**文章编号：**

# 基于混合隐马尔科夫模型的中文原子事件抽取

**摘要：**针对类型相关事件抽取无法全面表达自然文本的问题，本文采用原子事件的概念来结构化表示文本，深入分析了已有事件抽取技术和隐马尔可夫模型，提出了基于混合隐马尔科夫模型（HHMM, Hybrid Hidden Markov Model）的中文原子事件抽取方法，其中混合隐马尔可夫模型包括二阶隐马尔科夫模型、二维隐马尔可夫模型以及混合 Viterbi 算法。实验结果表明该方法能够有效地从非结构化文本中抽取出原子事件语义元素，模型的准确率、召回率与 F 值分别为 57.89%、63.43%与 60.53%。

**关键词：**原子事件抽取；二阶隐马尔可夫模型；二维隐马尔可夫模型；混合隐马尔科夫模型；

**中图分类号：**TP391　　**文献标识码：**A

## Hybrid Hidden Markov Model Based Chinese Atomic Event Extraction

**Abstract:** The type-dependent event extraction cannot express natural language text comprehensively, and the concept of atomic event is used to express text in this paper. A Chinese atomic event extraction method based on Hybrid Hidden Markov Model (HHMM) is proposed by analyzing existing event extraction and Hidden Markov Model (HMM). The HHMM includes a second-order HMM, a two-dimensional HMM and a hybrid Viterbi algorithm. The experimental result shows that this hybrid model can extract atomic events from an unstructured text effectively, and the precision, recall and F-score are 57.89%, 63.43% and 60.53% respectively.

**Key words:** Atomic Event Extraction; Second-Order HMM; Two-Dimensional HMM; Hybrid HMM

## 1 引言

事件抽取属于信息抽取领域，研究如何将自然语言文本表示为事件形式，一个事件的结构可以表示为：“[何时] [何地]，[谁] 对 [谁] 做了 [何事]”。随着互联网上的文本信息呈井喷式增长，事件抽取已成为自然语言处理领域的研究热点。事件抽取技术已被广泛应用于各个领域，如文本蕴含[1]、信息检索[2]、股票价格预测[3]、指代消解[4]以及社区问答[5]等。

目前大部分关于事件抽取的研究都是基于 MUC、TDT、ACE 与 TAC 等评测任务，由于这些评测任务对事件类型进行了限定，传统的事件抽取方法将事件识别作为分类问题来处理，通过机器学习或者事件模版的方法从文本中挖掘某领域、某类型或某主题的事件信息。为使计算机全面透彻地理解某一段落、篇章甚至文档，仅抽取类型相关的主要事件是不够的，还要抽取其中的附属原子事件[6]。

例 1 选自 ACE[1] 事件抽取评测任务所用语料，其中类型相关的事件如图 1 所示。

**例 1：**

俄罗斯总统 6 号早上也已经派遣外交部长[伊万诺夫 Artifact] [前往 Anchor] [当地 Destination] ，希望协助化解这次南斯拉夫的政治危机。


图 1. 例句 1 事件图

从图 1 可以看出，尽管 ACE 事件抽取评测任务中定义了 8 种类型以及 33 种子类型的事件，使用这种事件类型相关的方式对文本进行结构化表示很难全面覆盖文本中的信息。通过事件类型无关的原子事件对文本进行结构化表示则能保留更为丰富的信息，目前的原子事件中有 6 种语义元素，分别是施事（Agent）、受事（Patient）、谓词（Predicate）、时间（Time）

[1] http://www.itl.nist.gov/iad/mig/tests/ace/2005/

和地点（Location）。使用原子事件结构化表示例句 1 的原子事件图如图 2 所示。

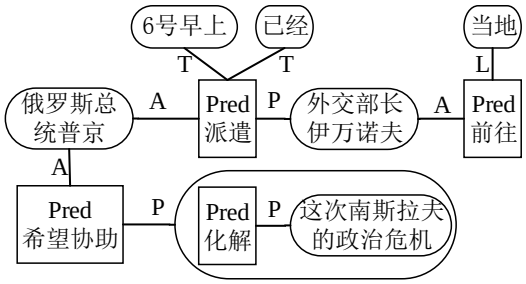


图 2. 例句 1 原子事件图

Boros 等[7]通过有限的先验领域知识训练得到神经网络模型，用于抽取特定领域的事件角色。Li 等[8]为避免事件触发词和事件论元抽取过程中的错误传播，利用谓词结构理论将事件触发词抽取模型和事件论元抽取模型进行联合，同时采用大量的全局特征抽取触发词与论元之间的依赖关系。Li 等[9] 根据中文触发词语义特征以及触发词之间的篇章一致性提出了两种推理机制来解决触发词识别过程中未登录词识别问题以及分词不当导致的登录词错误识别问题。Chen 等[10]为解决 Li[9]方法中错误传播问题，使用联合模型进行事件抽取。Zhou 等[11]提出了基于贝叶斯模型的“Latent Event Model”，用来从社会媒体中抽取结构化事件。Llorens 等[12]通过 CRF 模型进行语义角色标注，并应用于 TimeML 的事件抽取，提升了系统性能。

隐马尔可夫模型是一种比较成熟的概率图统计模型，近年来依然被广泛应用到各个领域。Cahyadi 等[13]使用连续隐马尔科夫模型对齐双语科技文献语料关键词列表，以此得到科技术语词典。Carter 等[14]受自适应拒绝抽样和启发式搜索的启发，通过优化低阶语言模型的手段来提升高阶隐马尔科夫模型的效果。Engelbrecht 等[15]将用户观点看作时间相关的连续过程，提出了一种基于 HMMs 的方法来预测用户对“Spoken Dialog System”质量的评价。Ramanath 等[16]提出了一种使用 HMMs 的无监督模型来对具有相似结构的文档进行章节对齐。

目前大部分事件抽取的研究都集中在类型相关的事件抽取上，关于事件类型无关的原子事件抽取的研究相对较少。本文对传统 HMM 进行深入分析，提出一种适用于原子事件抽取的混合隐马尔可夫模型 HHMM，首先在一阶 HMM 的基础上进一步考虑历史状态对当前状态转移和观察值发射的影响，将其扩展到二阶 HMM，然后在二阶 HMM 的基础上进一步考虑高维特征在原子事件抽取过程中的作用，将其扩展到二阶二维 HHMM，最后利用混合 Viterbi 算法进行原子事件抽取，实验结果表明将 HHMM 应用于原子事件抽取是有效的。

## 2 混合隐马尔科夫模型 HHMM

混合隐马尔科夫模型 HHMM 包含语料预处理、二阶 HMM 模型参数训练、二维 HMM 模型参数训练和原子事件抽取四部分，模型框架如图 3 所示。

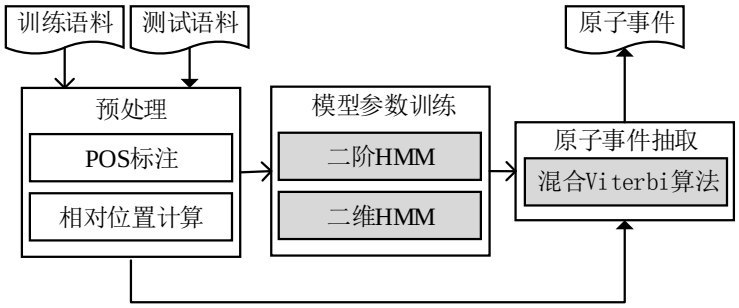

图 3 HHMM 模型框架图

在语料集预处理阶段，首先将语料集中的文本切分为句子，然后使用 LTP-Cloud[2]对句子进行中文分词和词性标注，再计算每个词在句子中的相对位置。训练语料的句子中标注的原子事件元素标签、POS 标签和相对位置会被映射到一个三维向量空间作为模型参数训练的输入。测试语料的句子中的 POS 标签和相对位置会被映射到一个二维向量空间作为原子事件抽取步骤的输入，测试语料中标注的原子事件元素标签作为评估模型效果的依据。在模型参数训练阶段，通过极大似然估计算法 MLE（Maximum Likelihood Estimation）使用训练语料来拟合二阶 HMM 和二维 HMM 的模型参数。在原子事件抽取阶段，通过混合 Viterbi 算法以及训练得出的模型参数对测试语料进行原子事件抽取。

## 2.1 HHMM 结构

在原子事件抽取问题中，隐藏状态序列是一个句子的原子事件元素标签，表示为 $S=\{s_1, s_2, ..., s_i, s_j, ..., s_T\}$，其中 T 表示序列的时间长度，$s_t$ 表示 t $(1\le t\le T)$时刻的隐藏状态；一个观察值序列是句子中的 POS 标签序列，表示为 $PO=\{po_1, po_2, ..., po_i, po_j, ..., po_T\}$，其中 $po_t$ 表示 t $(1\le t\le T)$时刻的 POS 标签；另一个观察值序列是句子中各个成分的相对位置序列，表示为 $RO=\{ro_1, ro_2, ..., ro_i, ro_j, ..., ro_T\}$，其中 $ro_t$ 表示 t $(1\le t\le T)$时刻在整个序列中所处的相对位置。

HHMM 可表示为八元组 $\lambda=(N,M,L,A,B^0,B,C,\pi)$，其中各个参数的含义如下。

（1）N：模型中隐藏状态的种类数目。在原子事件抽取问题中，N 代表原子事件元素种类数目，原子事件元素种类集合可表示为 $E=\{e_1, e_2, ..., e_N\}$。

（2）M 和 L：模型中两类观察值的种类数目。在原子事件抽取问题中，M 代表 POS 标签种类数目，POS 标签种类集合可表示为 $P=\{p_1, p_2, ..., p_M\}$；L 代表相对位置种类数目，相对位置种类集合可表示为 $R=\{r_1, r_2, ..., r_L\}$。

（3）$A=\{a_{ijk}\}$：状态转移概率矩阵。$a_{ijk}$ 表示从已知历史原子事件元素对$(e_i, e_j)$转移到原子事件元素 $e_k$ 的概率。

（4）$B^0=\{b_{j(k)}\}$和 $B=\{b_{ij(k)}\}$：POS 标签观察值的发射概率矩阵。$b_{j(k)}$表示序列初始状态为 $e_j$ 时，其初始 POS 标签观察值为 $p_k$ 的概率；$b_{ij(k)}$表示当前隐藏状态为 $e_j$ 且前一隐藏状态为 $e_i$ 时，当前 POS 标签观察值为 $p_k$ 的概率。

（5）$C=\{c_{j(k)}\}$：相对位置观察值的发射概率矩阵。$c_{j(k)}$表示序列处于状态 $e_j$ 时，当前相对位置观察值为 $r_k$ 的概率。

（6）$\pi=\{\pi_i\}$：初始状态分布矩阵。$\pi_i$ 表示隐藏状态序列以原子事件元素 $e_i$ 作为初始状态的概率。

## 2.2 二阶 HMM

传统一阶 HMM 在当前状态进行转移时没有考虑历史状态对未来状态的影响，受 N-gram 模型启发，二阶 HMM 在当前状态进行转移时，将历史状态对其状态转移的影响因素考虑进去，同时二阶 HMM 认为历史状态对当前状态对应观察值的发射概率也会产生影响。一阶 HMM 和二阶 HMM 的结构分别如图 4 和图 5 所示。

[2] http://www.ltp-cloud.com/

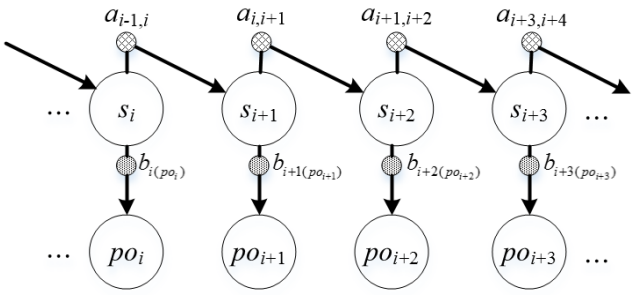


图 4 一阶 HMM 结构示意图

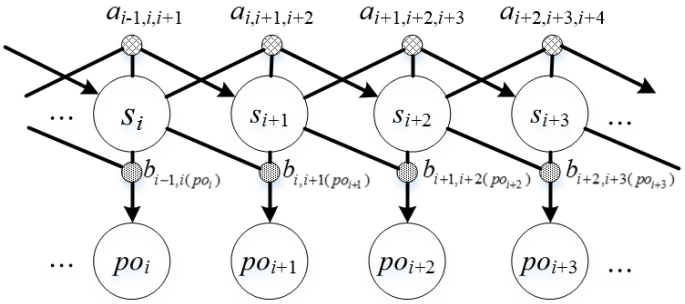


图 5 二阶 HMM 结构示意图

在二阶 HMM 中，模型参数初始状态分布矩阵，状态转移矩阵和 POS 标签观察值分布矩阵可以通过 MLE 算法从完全标注的训练语料中训练得出。根据二阶 HMM 定义以及 MLE 算法流程，训练二阶 HMM 模型参数的核心公式如下。

初始状态概率可由公式（1）得出。

$$\pi_i = \frac{c(e_i)}{\sum_{j\in[0,N]} c(e_j)} \quad (i \in [1,N]) \qquad (1)$$

其中 $c(e_i)$ 表示训练语料中以状态 $e_i$ 为初始状态的序列总数，$\sum_{j\in[0,N]} c(e_j)$ 表示训练语料中序列总数。

状态转移概率可由公式（2）得出。

$$a_{ijk} = P(s_t = e_k \mid s_{t-1} = e_j, s_{t-2} = e_i) = \frac{c(e_i, e_j, e_k)}{\sum_{l\in[1,N]} c(e_i, e_j, e_l)} \quad (i, j, k \in [1,N], t \in [2,T]) \qquad (2)$$

其中 $c(e_i, e_j, e_k)$ 表示当 $t$ 时刻状态为 $e_k$，$t$-2 和 $t$-1 时刻状态分别为 $e_i$ 和 $e_j$ 的次数，$\sum_{l\in[1,N]} c(e_i, e_j, e_l)$ 表示当 $t$ 时刻状态为任意状态，$t$-2 和 $t$-1 时刻状态分别为 $e_i$ 和 $e_j$ 的次数总和。

POS 标签观察值分布概率可以由公式（3）和公式（4）得出。

$$b^{0}_{j(k)} = P(po_1 = p_k \mid s_1 = e_j) = \frac{c(e_j, p_k)}{\sum_{l\in[1,M]} c(e_j, p_l)} (j \in [1,N], k \in [1,M]) \qquad (3)$$

其中 $c(e_j, p_k)$ 表示 $t$=1 时，状态为 $e_j$ 并且 POS 标签为 $p_k$ 的次数，$\sum_{l\in[1,M]} c(e_j, p_l)$ 表示 $t$=1 时，状态为 $e_j$，POS 观察值 $p_l$ 为任意标签的总次数。

$$b_{ij(k)} = P(po_t = p_k \mid s_t = e_j, s_{t-1} = e_i) = \frac{c(e_i, e_j, p_k)}{\sum_{l \in [1,M]} c(e_i, e_j, p_l)} \quad (4)$$

$$(i, j \in [1, N], t \in [2, T], k \in [1, M])$$

其中 $c(e_i, e_j, p_k)$ 表示 $t$ 时刻状态为 $e_j$ 且 $t$-1 时刻状态为 $e_i$ 时，t 时刻的 POS 标签观察值为 $p_k$ 的次数，$\sum_{l \in [1,M]} c(e_i, e_j, p_l)$ 表示 $t$ 时刻状态为 $e_j$ 且 $t$-1 时刻状态为 $e_i$ 时，t 时刻的 POS 标签观察值 $p_l$ 为任意标签的总次数。

## 2.3 二维 HMM

二阶 HMM 在传统 HMM 的基础上进一步考虑了历史状态对状态转移概率和观察值发射概率的影响，鉴于中文是一种灵活的语言，词义信息仅属于中文语义信息的一个层面，仅通过 POS 标签这一维度来进行原子事件抽取是不够的，比如原子事件中的施事和受事就很可能拥有相同 POS 标签，但是它们在一个事件中所处的地位不同，这一点可以从它们在句子中的位置来体现，例如施事元素一般出现在句子的前部，而受事元素一般出现在句子的中部或者后部，因此本文在应用 POS 语义标签进行原子事件元素识别过程中，还从中文句子结构中各个原子事件元素相对位置的层面进行考虑。二维 HMM 的结构如图 6 所示。

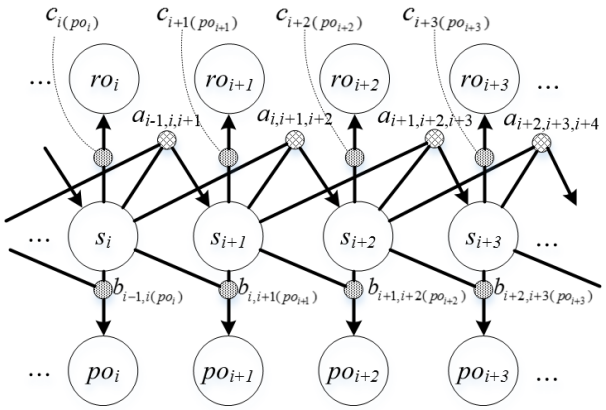


图 6 二维 HMM 结构示意图

从图 6 中可以看出，加入了句子结构信息的二维 HMM 是在二阶 HMM 基础上进行扩展的，因此只需要训练相对位置观察值矩阵即可，该矩阵同样通过 MLE 算法从训练语料中

得出，本文将原子事件元素在句子中的相对位置划分为十种，分别对应数字 1-10，一个原子事件元素在句子中的相对位置可以通过公式（5）得出。

$$ro_i = 10 * i / T \quad (i \in [1, T]) \qquad (5)$$

相对位置观察值分布概率可以由公式（6）得出。

$$c_{j(k)} = P(ro_t = r_k \mid s_t = e_j) = \frac{c(e_j, r_k)}{\sum_{l \in [1,L]} c(e_j, r_l)} (j \in [1, N], t \in [1, T], k \in [1, L]) \qquad (6)$$

其中 $c(e_j, r_k)$ 表示 $t$ 时刻状态为 $e_j$ 且相对位置观察值为 $r_k$ 的次数，$\sum_{l \in [1,L]} c(e_j, r_l)$ 表示 $t$ 时刻状态为 $e_j$ 且相对位置观察值 $r_l$ 为任意位置的总次数。

## 2.4 混合 Viterbi 算法

在混合 Viterbi 算法中，需要定义三个变量，即 Viterbi 变量 $\delta_1(i)$、Viterbi 变量 $\delta_{t,t+1}(j,k)$ 和回溯变量 $\varphi_{t,t+1}(k)$。其中 $\delta_1(i)$ 表示已知 POS 标签观察序列和相对位置观察序列的前提下，$t$=1 时状态为 $e_i$ 的概率；$\delta_{t,t+1}(j,k)$ 表示已知 POS 标签观察序列和相对位置观察序列的前提下，$t$ 时刻状态为 $e_j$ 且 t+1 时刻状态为 $e_k$ 的最大概率；$\varphi_{t,t+1}(k)$ 为回溯变量，用于记录 t+1 时刻的状态 k 的前一个历史状态。混合 Viterbi 算法中用到的核心公式（7）-（11）如下。

$$\delta_1(i) = \log(\pi^0{}_i) + \log(b^0{}_{i(po_1)}) + \mu \log(c_{i(ro_i)}) \quad i \in [1, N] \qquad (7)$$

$$\delta_{1,2}(i, j) = \max_{i, j \in [1,N]} (\delta_1(i) + \log(a_{ij})) + \log(b_{ij(po_2)}) + \mu \log(c_{j(ro_2)}) \quad i, j \in [1, N] \qquad (8)$$

$$\delta_{t,t+1}(j,k) = \max_{i, j, k \in [1,N], t \in [2, T-1]} (\delta_{t-1,t}(i, j) + \log(a_{ijk})) + \log(b_{jk(po_{t+1})}) + \mu \log(c_{k(ro_{t+1})}) \quad j, k \in [1, N], t \in [2, T-1] \qquad (9)$$

$$\varphi_{1,2}(j) = \arg\max_{i \in [1,N]} (\delta_1(i) + \log(a_{ij})) \quad j \in [1, N] \qquad (10)$$

$$\varphi_{t,t+1}(k) = \arg\max_{j} (\delta_{t-1,t}(i, j) + \log(a_{ijk})) \quad i, j, k \in [1, N], t \in [2, T-1] \qquad (11)$$

其中系数 μ 用来调整相对位置特征的影响程度，其他各个参数含义与前文相同。算法 1 描述了混合 Viterbi 算法的流程。

**算法 1** 混合 Viterbi 算法

**输入：** POS 标签观察序列 $PO=\{po_0, po_1, ..., po_T\}$
相对位置观察序列 $RO=\{ro_0, ro_1, ..., ro_T\}$
HHMM $\lambda = (N, M, L, A, B^0, B, C, \pi)$

**输出**: 最优的原子事件元素序列 $S^*=(s_1^*, s_2^*, ..., s_T^*)$

//初始化
**for** $i$ **from** 1 **to** $N$
　根据公式（7）计算 $\delta_1(i)$;
**end for**
**for** $j$ **from** 1 **to** $N$
　**for** $i$ **from** 1 **to** $N$
　　根据公式（8）计算 $\delta_{1,2}(i, j)$;
　　根据公式（10）计算 $\varphi_{1,2}(j)$;

**end for**

**end for**

//递归

**for** *t* **from** 2 **to** *T*-1

**for** *k* **from** 1 **to** *N*

**for** *i* **from** 1 **to** *N*

**for** *j* **from** 1 **to** *N*

根据公式（9）计算 $\delta_{t,t+1}(j,k)$;

根据公式（11）计算 $\varphi_{t,t+1}(k)$;

**end for**

**end for**

**end for**

**end for**

//搜索最优原子事件元素序列

$s_T^* = \arg\max_j[\varphi_{T-1,\mathrm{T}}(j)]$;

**for** *t* **from** *T*-1 **to** 1

根据回溯关系 $s_t^* = \varphi_{t,t+1}(s_{t+1}^*)$ 计算 t 时刻的状态;

**end for**

# 3 实验结果与分析

## 3.1 实验背景

虽然已有大量关于事件抽取的相关评测会议与相应的评测语料，但是这类会议主要关注特定类型事件模版填充以及事件要素识别，其中 ACE 评测任务最高 F 值为 53.9%[17]。目前，国内外关于事件类型无关的事件抽取研究较少，且尚未发现可供公开评测的语料库。本文使用的语料来自 NITCIR-9 中 RITE 任务，其中包含 1414 个句子，我们在该语料的基础上进行人工标注，语料中原子事件元素分布如表 1 所示。人工标注具体流程如下：

（1） 将标注人员分为 3 组，分组之后独立标注全部语料，仅允许组内讨论；

（2） 标注完之后，进行一致性检查。对于一个句子，如果三组标注结果一致，则认为标注结果正确；

（3） 对于不一致的标注结果，全部标注人员投票决定；

表 1 NTCIR-9 RITE 语料中原子事件元素标注结果

| 类型 | 原子事件元素 | 数量 |
|---|---|---|
| 0 | 施事 A(Agent) | 1571 |
| 1 | 谓词 Pred(Predicate) | 861 |
| 2 | 受事 P(Patient) | 2569 |
| 3 | 时间 T(Time) | 449 |

| | | |
|---|---|---|
| 4 | 地点 L(Location) | 443 |
| 5 | 非原子事件成分 N(Not an atomic event element) | 2310 |

在对 HHMM 进行评测时，使用信息抽取系统中常用的准确率(P)，召回率(R)和 F 值（*F*1-score)来评估。我们通过计算每一类原子事件元素的准确率、召回率和 F 值，来评估 HHMM 对每一类原子事件元素的抽取效果；通过在不统计 N 元素的前提下，评估 HHMM 用于原子事件抽取的整体效果。

## 3.2 实验结果

本文采用的基准模型为传统 HMM，由于 HHMM 与基准模型相比主要在于考虑了历史状态影响因素的二阶 HMM 和加入了句子结构信息的二维 HMM，在实验过程中主要集中于这两方面进行对比。表 2 展示了系统的整体效果。

表 2 HHMM 的整体效果

| 方法 | P | R | F |
|---|---|---|---|
| 传统 HMM | 54.64% | 58.74% | 56.61% |
| 二阶 HMM | 58.14% | 58.88% | 58.51% |
| 二阶二维 HHMM | 57.89% | 63.43% | 60.53% |

从表 2 中可以看出 HHMM 的整体效果，传统 HMM 在事件类型无关的原子事件元素抽取中取得了 56.61%的 F 值；之后通过将一阶 HMM 扩展到二阶 HMM，加入了当前状态的历史状态对转移概率和发射概率的影响，F 值取得了 1.9%的提升；再通过增加特征维度，在考虑 POS 标签的同时，引入句子结构的相对位置特征，将一维 HMM 扩展到二维 HMM，再与基准模型相比，F 值取得了 3.92%的提升。

传统 HMM 与二阶 HMM 对每一类原子事件元素的抽取效果如表 3 所示。

表 3 传统 HMM 与二阶 HMM 的效果

| 模型 | 传统 HMM | | | 二阶 HMM | | |
|---|---|---|---|---|---|---|
| % | P | R | F | P | R | F |
| A | 61.77 | 57.80 | 59.72 | 60.96 | 63.91 | 62.40 |
| Pred | 54.89 | 63.18 | 58.75 | 59.34 | 60.16 | 59.75 |
| P | 49.72 | 64.54 | 56.17 | 55.70 | 59.67 | 57.62 |
| T | 78.30 | 59.47 | 67.59 | 72.11 | 61.02 | 66.10 |
| L | 36.02 | 17.16 | 23.24 | 41.52 | 25.96 | 31.94 |
| N | 73.51 | 58.40 | 65.09 | 62.58 | 60.39 | 61.47 |
| Overall | 54.64 | 58.74 | 56.61 | 58.14 | 58.88 | 58.51 |

从表 3 可以看出，随着二阶 HMM 考虑了历史状态的影响，原子事件元素 A、Pred、P 和 L 的 F 值得到了不同程度的提升，表明考虑历史状态的影响对原子事件抽取整体效果的提升是有效的，也进一步说明语言是一种十分严谨的信息表达方式，在临近范围内的各个原子事件成分之间都会有一定的联系，传统 HMM 是无法考虑这种联系的，而二阶 HMM 引入历史状态的影响因素，在一定程度上还原了临近范围内原子事件成分之间的联系。与其他原子事件元素相比，T 和 N 效果反而下降了，通过对表 4 中 T 和 N 的抽取结果的深入分析，可以得知出现该现象的原因在于将传统 HMM 扩展到二阶 HMM 之后，被抽取的 T、N 元素的总数以及正确抽取的数量都分别有所提升，因此召回率得到提升而准确率却降低，由于准确率降低幅度比召回率提升幅度更大，最终导致 F 值下降。

表 4 原子事件元素 T 和 N 抽取结果

| | T | | N | |
|---|---|---|---|---|
| | 传统 HMM | 二阶 HMM | 传统 HMM | 二阶 HMM |
| 语料标注 | 449 | 449 | 2310 | 2310 |
| 模型抽取 | 341 | 380 | 1835 | 2229 |
| 正确抽取 | 267 | 274 | 1349 | 1395 |

二维 HMM 在二阶 HMM 的基础上，顾及了这样一个事实，一个句子中的不同成分所处位置不同，但是同一类成分的位置是相似的。本文没有采用传统基于规则的句式结构的概念，因为基于规则的句式结构表示句式的能力是有限的，本文采用相对位置这一统计概念来表示不同成分在句子中的位置信息。此时的 HHMM 同时采用了词义特征和句式特征，本文认为词义特征比句式特征具有更丰富的语义信息，因此考虑词义信息是很有必要的，为解决设定句式特征在原子事件抽取过程中的影响程度这一问题，在混合 Viterbi 算法中引入了用于调节相对位置特征影响程度的系数 μ，为找到合适的系数 μ 的取值，在 0-1 范围内，进行了以 0.1 为步长的单一变量实验，实验结果如图 7 所示。

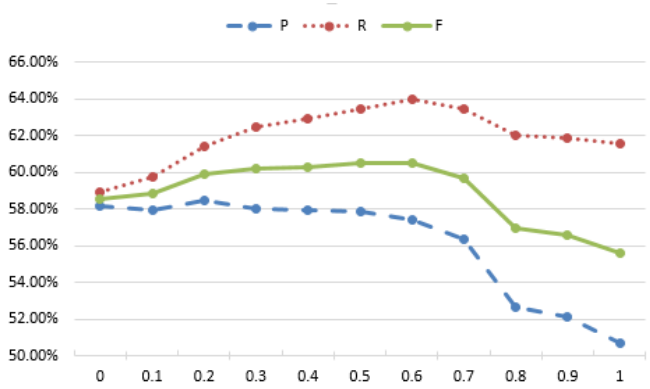


图 7 不同系数 μ 条件下二维 HMM 的效果

从图 7 中可以看出，随着系数 μ 的增大，准确率呈下降趋势。当系数处于 0 和 0.6 之间时，召回率呈上升趋势，当系数处于 0.6 和 1 之间时，召回率亦呈下降趋势。综合考虑准确率和召回率的 F 值在 μ 为 0.5 时取得最好的效果，为 60.53%。从上述实验结果可以得出至少以下两个结论：一方面，相对位置特征影响程度为 50%时，系统的整体效果最好，当 μ 为 0 时，相当于没有考虑句式特征，而当 μ 为 1 时，过多的引入句式特征的影响程度，导致系统整体效果的下降；另一方面，随着系数 μ 的增加，准确率和召回率的趋势基本上是相反的，出现这种现象的原因可以通过表 5 得知，当系数 μ 越来越大时，模型抽取 N 元素的数目不断减少，也就是说模型抽取其他元素的数目不断增加，结合图 7 中的数据可知，上述情况最终引起系统整体效果的召回率提升而准确率下降。

表 5 系数 μ 为 0 和 0.5 时 N 元素抽取结果

| 系数 μ | 0 | 0.5 |
|---|---|---|
| 语料标注 | 2310 | 2310 |
| 模型抽取 | 2229 | 1704 |
| 正确抽取 | 1395 | 1300 |

## 4 结论

本文提出了一种用于原子事件抽取的 HHMM，与事件抽取相比，原子事件抽取可以更全面的将自然语言文本结构化表示。在 HHMM 中，首先在一阶 HMM 的基础上进一步考虑历史状态对状态转移概率和观察值发射概率的影响，扩展到二阶 HMM，然后在考虑 POS 标签观察特征的同时，引入相对位置特征，将一维 HMM 扩展到二维 HMM，最终得到用于原子事件抽取的 HHMM。实验结果表明，HHMM 在解决事件类型无关的原子事件抽取的问题上是有效的。

在下一步工作中，一方面可以在统计模型中考虑更加丰富有效的语义特征来提升原子事件抽取效果，或者采用深度学习自动提取有效特征、训练模型参数来进行原子事件抽取，另一方面，可以在自动抽取原子事件之后对原子事件进行聚类分析，为其添加更为丰富的语义信息，以便计算机更为深层的理解文本。

## 参考文献


[1] 刘茂福, 李妍, 姬东鸿. 基于事件语义特征的中文文本蕴含识别[J]. 中文信息学报, 2013, 27(5):129-136.

[2] Goran Glavaš, Jan Šnajder. Event-Centered Information Retrieval Using Kernels on Event Graphs[J]. Graph-Based Methods for Natural Language Processing, 2013: 1-5.

[3] Xiao Ding, Yue Zhang, Ting Liu, et al. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation[C]. Proc. EMNLP 2014: 1415-1425.

[4] Heeyoung Lee, Marta Recasens, Angel Chang, et al. Joint Entity and Event Co-reference Resolution Across Documents[C]. Proc. EMNLP 2012: 489-500.

[5] Liqiang Nie, Meng Wang, Gao Yue, et al. Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information[J]. IEEE Transactions on Multimedia, 15(2): 426-441, 2013.

[6] 刘茂福, 胡慧君. 基于认知与计算的事件语义学研究[M]. 北京:科学出版社. 2013: 45-67.

[7] Boros E, Besançon R, Ferret O, et al. Event Role Extraction using Domain-Relevant Word Representations[C]. Proc. EMNLP 2014: 1852-1857.

[8] Qi Li, Heng Ji and Liang Huang. Joint Event Extraction via Structured Prediction with Global Features[C]. Proc. ACL 2013:73-82.

[9] Peifeng Li, Guodong Zhou, Qiaoming Zhu, et al. Employing Compositional Semantics and Discourse Consistency in Chinese Event Extraction[C]. Proc. EMNLP 2012:1006-1016.

[10] Chen Chen, Vincent Ng. Joint Modeling for Chinese Event Extraction with Rich Linguistic Features[C]. Proc. COLING 2012:529-544.

[11] Deyu Zhou, Liangyu Chen and Yulan He. A simple Bayesian modelling approach to event extraction from Twitter[C]. Proc. ACL 2014:700-705.

[12] Hector Llorens, Estela Saquete, Borja Navarro-Colorado. TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles[C]. Proc. CCL 2010: 725-733.

[13] Denny Cahyadi, Fabien Cromieres, Sadao Kurohashi. Constrained Hidden Markov Model for Bilingual Keyword Pairs Alignment[C]. Proc. of the 10th Workshop on Asian Language Resources, 2012:85-94.

[14] Simon Carter, Marc Dymetman, Guillanume Bouchard. Exact Sampling and Decoding in High-Order Hidden Markov Models[C]. Proc. EMNLP 2012:1125-1134.

[15] Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, et al. Modeling User Satisfaction with Hidden Markov Models[C]. Proc. ACL 2009:170-177.

[16] Rohan Ramanath, Fei Liu, Norman Sadeh, et al. Unsupervised Alignment of Privacy Policies using Hidden Markov Models[C]. Proc. ACL 2014: 605-610.

[17] Peifeng Li, Qiaoming Zhu, Hongjun Diao, et al. Joint Modeling of Trigger Identification and Event Type Determination in Chinese Event Extraction[C]. Proc. COLING 2012:1635-1652.