# 基于条件随机场与时间词库的中文时间表达式识别*

吴琼，黄德根

（大连理工大学 计算机科学与技术学院，辽宁 大连 116024）

**摘要：**提出一种统计与规则相结合的时间表达式识别方法。首先，通过分析中文文本中时间表达式的词形、词性和上下文信息，采用条件随机场识别时间单元而非时间表达式整体，避免了中文时间表达式边界定位不准确的问题；之后，从训练语料中自动获取候选触发词，并依据评价函数对候选触发词打分，筛选出正确的触发词完善触发词库；最后根据时间触发词库与时间缀词库，制定规则对时间表达式边界进行定位。实验结果显示开式测试 F1 值达到 98.31%。

**关键词：**CRF；规则；时间触发词；时间缀词

**中图分类号：**TP391　　**文献标识码：**A

# The Exploration of Temporal Information Extraction Based on CRF and Time Thesaurus

WU Qiong，HUANG Degen

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China)

**Abstract:** This paper proposes a generic algorithm for time expression recognition task based on combining rules with statistics. By analyzing a set of linguistic features of time expression in text such as lexical features and context information, using Conditional Random Fields (CRF) recognize time unit rather than time expression, avoiding the boundary localization problems in Chinese time expressions; automatically obtain the candidate trigger words through the test corpus, score the candidate trigger words based on evaluation function, filter out the right trigger word to perfect trigger thesaurus; set rules for the time expression boundary localization based on time trigger thesaurus and time affix word thesaurus. Our experimental results show that the F1 value reaches 0.9831 on open test.

**Key words:** CRF; rule; Time trigger; Time affix word

## 1 引言

时间表达式的识别在中文信息处理中有重要的作用和意义，时间表达式识别可为问答系统提供基本的素材，可以用于机器翻译、事件跟踪、信息检索中相关时间的定位，还可以用于定位事件发生的时间及回答时间相关的问题等。

1995 年，信息理解会议（Message Understanding Conference）首次将时间表达式的识别作为命名实体识别的一个子任务。2004 年，美国技术标准局（NIST）举办了第一届时间表达式识别与归一化（Time Expression Recognition and Normalization, TERN）的评测，随后，ACE2005（Automatic Content Extraction）和 SemEval2007（Semantic Evaluations）也将时间表达式评测纳入自己的任务中。

时间表达式识别的常用方法有两种：一种是基于规则的方法[1]，如文献[2]通过建立一些语法规则和补充限定条件，用规则匹配的方式识别时间表达式；文献[3]将时间信息划分为一系列的“时间基元”，使用启发式规则抽取时间表达式，再利用错误驱动方法对规则库进行剪枝，提高规则抽取的正确率；文献[4]提出一种基于正则表达式的 TIMEX2[5][6]中文时间短语边界识别方法。由于时间表达式具有一定的稳定性，直到现在，仍有人用规则的方法识

**收稿日期：** 2014-7-28　**定稿日期：**

**基金项目：**国家自然科学基金资助项目（61173100，61173101，61272375）

别时间表达式。规则[7]方法优点是使用简单，正确率高；缺点是很难将所有时间表达式完全覆盖，而且制定规则需要大量的人工，领域适应性较差。

另一种是基于机器学习的方法。这类方法一般借助于统计模型，常用的统计模型有：条件随机场(CRF)[8]和条件最大熵[9]。条件最大熵的方法优点是能够将各种特征在同一框架内刻画，不需要特征独立性假设，缺点是时空复杂度大，耗费资源；CRF[10]方法能找出全局最优解，可充分利用上下文的信息，但是它的结果好坏过分依赖于训练语料的质量，还存在数据稀疏和词序依赖的问题。文献[11]将中文时间短语分为日期型和事件型两类，利用 CRF 加入任意特征表达长距离的上下文依赖信息的能力，解决了时间短语词数较少时的噪声过大问题。

除此之外，文献[12]提出基于依存分析和错误驱动的方法来识别时间表达式，解决了长距离依赖的问题。文献[13]将浅层语义分析中的语义角色标注加入中文时间词识别中，在 CRF 训练中达到了较好的识别效果。

当句子中存在时间短语时，机器翻译的效果通常不太理想，若能将时间表达式提前识别，并作为一个整体去翻译整句，不仅能降低句法分析的复杂度，而且机器翻译的效果能得到改善。

例如，在句子“土地基金於一九八六年根据中英联合声明的规定成立，订明由联合声明於一九八五年五月二十七日生效当天起至一九九七年六月三十日止”中，存在三个时间表达式：

第 1 个时间表达式：一九八六年（英文 1986）

第 2 个时间表达式：一九八五年五月二十七日（英文 May 27, 1985）

第 3 个时间表达式：当天起至一九九七年六月三十日（英文 the day until June 30, 1997）

上述句子用 Google 翻译得到：Land Fund in accordance with the provisions of the Joint Declaration was established in 1986, stipulates that the joint statement on May 27, 1985 ended June 30, the day of the entry into force until 1997.

若将“一九八五年五月二十七日”和“当天起至一九九七年六月三十日”作为一个时间表达式整体，再翻译则变成：Land Fund in accordance with the provisions of the Joint Declaration was established in 1986, Stipulates the joint statement entry into force on May 27, 1985 ended the day until June 30, 1997.

再如“为了投票，他们排队长达三、四个小时。”Google 的翻译是：In order to vote, they waited three hours and four hours.将时间表达式“三、四个小时”作为整体识别后再翻译的话，得到正确的结果：In order to vote, they waited three or four hours.

## 2 基本概念及问题分析

### 2．1 时间表达式

TIDES 2003 Standard for the Annotation of Temporal Expressions 于 2004 年 4 月份发布的中文补充版[14]对时间表达式的定义是：时间表达式是时间单元的一个序列。在本文中，时间表达式可认为是时间触发词与时间缀词的组合。SemEval2010（Semantic Evaluations）task 13 首次将时间表达式的类型作为识别内容，将时间表达式分为以下 4 类：

- DURATION：例如：two weeks；
- SET：例如：every Monday morning；
- TIME：例如：at 2:45p.m；
- DATE：例如：January 27,1920，yesterday。

结合中文时间表达式的特点与翻译的需要，本文在以上四类的基础上进行了修改，将时间表达式分为了以下 7 类，即 DURATION 类、SET 类、TIME 类、DATE 类、LUNAR 类、FUZZY

类、RELATIVE-TIME 类，其中 DURATION 和 SET 与 SemEval2010 task 13 定义的时间类型相同，而 TIME 和 DATE 进行了修正，新增加 LUNAR、FUZZY、RELATIVE-TIME，具体如下：

- TIME 类：修改为一天中的某个具体时间点，如“4 点半”；
- DATE 类：修改为年月日这类的标准时间，如“2013 年 9 月 1 日”；
- LUNAR 类，表示中国传统节气，包括中国农历、节日等各种传统说法，如“国庆黄金周”、“大年初一”；
- FUZZY 类，表示模糊时间，如“数十年”；
- RELATIVE-TIME 表示相对时间，是相对于 DATE 来说的，不能具体到某一天的时间就称为相对时间，如“明天”、“下午”。

以上 7 类就是本文识别的时间表达式范畴，超出其中范围的时间相关的式子，不在本文识别范围内。

### 2．2 时间单元

本文中的时间单元是指时间表达式的最小组成单位。比如：“2014 年 5 月 2 日下午 3 点”这个时间表达式中，包含有 2014 年、5 月、2 日、下午、3 点这五个时间单元。

之所以选择时间单元作为 CRF 标注目标，是因为时间表达式是由时间单元组成的，时间单元之间搭配相对松散，没有很强的先后依赖关系，而时间表达式的形式多样。因此，时间单元的格式相对时间表达式比较固定，抽取难度小，准确度高。

### 2．3 时间触发词

时间触发词是判断一个短语是否是时间短语的关键词。通常是一个时间单位，表示时间概念，如“月”、“日”。根据识别的需要，本文将触发词分为两类，一类是独立触发词，另一类是数字触发词。

独立触发词指的是，单独存在就能表示时间的这样一类触发词。例如：“下午”、“昨天”等词语，不需要上下文信息，本身就可作为时间表达式。

数字触发词可细分为数字前缀触发词与数字后缀触发词，其中，当数字前缀触发词的前面为数字时，则数字与数字前缀触发词一起构成一个完整的时间单元（如“21 世纪”、“2008 年”），而其单独存在时不具有时间意义；数字后缀触发词的概念与数字前缀触发词类似，不同的是数字是在触发词的后面，如“星期三”。

### 2．4 问题分析

时间表达式识别作为自然语言处理的一个分支，其识别的难点是歧义问题。时间表达式是由时间触发词触发，但是，并非含有触发词的表达式就一定是时间表达式。例如，“后天因素”中的“后天”是时间触发词，但是“后天因素”这个词并不是时间表达式。再比如，“6 分”可以是时间表达 6 分钟，也可能是得分 6 分。具体的含义要看上下文。这类问题都是时间表达式的识别歧义问题。

本文针对这类歧义问题，利用规则进行限定，查看其上下文信息，判断其是否为时间表达式。

## 3 时间表达式识别模型

中文时间表达式识别模型主要包含以下两部分：

1) CRF 生成特征模板部分：输入训练集，选取 CRF 的模板特征，自动生成对应的特征模板。
2) 规则处理部分：该部分主要是对 CRF 没识别的时间单元进行补充识别，并确定时间表达式边界。具体分为以下六个步骤：

第 1 步：对测试语料的格式进行预处理，将语料格式转化为 CRF 要求的格式；

第 2 步：对语料进行分词、词性标注[1]；

第 3 步：使用第一部分生成的 CRF 特征模板对测试语料进行测试，得到标注出时间单元的语料；

第 4 步：对 CRF 模型标注后的结果进行处理，去除错误的时间单元，自动获取时间单元中的时间词生成候选触发词表，通过评价函数对候选触发词进行除杂，将正确的触发词分类放入相应的触发词库中；

第 5 步：考虑到 CRF 对于训练语料具有很强的依赖性，对某些不常用的时间触发词的标注效果不理想，因此，制定规则，对语料中的时间触发词进行补充标注；

第 6 步：利用相邻时间单元合并的原则合并时间单元，由于时间表达式的上下边界通常是由时间缀词修饰，因此，借助时间前缀词库和时间后缀词库，确定时间表达式的上下边界；

第 7 步：对标注的时间表达式进行筛选，去除其中错误的标注，得到正确的时间表达式。

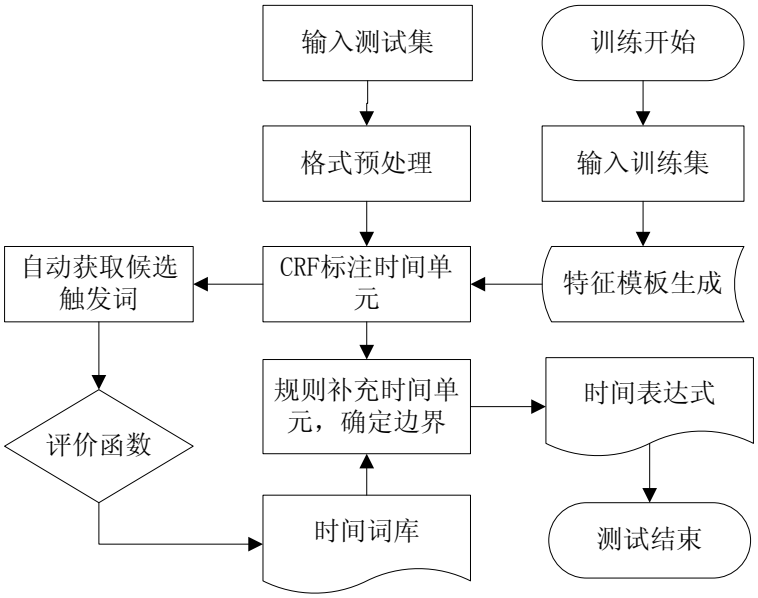


图 1　中文时间表达式识别模型

### 3．1 基于 CRF 的时间单元标注

CRF 是一种基于统计的无向图模型，它定义了在给定观察序列条件下，计算整个标注序列的单一联合概率分布。Lafferty 等人定义 CRF 为指数形式分布，这就使得不同状态下的不同特征的权值可以相互平衡。给定观察序列 X={$x_i$}(i=0, 1, …, n)和状态序列 Y={$y_i$}(i=0, 1, …, n)，线性链的 CRF 定义序列 Y 的条件概率为：

$$p(Y \mid X) = \frac{1}{z(X)} \exp\left( \sum_{i-1}^{n} \sum_{j-1}^{m} \lambda_j f_j(y_{i-1}, y_i, X, i) \right) \quad (1)$$

其中，$z(X)$是归一化因子；n 表示给定词序列的长度；$f_j(y_{i-1}, y_i, X, i)$是特征函数，既可以表示无向图边的转移特征$e(y_i - 1, y_i, X, i)$，也可以表示节点的状态特征$v(y_i, x, i)$；$\lambda_j$是第 j 个特征函数的权重系数。时间表达式识别问题可以归结为序列标注问题，其任务是给定观察序列 x 的条件下，估计产生标注序列 y 的条件概率有多大。而 CRF 模型具有强大的特征描述能力和特有的克服标注偏置问题的能力，它可以非常容易地将观察序列中的任意特征

[1] 分词工具使用的是大连理工大学自然语言处理实验室的 NiHao 分词系统

加入到模型中，表达长距离依赖的上下文依赖信息，从而确定标注时间单元的左右边界。

基于 CRF 的时间单元识别模块的具体流程见图 2。其中，训练集采用的是 2000 年的人民日报，共 26 万多条词，人工标注时间单元。本文特征选取的是：当前词的词形、词性；前一个词的词形、词性；后一个词的词形、词性。

前人也有尝试过使用 CRF 方法来标注时间表达式，但是标注的对象是整个时间表达式而非时间单元。由于时间缀词的不定性，导致标注的时间表达式容易发生边界错误。本文采用 CRF 方法标注时间单元，与时间表达式相比，粒度更小，而且，时间单元的格式较时间表达式更稳定，识别效果有一定的提高。

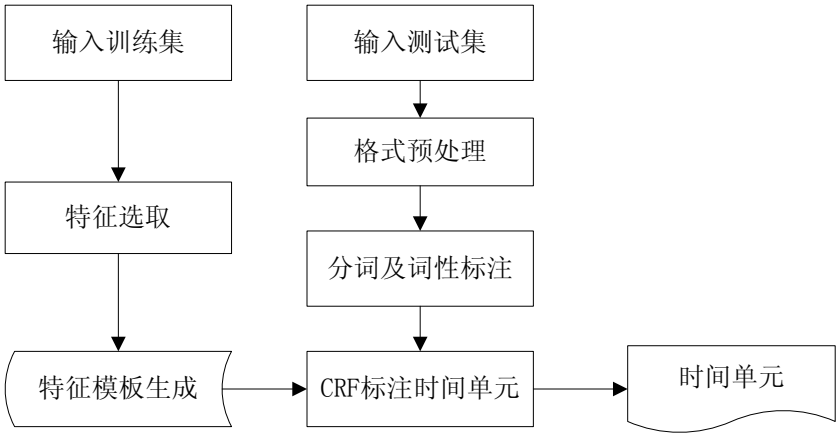


图 2　CRF 识别时间单元过程

### 3．2 基于时间词库的规则方法

由于 CRF 的局限性，当遇到训练语料中很少出现的时间表达式类型，其识别效果很差，甚至不能识别。时间表达式边界的识别效果也不好。为此，本文通过以下两个方面对其进行补充修改。

对于训练语料中很少甚至不存在的时间单元，本文采取的方法是，构建时间触发词库，对其进行补充标注。初始的时间触发词库由人工构建，由于时间表达式的识别效果一定程度依赖于时间触发词库的规模，而人工完善的耗费太大，因此，，利用转换规则自动获取时间单元中的候选触发词，生成候选触发词表。候选触发词表中含有很多错误的候选词，如若直接放入触发词库会极大地降低识别结果的正确率，因此需要对候选触发词进行除杂。本文通过引入评价函数 Score($T_i$)来对候选触发词进行打分，设置 λ 阈值进行筛选。获得的触发词根据类型分为数字触发词与独立触发词，分别放入对应的触发词库，达到完善时间触发词库的目的。

研究发现，时间表达式的边界一般是时间缀词，时间缀词对时间表达式的作用范围起到限定的作用。因此，我们构建了时间前缀词库和时间后缀词库，通过查看时间单元的前后词是否是时间缀词来确定时间表达式的边界。

本文使用的时间触发词库有两个：独立触发词库和数字触发词库。设立两个触发词库的原因是，能有针对性的处理不同类型的时间单元，有利于下面的筛选环节。

基于 CRF 进行时间单元的标注后，再通过规则识别时间表达式，具体过程如下：

1) 将 CRF 标注时间单元中部分错误的标注去除；
2) 利用转换规则自动获取时间单元中的候选触发词，生成候选触发词表；
3) 候选触发词 $T_i\{i=0,1,2,\ldots n\}$，将候选触发词表加入后的系统识别结果与语料的正确结果进行对比，每个候选触发词的得分计算方法是：

$$Score(T_i) = \frac{Ture(T_i)}{Ture(T_i) + False(T_i)} \quad (2)$$

其中，True($T_i$)表示候选触发词 $T_i$ 在测试语料标注中正确的个数，False($T_i$)表示错误的个数。设置阈值 λ，将得分小于 λ 的候选触发词从表中删去，具体过程见图 3。

4) 合并相邻的时间单元，并根据时间前缀词库与时间后缀词库，确定时间表达式的边界。

5) 设置限制条件，去除错误的时间表达式，得到正确的时间表达式。

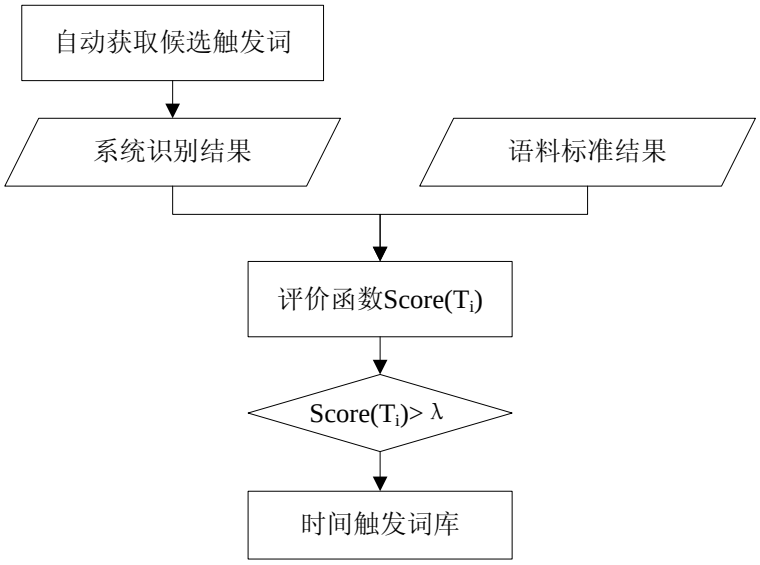


图 3　自动获取时间触发词过程

## 4 实验结果及分析

### 4.1 特征选取

CRF 模型使用 CRF++-0.54[2]工具包获得。CRF 的训练语料是纯人工标注，因此，训练效果比较好。据文献[9]统计，近 49%的时间表达式为一个独立的时间单元；26%的时间表达式是由两个时间单元构成；21%的时间表达式为 3 个时间单元；2.3%的为 4 个时间单元；1.7%为 5 个以上时间单元组成。因此，选取了以下两个模板特征，分别做了开式测试实验：

表 1 特征选取细节

| 特征 1 | 特征 2 |
|---|---|
| 当前词的词形、词性 | 当前词的词形、词性 |
| 前一个词的词形、词性 | 前一个词的词形、词性 |
| 后一个词的词形、词性 | 后一个词的词形、词性 |
| | 前两个词的词形、词性 |
| | 后两个词的词形、词性 |

为了测试上述两组特征哪个的效果更好，我们做了两组开式测试实验，测试语料是 2011 年的国际新闻，测试语料中共含有 1026 个时间单元，实验结果见图 4.

从图 4 可以看出，特征 1 模板的标注效果更好。本文最终选取的特征为：当前词的词形、词性，前一个词的词形、词性，后一个词的词形、词性。

[2] Available at http://crfpp.googlecode.com/svn/trunk/doc/index.html

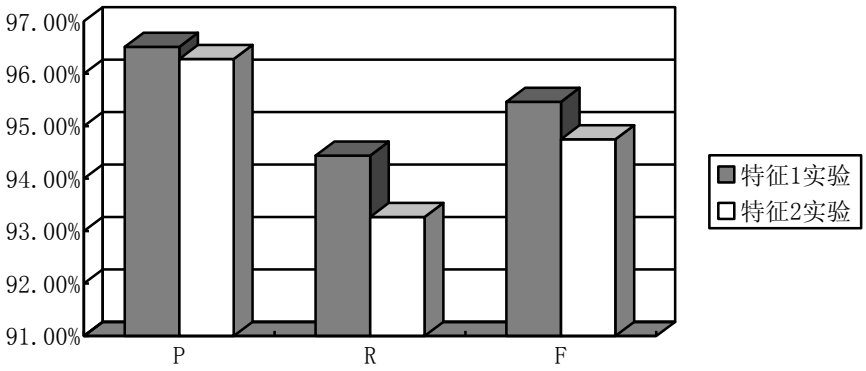


图 4　不同特征识别时间单元效果

**4．2 实验结果**

本文采用的 CRF 训练语料是 2000 年的人民日报新闻，共含有 24 万条词；测试语料为 2011 年国际新闻，共含有 18 万条词，其中时间表达式 1954 个。为了比较该方法的效果，我们分别做了两个基线系统：基于 CRF 识别时间表达式与基于规则识别时间表达式。在实验数据集一致的情况下，结果如表 2 所示。

表 2　时间表达式识别结果对比

| 识别方法 | P | R | F-measure |
|---|---|---|---|
| CRF 方法 | 91.78% | 85.67% | 88.62% |
| 规则方法 | 95.00% | 93.88% | 94.44% |
| 本文方法 | 98.41% | 98.21% | 98.31% |

从表 2 中的实验结果可以看出，本文采用的 CRF 与规则相结合的方法，比单纯用 CRF 方法或规则方法的识别效果好。CRF 方法识别较复杂的时间表达式时容易发生边界错误，如时间表达式“3、4 小时”，CRF 识别的结果为“4 小时”；并且，CRF 识别结果中存在很多歧义问题导致的错误，如 CRF 标注的表达式“4.1 分”表示得分 4.1，并不是时间表达式；此外， CRF 方法在训练语料比较单一，类型不丰富的情况下识别结果召回率低。文献[11]将时间词表、词性标注、位置信息等作为特征，采用 CRF 方法对中文时间表达式进行识别，日期型时间短语的识别结果 F 值为 95.70%；文献[13]采用了语义角色标注之后 CRFs 识别的方法，识别结果 F 值达到 85.6%。与它们的实验结果相比，本文使用 CRF 识别时间单元，减小了识别粒度；增加了 CRF 错误标注去除部分，进一步提高 CRF 识别结果的精度；再制定了规则补充识别时间单元，提高了召回率；添加限制条件去除有歧义时间表达式，进一步提高识别精度，最后达到了 F 值为 98.31%的较好效果。

规则的方法，精确率较高，但是召回率取决于规则的完善度，规则完善则召回率高，反之，则召回率低。制定规则需要大量人工，且领域适应性较差，对于不常见的时间词识别效果不好。文献[9]采用规则方法识别时间单元，再根据就近结合时间单元原则识别时间表达式，识别结果 F 值为 84.67%。单纯采用规则方法识别时间表达式，容易产生错误规则，导致精确率不高，因此，本文采用统计方法识别时间单元，并且添加了时间表达式错误处理部分，进一步提高了识别时间表达式的精确率。

本文在衡量两种方法的利弊后，采用两者结合的办法。在 CRF 识别时间单元的基础上，根据触发词库制定规则，补充识别时间单元，能有效地提高识别结果的召回率，弥补 CRF 由于训练语料不全面导致的召回率低的问题，还能通过完善触发词库较好的识别低频时间单元。识别的效果一定程度上受时间触发词库的规模影响，本文采用规则自动获取训

练语料中的时间单元作为候选触发词，通过评价函数筛选之后，加入到触发词库中，自动获取触发词能节省人工，不断更新、完善触发词库，提高识别的召回率。在时间单元正确识别的基础上，根据时间缀词库，制定规则识别时间表达式的边界，能有效地解决时间表达式边界识别的难题。

## 5 结论及下一步工作

本文在 TempEval2 基础上，结合中文时间表达式实际情况，对时间类型进行补充、修改，使得时间表达式类型更符合中文的特点。利用 CRF 与规则相结合的方法，识别出中文时间表达式。在新闻领域的 F1 值达到 98.31%，取得了不错的效果。时间单元的格式相对固定，转化应用的领域时，只需要较少的变动（修改相应的时间触发词库与时间缀词库）就能适用于其他领域，具有较好的可移植性。

基于测试语料自动获取时间触发词的方法可以提高时间表达式识别的召回率，但是，不可避免的会带来一些杂质，降低识别的正确率。评价函数能有效地去除一些明显的杂质，阈值 λ 的设置需要经过反复试验，过高容易过滤掉很多正确的触发词，过低则会极大降低识别的正确率。另外，评价函数只能作用于有正确标准答案的测试语料，而这部分的语料资源较少，导致自动获取的触发词较少，这也是本文不足的地方之一，需要进一步加以改进。

并非所有的时间表达式都含有触发词，存在不含有触发词的时间表达式，该类时间表达式的识别只能通过分析其语义确认，如“2013 底”。本文对于该类不含有触发词的时间表达式识别效果不好。下一步，我们将研究如何识别这类的时间表达式及中文时间表达式的规范化问题。

## 参考文献


[1]高霄云，杨建林.基于规则的中文时间词和数词的自动识别算法[J].现代图书情报技术，2007(3):46-50.

[2] Mingli Wu,Wenjie Li,Qin Lu,Baoli Li.A Chinese Temporal Parser for Extracting And Normalizing Temporal Information[C]//International Joint Conference on Natural Language Processing (IJCNLP),2005,Volume 3651:694-706.

[3]乌桐，周雅倩，黄萱菁，等.自动构建时间基元规则库的中文表达式识别[J].中文信息学报,2010，24(4):3-10.

[4]林静，曹德芳，苑春法.中文时间信息的 TIMEX2 自动标注[J].清华大学学报（自然科学版）,2008,48(1):117-120.

[5] Ferro L.Gerber L.Mani I.et al.TIDES 2003 Standard for fhe Annotation of Temporal Expressions[EB/OL].2003-09. http://timex2.mitre.org.

[6] Ferro L.Gerber L.Mani I.et al.TIDES 2005 Standard for fhe Annotation of Temporal Expressions[EB/OL].2005-09.http://timex2.mitre.org.

[7] PawelMaqur ,Robert Dale. A Rule Based Approach to Temporal Expression tagging[C]//Proceeding of the International Multiconference on Computer Science and Information Technology.2007,293-03.

[8]赵紫玉,徐金安，张玉洁，等.规则与统计相结合的日语时间表达式识别[J].中文信息学报,2013，27(6):192-200.

[9]李君婵，谭红叶，王凤娥.中文时间表达式及类型识别[J].计算机科学,2012，39(11A)：191-211.

[10] David Ahn,Sisay Fissaha Adafre,Maarten De Rijke.Towards Task-Based Temporal Extraction and Recognition[C]//Proceedings Dagstuhl Workshop on Annotating, Extracting, and Reasoning about Time and Events,2005.
[11]朱莎莎，刘宗田，付剑锋，等.基于条件随机场的中文时间短语识别[J].计算机工程，2011，37(15):164-167.
[12]贺瑞芳，秦兵，刘挺，等.基于依存分析和错误驱动的中文时间表达式识别[J].中文信息学报,2007，21(5):36-40.
[13]刘莉，何中市，邢欣来，等.基于语义角色的中文时间表达式识别[J].计算机应用研究,2011，28(7):2543-2545.
[14] Gerber L,Huang S,Wang X. Standard for fhe Annotation of Temporal Expressions, Chinese supplement draft[EB/OL].2004-04. //timex2.mitre.org.

**作者简介：**吴琼（1988—），女，硕士，主要研究方向为自然语言处理与机器翻译。Email: wuqiong2012@mail.dlut.edu.cn；黄德根（1965—），男，博士，教授，博士生导师，中国中文信息学会高级会员，主要研究方向为自然语言处理与机器翻译。Email: huangdg@dlut.edu.cn。