汉语分词系统

陈啸、余羿辰

哈尔滨工业大学

2919730935@qq.com
2446172093@qq.com

摘要. 本次实验目的是对汉语自动分词技术有一个全面的了解，包括从词典的建立、分词算法的实现、分词结果性能评价和优化等环节。本次实验用到了以下知识：基本编程能力，如文件处理和数据统计能力、相关的查找算法和数据结构实现能力，如双 trie 树和 Hash 列表、语料库的相关知识、正反向最大匹配分词算法、分词性能评价的常用指标，如准确率和召回率、N 元语言模型的相关知识、马尔可夫模型和隐马尔可夫模型的相关知识。

1 绪论

汉语中词是最小可独立活动的有意义的语言成分，汉语以字为单位，有别于西方语言，词与词之间没有空格之类的标志指示词的边界，因此需要汉语言分词系统。而同时分词问题为中文文本处理的基础性工作，分词好坏对后面的中文信息处理起关键作用。

目前中文分词存在以下难点：分词规范、词定义不明确、歧义切分、交集型切分问题和多义组合型切分歧义等；同时未登录词识别问题也对中文分词提出了更高的要求。对于大规模真实文本来说，已有的词表中没有收录的词，和已有训练语料中未曾出现过的词对分词的精度的影响远超歧义切分。

2 相关工作

中文分词指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

现有的分词算法可分为三大类：基于字符串匹配、基于理解和基于统计的分词方法。

基于字符串匹配的分词方法：又叫做机械分词方法，它是按照一定策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行配，若在词典中找到某个字符串，则匹配成功（识别出一个词）：目前存在正向最大匹配法、逆向最大匹配法、最少切分和双向最大匹配法分词方法。

基于统计的分词方法：给出大量已经分词的文本，利用统计机器学习模型学习词语切分的规律（称为训练），从而实现对未知文本的切分。例如最大概率分词方法和最大熵分词方法等。随着大规模语料库的建立，统计机器学习方法的研究和发展，基于统计的中文分词方法渐渐成为了主流方法。主要统计模型：N 元文法模型，隐马尔可夫模型，最大熵模型，条件随机场模型等。

实验实现了所有功能，具体内容如下：

a. 词典的构建
b. 正反向最大匹配分词的实现和效果分析
c. 基于机械匹配的分词系统的速度优化
d. 基于 MM 的一元、二元文法分词
e. 基于 HMM 的未登录词识别

3 实验内容及过程

3.1 词典的构建[陈啸/余羿辰]

词是词典的基本单位，但是不是全部的词都可以加入词典。基于这样的一个事实：分词词典的构建是要服务于分词任务的，或提高算法运行速度、或提高分词的准确率，在词典的构建中采用了这样几条策略：

• 选择性，通过进一步实验测试，选择将人名加入词典、不将未出现过的单字词加入词典，不将重复的词加入词典。

• 易读性，词典是 GB2312 格式的，字符串形式的，关键信息按行分布的。这样做可以方便实验第 2 部分的最少代码量实现，同时可以便于通过观察词典中的词结构实时修改词典中词的放置策略，以提高算法性能。

• 有序性，词典的内容是高度排序的，通过简单的词典中词的排序，可以进一步实现复杂度更低的查找匹配算法，便于实验第 4 部分的进一步处理。

• 有效性，离线词典中的词全部来源于标准训练集的一部分，同时词典中的词是唯一的，这样可以极大的减小词典的在线数据结构的空间复杂度。

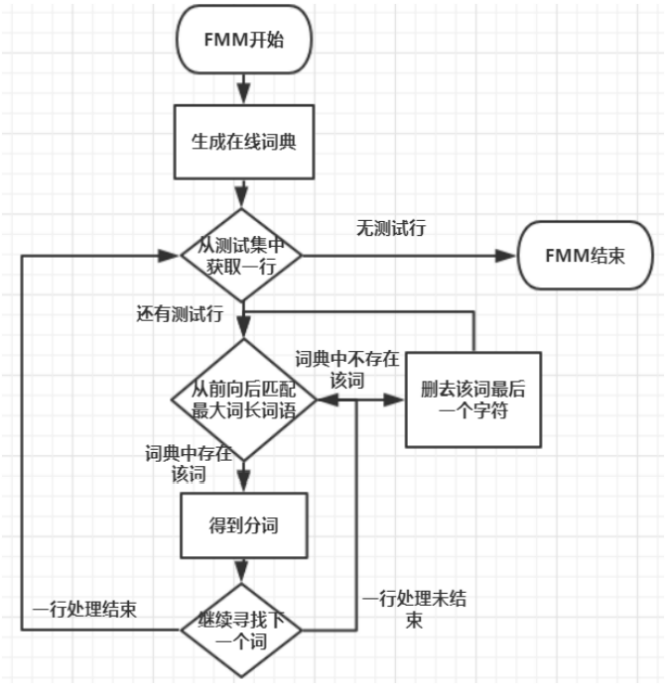
注：因为本实验两位同学确实是共同完成的，所以 3.1-3.4 节两位同学实现方式确实是一致的，所以不重复叙述了。

3.2 正反向最大匹配分词[陈啸/余羿辰]

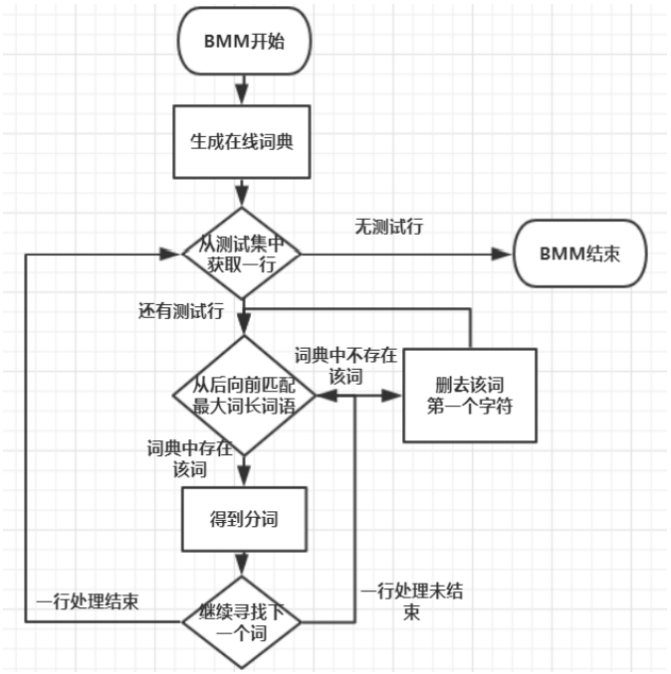
正反向分词的最大匹配实现是基于一个基本思想：找到一句文本中最大词长以下的尽可能长的词序列，即长词优先原则，匹配结果就是分词结果。

• FMM 代码逻辑，通过读取词典文件生成在线词典的数据结构，开始分词；将测试文本的每行作为循环内容，不断将该行减去一个得到的分词，得

到分词的条件是该词在词典中存在且该词后面没有字可以和该词组合形成一个新的词典中的词，直到该行完全分词，继续下一行。



• BMM 代码逻辑，与 FMM 不同的是，BMM 采用从一行测试文本的末尾开始分词，找到最长的在词典中的词，直到该行分词结束。



3.3 正反向最大匹配分词效果分析[陈啸/余羿辰]

该部分代码的基础是分词效果评价公式，通过逐行比较标准答案文本和分词文本得到分词效果的评价。主要是统计标准答案的总词数、分词的总词数以及正确分词数。

分词性能分析：

• 分词性能差异，从不同规模和不同方式生成的的随机训练集和对应互补测试集看 FMM 和 BMM 的分词效果存在一定的性能差异的，且均是 BMM 效果较好。

• 分词精度差异，汉语独特的构词特点使得逆向最大匹配分词的性能更优。直观来看，对于汉语的一个词语来说，在词后加一个字仍能构成新词概率较高，而在一个词前加上一个字还可构成新词概率较低，这就使得逆向最大匹配得到的最长词为准确分词结果的可能性较高。例如，“当时间的脚步”，正向最大匹配会因为在寻找最长词的时候失去本意选择“当时”作为第一个词，划分错误，而逆向最大匹配可正常分词。

• 汉语特点引起的机械分词差异，汉语的构词法和侧重词语偏句后的特点决定了一般 BMM 分词效果较好。无论是 FMM 还是 BMM 都是对已知词语的一个判断。但是在进行已知词语判断的过程中会出现“伪最长词”的情况，即匹配的一个失去原意的最长词会导致后面的词语因为匹配不到对应词语而出

现单字或者说出现错误词。但汉语一句话一般侧重的词语在句子的后面（有别于英语），如“教育部门”落脚点在“部门”，而“教育”是形容该机构的一个属性，“教育部”是由“教育部门”构词形成的相应机构。汉语的构词法使得两个词语可以有相同的前缀且表意相近，从而使得正向匹配分词可能落入这样构词特点的“陷阱”。

3.4 基于机械匹配分词系统的速度优化[陈啸/余羿辰]

实验第 4 部分，代码实现在两个方面超过了实验的要求：一、不仅对 FMM 进行了分词性能优化也对 BMM 进行了相应的性能优化。二、分词时间的优化结果远超底线要求，时间优化了至少 300 倍。

传统机械匹配分词天然缺点是匹配次数过多。对于一个 26 最长词的分词系统，得到一个分词平均需匹配 23 次（假设在一个待分词系统中平均词长为 3），这极大地提高了时间复杂度。而这里由于单词按照字典序具有有序性这一性质，因此可以使用二分搜索来解决这一问题。按照组成词组的单词数量来进行分类，对于单词表中的每一个词，我们将其放到其长度对应的数组的末尾。由于单词表已经有序，所以按先后顺序放入的词组也是有序的。对于每次查询，我们按照词组长度由小到大的顺序，在数组中进行二分查找，直到查询到该词组，或发现该词组不存在。由于单词表的性质，该结构的插入复杂度为 O(1)，查询复杂度为 O(LlogN)。

进一步优化分词速度的方式：用 Trie 树储存节点，内部子节点使用 Hash 思想对 List 进行操作。分词速度优化的关键是减少查找词是否在词典中的搜索次数，减少最长匹配实现过程中的匹配次数。Trie 前缀和后缀树分别减少词前缀和后缀匹配的次数、Hash 思想减少词查找次数。

3.5 基于统计语言模型分词系统实现[陈啸]

该部分实现了所有的功能，包括基于 MM 的二元文法分词、基于 HMM 的未登录词识别。

一阶马尔可夫链即二元文法的假设是一个词语的出现仅依赖于前一个词语。一元文法的假设是一个词出现的概率与其余词无关，这显然是一个过强假设，一元文法对上下文的信息利用是十分不足的。二元文法的假设相对弱化了一元文法的强假设，使得该模型从原理上讲更加的合理化。同时二元文法可以利用更多上下文信息，从而进一步提高性能。

- 二元文法算法思想，首先读取离线词典建立在线词典结构数据结构，保存所有词对应的前缀词及其对应的词频；根据建立的数据结构，不断处理测试文本的每一行，计算得到此行的 DAG，随即对该行进行最大概率分词，计算概率最大路径。值得注意的是计算概率最大路径的时候需要综合考虑一

个词的在某处出现的概率最大仍然需要考虑该词的前词情况，因为二元文法的两个相邻词是条件相关的。

• 条件相关最大概率求解，二元文法在条件相关的最大概率路径求解上与一元文法有所区别。二元文法从前向后依据前词及其对应的组合概率生成每一个字可能的最大概率分词结果，同时需要保存该词对应的前词及其概率结果。不断向后对每个字组成的词进行填表规划，最终生成最大概率结果。

• 0 概率平滑处理。采用对数概率的处理方式避免浮点数下溢，概率结果计算方式变为相应的加法操作，对于可能出现条件概率出现 0 概率的情况，特意对全部词的条件概率求解采用加 1 平滑处理，分母整体加上总词数，分子词频加 1。

HMM 算法对于未登录词的识别具有一定的特点和优势。HMM 算法通过对训练集文本的状态‘BMES’进行标注和统计得到 HMM 模型参数 λ=(Π,A,B)三元组，并根据训练得到的参数利用 Viterbi 算法对测试集的每一行进行寻找最有可能产生观测事件序列的维特比路径——隐含状态序列，如下图公式所示：

$$\delta_t(i)=\max_{i_1,i_2,\cdots,i_{t-1}} P(i_t=i,i_{t-1},\cdots,i_1,o_t,\cdots,o_1\mid\lambda),\quad i=1,2,\cdots,N$$

$$\begin{aligned}\delta_{t+1}(i)&=\max_{i_1,i_2,\cdots,i_t} P(i_{t+1}=i,i_t,\cdots,i_1,o_{t+1},\cdots,o_1\mid\lambda)\\&=\max_{1\le j\le N}[\delta_t(j)a_{ji}]b_i(o_{t+1}),\qquad i=1,2,\cdots,N;\ t=1,2,\cdots,T-1\end{aligned}$$

• 生成在线词典结构，HMM 模型使用二元文法的词典生成在线数据结构集合，用于判断一个词是否在词典中。

• 算法设计思想，从参数文本中读取 HMM 模型训练的参数 Π：初始状态概率，A：状态转移概率和 B：发射概率；不断处理每一行，然后利用 Viterbi 算法输出该行测试文本最大可能的分词隐藏状态；然后对标注结果按照 B、E 和 S 界定一个词的始末位置，从而对整个文本进行状态还原，得到分词结果。

二元文法对于未登录词的处理是十分欠缺的，而 HMM 模型则可以很好地解决这个问题。可在二元文法进行搜索分词空间的最大概率路径时，利用 HMM 思想识别出未登录词，并将其作为分词空间的一个新的分词路径，适当增加其权重值；最后对新的解空间的路径进行统一的最大化概率求解，得到含有未登录词的新的分词路径。

4 实验结论[余羿辰]

实验通过简短的代码实现了复杂度很大的机械匹配分词，后续算法对查找结构和存储结构进行调整和优化通过减少查找次数和减少匹配次数实现了时间性

能上 300 倍的优化；相应的评价代码也为整个实验的分词系统性能表现提供了数据支持和比较。

同时不同于机械匹配分词的思想，统计语言模型分词系统的实现从概率上对汉语言分词提供了新的解决方案。

未登录词的问题仍然无法通过简单的 MM 算法解决，利用 HMM 的未登录词识别的思想对 MM 模型生成的分词解空间进行扩充可以得到很好的算法性能。

5 参考文献

1. L.Rabiner, B.Juang. An introduction to hidden Markov models. *10.1109/MASSP.1986.1165342*.
2. LEGGETTER. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. 10.1006/csla.1995.0010.
3. Crouse, M.S，Nowak, R.D. Wavelet-based statistical signal processing using hidden Markov models. 10.1109/78.668544.