**基于NLP的文字指令分析**

小组成员：

陈柳杨 2022211043 （队长）

左浩渝 2022211040

王若竹 2022210985

## 1 实验任务背景及目的

**1.1项目背景**

当今社会是一个快节奏时代：繁重的工作、纷繁多变的社交场合、紧张拥堵的通勤道路……种种因素让人们的时间变得更加紧凑而宝贵。而在这样一个快节奏时代，人们同时每天总会为穿衣发愁，多变的天气、不同的出差地点、不懂时尚搭配的苦恼……一套小小的衣服穿搭却难倒了不少人。本项目将依托图像处理与机器学习技术，旨在开发一款智能化根据实际情况，在保障冷暖的同时，针对用户身材、风格偏好等进行个性化穿衣搭配推荐的软件。而本次实验将完成该项目的指令分析部分，优先解决了分词问题。

**1.2研究背景**

自然语言处理（Natural Language Processing，NLP）作为人工智能领域的一个重要分支，致力于使计算机能够理解、解释和生成人类语言。在NLP的众多任务中，分词作为最基本且至关重要的步骤之一，对于机器对文本的准确理解和处理至关重要。分词的质量直接影响到后续任务的性能，如文本分类、情感分析和机器翻译等。因此，开发高效、准确的分词算法对于提升NLP系统整体性能具有重要意义[1]。

在现今的技术水平下，已经存在多种分词算法，例如基于规则的方法、统计方法和深度学习方法，但这些算法在处理特定语境下的文本时仍然存在一些局限性。

（1）基于规则的分词算法： 这类算法依赖于语言学家设计的规则集，通过规则匹配来进行分词。然而，在面对复杂的语法结构和不断变化的语言规则时，这些算法可能表现不佳。

（2）统计方法： 统计方法基于语料库的频率信息，常用的有HMM（Hidden Markov Model）和CRF（Conditional Random Fields）等。尽管在一些应用中表现良好，但对于处理未知领域的文本，统计方法可能面临一定的限制。

1. 深度学习方法： 近年来，随着深度学习的兴起，采用神经网络进行分词的方法也得到了广泛应用。然而，这些方法往往需要大量的标注数据和强大的计算资源，对于资源有限的场景可能不太适用[2]。

因此，有必要进一步研究和开发新的分词算法，或者通过模仿和改进现有工具以提高分词的准确性和适应性。

**1.3实验目的**

本实验的主要目标是基于HMM（Hidden Markov Model）算法对jieba库进行模仿。通过这种方式，我们旨在拟合以至改进jieba库的分词效果，使其更好地适应不同语境，并提高中文分词的整体性能[3] 。

除此之外，我们还将使用TextRank算法对分词后的文本中的部分信息进行了分类处理。

## 2 相关技术介绍

2.1 隐马尔科夫模型（Hidden Markov Model，HHM）

2.1.1 马尔科夫链

在任一时刻t，观测变量的取值仅依赖于当前的状态变量，与其他时刻的状态变量及观测变量无关；同时，当前的状态值仅依赖于前一时刻的状态.

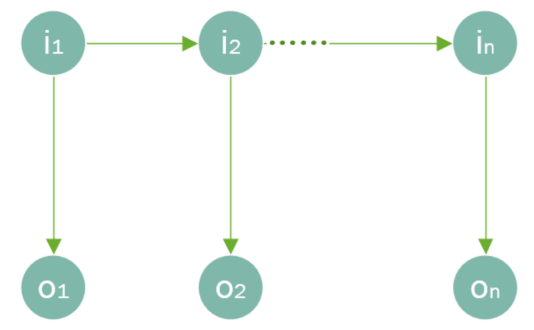


Fig. 1 Markov Chain

图 1马尔科夫链

2.1.2 模型结构

HMM的联合概率分布为:

2.1.3 变量

1)状态变量，表示第t时刻的状态，假定状态是隐藏的、不可被观测的.

2)观测变量，表示第t时刻的观测值.

3)状态I的枚举值.

2.1.4 参数

1)状态转移概率A：.

2)输出观测概率B：.

3)初设状态概率π：.

2.1.5 三个基本问题

1)概率计算问题。给定模型和观测序列，计算观测序列O出现的概率.

2)学习问题。已知观测序列，估计模型参数，使得观测序列概率最大.

3)预测问题。给定模型和观测序列，求对给定观测序列条件概率最大的隐藏状态.

2.1.6 概率计算

1)前向算法

给定隐马尔科夫模型，至时刻t的观测序列为，且状态为的概率为前向概率:

将 HMM 的参数代入，得到

2)后向算法=

给定隐马尔科夫模型，满足时刻t的状态为的条件下，从时刻t+1到T的观测序列为的概率为后向概率:

将 HMM 的参数代入，得到

2.1.7 学习问题

1. 有监督学习

在选取的数据集有标注时，我们可以采用统计的方法直接估计 ，具体公式如下：

其中 表示满足 的状态对个数， 表示满足 的状态对个数， 表示 的状态数，且要求 是合法的开始状态。

1. 无监督学习

考虑如果我们手里有的数据集只有没有标注的数据集，则我们可以通过Baum-Welch算法得到 的最大似然估计，算法执行的标准过程如下：

1. 根据初始的 求出在当前模型下的

2. 定义

经转化易得

3. 定义最大似然函数

其中 表示当前模型， 表示模型的最大似然估计，则我们的最终目标就是求取使得上式最大的 。

将 拆分得到

因为拆分得到的三个部分相互独立，我们可以通过使各部分均取到最大值的方式使得最大似然函数的值最大。

对各部分应用拉格朗日函数，求解在有限区域内的函数极值，得到极值点为

将上述公式用 表示，得到

4. 重复上述过程，直到函数收敛

* 1. 预测问题

考虑到人们正常的说话顺序与思考逻辑，我们考虑使用前向概率计算 ,具体使用的算法为 Viterbi 算法，执行过程如下：

定义 表示第 个时刻状态为 的前向概率，即 ，则有状态转移公式

最终最大的前向概率就是 ，其中 表示所有合法的结束状态的集合。由于要求取的是概率最大的 序列，故可以在转移过程中记录最佳转移点，在得到最大的概率后从最后一个状态开始往前递归回溯，得到概率最大的整个序列。

2.2 预测与词典结合的语句分割算法

由于预测的精确度有限，我们考虑使用类似于 jieba 库的方法，将预测与词典相结合，优先尝试使用词典对原句进行分割，具体算法类似 Viterbi 算法，利用后向概率计算出概率最大的分割路径，然后从前往后对原句进行分割。但由于过程中可能出现没有出现过的字词，我们对这种词应用 HMM 模型对其划分方式和词性标记等进行预测，最终得到语句分割的结果

2.3 关键字提取算法 TextRank（改进版）

2.3.1 算法原理

TextRank 算法是基于 PageRank 衍生出来的一种无监督关键字提取算法，其主要原理来源于概率问题中的随机游走问题，本质上是在计算一个在图上随机游走的人最终处于每个节点的概率。假设给定的图是一个连通图，根据随机游走问题的时齐性、马氏性与遍历性，我们可以得知随机游走问题的结果必定收敛于一个固定的值。因此我们可以通过迭代加深的方式得到该随机游走的人最终处于每个节点的概率。

2.3.2 具体实现

1)建图

首先，我们进行关键词排序的本质上是一个词语序列，所以我们需要先根据词语序列构建出词语之间关系的图结构，才能在其上应用 TextRank 算法。一般而言，我们定义两个单词之间有联系当且仅当两个词语之间的字数小于等于一个给定值 。另外，对于一些特殊类型的词语对，我们可能需要特殊定义他们的边权，如名词和形容词之间的边边权可能就需要设大一点，表明两者之间的联系比较紧密。

2)概率转移定义

定义 表示随机游走的人行走时长 后处于节点 的概率，则该人接下来一步的游走方式有两种：①该词与关键词联系不大，随机选择一个新的词 或者 ②该词与关键词联系大，走向与其相关的词。将每个词与关键词的联系用 表示，则概率转移的公式可以表示为如下形式：

将上述操作矩阵化，可得到

其中 ，

，,

多次迭代即可，一般而言迭代 次即可收敛

3)每个词与关键词联系的定义

为了更好的规范化上述 的大小，我们队每个词语定义了一些附加属性，具体表格如下：

## 3 结论

由于相关数据的缺乏，我们无法使用大规模数据集对模型的效果进行评估。因此我们采用随机生成一些请求语句的形式测试模型的运行结果。运行结果如下：

可以看出，。。。

在未来，我们打算。。。

## 参考文献

备注：网页引用生成于 https://www.citationmachine.net/apa/cite-a-website