则表达式对所有的数据进行了格式错误检测。最终获得的数据集合效果如表4-7

### 4.2.3 特征模板设计

#### 4.2.3.1 原子特征

原子特征可以描述当前词和前后连续两个词的词性和词形，能够表达有限的上下文信息。内容如表4-8

表 4-6 分词及标注后结果

| 分词及标注词性后文件内容 |
| --- |
| 一 m |
| 个 q |
| 工 n |
| 厂 zg |
| 有 v |
| 若 c |
| 干 v |
| 个 q |
| 车 zg |
| 间 f |

#### 4.2.3.2 组合特征

通过搭配词性和词形以及组合多词等手段，构造符合的非线性特征，形成了组合特征。内容如表4-9

最后，将数据集拆分成训练集合和测试集合，使用训练集合对CRF++进行训练，并测试得到机器标注结果。如4-10

## 4.3 本章小结

本章对于CRF++软件的编译安装以及基本使用方法进行了简述，并结合实例阐明了CRF++软件使用的特征模板的运作原理，根据原理与初等数学中数学命名实体的一些规律，制定了原子特征模板和组合特征模板。并完成了训练和测试的工作，最终得到了既包含人工数据也包含机器标注的初等数学概率与统计题数学命名实体的测试数据。

表 4-7 人工标注后的数据

| 标注命名实体后文件内容 |
| --- |
| 某 r O |
| 单 n O |
| 位 q O |
| 有 v O |
| 职 ng B-NAM |
| 工 n E-NAM |
| 7 x B-NUM |
| 5 x I-NUM |
| 0 x E-NUM |
| 人 n S-UNI |
| ， x O |
| 其 r O |
| 中 f O |
| 青 ns B-NAM |
| 年 m I-NAM |
| 职 ng I-NAM |
| 工 n E-NAM |
| 3 x B-NUM |
| 5 x I-NUM |
| 0 x E-NUM |
| 人 n S-UNI |

表 4-8 原子特征模板

| 本文所用的部分特征模板 |
| --- |
| U00:%x[-2,0] |
| U01:%x[-1,0] |
| U02:%x[0,0] |
| U03:%x[1,0] |
| U04:%x[2,0] |
| U05:%x[-2,1] |
| U06:%x[-1,1] |
| U07:%x[0,1] |
| U08:%x[1,1] |
| U09:%x[2,1] |

表 4-9 组合特征模板

| 本文所用的部分特征模板 |
| --- |
| U10:%x[0,0]/%x[0,1] |
| U11:%x[0,0]/%x[-1,0] |
| U12:%x[0,0]/%x[1,0] |
| U13:%x[0,1]/%x[-1,0] |
| U14:%x[0,0]/%x[1,1] |
| U15:%x[-1,0]/%x[-1,1] |
| U16:%x[-1,0]/%x[-2,0] |
| U17:%x[-2,0]/%x[-2,1] |
| U18:%x[1,0]/%x[2,0] |
| U19:%x[-1,0]/%x[1,0] |
| U20:%x[1,0]/%x[0,1] |
| U21:%x[-2,1]/%x[-1,1] |
| U22:%x[-2,1]/%x[0,1] |
| U23:%x[-1,1]/%x[0,1] |
| U24:%x[-1,1]/%x[1,1] |
| U25:%x[0,1]/%x[1,1] |
| U26:%x[0,1]/%x[2,1] |
| U27:%x[1,1]/%x[2,1] |

表 4-10 机器标注结果

| 机器标注结果输出文件部分内容 |
| --- |
| 某 r O O |
| 校 ng O O |
| 有 v O O |
| 高 a B-NAM B-NAM |
| 级 q I-NAM I-NAM |
| 教 v I-NAM I-NAM |
| 师 ng E-NAM E-NAM |
| 2 x B-NUM B-NUM |
| 6 x E-NUM E-NUM |
| 人 n S-UNI S-UNI |
| ， x O O |
| 中 f B-NAM B-NAM |
| 级 q I-NAM I-NAM |
| 教 v I-NAM I-NAM |
| 师 ng E-NAM E-NAM |
| 1 x B-NUM B-NUM |
| 0 x I-NUM I-NUM |
| 4 x E-NUM E-NUM |
| 人 n S-UNI S-UNI |

# 第五章 实验结果与评测分析

## 5.1 相关术语解释

### 5.1.1 正类与负类

在命名实体标注任务中，可以将需要预测的元素分为两类，正类与负类。举个例子，比如在“我要努力学习数学知识。”，若“数学知识”一词是我们需要的命名实体，那么此词为正类，其他为负类。

训练模型将会对输入的特征向量进行预测，训练模型所预测出的结果最终将分为四中。即：（1）真正（True Positive，简称TP），含义为本身为正类，同时被预测为正；（2）真负（True Negative，简称TN），含义为本身为负类，同时被预测为负；（3）假正（False Positive，简称FN），含义为本身为负类，但却被模型预测为正类；（4）假负（False Negative），含义为本身为正类，但却被预测为负类。

当我们定义了这些数值时，便可以方便地定义一些统计量，这些统计量可以较好地评价一个模型的预测效果。

### 5.1.2 评测指标

精确率（precision）、召回率（recall）、准确率（accuracy）和F值是评测中文命名实体识别系统性能的指标，也是本文采取的评测指标。如果我们设定真正为$TP$，真负为$TN$，假正为$FP$，假负为$FN$，可以对四种评测值进行定义：

（1）精确率$P$：精确率针对某一预测结果而言，它表示预测为正的样本里面有多少的真正的正样本。预测为正的样本，显然有真正和假正，公式定义为：

$$P = \frac{TP}{TP + FP} \tag{5-1}$$

在实体标注中，可以看成正确识别的命名实体个数与识别到所有的命名实体的比值。

（2）召回率$R$：召回率则是表示样本中的正类有多少被预测正确了。其中包含了真正和假负。和精确率相比，其关注的是我们原来的样本而非预测结果。公式定义为：

$$R = \frac{TP}{TP + FN} \tag{5-2}$$

（3）F值：F值公式定义如下：

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 P + R} \tag{5-3}$$

是综合了精确率和召回率两个值进行评估的办法，同时考虑了$P$和$R$两个值，其中$\beta$是一个权重，决定比较精确率和召回率的时候更加侧重精确率还是召回率，通常设定为1或0.5。在本文，我们选择1，即同等地重视两种指标。

（4）准确率$A$：准确率是无论负类还是正类都考虑在内，所有预测正确的结果与所有样本的比值。公式定义如下：

$$A = \frac{TP + TN}{TP + FN + FP + TN} \tag{5-4}$$

## 5.2 评测过程与结果

我们使用了CoNLL-2000工具，其提供了一个perl语言脚本conlleval.pl，这时专门用于统计命名实体识别系统性能评测指标的软件。将实验获得的测试文档输入重定向到此脚本，便可统计出评测指标。

我们的最终结果，总体来看，准确率达到了90.99%，精确率为89.74%，召回率为82.21%，F值为85.81%。最终输出文件如表5-1所示，由此可以得出不同命名实体识别效果如表5-2：

表 5-1 conlleval.pl输出文件

| 统计结果输出文件内容 |
| --- |
| ssed 3617 tokens with 596 phrases; found: 546 phrases; correct: 490. |
| accuracy: 90.99%; precision: 89.74%; recall: 82.21%; FB1: 85.81 |
| NAM: precision: 77.83%; recall: 65.57%; FB1: 71.17 230 |
| NUM: precision: 98.31%; recall: 96.69%; FB1: 97.49 178 |
| UNI: precision: 98.55%; recall: 95.77%; FB1: 97.14 138 |

表 5-2 评测指标结果

| 命名实体种类 | 精确率 | 召回率 | F值 |
|---|---|---|---|
| NAM | 77.83% | 65.57% | 71.17% |
| NUM | 98.31% | 96.69% | 97.49% |
| UNI | 98.55% | 95.77% | 87.14% |

## 5.3 结果分析

可以看出，结合了原子模板和组合模板后，对于数字实体和单位实体的识别效果较好，但对于名字实体的识别效果相对较差。我们认为可以使用一些人工启发式的规则识别测试样本运用统计方式未识别出来的实体[2]或提供高层条件随机场支持来识别为识别出的正类[3]。

## 5.4 本章小结

本章首先详细介绍了与评价中文命名实体标注系统性能评测指标相关的重要概念。然后使用工具得到了评测指标值，并对评测指标进行简单分析。可以看出，使用组合模板和原子模板结合的情况下，在数字实体和单位实体的识别效果是较好的，但在名字命名实体的识别上仍有相当大的研究空间。

# 第六章 总结与展望

## 6.1 总结

本文在预设的初等数学概率与统计题自动求解工作的背景下，分析得出解题的首要工作为计算机对数学题意的语义理解。对于语义的理解的关键在于关键信息的提取，而对于初等数学概率与统计题，命名实体即使其中最关键的信息点。

本文对于多个可以用于命名实体识别的概率图模型进行了简要的描述和比较，最终确定了条件随机场对于此工作的适用性和优越性。本文的创新点如下：

（1）以自动解题目的为指导下对于初等数学概率与统计题目中数学命名实体的确定。本文结合实例，分析了初等数学概率与统计题的解题过程，结合解题过程，确定了初等数题解题过程中所需的关键条件为一些命名，及与命名对应的数字与单位。并确定了为了能够达到解题，需要确定的关键条件。并由此确定了所需的标注集合为名字实体，数字实体以及单位实体。

（2）基于CRF原理的适用于中文初等数学命名实体标注的特征研究：首先充分地分析了条件随机场模型中特征函数的含义以及CRF++中特征模板生成特征函数的规律后，根据中文初等数学命名实体的词性，词形以及组合多词的方法构建了组合特征。提升了识别效果。

## 6.2 展望

本文通过人工标注命名实体语料数据训练概率模型的方法来实现识别初等数学命名实体。而单层条件随机场的识别效果有限，尤其是初等数学名字实体构词规则更复杂，识别效果较差。我们认为，结合一些构词规则对未识别出的实体进行二次识别[1]，或使用层叠条件随机场可能会有更好的效果[3]。在今后的研究工作，会考虑借鉴这些方法。而分类机器学习模对于数据的人工标注工作量巨大，而且不可能将所有数据加入训练集合，今后可能考虑研究半聚类的算法，以减轻人工工作。

# 参考文献


[1] 潘正高. 基于规则和统计相结合的中文命名实体识别研究[J]. 情报科学, 2012:708–786

[2] 何炎祥, 罗楚威;胡彬尧. 基于CRF和规则相结合的地理命名实体识别方法[J]. 计算机应用与软件, 2015, 32(1)

[3] 周俊生;戴新宇;尹存燕;陈家骏, 周俊生. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006, 34(5):804–809

[4] 周志华. 机器学习[M]. 清华大学出版社, 2016

[5] 鞠久朋;周国栋, 周国栋, 张伟伟,等. CRF与规则相结合的地理空间命名实体识别[J]. 计算机工程, 2011, 37(7):210–215

[6] 王根, 赵军. 基于多重冗余标记CRFs的句子情感分析研究[J]. 中文信息学报, 2007, 05

[7] 洪铭材, 张阔, 唐杰, et al. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学, 2006, 10

[8] 丁德鑫, 曲维光, 徐涛, et al. 基于CRF模型的组合型歧义消解研究[J]. 2008, 04

[9] 孙虹, 陈俊杰. 双层CRF与规则相结合的中文地名识别方法研究[J]. 计算机应用与软件, 2014, 11

[10] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 10

[11] 郑家恒, 张辉. 基于HMM的中国组织机构名自动识别[J]. 计算机应用, 2002, 11

[12] 薛翠芳, 郭炳炎. 汉语文本特征词的抽取方法[J]. 情报学报, 2000, 03

[13] 黄德根, 杨元生, 王省, et al. 基于统计方法的中文姓名识别[J]. 中文信息学报, 2001, 02

[14] 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, 27(2):88–94

# 致 谢

感谢我的指导老师钟秀琴老师，在本实验过程中给予我悉心指导，从理论学习到实践结束中过程遇到很多困难都是老师给予我鼓励与指引，使我能够克服重重困难，完成任务，特别在此谨向钟老师致以诚挚的谢意和崇高的敬意。谢谢! 感谢计算机科学与工程学院四年来对我的辛苦培育，让我在大学这四年来获得了成长，为我提供了良好的学习环境、生活环境。感谢老师们四年来对我无微不至的关怀和指导，让我得以在这四年中学到很多有用的知识。在此，我还要感谢在班里同学和朋友，感谢你们在我遇到困难的时候帮助我，给我支持和鼓励，感谢你们。

外文资料原文

# 1 An Introduction to Conditional Random Fields for Relational Learning

**Charles Sutton**
Department of Computer Science
University of Massachusetts, USA
casutton@cs.umass.edu
http://www.cs.umass.edu/~casutton

**Andrew McCallum**
Department of Computer Science
University of Massachusetts, USA
mccallum@cs.umass.edu
http://www.cs.umass.edu/~mccallum

## 1.1 Introduction

Relational data has two characteristics: first, statistical dependencies exist between the entities we wish to model, and second, each entity often has a rich set of features that can aid classification. For example, when classifying Web documents, the page's text provides much information about the class label, but hyperlinks define a relationship between pages that can improve classification [Taskar et al., 2002]. Graphical models are a natural formalism for exploiting the dependence structure among entities. Traditionally, graphical models have been used to represent the joint probability distribution $p(\mathbf{y}, \mathbf{x})$, where the variables $\mathbf{y}$ represent the attributes of the entities that we wish to predict, and the input variables $\mathbf{x}$ represent our observed knowledge about the entities. But modeling the joint distribution can lead to difficulties when using the rich local features that can occur in relational data, because it requires modeling the distribution $p(\mathbf{x})$, which can include complex dependencies. Modeling these dependencies among inputs can lead to intractable models, but ignoring them can lead to reduced performance.

A solution to this problem is to directly model the conditional distribution $p(\mathbf{y}|\mathbf{x})$, which is sufficient for classification. This is the approach taken by *conditional random fields* [Lafferty et al., 2001]. A conditional random field is simply a conditional distribution $p(\mathbf{y}|\mathbf{x})$ with an associated graphical structure. Because the model is

图 -1

conditional, dependencies among the input variables $\mathbf{x}$ do not need to be explicitly represented, affording the use of rich, global features of the input. For example, in natural language tasks, useful features include neighboring words and word bigrams, prefixes and suffixes, capitalization, membership in domain-specific lexicons, and semantic information from sources such as WordNet. Recently there has been an explosion of interest in CRFs, with successful applications including text processing [Taskar et al., 2002, Peng and McCallum, 2004, Settles, 2005, Sha and Pereira, 2003], bioinformatics [Sato and Sakakibara, 2005, Liu et al., 2005], and computer vision [He et al., 2004, Kumar and Hebert, 2003].

This chapter is divided into two parts. First, we present a tutorial on current training and inference techniques for conditional random fields. We discuss the important special case of linear-chain CRFs, and then we generalize these to arbitrary graphical structures. We include a brief discussion of techniques for practical CRF implementations.

Second, we present an example of applying a general CRF to a practical relational learning problem. In particular, we discuss the problem of *information extraction*, that is, automatically building a relational database from information contained in unstructured text. Unlike linear-chain models, general CRFs can capture long distance dependencies between labels. For example, if the same name is mentioned more than once in a document, all mentions probably have the same label, and it is useful to extract them all, because each mention may contain different complementary information about the underlying entity. To represent these long-distance dependencies, we propose a *skip-chain CRF*, a model that jointly performs segmentation and collective labeling of extracted mentions. On a standard problem of extracting speaker names from seminar announcements, the skip-chain CRF has better performance than a linear-chain CRF.

## 1.2 Graphical Models

### 1.2.1 Definitions

We consider probability distributions over sets of random variables $V = X \cup Y$, where $X$ is a set of *input variables* that we assume are observed, and $Y$ is a set of *output variables* that we wish to predict. Every variable $v \in V$ takes outcomes from a set $\mathcal{V}$, which can be either continuous or discrete, although we discuss only the discrete case in this chapter. We denote an assignment to $X$ by $\mathbf{x}$, and we denote an assignment to a set $A \subset X$ by $\mathbf{x}_A$, and similarly for $Y$. We use the notation $\mathbf{1}_{\{x=x'\}}$ to denote an indicator function of $x$ which takes the value 1 when $x = x'$ and 0 otherwise.

A graphical model is a family of probability distributions that factorize according to an underlying graph. The main idea is to represent a distribution over a large number of random variables by a product of local functions that each depend on only a small number of variables. Given a collection of subsets $A \subset V$, we define

图 -2

随机数据有两个特点：第一，我们希望能够标注的实体之间存在统计上的依赖性。第二，每一个实体通常有丰富的特征来辅助分类。例如。当我们分类web文件时，网页中的文本提供了关于分类标签的大量信息，但是超链接定义了网页之间的联系，这可以优化分类图模型是一种利用实体间的依赖结构的自然形式。传统上，图模型用于表示联合分布律p(x,y)，其中变量y表示我们希望预测的实体的属性，输入变量x表示我们观测到的实体的特征。但是在应用大量的关联数据的局部特征进行联合概率建模会有困难，因为这要求对分布p（x）建模，其中包含了纷杂的依赖性。在输入中对这些依赖性建模会产生难以控制的模型，但是忽略他们会降低效果。解决这个问题就是直接对条件概率p（y—x）建模，能够进行充分的分类。这是crf一文中提出的方法。条件随机场是一个简单的条件概率p（y—x）结合了一个相关的图结构。因为模型是随机的，输入变量x之间的依赖性并不需要被精确地表示出来，提供了输入的全局特征使用。例如，在自然语言处理任务中，有用的特征包括邻近的词、二元语法、前缀、后缀、在特定领域里的词汇、以及像来源于Wordnet的语义信息。最近，随着在文本、生物信息学以及计算机视觉等领域的成功应用，人们对于CRF的兴致越来越为浓厚。

这一章分为两个部分：

第一部分中我们提出了目前训练的说明以及条件随机场的推理技术。我们探讨重要的线性链条件随机场特例，并将其推广到任意的图结构上。也包括一段简短的讨论关于实际CRF应用所需的技术。

第二部分中我们提供了一个将综合CRF应用于关系学习问题中的例子。特别的，我们探讨了信息抽取的问题，即自动的从无结构的语料中建立有关联的数据库。与线性链CRF不同，综合CRF能得到标签之间远距离的依赖性例如，如果同样的名字在一个文件中提到了超过一次，所有的提及可能会获得相同的标签，并且这对于对他们全部的提取很有用处，因为每次提及也许会包涵关于潜在实体不同的互补的信息。为了实现这种远距离依赖性，我们拟采用跳跃链CRF，它会同时对提到的内容进行分隔与集合的标注。在一个标准的从研讨会声明提取发言者名字的问题中，跳跃链CRF相比于线性链CRF会有更好的表现。