中文语句分词的研究与实践

分别基于最大匹配法和隐马尔可夫模型

16337327-郑映雪　16337341-朱志儒　16337347-邹元昊

July 2019

摘要

不同于英文分词基于空格的简便性，中文语句的分词原理非常复杂不便。中文分词在生活中有广泛的应用，如搜索引擎、提取关键词等。因此，研究如何更好地实现中文分词具有重要的意义。**本文对中文语句分词的方法进行了探究，小组成员使用两种方法——匹配法和隐马尔可夫模型进行了中文分词实践。**

1　引言

1.1　中文分词简介

在英文文本中，空格为显著的分词标志。中文语句则由连续的一串汉字组成，词与词之间常常没有显著的分词标志，但是在个人理解中，中文语句中最有意义的语言成分是单个或多个汉字代表的词语。在中文信息处理中，中文分词操作是非常重要的一个组成部分，主要被应用在自动检索、分类，文本的校对，机器翻译等方面。

随着网络上中文网页和信息的增加，中文分词的急迫需求催生了国内一批优秀的中文分词工具，如结巴分词、百度 NLP、阿里云 NLP 等。

1.2　常用方法

首先说明中文分词需要面临的问题 [1]：

(1) 切分歧义。定义汉字串 AJB 被称作交集型切分歧义，如果满足 AJ、JB 同时为词（A、J、B 分别为汉字串）。此时汉字串 J 被称作交集串。例如，在“二者结合成一个完整的整体”一句中，“结合”和“合成”都为一个自然意义上的词语，此时就会出现切分歧义。

(2) 未登录词。大致包含两大类：主要包含新涌现的词语等以及专有名词。未登录词有可能不能预期的特点。

切分歧义的解决要从检测和消除入手。检测采用最大匹配法，也是我们进行实践的一种方法。最大匹配法又分为正向和逆向，具体的算法会在算法部分阐述。消除的方法实践历经了很长的时间，前后提出了语素、音节、语句重组等方法来消除切分歧义。随着时间的推移，解决切分歧义的方法越来越多，如 LSTM 训练等。

未登录词的处理基于一个较大规模的语料库，根据无监督学习的方法生成新的词表，再由人工筛选加入原有词表中。除此之外，还可以根据词语与整个句子的关联程度判断置信程度，从概率的角度来解决。

由于小组能力有限，我们仅仅就“切分歧义”这个主要分词问题入手进行探究。

1.3 解决思路

我们准备用两种方法解决切分歧义的问题。

首先是最大匹配法，由于正向的最大匹配法准确率较低，所以我们考虑使用逆向和双向的最大匹配法。

为了追求更多的分词方法，我们使用了隐马尔科夫模型（HMM），采用有监督学习来训练一个分词机。

2 相关工作

上文提及，中文分词主要涉及的问题有两个——切分歧义和未登录词。在解决这两个问题方面，我国的科学研究人员已经有了很大的进展。

在切分歧义方面，主要有了以下的解决办法：

早在 1997 年，清华大学的黄昌宁便提出，制定一份汉语通用词表并使其配合于分词规范是十分必要的 [2]。自 2003 年 7 月开展的首届国际中文分词活动 Bakeoff [3] 以来，切分歧义的进展十分迅速。除了使用最大匹配和隐马尔可夫模型，近些年来更有了一些优化的解决办法。如使用 LSTM 网络采用不同词位标注集，并加入预先训练的字嵌入向量 [6]；哈尔滨工业大学信息检索研究所提出了基于感知器进行中文分词增量训练的方法 [7]；同济大学中德学院提出了使用 MMSEG 算法的中文分词技术 [8] 等。

在处理未登录词方面，主要有了以下的解决办法：

中文分词研究初期，处理未登录词主要是基于决策树、动态规划等基本方法。近些年来，处理未登录词也有了很多新的成果：首先，可以使用 PMIk 算法（互信息的改进算法）从大规模语料中自动识别 2 n 元网络新词 [4]；可以使用卡方统计量以及边界熵来强化字标注分词方法以处理未登录词 [5]；可以基于汉语词典对未登录词的语义进行预测 [9]；可以将未登录词编成向量输入神经网络，同时可以探究未登录词对文本整体意思的影响等 [10]。

3 问题定义

本次实验需要就解决的是中文分词中最主要的问题——歧义切分。如：

“提高人民生活水平”这一语句在自然理解上的分句法为：提高、人民、生活、水平。但是就词汇匹配来说，还可以分为：提、高人、民生、活水、平。

因此，算法的输入为一个未分隔的完整的中文语句，输出为该语句分词的结果组成的列表。如：

输入：提高人民生活水平

输出： [提高，人民，生活，水平]

4 算法描述

4.1 使用最大匹配法（Maximum Matching）

最大匹配法的算法原理为，采取字典中最大长度的单个词语作为切分长度，并且使用该长度切割语句，再依据某一个方向在字典中寻找匹配词语，如未寻找到，则去掉最后一个字符继续匹配，直到不可删除为止。最大匹配法包括正向最大匹配法、逆向最大匹配法和双向最大匹配法，实验证明，正向最大匹配法的正确率不如逆向最大匹配法和双向最大匹配法[11]，所以在本次实验中，我们直接跳过了正向最大匹配法的实现，直接选择实现逆向最大匹配法和双向最大匹配法。

4.1.1 逆向最大匹配法（BMM）

逆向最大匹配法依据从右往左的方向应用最大匹配的思想，将字符与字典进行匹配判定，若匹配成功，则将这个字符切分出来，作为一个词的切分结果；若匹配不成功，则将这个字符的最左边的一个字去掉，再进行匹配。

如“南京市长江大桥”：

1. 假设词典中最大元素长度为 5，则取出后 5 个字“市长江大桥”，发现字典中无匹配元素。

2. 将切割元素的最左元素“市”去掉，发现词典中有匹配元素，此时进行切割；

3. 对剩余部分进行上述动作，得到的结果为：南京市、长江大桥。

Algorithm 1 阐述了这个方法在本实验中的应用。

4.1.2 双向最大匹配法（TMM）

双向最大匹配法是综合了正向和逆向匹配，将二者得到的结果进行比较来得到正确的分词结果。研究表明，中文中 90%左右的句子，用两种方法都能得到相同的结果；9%的句子中，两种方法必有一个是正确的；剩下的不到 1.0%的句子，两种方法都是得到错误的答案[12]。

如“南京市长江大桥”，正向和逆向就是不相同的，但是逆向的结果是正确的。**Algorithm 2** 阐述了这个方法在本实验中的应用。

Algorithm 1 逆向最大匹配法

Input: 训练集已切分好的句子；待分隔的句子；

Output: 句子的切分结果；

1: 根据训练集中已切分的句子中的词语构建词典，并得到词典中最长词条的长度 m；
2: 从右到左将待切分句子的 m 个字符作为匹配字符；
3: 将匹配字符与词典中的元素进行匹配：
    如果匹配成功，则将该匹配字符作为一个词从句子中切分出来；
    如果匹配失败，则去掉该匹配字符的第一个字再进行匹配；
    如果匹配字符只有一个字，则将该字作为一个词从句子中切分出来;
4: 重复上述过程，直到句子整个被切分为词语为止，然后选择分词数较少的结果。

Algorithm 2 双向最大匹配法

Input: 训练集已切分好的句子；待分隔的句子；

Output: 句子的切分结果；

1: 根据训练集中已切分的句子中的词语构建词典，并得到词典中最长词条的长度 m；
2: 从左到右将待切分句子的 m 个字符作为匹配字符；
3: 将匹配字符与词典中的元素进行匹配：
    如果匹配成功，则将该匹配字符作为一个词从句子中切分出来；
    如果匹配失败，则去掉该匹配字符的第一个字再进行匹配；
    如果匹配字符只有一个字，则将该字作为一个词从句子中切分出来;
4: 重复上述过程，直到句子整个被切分为词语，从而得到正向切分的结果;
5: 进行逆向最大匹配算法，得到逆向切分的结果，比较两者的结果，从而得到双向匹配的分词结果。

4.2 使用隐马尔可夫模型（Hidden Markov Model）

4.2.1 模型简介

隐马尔可夫模型 (HMM) 是一种概率机器学习过程, 是一个强有力的模型。HMM 为一个二重的马尔可夫随机过程，它具有状态转移概率的马尔可夫链和输出的观测值。HMM 包含一个可观察层和一个隐藏层。其中隐藏层是一个有限状态机（马尔可夫过程），每个状态转移具有概率 [13]。

HMM 由一个五元组 $\lambda = (S, O, A, B, \Pi)$, 简单记为 $\lambda = (A, B, \pi)$, 其中 $S = (s_1, s_2, ..., s_n)$ 为状态序列，$O = (v_1, v_2, ...v_m)$ 观察值序列。其余分别为：

$A = \{a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \leq i, j \leq N\}, a_{ij} \geq 0, \sum_{j=1}^{N} a_{ij} = 1;$
$B = \{b_j(k) = P(o_t = v_k | q_t = s_j), 1 \leq i \leq N, 1 \leq k \leq M\} b_j(k) \geq 0, \sum_{k=1}^{M} b_j(k) = 1$;
$\Pi = \{\pi_i = P(q_1 = s_i), 1 \leq i \leq N\} \pi_i \geq 0, \sum_{i=1}^{N} \pi_i = 1$ [14]。

那么，什么问题可以用 HMM 解决呢? 答案是，有一类随机现象，在已知情况的条件下，未来的时刻只与现在有关，而与过去无关，这时候就可以应用 HMM 来预测问题。

4.2.2 模型应用

在中文分词中，我们就可观察的序列寻找一个最可能的隐藏状态序列，正好可以通过建立 HMM 模型来实践。我们将输入语句编成一个状态序列，以句子的每一个字属于哪一个状态为概率，然后得到初始状态概率，随后建立状态转移概率矩阵和观测概率矩阵。在分词中，我们将状态集设置为 $\{B, E, M, S\}$，其中 B 代表单词第一个汉字,E 代表单词最后一个汉字,M 代表单词中间的汉字,S 代表单个汉字组成的单词。

在假设 s_i 只能由 s_{i-1} 决定、v_i 只能由 v_{i-1} 决定时，可以利用 Viterbi 算法 [15] 找出目标概率最大值。

我们之前知道，根据动态规划原理，两个节点的最优路径，节点之间可能的部分路径也一定是最优的。使用 Viterbi 算法进行如下操作：从时刻 t=1 开始，递推地计算在时刻 t 时、状态为 i 时的各条部分路径的最大概率，直至得到最终结果。之后，再由终点开始，由后向前逐步求得各个结点，最终得到最优路径。**Algorithm 3** 阐述了本实验中的具体应用。

Algorithm 3 使用 HMM 进行中文分词

Input: 训练集已切分好的句子；待分隔的句子；

Output: 句子的切分结果；

1: 通过句子的第一个字属于 $\{B, E, M, S\}$ 这四种状态的概率生成初始状态概率表；
2: 通过计算每个状态之后出现不同状态的次数构建状态转移矩阵 A；
3: 通过计算在每个状态时汉字出现的次数构建观测矩阵 B；
4: 通过 Viterbi 算法计算隐含状态序列得到每个汉字的隐含状态 $\{B, E, M, S\}$
$\delta_{t+1}(i) = \max\ [\delta_{t(j)} \cdot a_{ji}] b_i(v_{t+1})$；
5: 通过动态规划可以得到最有可能产生观测时间的路径，通过结果反推就可以得到整条状态路径；
6: 通过得到的状态路径划分单词，得到结果。

5 实验

5.1 数据集介绍

本次实验采用的数据集有两组，一组是中文自然语言处理里的经典数据集——北大计算语言学研究所和富士通研究开发中心有限公司共同制作的，以《人民日报》为素材的标注语料库，一组是微软亚洲研究院中文分词语料。两组数据集的训练集和测试集的格式均相同。其中，北大语料库共有训练集语句 19056 条，测试集语句 1944 条；微软语料库共有训练集 86924 条，测试集语句 3985 条。

在实验中，使用最大匹配法时我们将两组数据集合并成为字典；使用 HMM 时我们考虑到不一样的数据集会造成概率的干扰问题，所以只使用了微软的语料库进行训练和测试。**图 1** 和**图 2** 分别为北大和微软两个训练集中单词长度的统计，我们可以看到大部分词语长度在 2-4 个汉字之间，但也存在着少量长度大于 10 的词语（大多是专有名词）。

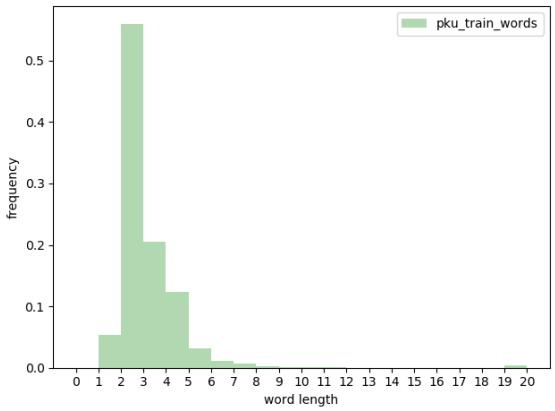

图 1: 北大语料库词语长度统计

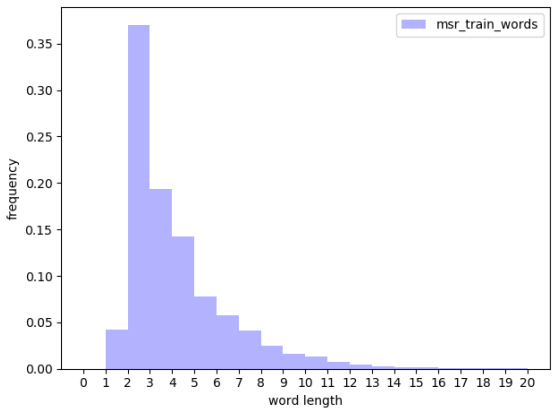

图 2: 微软语料库词语长度统计

5.2 实验结果及分析

我们将测试集的分词结果与数据集自带的测试集答案进行了对比。将分词效果分为完全划分正确和部分词语正确，并将分词效果分成了不同的维度：如 1 即为整个句子分词都正确，0.9 即为百分之 90 的词语分词正确，0.8 即为百分之 80 的词语分词正确等。

图 3 和**图 4** 为两种方法的实验结果数据分布图及具体数据。其中，横坐标为划分正确的词语比例，纵坐标为该划分效果下正确的语句占比。

图 3 为两种最大匹配法的实验结果，从中可以得出：逆向匹配法已经足够挑选出更加符合自然语言习惯的分词结果，但是考虑到存在着一些特殊情况，即正向匹配法的结果更符合常理的情况，所以进行双向匹配法来保证万一。实验结果证明，双向匹配法准确度确实比逆向匹配法要高一点点，这里面应该是包含了正向匹配更加优越的少量特殊情况。整体来看，要把一个长句完全分词正确，准确率并不高，但是只要稍微降低一下分类的要求，准确率就有明显的上升。

图 4 为使用 HMM 的实验结果，从中可以得出：同最大匹配法一样，在考虑完全分词正确的时候，准确率并不高，但是在降低了每句话分类需求的时候，正确率大幅上升。

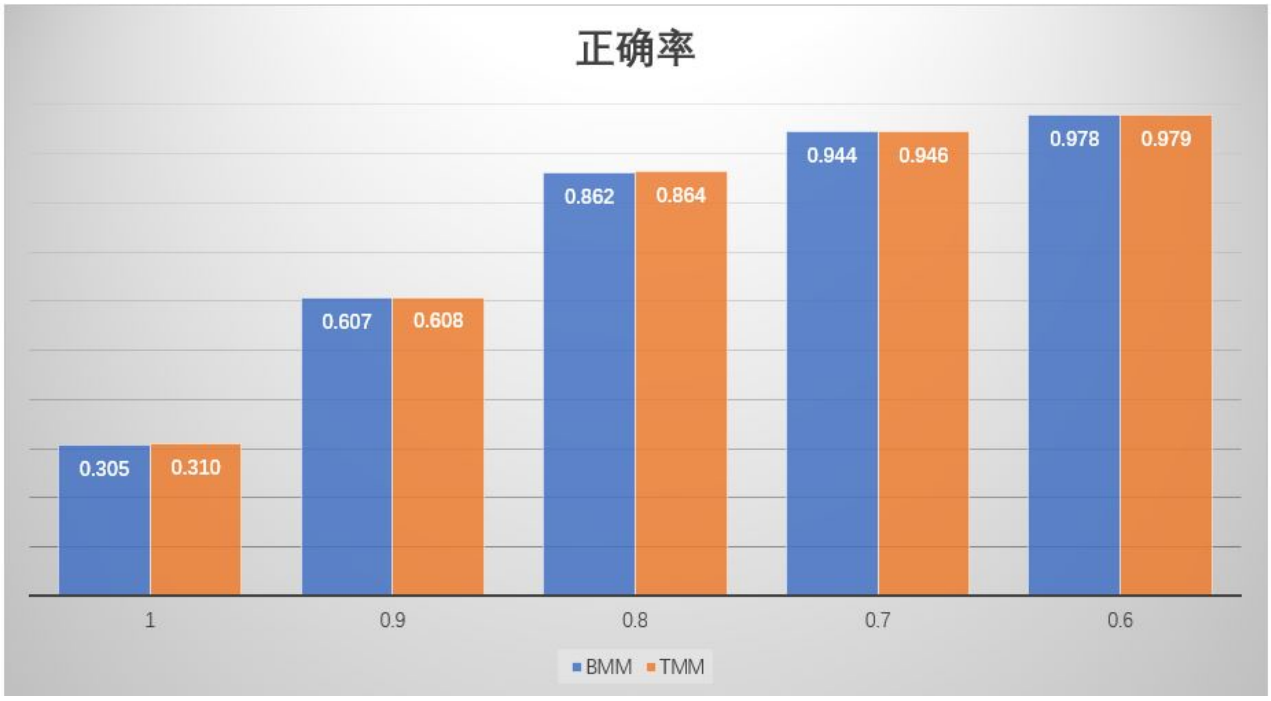

图 3: 最大匹配法的实验结果，其中横坐标为划分正确的词语比例，纵坐标为该划分效果下正确的语句占比

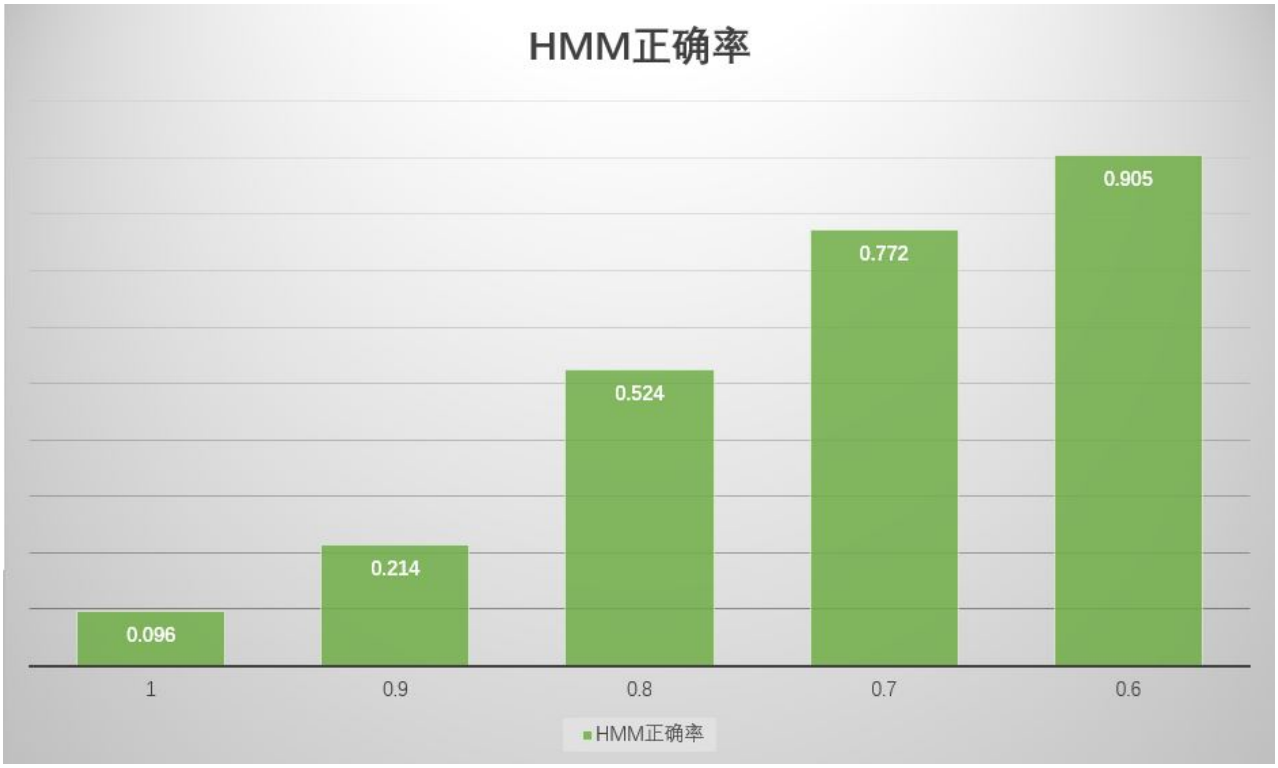

图 4: HMM 的实验结果，其中横坐标为划分正确的词语比例，纵坐标为该划分效果下正确的语句占比

6 结论

本次实验是我们小组三人在中文分词这一与我们生活息息相关的项目的初探。考虑到老师的课堂上花了很久的时间给我们讲解隐马尔可夫模型，因此我们怀着探索新的知识和实践课堂所学的初衷，尝试了两种最主流的方法来实现中文分词。其实从实验结果就可以看出，我们的效果并没有市面上主流分词产品好，这说明我们还存在着不足，比如对匹配法地词典进行结构优化，对 HMM 模型进行优化等，以后还有一些尝试等着我们去完成。

总之，这次大作业给了我们一个自由探索的机会，以前我们从来没有做过中文的自然语言处理，这次实践过后更是体会到了中华文化的博大精深导致中文分词探索的不易。在了解相关工作时，我们在查找论文时感受到从本世纪初开始，一代代研究人员便在不断探索中文分词的工作。在实践上，我们没有好高骛远，而是选择对老师重点讲解的算法，结合与生活相关的项目进行了尝试，实践更是加深了印象，受益匪浅!

参考文献

[1] 孙茂松, 邹嘉彦. (2001). 汉语自动分词研究评述 (Doctoral dissertation).

[2] 黄昌宁. ” 中文信息处理中的分词问题.” 语言文字应用 1 (1997): 74-80.

[3] Sproat, Richard, and Thomas Emerson. "The first international Chinese word segmentation bakeoff." Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003.

[4] 杜丽萍, et al. " 基于互信息改进算法的新词发现对中文分词系统改进." 北京大学学报(自然科学版) 52.1 (2016): 35-40.

[5] 韩冬煦, 常宝宝. 中文分词模型的领域适应性方法. 计算机学报. 2015;38(2):272-81.

[6] 任智慧, et al. " 基于 LSTM 网络的序列标注中文分词法." 计算机应用研究 34.5 (2017): 1321-1324.

[7] 韩冰, et al. " 基于感知器的中文分词增量训练方法研究." 中文信息学报 29.5 (2015): 49-54.

[8] 张中耀, et al. " 基于 MMSEG 算法的中文分词技术的研究与设计." 信息技术 40.6 (2016): 17-20.

[9] 尚芬芬, et al. " 基于《现代汉语语义词典》的未登录词语义预测研究." 北京大学学报 (自然科学版) 1 (2016): 10-16.

[10] Luo, Shang-Bao, Ching-Hsien Lee, and Kuan-Yu Chen. " 未登[illegible]詞之向量表示法模型於中文機器[illegible]讀理解之應用 (An OOV Word Embedding Framework for Chinese Machine Reading Comprehension)[In Chinese]." Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018). 2018.

[11] 丁振国, 张卓, and 黎靖. " 基于 Hash 结构的逆向最大匹配分词算法的改进." 计算机工程与设计 29.12 (2008): 3208-3211.

[12] Sun,M.S.and Benjamin K. T. 1995. Ambiguity resolution inChinese word segmentation. Proceedings of the 10th AsiaConference on Language,Information and Computation, 121-126.Hong Kong.

[13] Information extraction with HMMstructures learned by stochastic optimization. Freitag D,McCallum A. Proceedingsof the Eighteenth Conference on Artificial Intelligence . 2000

[14] 韩普, and 姜杰. "HMM 在自然语言处理领域中的应用研究." 计算机技术与发展 2 (2010): 245-248.

[15] Forney, G. David. "The viterbi algorithm." Proceedings of the IEEE 61.3 (1973): 268-278.