**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隐马尔科夫模型（Hidden Markov Model，HHM）[9]

2.1 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。

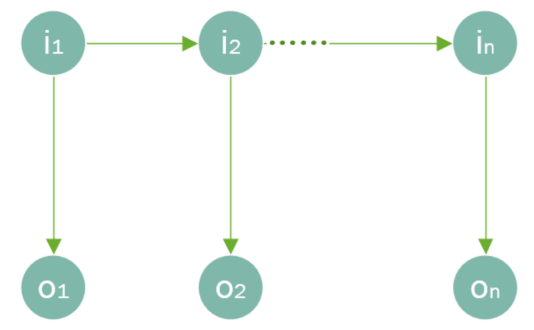


Fig. 1 Markov Chain

图 1马尔科夫链

2.2 模型结构

HMM的联合概率分布为:

2.3 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。

2.4 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：.

2.5 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。

2.6 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

2.7 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[8]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

1. 重复上述过程，直到函数收敛

但由于上述方法在数据集语句长度过长时， 容易极速减小到超出小数的可表示范围，导致出现概率为 的现象并发生运行错误，所以我们选择使用自己制作的小型数据集进行训练，从而保证模型的正确运行。

除此之外，由于我们采用的数据集含有多条语句，所以我们对原演化公式进行了拓展，最终使用的全部公式如下：

由原数据集的各行语句都是相互独立的，我们可以证明上述公式正确性。

最后，作为循环结束的标准，我们定义了两个相关量：loss，表示模型参数前后变化的大小，当其小于一定值（err=0.01）时终止循环；loop\_lim，当循环次数超过上界时同样终止循环。由此无监督学习算法整体完成。

2.8 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[7]，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。

**3 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[4] 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。

**4 关键字提取算法 TextRank（改进版）**

4.1 算法原理

TextRank[6] 算法是基于 PageRank[5] 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。

4.2 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[[1]](#footnote-1)：

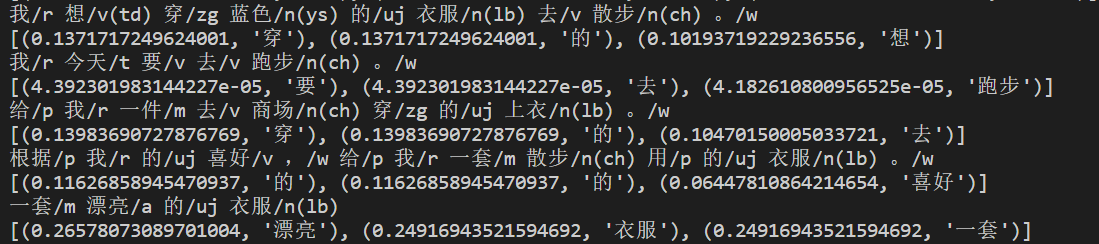
将上述操作矩阵化，可得到

其中 ，

，,

多次迭代即可，一般而言迭代 次即可收敛[6]

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，我们在北京大学语料库加工规范[10]对词性的定义的基础上将词性进行了简化分类。除此之外，我们还对每个词语都定义了一些附加属性，具体表格见附录B。

## 5 结论

本次实验使用的训练数据集是“人民日报词性标注语料数据（199801）”[11]和自造的小型数据集（self\_made.txt），执行的内容包括语句的分割、词性标记与关键词提取，其中词性标记参考“北京大学语料库加工规范[10]”。另外，我们将本项目的代码进行了开源，具体可参考[gitee](https://gitee.com/sherlocked-hzoi/machine_learning_nlp)。

我们进行语句分割和标记的模式主要有四种：①用未分割的语料以字为单位训练无监督模型，②用已分割并带有词性标注的语料以字为单位训练有监督模型，③用已分割且不带有词性标注的语料以词为单位训练无监督模型，④用已分割且带有词性标注的语料以词为单位训练有监督模型，得到的训练结果使用 Levenshtein distance[12] 进行评估，具体的公式如下：

其中pred表示预测结果，orig表示正确结果，len(orig)表示正确结果的长度。

依照上述公式得到各项结果的正确率如下：（由于评估时选用的是随机数据，所以结果具有一定的不确定性）

1. 用未分割的语料以字为单位训练无监督模型
   1. 分词准确率：1.0；
   2. 词性标注准确率：0.771410256410257
   3. 语句匹配准确率：0.9735950206714822
2. 用已分割并带有词性标注的语料以字为单位训练有监督模型
   1. 分词准确率：1.0
   2. 词性标注准确率：0.8116757099110039
   3. 语句匹配准确率：0.979940855951303
3. 用已分割且不带有词性标注的语料以词为单位训练无监督模型
   1. 分词准确率：0.6888888888888891
   2. 词性标注准确率：0.7403846153846158
   3. 语句匹配准确率：0.8939552707412531
4. 用已分割且带有词性标注的语料以词为单位训练有监督模型
   1. 分词准确率：0.7004992106690316
   2. 词性标注准确率：0.8297651482504423
   3. 语句匹配准确率：0.8897645126642648

由此可见，一般选用以字为单位进行训练可以得到更优秀的效果

另外，由于关键词的定义我们并没有得出一个十分统一的标准，因此我们选择通过使用几个测试输出观察结果的形式来评估模型的效果，最终的输出结果如下：

尽管目前的结果还可以接受，我们的算法还有很多的可改进之处：

1)我们当前的无监督学习方法由于精度问题存在无法使用大规模数据集训练的缺陷，虽然对于实际应用没有过大的影响，但增大数据集的数据量无疑是一种模型的优化方向，所以也许存在一种方式，能够使得这种模型应用于更大规模的数据集训练，这也是我们之后要研究的课题之一

2)尽管在表面上算法的准确率是可以接受的，但这很大程度上是依赖于词典的加持，模型本身的预测效果并不是很理想（实际上词性预测的正确率只有不到50%，甚至在基于单词进行预测时正确率仅有不到10%），所以未来我们考虑使用能够通过增大数据集数据量提高预测精度的深度学习模型来获得更好的模型预测效果，提高预测部分的精确度

3)关于附加属性的定义其实我们只应用于了很小一部分的已知词汇，大部分的词汇都处于一种没有标注的状态，未来我们打算更加细化这一方面的属性定义，从而辅助关键词提取的效果，使得关键词提取的效果更加精确，甚至考虑将其作为一种词性加入预测模型进行训练，这样也许就可以使用较少的标注数据得到较为可靠的预测结果，这也是我们之后的主要研究方向

## 参考文献

[1] Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

[2] Zhang, Y., & LeCun, Y. (2015). Text Understanding from Scratch. arXiv preprint arXiv:1502.01710.

[3] Fuyuan Hu, Haozhen Wu, Baijun Wu, Junhui Li. (2019). Jieba: Chinese Text Segmentation. GitHub Repository. https://github.com/fxsjy/jieba

[4] A. Sun, 《fxsjy/jieba》. 2023年12月12日. 见于: 2023年12月12日. [在线]. 载于: https://github.com/fxsjy/jieba

[5] 《PageRank》, *Wikipedia*. 2023年12月9日. 见于: 2023年12月12日. [在线]. 载于: https://en.wikipedia.org/w/index.php?title=PageRank&oldid=1189073599

[6] R. Mihalcea和P. Tarau, 《TextRank: Bringing Order into Texts》, Conference on Empirical Methods in Natural Language Processing (EMNLP), July 25-26, 2004. Barcelona, Spain. 见于: 2023年11月28日. [在线]. 载于: https://digital.library.unt.edu/ark:/67531/metadc30962/

[7] 《Viterbi algorithm》, *Wikipedia*. 2023年10月29日. 见于: 2023年12月12日. [在线]. 载于: https://en.wikipedia.org/w/index.php?title=Viterbi\_algorithm&oldid=1182487909

[8] 《Baum–Welch algorithm》, *Wikipedia*. 2023年10月31日. 见于: 2023年12月12日. [在线]. 载于: https://en.wikipedia.org/w/index.php?title=Baum%E2%80%93Welch\_algorithm&oldid=1182878935

[9] L. R. Rabiner, 《A tutorial on hidden Markov models and selected applications in speech recognition》, *Proc. IEEE*, 卷 77, 期 2, 页 257–286, 2月 1989, doi: 10.1109/5.18626.

[10] S. Yu, H. Duan和Y. Wu, 《Corpus of Multi-level Processing for Modern Chinese》. Peking University Open Research Data Platform, 2018年3月8日. doi: 10.18170/DVN/SEYRX5

[11] 数据集市-www.shujujishi.com, 《人民日报词性标注语料数据（199801）》. 见于: 2023年12月12日. [在线]. 载于: http://shujujishi.com/dataset/a27b9a15-24ab-4dab-aa2b-7e073458973c.html

[12] 《Levenshtein distance》, *Wikipedia*. 2023年12月14日. 见于: 2023年12月17日. [在线]. 载于: https://en.wikipedia.org/w/index.php?title=Levenshtein\_distance&oldid=1189858598

Appendix A. 词性表

|  |  |  |
| --- | --- | --- |
| Ag | 形语素 | 形容词性语素。形容词代码为 a，语素代码ｇ前面置以 A。 |
| a | 形容词 | 取英语形容词 djective 的第 1 个字母。 |
| ad | 副形词 | 直接作状语的形容词。形容词代码 a 和副词代码 d 并在一起。 |
| an | 名形词 | 具有名词功能的形容词。形容词代码 a 和名词代码 n 并在一起。 |
| Bg | 区别语素 |  |
| b | 区别词 | 取汉字“别”的声母。 |
| c | 连词 | 取英语连词 conjunction 的第 1 个字母。 |
| Dg | 副语素 | 副词性语素。副词代码为 d，语素代码ｇ前面置以 D。 |
| d | 副词 | 取 adverb 的第 2 个字母，因其第 1 个字母已用于形容词。 |
| dc | 程度副词 |  |
| df | 否定副词 |  |
| e | 叹词 | 取英语叹词 exclamation 的第 1 个字母。 |
| f | 方位词 | 取汉字“方” 的声母。 |
| g | 语素 | 绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。由于实际标注时，一定标注其子类，所以从来没有用到过 g。 |
| h | 前接成分 | 取英语 head 的第 1 个字母。 |
| i | 成语 | 取英语成语 idiom 的第 1 个字母。 |
| ia | 形容词功能成语 |  |
| ib | 区别词功能成语 |  |
| id | 副词功能成语 |  |
| in | 名词功能成语 |  |
| iv | 动词功能成语 |  |
| j | 简称略语 | 取汉字“简”的声母。 |
| ja | 形容词功能简称 |  |
| jb | 区别词功能简称 |  |
| jd | 副词功能简称 |  |
| jn | 名词功能简称 |  |
| jv | 动词功能简称 |  |
| k | 后接成分 |  |
| l | 习用语 | 习用语尚未成为成语，有点“临时性”，取“临”的声母。 |
| la | 形容词功能习用语 |  |
| lb | 区别词功能成语 |  |
| ld | 副词功能习用语 |  |
| ln | 名词功能习用语 |  |
| lv | 动词功能习用语 |  |
| m | 数词 | 取英语 numeral 的第 3 个字母，n，u 已有他用。 |
| mq | 数量词 | 在语法信息词典中归入数词库的数量短语。 |
| Ng | 名语素 | 名词性语素。名词代码为 n，语素代码ｇ前面置以 N。 |
| n | 名词 | 取英语名词 noun 的第 1 个字母。 |
| nr | 人名 | 名词代码 n 和“人(ren)”的声母并在一起。 |
| nrf | 姓 |  |
| nrg | 名 |  |
| ns | 地名 | 名词代码 n 和处所词代码 s 并在一起。 |
| nt | 机构团体 | “团”的声母为 t，名词代码 n 和 t 并在一起。 |
| nx | 非汉字串 |  |
| nz | 其他专名 | “专”的声母的第 1 个字母为 z，名词代码 n 和 z 并在一起。 |
| o | 拟声词 | 取英语拟声词 onomatopoeia 的第 1 个字母。 |
| p | 介词 | 取英语介词 prepositional 的第 1 个字母。 |
| Qg | 量语素 |  |
| q | 量词 | 取英语 quantity 的第 1 个字母。 |
| qb | 不定量词 |  |
| qc | 成形量词 |  |
| qd | 度量词 |  |
| qe | 个体量词 |  |
| qj | 集体量词 |  |
| ql | 倍率量词 |  |
| qr | 容器量词 |  |
| qt | 时量词 |  |
| qv | 动量词 |  |
| qz | 种类量词 |  |
| Rg | 代语素 |  |
| r | 代词 | 取英语代词 pronoun 的第 2 个字母 因 p 已用于介词。 |
| rr | 人称代词 |  |
| ry | 疑问代词 |  |
| ryw | 谓词性疑问代词 |  |
| rz | 指示代词 |  |
| rzw | 谓词性指示代词 |  |
| s | 处所词 | 取英语 space 的第 1 个字母。 |
| Tg | 时语素 | 时间词性语素。时间词代码为 t 在语素的代码 g 前面置以 T。 |
| t | 时间词 | 取英语 time 的第 1 个字母。 |
| tt | 专名时间词 | 用于标注中国历史朝代的时间词。 |
| u | 助词 | 取英语助词 auxiliary 的第 2 个字母 因 a 已用于形容词。 |
| ud | 助词“的” |  |
| ue | 助词“得” |  |
| ui | 助词“地” |  |
| ul | 助词“了” |  |
| uo | 助词“过” |  |
| us | 助词“所” |  |
| uz | 助词“着” |  |
| Vg | 动语素 | 动词性语素。动词代码为 v。在语素的代码 g 前面置以 V。 |
| v | 动词 | 取英语动词 verb 的第一个字母。 |
| vd | 副动词 | 直接作状语的动词。动词和副词的代码并在一起。 |
| vi | 不及物动词 |  |
| vl | 联系动词 |  |
| vn | 名动词 | 指具有名词功能的动词。动词和名词的代码并在一起。 |
| vq | 趋向动词 |  |
| vu | 助动词 |  |
| vx | 形式动词 |  |
| w | 标点符号 |  |
| wd | 逗号 |  |
| wf | 分号 |  |
| wj | 句号 |  |
| wk | 括号 |  |
| wky | 右括号 |  |
| wkz | 左括号 |  |
| wm | 冒号 |  |
| wp | 破折号 |  |
| ws | 省略号 |  |
| wt | 叹号 |  |
| wu | 顿号 |  |
| ww | 问号 |  |
| wy | 引号 |  |
| wyy | 右引号 |  |
| wyz | 左引号 |  |
| x | 非语素字 | 非语素字只是一个符号，字母 x 通常用于代表未知数、符号。 |
| y | 语气词 | 取汉字“语”的声母。 |
| z | 状态词 | 取汉字“状”的声母的前一个字母。 |
| Mg |  |  |
| vvn |  |  |
| na |  |  |
| Yg |  |  |

Appendix B. 附加属性表

|  |  |
| --- | --- |
| 标签 | 含义 |
| 1b | 类别(比如短裤、卫衣) |
| ys | 颜色 |
| fg | 风格 |
| ml | 面料 |
| jj | 季节 |
| xj | 细节(领口形状、裤腰高低、是否加绒等) |
| ch | 场合 |
| pp | 品牌 |
| td | 态度 |

1. 这里的公式表示曾存在歧义，这里使用Wikipedia[5] 中提供的写法 [↑](#footnote-ref-1)