规则与统计相结合的日语时间表达式识别

赵紫玉，徐金安，张玉洁，刘江鸣

（北京交通大学 计算机与信息技术学院，北京 100044）

摘要：本文提出了一种基于自定义知识库强化获取规则集，以及规则与统计模型相结合的日语时间表达式识别方法。在按照 Timex2 标准对时间表现进行细化分类的基础上，我们结合日语时间词的特点，渐进地扩展重构日语时间表达式知识库，实现基于知识库获取的规则集的优化更新，旨在不断提高时间表达式的识别精准度。同时，融合 CRF 统计模型提高日语时间表达式识别的泛化能力。实验结果显示开放测试 F1 值达 0.8987。

关键词：知识库；规则集；统计模型

中图分类号：TP391　　**文献标识码：**A

Japanese Time Expression Recognition by Combining Rules with Statistics

ZHAO Ziyu，XU Jin'an，ZHANG Yujie，LIU Jiangming

（School of Computer and Information Technology，Beijing Jiaotong University，Beijing 100044，China）

Abstract: Based on the knowledge base we defined, this paper presents a Japanese time expression recognition method through combining rules set strengthened by knowledge base with statistical model. According to the Timex2 standards' granular classification on time, we progressively expanded and reconstructed the knowledge base given the Japanese time characteristic, and then achieved rules set optimization and update, in order to increase recognition accuracy. Simultaneously, we fused CRF model to enhance the generalization ability of Japanese time expression recognition. Our experimental results show that the F1 value reaches 0.8987 on open test.

Key words: Knowledge base; Rules set; Statistical model

1 引言

时间表达式是句子中的重要成分，是关键信息的载体。时间表达式的抽取和处理是当前自然语言处理中的一个重要研究方向。正确识别时间表达式具有重要的意义。

近年来，时间表达式的识别和规范化在事件跟踪，时间关系推理，时序定位等方面的应用越来越多，不仅可以提高分词、句法分析的精度，还可改善机器翻译、信息抽取、文本摘要、对话系统的性能。比如，在机器翻译中，可以使译文更加流畅[1]；在多文档自动摘要中，可以对文档信息进行时序排序[2]；在自动问答系统中，可以用于回答“多久，何时”等与时间相关的问题。

时间表达式识别与规范化研究，最早是 1995 年信息理解会议（Message Understanding

基金项目：科技部国际科技合作计划（K11F100010）；中央高校基本科研业务费专项资金（2010JBZ2007）；北京市重点学科共建项目(计算机应用技术)；中国科学院计算技术研究所智能信息处理重点实验室开放课题(IIP2010-4)；北京交通大学人才基金(2011RC034)

作者简介：赵紫玉(1987—)，女，硕士研究生，主要研究方向为自然语言处理；徐金安(1970—)，通讯作者，男，副教授，硕士生导师，主要研究方向为自然语言处理和机器翻译；张玉洁(1961—)，女，教授，硕士生导师，主要研究方向为自然语言处理、机器翻译和文本大数据处理；刘江鸣(1989—)，男，硕士研究生，主要研究方向为自然语言处理。

Conference, MUC）把时间表达式的识别作为命名实体识别的一个子任务。在美国国家技术标准局（NIST）于 2004 年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。TERN 评测有TER(Time Expression Recognition)和 TEN(Time Expression Normalization)两个子任务，前者的任务是识别时间表达式边界，而后者是按照 Timex2 规范标注时间表达的属性值，即进行时序语义标注。目前，TERN 评测涉及阿拉伯语、英语和汉语等，对于韩语、法语和西班牙等语言也有人进行初步的探索性研究，但是针对日语时间表达式识别的评测工作比较少。

时间表达式识别方法一般可以分为以下两类。

一类是基于规则的方法[3]，该类方法一般通过分析短语内部的构成规律和短语外部的约束信息来识别时间表达式。香港理工大学的李文婕[4]等人做了比较具有代表性的尝试，文章提出建立一些语法规则和补充限定规则，通过规则匹配方式识别时间表达式。传统方法认为时间表达式的表现形式比较规范，倾向于采用规则的方法来做识别任务，但是规则的撰写耗时耗力，对具体领域的依赖性强，可移植性较差，而且构建的规则往往会有粒度过粗的缺点。

另一类是基于机器学习的时间序列标注方法[5]，该方法主要包括隐马尔可夫模型，最大熵模型和条件随机场模型。D.Ahn[6]和K.Hacioglu[7]都做了尝试，他们首先将语料进行预处理，接着有选择地抽取特征，建立特征向量，通过预选的分类器（CRF或者SVM）训练模型，然后用训练好的模型标注测试语料的时间表达式。这类方法最大的特点是可以充分利用已标注上下文信息，使得识别召回率较高，而且无需消耗太多的人力，但是受限于日语时间表达式语料规模的局限性和质量，训练语料容易存在数据稀疏问题，也无法充分利用时间表达式格式相对规范的特点，使得基于机器学习的序列标注方法难以充分发挥它的优势。

本文针对传统方法的优缺点，提出了统计与规则相结合的日语时间词识别方法。该方法不仅可提高时间表达式识别的精准度和召回率，同时可提高日语时间表达式识别的泛化能力和领域适应能力，从而节约人工成本。

另外，通过研究日语时间表达式的边界识别方法，按照 Timex2 标注方案对时序表达式类别的描述，本文将日语时间表达式分为七个类别，并为这七个类别预先定义日语时间词触发词表等知识库，建立人工启发式规则模板，最后提出基于知识库强化的规则集和统计模型相结合的识别方法，这样既有效地利用了上下文信息，又达到了较高的自动化程度。实验结果验证了提出方法的有效性。

本文结构安排如下：第 2 节论述日语时间表达式类型的基本概念及问题分析；第 3 节介绍日语时间表达式识别系统结构以及我们提出的基于知识库强化获取规则集和规则与统计相结合的识别方法的主要思想；第 4 节阐述系统实验，相关的评测方法和评测结果，并进行结果分析。最后，总结全文并提出未来工作。

2 基本概念及问题分析

2．1 时间表达式

时间是频繁使用的词类，日语和汉语在时间的使用上有很多相同和相似之处。参照 Timex2 中关于中文时间表达式的描述，本文将日语时间表达式定义为由一个或多个时间基类组成的时间短语，即时间表达式为时间基类的序列，如“平成 14 年 6 月 1 日”此时间表达式由 2 个时间基类集合而成：“平成 14 年”、“6 月 1 日”。“时间基类”，即基本时间类型，本文提出 7 种基本时间类型，是构成时间表达式的最小组成类型。由此可定义日语时间表达式为

$$T_e = (t_1, t_2, \ldots, t_i, \ldots, t_m)\ (\mathrm{m} >= 1), \tag{1}$$

其中，T_e为一个时间表达式，它是 m 元组；其中$t_1, t_2, \cdots, t_i, \cdots, t_m$是 m 个独立的时间基类。

2．2 时间基类

本文参照 Timex2 标注方案对时间表达式类别给出的描述，将日语时间表达式分为绝对时间（Absolute Time），相对时间（Relative Time），段时间（Duration），集合时间（Set-denoting Time），事件触发时间（Event-anchored Time），文化相关时间（Culturally -determined Time），不特定时间（Fuzzy Time）7 类时间基类。具体描述说明如下表 1。

表 1　时间基类

时间基类名称	时间基类描述
绝对时间	指一个具体固定且与时间的推移无关的时间词，如“2013 年 5 月 27 日”
相对时间	指随着时间的推移，所指的时间产生变化的时间词，如“一昨年、夕方”
段时间	指一个持续的时间段，如“一週間”
集合时间	指某一类时间的集合，如“梅雨期間、每日”
事件触发时间	指跟特定事件相关的时间词，如“天和元年、平成 11 年”
文化相关时间	指日本文化纪念日的时间词，如“建国記念の日、公休日”
不特定时间	指日本二十四节气、七十二候、季节、六曜历法、天干地支等时间词，如“立春、雨水”

2．3 日语时间表达式知识库

知识库是关于某一项领域的陈述性知识、过程性知识和策略性知识的集合[8]。在该集合中各类知识通过一定的表示方法表示，并建立相互之间的联系。它与数据库的区别就是知识库中不但包含了大量的简单事实，还包含了规则、过程型知识和策略性知识。从存储知识的角度来看，知识库以描述型方法来存储和管理知识。

相比Part-Of-Speech (POS)，大部分的知识库系统更多的是依赖于浅层句法分析技术，应用正则表达式或语言模式，以及适当检查名称列表。这些系统中有一些处理分析深层语义，这种方法已被证明性能杰出[9]。

这里我们总结的知识库包括，日语时间触发词知识库、日语时间表达式边界知识库、日语时间表达式规则关键词知识库，日语月份的多种表示法知识库、以及基于日语维基百科的知识库，如表 2 至表 6 所示。

表 2　日语时间触发词知识库

触发词类型	触发词个数	触发词示例
绝对时间触发词	7	時、分、秒
相对时间触发词	122	夜明け、朝午前
段时间触发词	19	秒間、分間、時間
集合时间触发词	2	期間、每
事件触发时间触发词	557	世紀、紀元、王朝、時代
文化相关时间触发词	115	の日、纪念日、诞生日
不特定时间触发词	148	暦、盛夏、夏、立春
总和	970	

表 3　日语时间表达式边界知识库

边界类型	边界个数	边界示例
助词	23	格助词、提示助词
日语标点符号	47	、。『』「」等
其他	1	約
总和	71	

表 4　日语时间表达式规则关键词知识库

关键词所属类别	关键词个数	关键词示例
触发词	970	7 类时间基类时间触发词
边界	71	助词、日语标点符号、其他
阿拉伯数字	10	0、1、2
中文简体数字	10	零、一、二
中文繁体数字	10	零、壹、貳
总和	1071	

表 5　日语月份的多种表示法知识库

月份	平假名表示法个数	示例
1 月	49	睦月、建寅月、孟春
2 月	34	衣更着、建卯月、仲春
3 月	45	弥生、建辰月、季春
4 月	35	卯月、建巳月、孟夏
5 月	32	皐月、建午月、仲夏
6 月	35	水無月、建未月、季夏
7 月	33	文月、建申月、孟秋
8 月	38	葉月、建酉月、仲秋
9 月	42	長月、建戌月、晚秋
10 月	36	神無月、建亥月、孟冬
11 月	8	霜月、建子月、仲冬
12 月	51	師走、建丑月、晚冬
总和	438	

表 6　基于日语维基百科的知识库

列表名称	词条举例
世界性节日	国際労働者の日、母の日
日本国节日	国民の祝日、建国記念日
基督教节日	クリスマス（圣诞节）、キリストのはりつけ（基督受难日）
二十四節気	立春、雨水
七十二候	東風解凍、黄鶯睍睆

2．4 问题分析

通过对大量日语语料分析，在日语时间表达式识别研究过程中，发现时间表达式的多样性问题及若干歧义现象：

1. 一般日语时间表达式中会包含标识时间表达式出现的触发词[10]，但是也存在不包含触发词的时间表达式；而且并非所有包含触发词的表达式都是时间，如“日中経済協会理事長”中的“日中”是相对时间的触发词，结合上下文可以看出此处的“日中”并不是时间表达式。因此单纯的基于触发词的规则方法不能准确地识别时间表达式。

2. 时间表达式由多个独立性较强的时间基类单元组成，时间基类为时间概念词。据统计[11]，近49％的时间表达式为一个独立的时间单元，如“先月”、“昨日”等；26％的表达式由两个时间单元构成，如“去年六月”是由“去年”和“六月”两个时间单元组成；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个以上的时间单元组成。另外，某些完整时间表达式中包含非时间概念词的时间单元，如“九時三分前”中的方位词“前”，由于其与时间概念词结合起来可表达完整的时间意义，因此这类非时间概念词也需要准确识别。

鉴于时间表达式的上述特点和难点，不能仅通过时间触发词等词形信息来制定规则，还应该结合知识库来强化规则集，并结合统计模型，提高识别准确率。因此，本文针对时间表达式的内部组成结构和时间基类单元的相对独立性，提出并构建基于知识库强化规则集和统计模型相结合的时间表达式识别系统。

3 规则与统计相结合的日语时间表达式识别

3．1 日语时间表达式识别系统结构

本文提出基于知识库强化获取规则集，以及规则集与统计模型相结合的识别方法。首先通过初始构建的日语时间表达式知识库强化获取规则集，训练统计模型；其次分别基于规则和统计两种方法进行日语时间表达式识别，并整合二者的识别结果；基于错误驱动学习算法[12]的思想，根据整合后的识别结果，学习初始构建的知识库与识别结果的差异来校正知识库及规则，其本身体现了重构知识库的触发环境。图 1 为本文提出的日语时间表达式识别系统流程图。

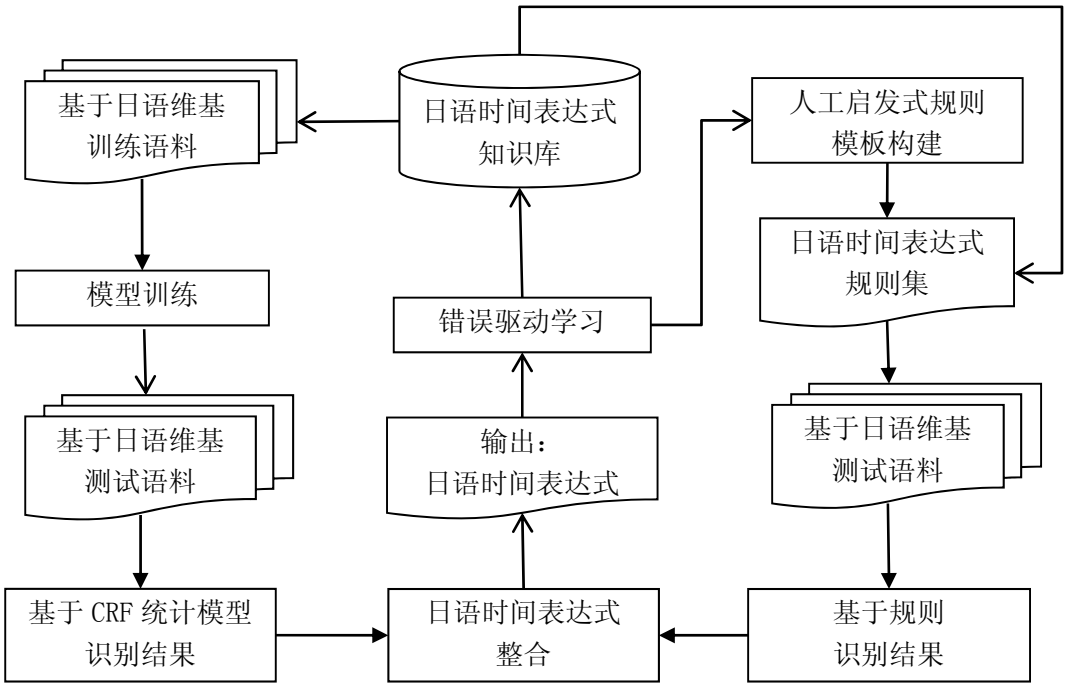

图 1　日语时间表达式识别流程图

3．2 基于条件随机场的日语时间表达式识别

时间表达式识别可以定义成序列的标记问题，即判断观察词是否属于预先定义的特征集合。目前，常用的序列标注模型主要有隐马尔科夫模型（HMM），最大熵模型（ME）和条件随机场模型（CRF）。HMM 一个最大的缺点就是由于其输出独立性假设，导致其不能考虑上下文的特征，限制了特征的选择。ME 解决了这个缺点，可以任意选择特征，但由于其在每一节点都要进行归一化，所以只能找到局部变量的最优值，同时存在标记偏置的问题（Lable Bias），即凡是训练语料中未出现的情况全部忽略掉。CRF 无独立性假设，可以任意选择特征，并且使用单一的指数族函数对整个观测序列的联合分布进行建模，可以求得全局最优解。因此，本文选用条件随机场进行时间表达式识别。

条件随机场（Conditional Random Field, CRF)是一种基于统计的序列标记识别模型，它由John Lafferty等[13]在2001年首次提出，模型的主要思想来源于最大熵模型。它是一种在给定输入节点(观察值)条件下，计算输出节点(标记)的条件概率的无向图模型，目标是在给定需要标记的观察序列条件下，使标记序列的联合概率达到全局最优。条件随机场模型(CRF)具有表达字串长距离依赖性和交叠性的能力, 能较好地学习新的领域知识[14], 所以采用CRF模型来识别日语时间表达式。条件随机场定义如下：

$$P(y \mid x)=\frac{1}{Z(X)}\exp\left(\sum_{i}\sum_{k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i}\sum_{k}u_k s_k(y_i,x,i)\right), \quad (2)$$

其中，$t_k(y_{i-1},y_i,x,i)$为转移函数, 表示观察序列和标记序列$i-1$和i时刻的特征; $S_k(y_i,x,i)$为状态函数,表示观察序列和标记序列在i时刻的特征; Z(X)为归一化因子; λ 和u为训练所得参数。

CRF 统计模型将日语时间表达式识别看作一个序列标注过程, 观察值为所有分析状态的集合（日语字符集合），基于由字构词的理念，利用词位信息来标记时间词，标记则是表示时间表达式“开始”，“中间”， “结尾”和“非／其他”位置的四种位置标签集合{B，I，E，O}，而对于不同种类的时间表达式，即本文提出的绝对时间、相对时间、段时间、集合时间、事件触发时间、文化相关时间及不特定时间，分别用 Absolute、Relative、Duration、Set、EventAnchored、CultureRelated 及 Fuzzy 标识，因此，时间表达式的识别过程即为字符在字串中的特征标记的过程。根据 BIEO 分类标记，本系统中的分类标注集说明如表 7 所示，表 8 为时间表达式的标注形式说明。

表7　四词位分类标注集

标注符号	符号说明
B-[时间基类]	时间表达式的开始位置
I-[时间基类]	时间表达式的中间位置
E-[时间基类]	时间表达式的结尾位置
O	非／其他位置

说明：时间基类是指 Absolute、Relative、Duration、Set、EventAnchored、CultureRelated、Fuzzy

表8　时间表达式标注形式

时间表达式标注形式	说明
B	词长为1的时间表达式
BE	词长为2的时间表达式
BI…IE	词长大于2的时间表达式

3．2．1　特征模板与特征

特征模板的设置对时间表达式的标注识别的好坏起到关键的作用,本文利用上下文信息，

从训练语料中获得字符特征，主要采用当前字和其前后两个字符及其词性信息作为特征。具体的特征模板的设置如表9所示，其中C代表当前字，S代表词性。

表9　特征模板

特征类型	特征模板
Unigram(一元)	$C_n, S_n,\ \mathrm{n} = -2, -1, 0, 1, 2$
Bigram(二元)	$C_nC_{n+1}, S_nS_{n+1},\ \mathrm{n} = -2, -1, 0, 1$

本文分析和研究日语时间表达式内部结构和上下文环境对其的影响程度，使用词形与日语形态素信息作为特征（表10）。

表 10　词法、句法特征

特征描述	说明
词形	词本身
形态素信息	包括词性，日语活用型等

3．2．2　识别算法

本文使用了 CRFs 开源的工具包 CRF++[15],该工具包的具体使用方法参考文献[16]。基于 CRF 统计模型识别算法流程如图 2 所示。

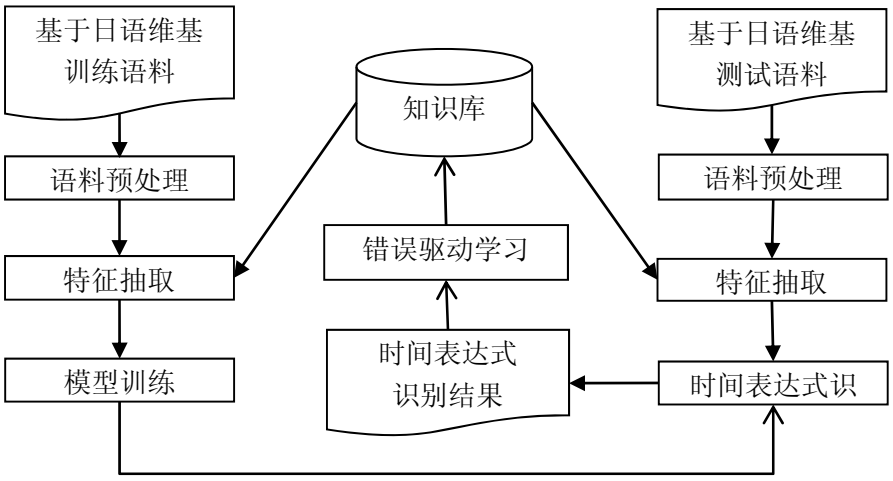

图 2　基于条件随机场的日语时间表达式识别

具体识别算法如下：

1. 语料预处理

对 XML 格式的日语维基语料进行解析清洗，去掉不需要的标签，保留完整的文本，去停词，同时对数字、百分号、货币等特殊实体进行整合。

2. 特征抽取

将对预处理后的观察序列进行特征抽取，上文详细介绍了 CRF 模型所使用的全部特征。

3. 模型训练

使用开源的 CRF 工具[15]完成参数训练过程。

4. 时间表达式识别

识别过程即解码过程。在测试语料上使用已经训练好的 CRF 模型识别日语时间表达式。

5. 错误驱动学习动态扩展重构知识库

根据识别结果与标准数据的差异，校正知识库。

3．3 基于知识库强化规则集的日语时间表达式识别

本文采用的规则集由自定义的人工启发式规则模板结合日语时间表达式知识库自动生

成。

3．3．1　人工启发式规则模板

文献[17]指出日汉两语完全相同的语言项并不多，仅占 8.6%，完全不同的也不多，占 21.9%。说明日汉时间词的相似程度较高。我们根据自己的日汉时间词研究经验构建一部分启发式规则，作为人工启发式规则模板。

根据我们对 TERN 任务的归纳，参照 TIMEX2 中文标注方案，日语时间表达式大体分为七中类型（见表 11）。针对每个不同类型，我们各自构建了人工启发式规则模板。基于日语时间表达式知识库，通过正则表达式匹配方式，这些规则集可以识别出大部分日语时间表达式。

表11　人工启发式规则模版和规则集示例

时间表达式类型	规则模板示例	规则示例（正则表达式）
绝对时间	[数字]+[绝对时间触发词]	\d+年\d+月\d+日
相对时间	[边界词][相对时间触发词]+[边界词]	を.*(夜明け|朝午前).*も
段时间	[数|何|半][段时间触发词]	数(秒間|時間)
集合时间	每[绝对时间出发词|集合时间触发词]	每(年|月|日)
事件触发时间	[事件触发时间触发词][数字]+年	平成\d+年
文化相关时间	[文化相关时间触发词][节日]	国際労働者の日
不特定时间	[不特定时间触发词]	立春|雨水|啓蟄|春分

3．3．2　识别算法

本文在人工规则模版上利用知识库信息，生成强化的规则集。利用正则表达式匹配方式，以句子为单位识别时间表达式。基于错误驱动学习的思想，根据识别结果与标准数据的差异，校正知识库。

3．4 规则和统计融合策略

基于规则和统计融合的时间表达式识别模型既可以获取训练语料的知识，又可以弥补统计模型的不足。一方面，基于规则的方法可以很好地表达语言的确定性现象，从而克服统计模型在此方面的缺点；另一方面，统计模型的泛化能力可以弥补基于规则方法的领域依赖性强和可移植性差等缺点。二者的结合可以达到很好的互补效果。因此，融合模型的建立已成为时间表达式识别过程中的关键问题。本文提出的融合策略如下：

1.以基于规则的识别结果为基础，观察未被识别到的时间表达式，通过错误驱动更新日语时间表达式知识库，并使用人工启发式的方法，对规则模版进行修正。最后以更新的知识库和修正的模版为基础重构规则集。

由于规则的泛化能力有限，观察规则识别错误的时间表达式。以此为依据修正规则。错误主要表现在两个方面：一方面，规则识别错误，例如“二日市”等专有地名，利用形态素分析信息（名詞-固有名詞-地域-一般），加强规则的限制；另一方面，规则识别召回率低，例如“卒業後しばらくして（毕业后一段时间）”，“から……（从……到）”等，总结时间表达式，提高规则的表现能力。

2.使用更新的知识库和规则模板处理语料，提高训练语料的质量。然后利用统计模型的泛化能力进行时间表达式的识别。

3.在规则与统计结果整合的过程中，采取贪心策略。最终的时间表达式为规则与统计同时覆盖到的片段的最长序列（公式 3）以及各自所识别到的时间表达式。

$$\text{TimeExpression} = \text{sequence}(\text{maxleft}(R, S):\text{maxright}(R, S)) \quad \text{if } R \cap S \neq \varnothing, \tag{3}$$

其中，maxleft(R, S)表示规则结果与统计结果最左边字符，maxright(R, S)表示规则结果

与统计结果最右边字符，sequence(i:j)函数表示从字符 i 到字符 j 的连续字符串

综上所述，本文提出的融合方法包括：一方面，系统采用错误驱动人工启发式的方法，利用知识库强化规则集，并融合统计模型泛化能力，识别时间表达式；另一方面，系统采用贪心策略，整合规则和统计的识别结果。因此，本文提出的规则和统计结合的日语时间表达式识别框架及方法，分别利用规则与统计的优点，旨在提高时间表达式识别系统性能。

4 实验及分析

4．1 实验语料

本文实验语料采用具有实时性的日语维基资源库。将该语料经过去标签、篇章分割、去不含时间词句子、标注时间表达式等预处理后，随机分为测试语料以及训练语料两部分，具体语料信息如表 12。

表 12　实验语料信息

实验数据	句子数	平均句长（字数）	时间表达式数
训练语料	236206	60.8	452498
测试语料	1691	57.8	3766

4．2 评测方法

本文使用 PRF 评测指标：识别的准确率(P)、召回率(R)和 F1值，和精度（Accuracy）计算公式分别为：

$$\text{准确率}(\mathrm{P})=\frac{\text{正确识别的时间表达式个数}}{\text{系统识别出的时间表达式个数}}\times 100\%\,, \tag{4}$$

$$\text{召回率}(\mathrm{R})=\frac{\text{正确识别的时间表达式个数}}{\text{测试语料中的时间表达式个数}}\times 100\%\,, \tag{5}$$

$$\mathrm{F}=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}\,, \tag{6}$$

在综合评测系统性能时， P 和 R 都要同时考虑，但同时比较 P 和 R 两个值，很难做到准确分析结果优劣，因此通常采用 F 值对系统进行评测。β是准确率 P 和召回率 R 的相对权重，在本文中二者同等重要，因此β取值为 1。

本文提出使用覆盖度（Coverage）指标评测系统正确识别的时间表达式覆盖 7 类时间单元的程度，其计算公式如下：

$$\text{覆盖度(Coverage)}=\frac{\text{正确覆盖的时间单元个数}}{\text{时间单元总数}}\,, \tag{7}$$

4．3 实验结果及分析

由于目前几乎没有中英论文做日语时间表达式识别的研究工作，故很难进行对比试验。本文对时间表达式识别系统进行多次实验，比较基于规则方法和基于统计方法的识别结果的差异。采用不同策略融合规则和统计的识别结果，使基于规则与统计想结合的日语时间表达式识别系统取得突出的识别效果。表 13 为基于规则，基于统计以及基于规则与统计融合的时间表达式识别结果。表 14 为时间基类识别的实验结果。

表 13　时间表达式识别结果

	P	*R*	F1
基于规则方法	0.8110	0.7534	0.7811
基于统计方法	0.9432	0.8383	0.8877
规则统计融合	0.9364	0.8639	0.8987

表 14　时间基类识别结果(Coverage)

	绝对时间	相对时间	段时间	集合时间	事件触发时间	文化相关时间	不特定时间
基于规则方法	0.8964	0.4005	0.8874	0.8902	0.9383	0.9478	0.4500
基于统计方法	0.9678	0.8816	0.9665	0.8140	0.9515	0.9395	0.1214
规则统计融合	0.9698	0.8816	0.9686	0.8963	0.9515	0.9560	0.5142

从表 13 可以看出基于规则与统计融合的识别方法优于单独使用规则或统计的方法，主要体现在召回率上；然而准确率有细微的降低，主要原因是融合算法目前按照时间表达式的字表面特征进行融合，还有很大的优化空间，此外，规则的不完备性可带来的噪声和语料自身的噪音也很难避免。表 14 说明基于规则的方法在有明显格式的时间基类（集合时间和文化相关时间）上表现较好，例如“期間”，“平成年間”和“国際労働者の日”。但是，基于规则的方法在相对格式复杂的时间识别上效果较差，特别是相对时间。一方面，相对时间识别规则中的边界词难以确定，本文暂使用日语的格助词及标点符号作为边界词；另一方面，统计模型识别的方法能够通过统计特征识别边界。因此基于统计的方法在相对时间基上的识别更为突出。

本文分析实验结果得出以下结论，无论是基于规则还是统计的方法在不特定时间上的识别效果尤其差，其中原因包括：

1. 相对训练数据稀少且难以保证其准确性。

2. 存在严重的词义歧义问题，例如“雨水”、“小雪”和“大雪”等，不仅作为二十四节气，且更为普遍地作为天气的自然现象词汇，其上下文信息极为类似。

综上所述，利用统计与规则的融合策略，弥补各自方法的不足。实验结果显示基于规则的方法和基于统计模型的方法，能够有效地识别日语时间表达式，并且基于规则与统计相结合的方法能够提高日语时间表达式的识别效果。

5 总结及未来的工作

本文提出一种知识库增强的基于规则和统计分析识别日语时间表达式的新方法，实现在训练语料规模匮乏的条件下，尽可能减少人工参与，使得系统在拥有较好的模型学习能力的同时高质量识别日语时间表达式，这是本方法的优点之一；但也会引起一些识别歧义问题，识别只包含一个词或知识库特征不明显的时间表达式带来的识别歧义问题更为突出，采用知识库增强规则集解决歧义问题又是本文的另一个优点。另外，可以进一步优化扩展重构知识库。实验证明这种方法可行。

今后，我们将在更多的日语数据集和领域上做日语时间表达式识别实验，并与日语论文中有关日语时间表达式识别的最好算法和识别结果作比较，进一步提高本文算法的泛化能力，使之适应更广泛的应用领域；同时，尝试更多有效的特征，提高统计模型的识别精度，特别针对不特定时间，使用深层语义特征提高识别效果；以及尝试运用错误驱动思想的规则筛选策略，达到自动学习规则，减少人力，提高识别性能和效率。在此之上，探索日语时间表达式如何高效地翻译成中文时间表达式，并应用于日中机器翻译系统中，旨在于提高日中机器翻译效果。

参考文献

1. 邬桐, 周雅倩, 黄萱菁, 等. 自动构建时间基元规则库的中文时间表达式识别[J]. 中文信息学报, 2010, 24(004): 3-10.
2. 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.
3. Pawel Maqur and 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.
4. Mingli Wu, Wenjie Li，Qin Lu，Baoli Li. A Chinese Temporal Parser for Extracting and Normalizing Tem—poral Information[C]//International Joint Conference on Natural Language Processing(IJCNLP)，2005. Volume 3651：694—706.
5. David Ahn,Sisay Fissaha Adafre,Maarten de Rijke.Recognizing and Interpreting Temproal Expressions in Open Domain Texts [J].Digital Information Management,2005,3(1):14-20.
6. 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.
7. Kadri Hacioglu，Ying Chen. Benjamin Douglas Automatic Time Expression Labeling for English and Chinese Text[C]//Computational Linguistics and Intelligent Text Processing(CfCLing)，2005，Volume 3406：548—559.
8. 刘成亮, 韩海伟. 知识库系统的原理及其在智能搜索引擎中的应用[J]. 电脑知识与技术, 2008, 8: 1512-1514.
9. Brun & Hag`ege, 2004; Brun & Hag`ege,2009; van Shooten et al., 2009.
10. ACE(Automatic Content Extraction)Chinese Annotation Gubdelines for TIMEX2(Summary). Version 1.2 20050610.
11. 李君婵, 谭红叶, 王风娥. 中文时间表达式及类型识别[J]. 计算机科学, 2012, 39(z3).
12. Brill, Eric. Transformation-based error-driven learning and natural language processing:A case study in part of speech tagging. Computational Linguistics,1995,21(4):543-565.
13. Lafferty J.McCallum A, Pereira F. Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data[J]. The Journal of Machine Learning Research,2001,ICML01:282-289.
14. He Y, Kayaal P M. Biological entity recognition with conditional random fields// Proceedings of AMIA Annual Symposium. Washington, DC, 2008: 293–297.
15. Kudo T.CRF++:Yet another CRF toolkit[OL].[2009-02-25]. http://crfpp.sourceforge.net/.
16. 廖先桃. CRF 理论、工具包的使用及在 NE 上的应用[OL]. [2010-06-05]. http://ir.hit.edu.cn/phpwebsite/index.php?module=documents&JAS_DocumentManager_op=viewDocument&JAS_Document_id=199.
17. 辛永芬. (2005). 日汉时间词对比分析及相关问题. 河南大学学报 (社会科学版), 3, 021.