基于双字特征的中文分词方法

首先，统计句子中相邻两个字的联合状态，将其视为一个特征元，一条 N 字的句子构成 N-1 个特征元。然后，利用相邻特征元之间共用一个字的状态，由此将 N-1 个特征构成一个有向图。最后，使用 Viterbi 算法走出一条最优状态路径得到最终的分词结果.针对歧义词和未登入特征元的问题，提出了三种策略：1）结合双字特征的上下文，考虑当前双字特征的前后一个字，分词精度提高了 1.5 个百分点；2）采用熵来衡量一个双字特征的不确定性，进而确定平滑处理程度，分词精度提高了 0.2 个百分点；3）针对训练集的不完备性，提出了备用字典方法和独立性假设，在备用字典完备的情况下，分词精度提高了 2 个百分点.在 MSR 数据集上进行实验，结果表明：该方法训练过程和分词速度快，整个过程只需要几分钟，在开放集上具有 96.05%的 F1 值.

关键字：中文分词，双字特征，特征状态转移图，最优路径，未登入词

中文分词是中文自然语言处理的基础，具有非常重要的理论和应用意义。在过去的几十年里，经过学者们的研究和探索，中文分词准确率得到了提升。特别是在使用了机器学习和基于统计的方法后，中文分词效果有了显著的进步。薛在[]中提出将中文分词看作是序列标注问题，使得分词方法由过去的基于词（或词典） 的逐渐转变为基于字的分词方法。隐马尔科夫模型，最大熵模型，条件随机场模型[]。这些方法都采用一个字的前后几个字的作为当前字的上下文信息。本文的方法则是将二个相邻的字看作一个特征元，不损失双字之间的信息，且把特征元的前后一个字作为当前双字特征元的上下文信息，这样不会损失当前字的上下文信息。经过笔者的统计，N 元字的组合情况不会使得模型复杂度成指数级增长。

1、双字特征

在字构词方法中字的状态分为四种{B,M,E,S}：B（与后一个字构词），E（与前一个字构词），M（与前后字构词），S（单独构词）。双字的合理联合状态有 8 种：BM，BE，MM，ME，EB，ES，SB，SS。双字特征是相邻两个字的联合状态。例如，“研究/生物学”的双字有：“研究”，“究生”，“生物”，“物学”。其双字特征是双字的联合状态特征:“BE”，“EB”，“BM”，“ME”。

自然语言处理的 N 元模型中提过，假设一个 5000 个不同汉字的训练集，直接考虑两个字的组合，则有 5000^2 情况。然而实际情况下相邻两个字共同出现是符合中文语境的，并非每个字后面都有 5000 种可能。在人民日报 170 万字的训练

集上，经过实际统计双字特征共 27 万，三字组合共 130 万，平均下来，每个字后面仅有 50 种可能。这是双字特征方法可计算的重要原因之一。同时在 PKU 测试集上，所有的双字共 17 万个左右，未登入的双字出现了 1.7 万个。这些未登入的双字的联合状态大部分集中在 EB,ES,SB,SS，也就是前后两个字不构词。而在整个状态序列中，这些未登入双字特征能在最优路径上通过已登入双字协调串联起来，增加未登入双字联合状态的可靠性。

2、**特征状态转移图**

根据一条句子相邻两个双字特征共用一个字的相同状态，那么相邻特征元的上一个特征元的后一字状态与下一个特征元的前一字状态相同。可以得到图 1 双字特征状态转移图。

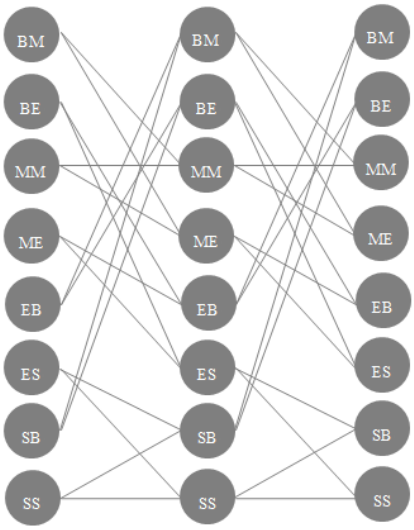


图 1. 双字特征状态转移图

为了用双字特征表示双字的联合状态概率，图中每个节点的权重为双字特征的 8 种联合状态的概率，采用最大似然估计求解概率参数。

$$P(T_i) = P(S_{i-1}, S_i \mid O_{i-1}, O_i) = \frac{(S_{i-1}, S_i)}{\sum_{S_i S_{i-1}} (S_{i-1}, S_i)} \qquad (2.1)$$

其中 S_iS_{i-1} 表示一个双字特征的联合状态，0_i0_{i-1} 表示一个双字。相邻两个特征单元通过相同的字具有相同的状态构成有向边。这样就形成了一条句子的有向图。

3、Viterbi 算法求解最优路径

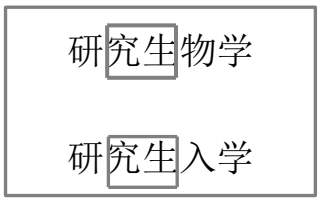
在上面构建的双字特征有向图之后，最大化一条句子状态序列的概率可以用最大化双字特征的乘积来表示。

$$\begin{aligned} &\max \prod_{i=1}^{N-1} P(T_i) \\ &\Leftrightarrow \max \log \sum_{i-1}^{N-1} P(T_i) \\ &\Leftrightarrow \min\left(-\log \sum_{i-1}^{N-1} P(T_i)\right) \end{aligned} \tag{3.1}$$

为了防止连乘下溢，使用对数将连乘转变成累加的形式。将最大化概率路径转化为求解最短路径问题。通过 Viterbi 算法求解最优路径得到最优状态序列后，最后根据状态 E 和 S 为词的结束标志进行分词得到最终的分词结果。

4、歧义词和未登入词处理

4.1 一个双字是句子中的一个局部信息，其联合状态与句子的上下文关系密切，如果忽略句子的上下文进行孤立统计，势必难以解决歧义性问题。



考虑双字特征的上下文信息可以从下面几点出发，$C_1, C_2, C_3, C_4, C_1C_2, C_2C_3, C_3C_4, C_1C_2C_3, C_2C_3C_4, C_1C_2C_3C_4$, 它们均是是 C_2C_3 双字特征的上下文信息，所以在 C_2C_3 双字特征的转移图中的权重有上下文信息确定。试验表明，这可以较好的解决局部歧义性问题。

4.2 借鉴“词表驱动”原则，如包含标点的双字特征，以及像“我们”，“虽然”这样的双字状态具有很强的确定性。在词表驱动原则下，这样的词严格按照词表划分，但按词表划分需要一个词表的维护，同时会降低分词效率。将这一思想推广，采用熵来衡量一个双字特征的不确定性，对于单一状态的双字特征必然熵为 0，这种情况下，严格按照词表驱动原则。同时，结合熵值的大小来确定平滑程度。

4.3 考虑到训练集的不完备性，针对未登入双字的情况，提出了一种备用字典的方法。在第一字典训练集上未出现的双字，可以在备用字典中进行查找。之所以不直接在两个字典上直接训练，是目前公开的训练集的切分标准有差异，直接在多个训练集上直接统计会导致参数的不一致性。对于极端情况下仍未登入的双字采用独立性假设。假设两个字之间独立，根据 Bayes 公式。其联合概率采用 4.1 式来计算。对于未登入字的情况，双字特征单元的概率由其中登入字确定，连续未登入字的情况采用“平滑”处理。

$$P(T_i) = P(S_{i-1}S_i \mid O_{i-1}O_i) = \frac{P(O_{i-1}O_i \mid S_{i-1}, S_i) \cdot P(S_{i-1}, S_i)}{P(O_{i-1}O_i)}$$
$$P(O_{i-1}O_i \mid S_{i-1}, S_i) = P(O_{i-1} \mid S_{i-1}, S_i) \cdot P(O_i \mid S_{i-1}, S_i) \quad (4.1)$$

5、实验结果

5.1 数据集和结果

考虑到句子的句首和句尾的特殊性，在每个句子的句首加上“#”，句尾加上“$”。对训练集中的全角半角进行统一采用半角符号，英语字母，公式，单位等字符均看作单个字。在公开的 PKU2005，MSR2005,CITYU2005 语料库进行实验。得到如下实验结果。

表 1. 开放集上双字

语料库	精度	召回率	F1 值
PKU	0.944843	0.946681	0.945761
MSR	0.941505	0.939967	0.940805
CITYU	0.934495	0.936604	0.935257

表 2. 双字结合词表驱动

语料库	精度	召回率	F1 值
PKU	0.944843	0.946681	0.945761
MSR	0.945505	0.949967	0.946805
CITYU	0.934495	0.936604	0.935257

表 3. 开放集三字上下文 F1 值

语料库	精度	召回率	F1 值
PKU	0.948	0.947	0.948
MSR	0.961	0.959	0.96
CITYU			

表 3. 开放集四字上下文 F1 值

语料库	精度	召回率	F1 值
PKU			
MSR			
CITYU			

表 4. 使用完备字典

语料库	精度	召回率	F1 值
PKU	0.972254	0.970289	0.971271
MSR	0.961407	0.970133	0.96575
CITYU	0.973527	0.970197	0.971859

表 5. 在封闭集上进行测试

语料库	精度	召回率	F1 值
PKU	0.997438	0.998467	0.997975
MSR	0.992675	0.995415	0.994043
CITYU	0.998377	0.99167	0.995012

对比表 1 和表 2 可以看出未登入双字对分词正确率有一定的影响，对比表 1 和表 3 可以看出训练语料与测试语料在词的分布上有一定的距离。其中 PKU 语料中含有大量的全角半角数字，英文字母，标点符号，甚至数学符号，法文单词以及拼写错误。考虑到这情况，可以在原始训练集上进行数据清洗来进一步提高分词正确率。

该方法在开放集上分词正确率与目前最好的分词方法相差两个百分点左右，在封闭集上，该方法的分词正确率是可观的。且该方法在分词速度上有明显的优势，整个训练加分词能在 60s 内完成。

5.2 其他方法对比

方法	PKU	MSR
Bi-LSTM	95.0	95.8
GRNN	95.8	96.2
Pei et al. (2014)	94.0	94.9
Chen et al. (2015)	96.1	96.2
DGRNN	96.1	96.3

6 不足与改进

不足与改进，该方法在训练集完备的情况下，分词正确率接近 99.97%。对于出现大量未登入词的情况，该方法还有待提高。而且该方法的状态转移图是建立

在相邻双字特征共用一个相同字的状态。后面可以采用条件随机场模型，将条件随机场中的基于状态的上下文信息转变为该方法中的基于字的上下文信息。这样方式从理论上来讲不损失当前字的上下文信息，而且经该方法统计证明基于字的上下文信息模型复杂度不会成指数级增长。