Pinyin Input Method Editor
Design Report

Yihong Gu
gyh15@mails.tsinghua.edu.cn
Department of Computer Science
Tsinghua University

1 Introction

报告分为三个部分：

- Language Model: 介绍所用的语言模型
- Search Algorithm: 介绍所用的搜索算法以及优化
- Experiments: 给出实验结果并作相关分析
- Conclusion and Furthor Work: 总结和指出可以改进的地方

2 Language Model

2.1 Probability Model

总体来说，我们使用以下语言模型：

$$\mathbb{P}(w_1 \cdots w_n) = \prod_{i=1}^{\min(n,m)} \mathbb{P}(w_i | w_{\max(i-m+1,1)} \cdots w_{i-1}) \tag{1}$$

我们把这个模型称为 m-gram 模型。
在这里面 w_i 表示第 i 个汉字，举个例子，取 $m = 2$：

$$\mathbb{P}(\text{清华大学}) = \mathbb{P}(\text{清})\mathbb{P}(\text{华}|\text{清})\mathbb{P}(\text{大}|\text{华})\mathbb{P}(\text{学}|\text{大}) \tag{2}$$

事实上，这里面我们没有考虑拼音的影响，那么，我们作最简单的假设，假设拼音和 $m-gram$ 独立并且条件分布是离散分布

$$\mathbb{P}(w_1 \cdots w_n | t_1 \cdots t_n) = \prod_{i=1}^{\min(n,m)} \mathbb{P}(w_i | w_{\max(i-m+1,1)} \cdots w_i - 1)\mathbb{P}(w_i | t_i) \tag{3}$$

我们让

$$\mathbb{P}(w_i|w_{i-m+1}\cdots w_{i-1}) = \frac{\#\{w_{i-m+1}\cdots w_m\}}{\#\{w_{i-m+1}\cdots w_{m-1}\}} \tag{4}$$

其中 $\#\{w_{i-m+1}\cdots w_i\}$ 为词组 $w_{i-m+1}\cdots w_i$ 在 corpus 中出现的频数，并且让 $\mathbb{P}(w_i|t_i)$ 为 1 当且仅当汉字 w_i 存在发音 t_i，否则为 0，我们也尝试了其他的模型 (均匀分布，按汉字的词频归一化的离散分布，但是发现实际上这些方法会引入大量噪声，实际效果并没有之前这种简单也不归一化的方法好，因为前一种方法让文本完全由 corpus 决定，不引入拼音造成的噪声)。

2.2 Frequency Count

在计算 $\#\{w_{i-m+1}\cdots w_i\}$ 的过程中，我们使用 sina 新闻 2016 作为 corpus，且把所有的 6763 个汉字作为 w_i 的字母表 Σ，把新闻正文中不属于 Σ 的部分作为分隔符，统计在 Σ 中的连续 m 个 token(中间不能有分隔符) 出现的次数。

由于总的次数过于多，我们考虑只保留部分 m-gram 的频数统计的结果，我们选取最大的 k，使得频数 $\geq k$ 的 m-gram 的频数之和大于总频数之和的 $100\sigma\%$，我们把 σ 称为 significance level，在这里我们取 $\sigma = 0.95$，最后我们保留频数 $\leq k$ 的 m-gram。

2.3 Probability Smoothing

首先，为了方便计算，我们同意使用概率取对数进行计算，这样原来的乘积就变成了求和。

由于词频很多时候都为 0，所以我们需要用对 $\log\mathbb{P}(w_i|w_{i-m+1}\cdots w_{i-1})$ 进行平滑处理。

我们下面考虑具体的处理过程 (递归处理)：

- 如果当前发现 $w_{i-m+1}\cdots w_i$ 和 $w_{i-m+1}\cdots w_{i-1}$ 的频数均非 0，那么就直接按照原式计算 $\log\mathbb{P}(w_i|w_{i-m+1}\cdots w_{i-1})$。
- 如果发现 $w_{i-m+1}\cdots w_{i-1}$ 的频数均为 0，并且 $m > 2$，计算 $m' = m-1$ 的结果 p_{m-1}，然后输出就是 $p_{m-1} - 100$，作为平滑处理的惩罚项。
- 如果发现 $w_{i-m+1}\cdots w_{i-1}$，并且 $m = 2$，计算 $m' = m-1$ 的结果 p_{m-1}，然后输出就是 $p_{m-1} - 2\times 10^8$，作为平滑处理的惩罚项。

另外，由于我们是 (要通过搜索) 需要最大化对数似然值，所以我们设置答案的下界为 -1×10^9，也就是说，像第三项的那种平滑处理不能超过 5 次。

3 Search Algorithm

有了 Langugae Model 后，我们的问题就转变成了最大化

$$w_1^*\cdots w_n^* = \mathrm{argmax}_{w_1\cdots,w_n}\mathbb{P}(w_1\cdots w_n|t_1\cdots t_n) \tag{5}$$

其中 $t_1\cdots t_n$ 是给定的拼音，同时 $w_1^*\cdots w_n^*$ 就是我们输出的结果。

我们考虑使用 A^* 算法来解决这个问题。

3.1 A^* Algorithm

我们把 $w_1\cdots w_i$ 称为一个状态 s_i，当 $i = n$ 的时候即到达终点，一个状态 s_i 的收益为 $v_i = \log\mathbb{P}(w_1\cdots w_i)$，我们需要最大化到达终点的收益 v_n。

服从 A^* 的记号，我们发现 $g(s_i) = v_i$，另外我们让 $h(s_i) = 0$，即可用 A^* 来优化。此时我们发现，这个问题实质上变成了一个最长路径问题，这时候的 A^* 也就等价于传统的 Dijkstra 算法。

3.2 Improvement

我们从以下一个角度来优化这个搜索过程：

SLF 优化

我们发现，是否对 OPEN 表排序 (即使用堆来维护 OPEN 表) 不影响时间消耗，所以我们不对 OPEN 表排序，这样的搜索算法就等价于传统的 SPFA 算法，我们沿用了 SPFA 算法的一个非常经典的优化手法 SLF 优化，即如果放入队尾的状态比放入目前队头的要优的话，把队头队尾的元素交换，这样可以使得效率提升 3 倍。

记忆化

我们发现，计算 local log probability ($\log \mathbb{P}(w_i|w_{i-m+1}\cdots w_{i-1})$) 非常消耗时间，所以我们对这一部分进行记忆化，这样效率也可以提升 1 倍。

4 Experiments

4.1 Toy data set - Sina News

我们随机选取了 sina 新闻 (2017.4.9) 的四篇不同文章的 11 个短语/句子，文章列表如下：

- 政府工作报告 7 次提及李克强为何再赠 4 字？
- 武汉最懒大学生：两周不收衣服鸟儿在内做窝
- 特朗普称叙化武袭击事件是“对人类的羞辱”
- 郎平: 女排备战奥运会培养新人已着眼下个周期

2-gram 的结果如下：

- 振兴实体经济是当前一个重要命题/振兴市体经济适当前一个重要命题
- 这方面的成功事例数不胜数/这方面的成功实力输部省属
- 很明显表达出两层意思/很明显表达出两个意思
- 他每天上完晚自习后要去健身房健身/他每天上万万字西后要去健身房间参
- 男大学生们普遍表示理解/南大学生们普遍表示理解
- 美国总统特朗普在记者会上讲话/美国总统特朗普在记者会上讲话
- 发生在叙利亚的针对无辜平民的化武袭击事件/发生在叙利亚的针对无辜平民的话务系及时间
- 自己将开始独立制订和执行球队的训练计划/自己将开始都理制定和支行求对的续联系化
- 中国队会继续培养新人/中国队会继续培养心人
- 最为引人注目的是中国影片/最为引人瞩目的是中国影片
- 遵照国际惯例和规则/遵照国际管理和规则

最后准确率为 76%，我们发现，他很难刻画一些长词/长句，比如” 国际惯例”、” 数不胜数”、” 是当前”、“上完晚自习”，他会把这些长词变成一些二字词语接龙，比如” 数不胜数” 变成了” 输部”+” 部省”+” 省属”。

3-gram 的结果如下：

- 振兴实体经济是当前一个重要命题/振兴实体经济是当前一个重要命题

- 这方面的成功事例数不胜数/这方面的成功实力数不胜数
- 很明显表达出两层意思/很明显表达出两个疑似
- 他每天上完晚自习后要去健身房健身/他每天上完晚自习后要去健身房间参
- 男大学生们普遍表示理解/南大学生们普遍表示理解
- 美国总统特朗普在记者会上讲话/美国总统特朗普在记者会上讲话
- 发生在叙利亚的针对无辜平民的化武袭击事件/发生在叙利亚的针对无辜平民的化物袭击事件
- 自己将开始独立制订和执行球队的训练计划/子即将开始都理制定和执行求对的续联系华
- 中国队会继续培养新人/中国队会继续培养新人
- 最为引人注目的是中国影片/最为引人注目的是中国影片
- 遵照国际惯例和规则/遵照国际惯例和规则

最后准确率为 86%，我们发现，他已经能够刻画一些四字词语，比如" 国际惯例"、" 数不胜数"、" 是当前"、"上完晚自习"。这是非常值得表扬的。

4-gram 的结果类似 3-gram，最后的准确率为 85%，没有明显的提升。

4.2 Toy data set - math

我们选取了夏道行的《实变函数与泛函分析 < 上 >》中的 10 个短语，作为第二个 toy data set，查看具体的效果，这里我们展示 n-gram 为 3 的效果。

- 虽然已经解决了建立新积分方法的首要问题/虽然已经解决了坚力新计分方法的首要问题
- 建立了较一般集上的测度理论/建立了较一般计上的策都理论
- 后面我们将称具有这种性质的函数为可测函数/[Can't Found Answer]
- 下面引入可测函数的概念/下面引入可测函数的概念
- 可测函数的有限可加性/可测汉书的有限可嘉兴
- 几乎处处收敛函数列的控制收敛定理/[Can't Found Answer]
- 证明积分与极限交换顺序/证明其分与其见交还顺序
- 再举一些控制收敛定理的应用/载具一些空置受联鼎立的影用
- 读者自己也可以列举并加以证明/读者自己也可以列车并加以证明 s
- 所以这两个函数几乎处处相等/所以这两个寒暑期护处处相等

最后准确率为 53%。值得欣喜的是，即使语料库中没有" 可测函数" 这个词语，他最后也能打出来，这与我们的平滑处理中的 2×10^8 的那一项密切相关。同时，我们可以发现，即使我们选取的都是数学书中比较贴近生活用语的句子，他的表现也不是非常好，比如" 列举并加以证明"、"函数几乎"这些分开来说得通的词语合起来却无法打出。

4.3 Overall Test Set Performance

我们考虑在整个测试集上的表现，几个 ngram 的表现分别如下：

- 2-gram:
- 3-gram: 73%
- 4-gram:

4.4 Samples

Well Done Samples

我们具体分析几个例子：

- 对染色体人工合成的工作给予了高度评价

我们查看其 local log probability:

- 对: 14.224
- 染: -200000000.000
- 色: -105.043
- 体: -100.540
- 人: -205.390
- 工: -205.242
- 合: -104.873
- 成: 0.000
- 的: -1.108
- 工: -104.803
- 作: -0.246
- 给: -5.245
- 予: -0.063
- 了: -1.134
- 高: -1.874
- 度: -0.018
- 评: -0.908
- 价: -0.002

我们发现，实际上没有“对染”这个 2-gram，所以我们引入了 smoothing 中的 2×10^8，让他能够断词，但是需要付出巨大代价 (能不断就不断)，同时注意这里的对的 local log likelihood 是正的这一点是为了方便计算，是直接对频数取 log 的结果，他和对频率取 log 之差一个常数，所以忽略了这个常数。

下面也是几个“出乎意料”的比较好的结果：

- 美女与野兽
- 深度神经网络对计算资源的消耗很大
- 北京的房价是否在透支年轻人的创造力
- 人文和工业工程必将会师决赛
- 人与人之间为什么要互相伤害呢

他们的句式都不是偏新闻的句式，但是效果都还不错。

Poor Done Samples

再看几个做得不好的例子：

- 拟 (你) 的世界会变得更精彩
- 请大家选择你觉得可疑 (可以) 的时间
- 读者自己也可以列车 (举) 并加以证明
- 现同期 (先统计) 大量真实与了 (语料) 中各个词出现的概率
- 我从未见过有如此后演舞池 (厚颜无耻) 之人

(1) 这主要是由于"P(世|拟的 > P(世|你的"，因为"你的"这样的太常见了，所以前面是"你的"后面接"世"的概率就会比较小。

(2) 这是由于 corpus 的保留的不合理导致的，实际原因就是"疑的时"保留在 language model 中但是"以的时"没有，实际上这两者都不应该被保留

(3) 列车 (举)：是由于多音字的混乱引起的，拼音为 lie ju，车有 ju 的音，但是在文本中却是列车 (che)，同时列车这个文本比列举出现得更多。

(4) 语料库不丰富导致的问题。

(5) 断词的问题，没有"此厚"和"耻之"。

同时，我们也注意到了这样的模型的不稳定性，看下面的例子：

- 人与人之间为什么要互相上海阿 (伤害啊)
- 人与人之间为什么要互相伤害呢
- 人与人之间为啥要互相伤害呢
- 仁与人之间怎么就不能互相伤害呢
- 人与人之间就是要和向上海阿 (互相伤害啊)

类似的语句有的却输出不是很好的结果。

5 Conclusion and Furthor Work

我们可以发现，这个框架的好坏基本上是由 language model 的好坏来决定的，事实上现在的这个 language model 不是一个很好的模型，关键在于现在的模型是由字决定的，这就会导致之前的一系列的问题。同样，现在的 corpus 的局限性也比较大，我们可以在之后考虑几个改进的地方

- 考虑更广泛的语言模型：可以让拼音起到一定作用解决多音字的影响，把词作为基本单位。
- 考虑使用更加好的 corpus，事实上，wiki 或者 baike 的效果应该会比新闻要好一点 (从英文的 word2vec 可以看出)。